Affective Language in Human and Machine

AISB 2008

AISB 2008 Proceedings Volume 2

GLoRiClass

aisb ABERDEEN - 2008

UNIVERSITY OF ABERDEEN

AISB 2008 Convention
Communication, Interaction and Social Intelligence

1st-4th April 2008

University of Aberdeen

Volume 2:
Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine

Published by
The Society for the Study of Artificial Intelligence and Simulation of Behaviour

http://www.aisb.org.uk/convention/aisb08/

ISBN 1 902956 61 3

Contents

The AISB'08 Convention: Communication, Interaction and Social Intelligence

As the field of Artificial Intelligence matures, AI systems begin to take their place in human society as our helpers. Thus it becomes essential for AI systems to have sophisticated social abilities, to communicate and interact. Some systems support us in our activities, while others take on tasks on our behalf. For those systems directly supporting human activities, advances in human-computer interaction become crucial. The bottleneck in such systems is often not the ability to find and process information; the bottleneck is often the inability to have natural (human) communication between computer and user. Clearly such AI research can benefit greatly from interaction with other disciplines such as linguistics and psychology. For those systems to which we delegate tasks: they become our electronic counterparts, or agents, and they need to communicate with the delegates of other humans (or organisations) to complete their tasks. Thus research on the social abilities of agents becomes central, and to this end multi-agent systems have had to borrow concepts from human societies. This interdisciplinary work borrows results from areas such as sociology and legal systems. An exciting recent development is the use of AI techniques to support and shed new light on interactions in human social networks, thus supporting effective collaboration in human societies. The research then has come full circle: techniques which were inspired by human abilities, with the original aim of enhancing AI, are now being applied to enhance those human abilities themselves. All of this underscores the importance of communication, interaction and social intelligence in current Artificial Intelligence and Cognitive Science research.

In addition to providing a home for state-of-the-art research in specialist areas, the convention also aimed to provide a fertile ground for new collaborations to be forged between complementary areas. Furthermore the 2008 Convention encouraged contributions that were not directly related to the theme, notable examples being the symposia on "Swarm Intelligence" and "Computing and Philosophy".

The invited speakers were chosen to fit with the major themes being represented in the symposia, and also to give a cross-disciplinary flavour to the event; thus speakers with Cognitive Science interests were chosen, rather than those with purely Computer Science interests. Prof. Jon Oberlander represented the themes of affective language, and multimodal communication; Prof. Rosaria Conte represented the themes of social interaction in agent systems, including behaviour regulation and emergence; Prof. Justine Cassell represented the themes of multimodal communication and embodied agents; Prof. Luciano Floridi represented the philosophical themes, in particular the impact of society. In addition there were many renowned international speakers invited to the individual symposia and workshops. Finally the public lecture was chosen to fit the broad theme of the convention – addressing the challenges of developing AI systems that could take their place in human society (Prof. Aaron Sloman) and the possible implications for humanity (Prof. Luciano Floridi).

The organisers would like to thank the University of Aberdeen for supporting the event. Special thanks are also due to the volunteers from Aberdeen University who did substantial additional local organising: Graeme Ritchie, Judith Masthoff, Joey Lam, and the student volunteers. Our sincerest thanks also go out to the symposium chairs and committees, without whose hard work and careful cooperation there could have been no Convention. Finally, and by no means least, we would like to thank the authors of the contributed papers – we sincerely hope they get value from the event.

Frank Guerin & Wamberto Vasconcelos

The AISB'08 Symposium on Affective Language in Human and Machine

The increasing awareness in HCI of the importance of considering emotions and other non-rational aspects of the human mind in computer interfaces has led to a recent surge of interest in the area known as Affective Computing. Potential applications of Affective Computing arise in areas such as healthcare and tutoring. Designing for affect also naturally arises in the development of embodied conversational agents. There are many significant unsolved problems in this area, including the modelling of emotions themselves (in terms of diagnosis and prediction) and the design of specialised hardware to facilitate the measurement or expression of affective state.

This symposium concentrates particularly on the ways in which emotion is communicated between humans through their written and spoken language. These ways may or may not suggest ways in which emotions might be communicated between humans and machines. Theories of language frequently ignore the fact that this system of communication has evolved as much for the expression and recognition of emotional state as for the conveying of factual information. The corresponding parts of AI, speech and natural language processing are only gradually waking up to the possibilities and challenges that the wider perspective opens up. This is reflected in the fact that, although there are general conferences in Affective Computing and Intelligent User Interfaces there is not yet a specialised forum for researchers to discuss how these issues relate to language.

In the symposium we particularly wish to consider the ways in which advances in fields such as Psychology and Linguistics can inform the implementations of Affective Computing that are beginning to emerge. How can knowledge of the ways that people express emotions to one another help to inform the development of affectively sensitive and effective computer interfaces using language?

The symposium takes place on 1st-2nd April 2008. Apart from the talks in which participants present their work there will also be panels/discussions. We are also very pleased to have Diane Litman of the University of Pittsburgh giving us an invited talk on "Detecting and Adapting to Student Uncertainty in a Spoken Tutorial Dialogue System".

Chris Mellish
University of Aberdeen

We also acknowledge the kind help of: Piero Cosi (University of Padova), Steliana Ivanova (Umbria Inc), Chiara Levorato (Dept. Sviluppo e Socialiazzazione, University of Padova), and Alessandro Valitutti (DiSCoF, University of Trento)

Attitude Display in Dialogue Patterns

Alessia Martalò[1], **Nicole Novielli**[1] **and Fiorella de Rosis**[1]

Abstract. We investigate how affective factors influence dialogue patterns and whether this effect may be described and recognized by HMMs. Our goal is to analyse the possibility of using this formalism to classify users' behavior for adaptation purposes. We present some preliminary results of an ongoing research and propose a discussion of open problems.

1 INTRODUCTION

Advice-giving is aimed at attempting to change, with communication, the behavior of an interlocutor in a given domain, by influencing his or her attitude (the system of beliefs, values, emotions that bring the person to adopt that behavior). Irrespectively of the application domain, this goal requires appropriate integration of two tasks: provision of general or interlocutor-specific information about aspects of the behavior that make it more or less 'correct', and persuasion to abandon a problem behavior, if needed, by illustrating negative long term consequences it entails and positive effects of revising it.

To be effective, advice-giving cannot be the same to all interlocutors. According to the Transactional Model, it should be adapted, first of all, to the *stage* at which the interlocutors may be located, in the process of passing from a 'problem' to a 'more correct' behaviour [1]: that is, to their beliefs, intentions and goals. In addition, the effect of the communication process will be conditioned by the *kind of processing* the interlocutor will make of information received. In this case, the Elaboration Likelihood Model helps in understanding how this processing is related, at the same time, to the Receiver's ability and interest to elaborate it [2]. In different situations of attention and interest, peripheral or central processing channels will be followed, each focusing on a particular kind of information, with more or less emotional features. The consequence of the two theories is that, in advice-giving dialogues, knowledge of the Receivers is essential to increase their information processing ability and interest, and therefore the effectiveness of advice-giving.

In previous papers, we discussed how the stage of change may be recognized and updated dynamically during the dialogue [3]. We also discussed how the user's 'social attitude' towards the advice-giver -in our case, an Embodied Conversational Agent (ECA)- could be recognized with a combination of language and speech [4]. In both cases, the unit of analysis was the individual user move: results were propagated in a dynamic probabilistic model, to progressively build an approximate image of the user.

In this article, we wish to discuss whether the user attitude reflects into the overall dialogue pattern rather than into the linguistic or the acoustic features of individual moves. We consider Hidden Markov Models (HMMs) as a candidate formalism to represent dialogue patterns and their relations with the user attitude. We also propose to apply this formalism in a stepwise recognition of this attitude that enables adapting the advice-giving strategy and the system behavior.

The paper is organized as follows. In Section 2, we clarify the aspects of the user attitude we intend to recognize and model: in particular, 'engagement'. In Section 3, after briefly introducing the kind of data on which our models were built, we describe how we applied HMMs to learn dialogue pattern descriptions for various user categories. In Section 4 we test the descriptive power of HMMs when trying to model the differences in dialogue dynamics between two classes of users, defined according to their background. Model testing is discussed in Section 5, while the topic of engagement recognition is dealt with in Section 6, before a brief analysis of related work (Section 7) and some final considerations about the limits of this ongoing study (Section 8).

2 WHICH ATTITUDE ASPECTS

Knowledge of the user characteristics is of primary importance when trying to build an effective persuasion strategy: this knowledge may be acquired by observing the users' behavior during the dialogue to build a dynamic, consistent model of their mind. This model can be used for adaptation purposes and should combine both affective and cognitive ingredients. Rather than considering emotions, we look at two aspects of affective interaction (social attitude and level of engagement) which are presumed to be key factors for the success of the dialogue [5].

2.1 Social attitude

With the term *social attitude* we intend *"the pleasant, contented, intimate feeling that occurs during positive interactions with friends, family, colleagues and romantic partners... [and] ... can be conceptualized as... a type of relational experience, and a dimension that underlines many positive experiences."* [6]. Researchers proposed a large variety of markers of social presence related to verbal and nonverbal behaviour [7,8,9]. By grounding on these theories, in a previous research we proposed a method to recognize social attitude in dialogues with an ECA by combining linguistic and acoustic analysis of individual user moves [4].

2.2 User engagement in advice-giving

Engagement is a quite fuzzy concept, as it emerges from analysis of the literature, to which researchers attach a wide range of related but different meanings. Sidner and Lee [10] talk about engagement in human-robot conversations as "*the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake"*. To other authors, it describes "*how much a participant is interested in and attentive to a conversation"* [11]. Pentland [12] engagement is a function of the level of

[1] Dept. of Informatics, Univ. of Bari, Via Orabona 4, 70126 Bari, Italy
Email: alessiamartalo@libero.it
{novielli,derosis@di.uniba.it}

involvement in the interaction, a concept especially addressed in the e-learning domain. Here, several researchers attempted to model the attitude of students in terms of their level of initiative [13, 14] or of how much a person is being governed by the preceding interaction rather than steering the dialogue [15].

Different definitions of engagement are meant to be coherent with application domain and adaptation purposes: some studies aim at implementing intelligent media switching, during human-human computer-mediated technology [11,16]; others [13] aim at tailoring interaction to the learner's needs. Our long-term goal is to implement a dialogue simulator which is able to inform and persuade a human interlocutor in a conversation about healthy dieting. We expect the level and kind of engagement in the system goals not being the same for all users, depending on their own goals and on how useful they perceive the interaction to be: we consider users to be 'highly engaged' when the system succeeds in involving them in its persuasion attempts. A lower level of engagement, on the contrary, is attributed to users who are only interested in the information-giving task

2.3 Attitude display and conversational analysis

This study is based on the assumption that affective phenomena influence the dialogue dynamics [10]. For this reason, we decided to model categories of users, by looking at differences in the dialogue pattern. Our assumption is supported by the usage that researchers do of *ad hoc* measures for conversational analysis, by taking into account several dimensions (linguistic and prosodic features) and units of analyis (phonemes, words, phrases, entire dialogues). Conversational turn-taking is one of the aspects of human behaviour that can be relevant for modeling social signalling [13]. Pentland [12] measures engagement by evaluating the influence that each person's pattern of speaking versus not speaking has on the other interlocutor's patterns. This is essentially a measure of who drives the conversational turn exchanges, which can be modelled as a Markov process. In a previous research, we dynamically estimated the probability value of social attitude [3] by looking at linguistic and acoustic evidences at the single user move level [4], to adapt the style of the next agent move. Detecting long lasting features of users (such as their level of engagement) can be seen as a further step towards long-term adaptation of agent's behaviour and strategy. Also, we believe that such features have a long-term impact on the overall behaviour of users. For this reason, we analyse complete dialogue patterns rather than individual *dialogue exchanges* [17]: rather than classifying the next user move, we want to predict their overall final attitude. By using the formalism of HMMs, we aim at representing differences in the whole structure of the dialogues among subjects with the kinds of engagement we mentioned above.

3 MODEL LEARNING

After becoming a very popular method in language parsing and speech recognition [18,19], Hidden Markov Models are, more recently, being considered as a formalism to be applied to dialogue processing with various purposes: to describe and classify dialogue patterns in various situations and to recognize the category to which new dialogues probably belong. This new application domain requires careful critical analysis, to understand the conditions under which successful studies can be performed. This paper is a contribution in this direction.

3.1 Corpus description

Our corpus includes 30 text-based and 30 speech-based dialogues with an ECA, collected with a Wizard of Oz (WoZ) study: overall, 1700 adjacent pairs (system – user moves). Subjects involved were equidistributed by age, gender and background (in computer science or humanities).

3.2 Corpus labelling

The corpus was labelled so as to classify both system and user moves into appropriate categories of communicative acts. These categories were a revision of those proposed in SWBDL-DAMSL (Switch Board Corpus - Dialogue Act Markup in Several Layers) [20]. The 86 moves the Wizard could employ (system moves) were organized into 8 categories (Table 1) by considering on one hand the DAMSL classification and on the other hand the frequencies with which they had been employed in the corpus.

Tag	Description
OPENING	initial self-introduction by the ECA
QUESTION	question about the user's eating habits or information interests
OFFER-GIVE-INFO	generic offer of help or specific information
PERSUASION-SUGGEST	persuasion attempt about dieting
ENCOURAGE	statement aimed at enhancing the user's motivation
ANSWER	provision of generic information after a user request
TALK-ABOUT-SELF	statement describing own abilities, role and skills
CLOSING	statement of dialogue conclusion

Table 1: Categories of Wizard moves

Similar criteria were applied to define the 11 subject move categories (Table 2):

Tag	Description
OPENING	initial self-introduction by the user
REQ-INFO	information request
FOLLOW-UP-MORE-DETAILS	further information or justification request
OBJECTION	objection about an ECA's assertion or suggestion
SOLICITATION	request of clarification or generic request of attention
STAT-ABOUT-SELF	generic assertion or statement about own diet, beliefs, desires and behaviours
STAT-PREFERENCES	assertion about food liking or disliking
GENERIC-ANSWER	provision of generic information after an ECA's question or statement
AGREE	acknowledgment or appreciation of the ECA's advice
KIND-ATTITUDE-SYSTEM	statement displaying kind attitude towards the system, in the form of joke, polite sentence, comment or question about the system
CLOSING	statement of dialogue conclusion

Table 2: Categories of subject moves

3.3 Dialogue representation

Formally [18,19], an HMM can be defined as a tuple: < S, W, π, A, B>, where

- $S = \{s_1, ... s_n\}$ is the set of states in the model,
- W is the set of observations or output symbols,

- π are a-priori-likelihoods, that is the initial state distributions: $\pi = \{\pi_i\}$, $i \in S$;
- $A = \{a_{ij}\}$, $i, j \in S$, is a matrix describing the state transition probability distribution: $a_{ij} = P(X_{t+1} = s_j \mid X_t = s_i)$;
- $B = \{b_{ijk}\}$, $i, j \in S$, $w_k \in W$, is a matrix describing the observation symbol probability distribution: $b_{ijk} = P(O_t = w_k \mid X_t = s_i, X_{t+1} = s_j)$.

In our models:

- *States* represent aggregates of system or user moves, each with a probability to occur in that phase of the dialogue.
- *Transitions* represent dialogue sequences: ideally, from a system move to a user move type and vice versa, each with a probability to occur (although in principle, user-user move or system-system move transitions may occur).

HMMs are learnt from a corpus of dialogues by representing the input as a sequence of coded dialogue moves. For example, the following dialogue:

T(S,1)= Hi, my name is Valentina. I'm here to suggest you how to improve your diet. Do you like eating?
T(U,1)=Yes
T(S,2)= What did you eat at breakfast?
T(U,2)=Coffee and nothing else.
T(S,3)=Do you frequently eat this way?
T(U,3)=Yes
T(S,4)= Are you attracted by sweets?
T(U,4)= Not much. I don't eat much of them.
T(S,5)= Do you believe your diet is correct or would you like changing your eating habits?
T(U,5)= I don't believe it's correct: I tend to jump lunch, for instance.

is coded as follows: (OPENING, GENERIC-ANSWER, QUESTION, STAT-ABOUT-SELF, QUESTION, GENERIC-ANSWER , QUESTION, STAT-ABOUT-PREFERENCES, QUESTION, STAT-ABOUT-SELF).

3.4 Setting the number of states in the model

In learning HMM structures from a corpus of data, one has first of all to establish the number of states with which to represent dialogue patterns [2]. This is a function of at least two factors: (i) the *level of detail* with which a dialogue needs to be represented, and (ii) the *reproducibility of the HMM* learning process, which may be represented in terms of *robustness of learned* structures.

The Baum-Welch algorithm adjusts the model parameters $\mu = (A, B, \pi)$ to maximize the likelihood of the input observations, that is $P(O|\mu)$. The algorithm starts with random parameters, and, at each iteration, adjusts them according to the maximization function. This algorithm is, in fact, very similar to the Expectation-Maximization (EM) algorithm and, like this, is *greedy*. That is, it does not explore the whole solution space (not exhaustive search) and can find a local maximum point instead than a global one: this is the reason why, as we will see later on, repeated applications of the algorithm to the same dataset may produce different results.

To establish the number of states with which to represent our models, we tested three alternatives: 6, 8 and 10 states. For each condition, we repeated q times the learning experiment with the same corpus of data, in identical conditions. Robustness of learning was evaluated from the following indices:

- *loglik values*: groups of HMMs with similar logliks were considered to be similar;
- *average differences* between the $q*(q-1)/2$ (HMM^i, HMM^j) pairs of HMMs, in the transition probabilities T^i, T^j and the observation probabilities E^i, E^j:

$$D(T^i,T^j) = \Sigma_{h,\,k\,=\,1,\ldots,n} \; | T^i_{h,k} - T^j_{h,k} | / n^2$$

$$D(E^i,E^j) = \Sigma_{h=\,1,\ldots,n;\; k\,=\,1,\ldots,m} \; | E^i_{h,k} - E^j_{h,k} | / n * m$$

where n denotes the number of states (6, 8 or 10) and m the number of communicative acts used in coding (19). Differences are computed after aligning the states in the two models. Our average difference is similar to the Euclidean measure of distance between pairs of HMMs that was proposed in [21]. It differs from the probabilistic measure proposed in [22], in which the distance between models is measured in terms of differences in observed sequences with increasing time.

States	Average (and variance) of D(Ti,Tj)	Average (and variance) of D(Ei,Ej)
6	.18 (.005)	.043 (.0003)
8	.041 (.001)	.014 (.00009)
10	.055 (.0005)	.022 (.00011)

Table 3: Comparison of HMMs with different n. of states

Table 3 shows the results of this analysis on the corpus of 30 text-based dialogues, after repeating q=10 times the learning experiment. HMMs with 8 states are the most 'robust' as they show the minimum average difference in transitions and observations. At the same time, they are quite easy to interpret, as they do not include 'spurious' states assembling system and user moves at the same time. HMMs with 10 states provide a more detailed dialogue description but are much less robust. We therefore selected the 8-state option for our following experiments. Robustness of learning measures how reproducible the learning process is. For instance, in the ten repetitions of the experiment on text-based dialogues, 7 over 10 HMMs had very similar loglikelihood values and low average differences of transitions and observations; they could therefore be considered as 'similar' and 'good' models. This result may be interpreted by saying that the probability of finding a 'good' model by repeating the learning experiment is .70. Although this is not a high value, we could notice, in our subsequent experiments, that other categories of dialogues were still less robust. In general, with the increasing complexity of dialogues in a category (as in the case of those collected with speech-based interaction), model learning becomes less robust. As we will see, this lack of robustness affects considerably the quality of our results.

4 DESCRIPTIVE POWER OF THE MODEL

To test the descriptive power of HMMs learnt from our corpus of data, we considered a user feature about which we had acquired some knowledge in a previous study. In that study we analysed the relationship between user background (in computer science -CS- or in humanities -HUM-) and 'social attitude' towards the ECA [4]. As we mentioned in the Introduction, the method of analysis employed was, in that context, language and speech processing of individual moves. Results of that study proved that users with a background in humanities displayed a different behavior in dialogues: they tended to have a 'warmer' social attitude towards the ECA, that was displayed with a familiar language, by addressing personal questions to the agent, by talking about self etc.

[2] We used HMM Matlab Toolbox: http://www.ai.mit.edu/~murphyk/Software/HMM/hmm.html

Figures 1 (a) and (b) show, respectively, the best 8-states HMMs for CS and HUM subjects. States Si correspond to aggregates of system moves: in 1a, an OPENING is associated with S1 with probability 1; a QUESTION to S2 with probability .88; a PERSUASION (p=.57), an OFFER-INFO (p=.20) or an ENCOURAGE (p=.14) with S3, etc. Interpretation of states Uj, to which user moves are associated, can be observed from the figure. Transitions between states in models 1a and 1b have a common core pattern, although with different probabilities: the path (S1, U1, S2, U2, S3, U3), the way back (U3, S2) and the direct link (S1, U3). Other transitions differ. Dissimilarities can be found also in the probability distributions of communicative acts associated with the phases of dialogue opening (S1, U1), question answering (S2, U2), system persuasion (S3, U3) and of a warm phase (S4,U4), in which the user displays a kind attitude towards the system in various forms. The following are the main differences between the models, in these phases:

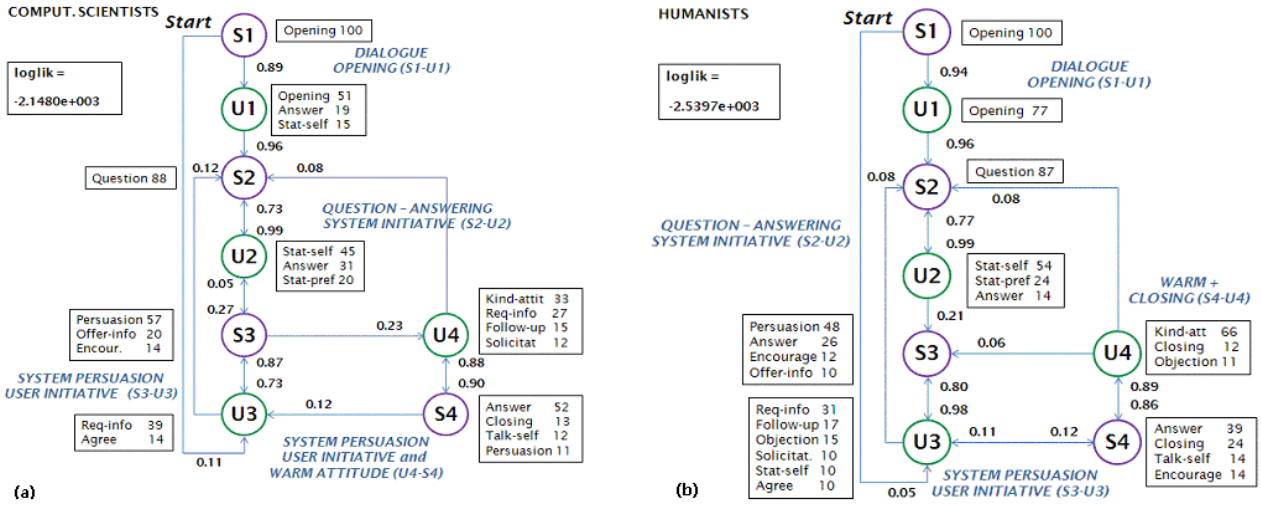


Fig 1: HMMs for subjects with a background in computer science (a) and humanities (b)

- *Question answering* (S2, U2)**:** the only difference, in this case, is that HUM subjects tend to be more specific and eloquent than CS ones, by producing more "statements about self", "statements about preferences" and less "generic answers".
- *Persuasion* (S3, U3)**:** in CS models, users may respond to persuasion attempts with information requests or declarations of consensus. They may enter, as well, in the warm phase (S3, U4 link). In the HUM model, after a persuasion attempt by S, U may stay in the persuasion phase (U3) by making further information requests, objections, solicitations or statements about self. In both models, question answering and persuasion may alternate (link U3, S2) in a more varied dialogue.
- *Warm phase* (S4, U4): although this phase exists in both models, communicative acts employed by U, once again differ: the 'objection' move of HUM is substituted by a 'solicitation' in CS's model. In this model, the state S4 may contain persuasion moves as well: hence, the whole phase may be called "system persuasion -user initiative- and warm attitude". The likelihood to produce a 'kind attitude' move (in U4) is, in the HUM model, twice than in the CS model.

These differences in the subjects behavior confirm the findings of our previous studies, by describing them in terms of dialogue structure rather than of individual moves' content.

5 MODEL TESTING

Before applying HMMs to reason about the differences among dialogues produced by different types of users, we needed to test the ability of learnt HMMs to classify correctly new cases.

5.1 Method

We describe the method to classify new cases in two classes: this can be easily extended to the case of p>2 classes. Given a corpus of dialogues classified in two sub-corpora (S-C1 and S-C2), of dimensions n and m, collected from two different categories of users (C1 and C2) according to a given target feature - for example interaction mode (text-based vs speech-based) or background (CS vs HUM) -:

a. Train two HMMs, respectively from n-1 cases from S-C1 and the m cases in S-C2, and call them HMM1 and HMM2.

b. Test the n-th case on HMM1 and HMM2 with the forward-backward algorithm, to compute the loglikelihoods:

loglik1 = log P(n-th case | HMM1)
loglik2 = log P(n-th case | HMM2)

c. Select the maximum between loglik1 and loglik2 to attribute the n-th case to C1 or C2.

d. Check the correctness of the classification (matching with an 'external reference').

e. Repeat from a. to d. by varying the training set according to the leave-one-case out approach.

f. Repeat from a. to e. (m-1 cases in S-C2 and n cases in S-C1).

g. Compute the recognition accuracy for HMM1 and HMM2 in terms of confusion matrix, precision and recall.

5.2. Recognizing the user background

By applying HMM analysis to users' background, we aimed at verifying whether any difference between the two typologies of users could be found, as well, in the dialogue patterns. Information about this feature was collected with a questionnaire

during the WoZ study. Here, it was taken as the 'external reference' in the testing phase. Table 4 shows the results of this analysis: a CS dialogue is recognized correctly in 77 % of cases, while a HUM dialogue is recognized correctly in 57% of cases. HUM dialogues tend to be confused with CS ones more frequently than the inverse.

	CS HMMs	HUM HMMs	Total
CS dialogues	**(23) .77**	(7) .23	30
HUM dialogues	(13) .43	**(17) .57**	30
Total	36	24	
Recall	.77	.57	
Precision	.64	.71	

Table 4: Confusion matrix for CS vs HUM dialogues

5.3. Stepwise recognition

Adaptation of the dialogue to the user goals and preferences requires recognizing them dynamically during the dialogue. In the testing method we described in the previous Section, on the contrary, the whole dialogue was submitted to testing. We therefore wanted to check the ability of our classification procedure to apply a stepwise recognition method on subsections of dialogue of increasing length. Given an average number n of dialogue pairs (system-user moves) considered in the training phase, we defined a 'monitoring' interval of t moves and applied the recognition method to parts of the dialogue of increasing length i*t, with i= 1, ... n/t. After every step, we checked whether the part of the dialogue examined was recognized correctly. What we expected from this analysis was to find an increase of recognition accuracy with the increasing monitoring time.

To check the validity of the method, once again we applied stepwise recognition to the distinction between CS and HUM dialogues, with a monitoring interval of t=5 pairs. The results we got were less positive than our expectation. In Table 5, results of stepwise recognition are classified in five categories, according to the consequences they entail on the quality of adaptation. The worst cases are that of 'steadily wrong' (22%) or 'up and down recognition' (15%): here, adaptation criteria would be always wrong, or would be changed several times during the dialogue, by producing an unclear and not effective advice giving strategy.

	CS	HUM	Total
Steadily correct recognition	14 (47%)	9 (30%)	23 (38%)
Initially wrong, then correct	2 (7%)	8 (27%)	10 (17%)
Steadily wrong recognition	7 (23%)	6 (20%)	13 (22%)
Initially correct, then wrong	1 (3%)	4 (13%)	5 (8 %)
Up and down recognition	6 (20%)	3 (10%)	9 (15 %)

Table 5: Stepwise recognition for CS vs HUM dialogues

In the cases in the 'initially correct, then wrong' category (8%) adaptation would become incorrect towards the end of the dialogue while, in the cases in the 'initially wrong, then correct' category (17%), the system might adopt correct adaptation criteria only towards the end of the dialogue. The only situation enabling a proper adaptation is that of 'steadily correct recognition' (38%). Notice that in HUM dialogues (which, as we said, are longer and more complex) it takes more time to the system to recognize properly the user category. We attributed this poor stepwise recognition ability to the limited robustness of both the learning and the testing procedure, due to the reduced dimension of our corpus. To test this hypothesis, we repeated the stepwise testing on the same dialogue with the same learned HMM, and applied 'majority agreement' as a criterion for recognizing the background at every step. We did this little check with 6 dialogues and 5 repeated tests, but found some improvement of results only in some of them .

6 RECOGNIZING ENGAGEMENT

In this section we present a possible application of the method described in Sections 3 to 5. In particular, we aim at testing whether HMMs can be employed to represent differences in the dialogue pattern of users which show different goals and levels of involvement in the advice-giving task.

In advice-giving dialogues two tasks are integrated: provision of specific information about aspects of the behavior that make it more or less 'correct', and persuasion to abandon a problem behavior. In a category of users, we found the typical attitude that Walton [23] calls of *examination dialogues*, in which *'one party questions another party, sometimes critically or even antagonistically, to try to find out what that party knows about something'*. Examination dialogues are shown to have two goals: the extraction of information and the testing of the reliability of this information: this testing goal may be carried out with critical argumentation used to judge whether the information elicited is reliable. We found this behaviour in some of our dialogues: we therefore named "information-seeking" (IS) the users asking several questions, either for requesting information or challenging the application, sometimes even right after the system's self introduction. In another category (AG), users seem to be more involved in the persuasion goal of advice-giving: they show a more cooperative attitude toward the system, by providing extra-information to the agent so as to build a shared ground of knowledge about their habits, desires, beliefs etc. Also, they react to the agent's suggestions and/or attempts of persuasion by providing a 'constructive' feedback in terms of objections, comments (either positive or negative) and follow-up questions. Finally, we have a third category of 'not engaged' (N) users who don't show any interest in any of the two mentioned tasks (information seeking or advice-giving); they rather give a passive and barely reactive contribution to the interaction, by mainly answering the system's questions, very often with general answers (eg. 'yes' or 'no'); their dialogues are usually shorter than the others and tend to be driven by the system (that sometimes seems to struggle to protract the interaction).

Distinguishing among the three levels of engagement is relevant for adaptation: IS users might be either helped in their information seeking goal or leaded by the system to get involved also in the advice giving task, by carefully choosing (or revising) the persuasion strategy [25]; AG users might perceive an increased satisfaction about the interaction if the agent is believable in playing the role of artificial therapist; N users represent a real challenge for the system: their attitude might be due to a lack of interest in the domain or to their being in the 'precontemplation stage' [1].

6.1 Corpus annotation

Two independent raters were asked to annotate the overall attitude of every user by using the labels N, IS and AG. The percentage of agreement (.93) and the Kappa value (.90) indicate a strong interrater agreement [24]. To classify the corpus by giving a final label to every dialogue, we asked the two raters to discuss about the cases for which they had given different

annotations. The resulting distribution of the corpus is skewed, which is an undesirable circumstance when the available set of data is not particularly wide: we will show how robustness of learning decreases if compared with the previous classification attempts, were the corpus was equally distributed.

6.2 HMM training and robustness

To evaluate how suitable are HMMs to model user engagement we repeated the robustness analysis described in 3.4, with 8-state models. Results (in Tab. 6) show that the robustness of the method is not the same for the three classes.

Subjects	Distribution (%)	Average (and variance) of D(Ti,Tj)	Average (and variance) of D(Ei,Ej)
N	.44	.05 (.003)	.03 (.0007)
IS	.28	.011 (.002)	.03 (.0003)
AG	.28	.15 (.002)	.06 (.0004)

Table 6: Robustness evaluation

In spite of the high interrater-agreement and of the good descriptive power of the HMMs, the analysis shows a lack of robustness, especially for AG models, due to the unequal distribution of the dataset. The restricted amount of available data is a major cause of this phenomenon, especially when the behaviour of users is extremely variable, as observed for the AG category. In fact, in the 10 repetitions of the learning experiment on these 28 dialogues, no groups of similar HMMs were found, and we got the highest average differences in transitions and observations. On the contrary, in the N learning experiment (44 cases) we had 6 similar models over 10 HMMs (similar likelihood values and low average differences in transitions and observations). Similarly, for the IS category, whose dialogues show a more regular structure than the AG ones, we found 7 over 10 similar models, even if the number of cases was the same as for the AG. This confirms our findings about the CS vs HUM classification experiments, by adding an extra insight due to the unequal distribution of the corpus.

	N	IS	AG	Total
N	(13) **.76**	(4) .24	(0) 0	17
IS	(1) .04	(22) **.85**	(3) .12	26
AG	(2) .12	(4) .24	(11) **.65**	17
Total	16	30	14	
Recall	.76	.85	.65	
Precision	.81	.73	.69	

Table 7: Confusion matrix for N, IS and AG

The results for the leave one case out validation (tab. 7), performed on the three classes of engagement, confirm, once again, that the higher is the variety of behaviour among users in a given class, the worse is the recognition performance (lowest value for both precision and recall, for AG users).

6.3 Descriptive power: classification of engagement

In spite of the issues highlighted in par. 6.2, HMMs still seem to meet our expectation about their ability of distinguishing among the three levels of engagement we wish to recognize. Figures 2 (a), (b) and (c) show, respectively, the best 8-states HMMs for N, IS and AG subjects.

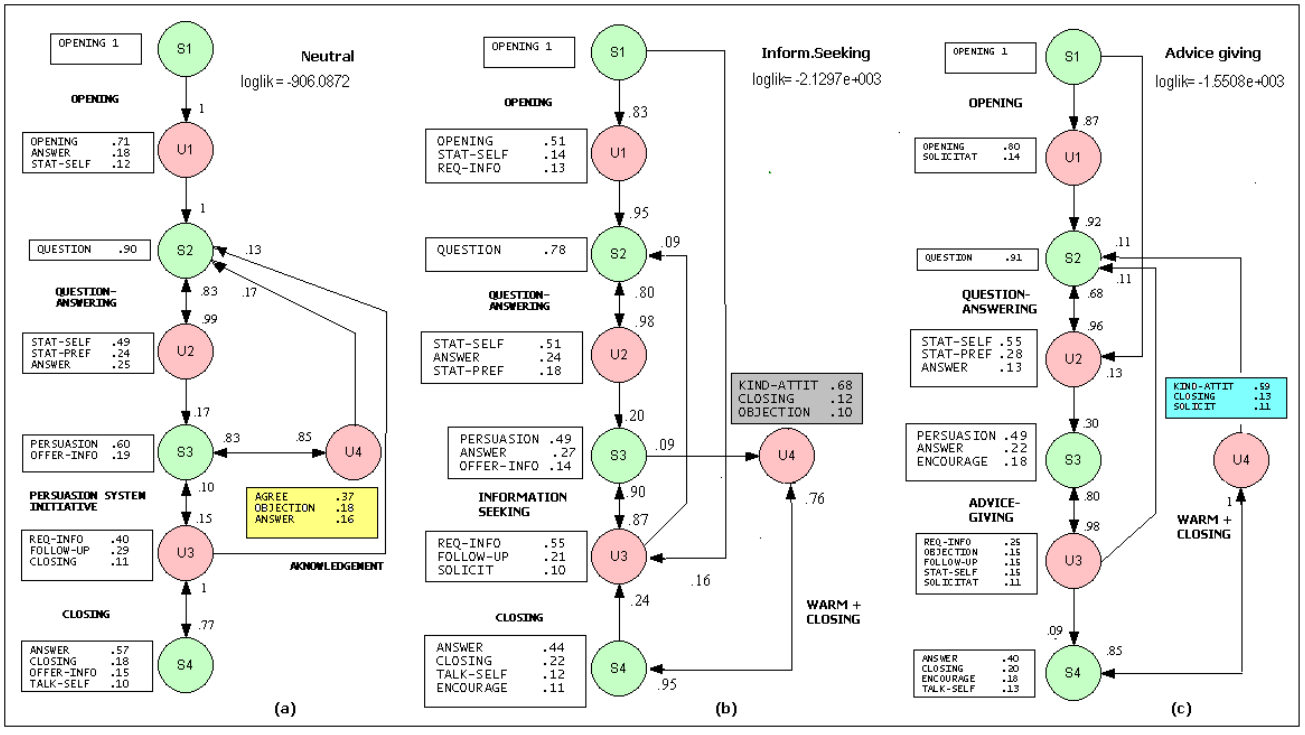


Fig 2: HMM for neutral (a), information seeking (b) and advice-giving (c) dialogues

The following are the main differences between these models:

- *Dialogue opening* (S1, U1): in the N model, U always reacts with an opening move to the self presentation of S while, in IS and AG models, there is some probability of directly entering the persuasion phase;
- *Question answering* (S2, U2): as hypothesized, IS and AG subjects tend to be more specific and eloquent than N ones, by producing more "statements about self", "statements about their preferences" and less "generic answers";
- *Persuasion* (S3, U3)**:** we named the persuasion phases according to the differences in the observable user categories of moves. In the N models, the users may respond to persuasion attempts with information requests, follow up questions and even with a closing move: so we named this phase '*persuasion* with *system initiative*'. IS users have the highest probability of performing a request of information and do not provide any kind of personal information ('*information seeking'* phase). In AG models, users are involved in an *advice-giving* phase: the probability of information requests is lower and the variety of reactions to system suggestions is wider, according to the users' goal of either enhancing the construction of a shared ground of knowledge about healthy eating, or giving a positive feedback to the ECA. The likelihood of entering the persuasion phase, core of the advice-giving process, after question answering, is higher in these models;
- *Warm phase* (S4, U4): in IS and AG models there is a high likelihood of observing a kind attitude, while N users mainly provide a feedback (either positive or negative) to the ECA's suggestion (*acknowledgement*). This can be seen as a cue of higher engagement in the interaction for IS and AG subjects. Also, contrary to IS and AG ones, in N models we notice that the probability of remaining in the persuasion phase (S3,U3) is lower than the probability of switching to the acknowledgement one; this could be seen as a proof of low level of engagement, probably due to a lack of interest in the interaction or in the domain itself.

The comments we just provided depict HMMs as a powerful formalism for differentiating among various categories of users. The lack of robustness of the method, though, suggests us to be cautious: we decided to describe the three best models (with best loglikelihoods) but the limited amount of training data and the huge variety in users' behavior, especially for the AG category, affect the reproducibility of the learning experiment and cause poor recognition performance. Our findings about the descriptive power of HMMs should therefore be validated by further investigation on larger corpora.

7 RELATED WORK

HMMs find their more natural and more frequent application domain in parsing and speech recognition problems. Their application to dialogue pattern description and recognition is more recent. Levin et al [26] were among the first authors to use this formalism in dialogue modeling. In their proposal, system moves are represented in states while user moves are associated with arcs. The costs of different strategies are measured in terms of distance to the achievement of the application goal (information collection in an air travel information system), and the optimal strategy is the minimal cost one. In [20], user moves are associated with states in a HMM-based dialogue structure, transitions represent the likely sequencing of user moves, and evidence about dialogue acts are their lexical and prosodic manifestations. Twitchell et al [27] proposed to use HMMs in classifying conversations, with no specific application reported.

The work with which our study has more in common is the analysis of collaborative distance learning dialogues in [28]. The aim, in this case, was to dynamically recognize when and why students have trouble in learning the new concepts they share with each other. To this purpose, specific turn sequences were identified and extracted manually from dialogue logs, to be classified as 'knowledge sharing episodes' or 'breakdowns'. Aggregates of student's acts were associated with the five states of HMMs learnt from this corpus. The overall accuracy of recognizing effective vs ineffective interactions was 74 %.

8 CONCLUSIONS AND OPEN PROBLEMS

There are, in our view, some new aspects in our study, from both the methodological and the result points of view. In previous studies, we combined linguistic and acoustic features of the user move to dynamically build an image of his/her social attitude towards the agent. In this article, we investigated whether and how it is possible to model the impact of user's attitude on the overall dialogue pattern. In particular, we aim at: (i) studying the suitability of the HMMs as a formalism to represent differences in the dialogue model among different categories of users, as differentiated by either stable user features (such as background) or long-lasting affective states (such as attitudes); (ii) highlighting the importance of evaluating the robustness of the trained HMMs before using them, to avoid the risk of building unreproducible models; (iii) proposing the usage of robustness metrics for assessing the role played by the size of the dataset used and how this affects the performance of recognition tasks.

We first tested the descriptive power of HMMs with a pilot experiment: two models were trained from our corpus of data to classify users on the basis of their background, a stable and objective user feature whose role and impact on interaction dynamics have been widely investigated in our previous research. We then extended the method to the recognition of three classes of users, who showed different levels of engagement in the advice-giving task.

Results present HMMs as a suitable and powerful formalism for representing differences in the structure of the interaction of subjects belonging to different categories (Sections 4 and 6). Still, we have to be cautious: all the models described in this paper are those who showed the best training likelihood. By using *ad hoc* metrics, we discovered a lack of robustness of the method that reduces the reproducibility of the learning experiment and lowers the recognition performance. We assumed that this is mainly due to the dimension of our corpus. Also, the complexity of dialogues and the huge variety in the behaviour of users of certain classes (e.g., people with background in humanities and AG users, especially when 'naturally' interacting via speech) play an important role in this sense. Especially when combined with a low cardinality of the class, these two factors are, in our opinion, the major causes of the reduction in the robustness of training. Future developments will involve the usage of a larger corpus of data, to achieve a final validation of the method.

Another open problem is how to use these models to dynamically recognize user attitudes (such as engagement), for

long-term adaptation of the agent's behaviour. In Section 5.3, we tested a possible stepwise approach to simulate the usage of HMMs during the interaction: the idea is to define/revise the ECA's dialogue strategy according to the predicted overall level of engagement, to prevent involved users to be unsatisfied or to try to enhance involvement of those users who show a lack of interest in the advice-giving task. Results of the stepwise simulation are not encouraging, again probably because of the limited amount of data we used for training. Poor recognition performances are obtained especially when dialogues belonging to the same class have particularly complex dynamics and there is high variability among them (e.g. HUM users). The results of the experiment show that a proper adaptation could be possible for only 38% of cases. In all the other cases, results of the recognition would lead the system to an unclear and uneffective persuasion strategy: whether this adaptation approach would be successful and would produce a significant increase of the level of engagement in the users is still not clear and should be further investigated. Researchers working on behavioral analysis [29] proposes a two layered approach combining Bayesian Networks with HMM models. This method enables integrating the HMM's ability of modeling sequences of states with the BN's ability of pre-processing multiple lower level input. In our case, HMMs learnt from dialogues about a particular category of users would be enriched by attaching to hidden states describing user moves a BN to process evidence resulting from linguistic analysis of this move. Our expectation is that the combination of the two probability distributions of HMM and bayesian models will improve the performance of the attitude recognition process. This approach would allow us to realize adaptation at two levels: the overall user attitude (HMM overall prediction) and the specific signs in dialogue moves (BN prediction).

REFERENCES

[1] J. Prochaska, C. Di Clemente and H. Norcross. In search of how people change: applications to addictive behavior. *Americal Psychologist*, 47, 1102-1114, 1992.

[2] R. E. Petty and J.T. Cacioppo. The Elaboration Likelihood Model of Persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology*. New York: Academic Press, 19, pp. 123-205, (1986)

[3] F. de Rosis, N. Novielli, V. Carofiglio and B. De Carolis. User modeling and adaptation in health promotion dialogs with an animated character. *Journal of Biomedical Informatics,* 39 (5), 514-531 (2006)

[4] F. de Rosis, A. Batliner, N. Novielli and S. Steidl. 'You are soo cool Valentina!' Recognizing social attitude in speech-based dialogues with an ECA. In: *Procs of ACII 2007*, Lisbon (2007)

[5] T. Bickmore and J. Cassell. Social Dialogue with Embodied Conversational Agents, in: J. van Kuppevelt, L. Dybkjaer, & N. Bernsen (Eds.), *Advances in Natural, Multimodal Dialogue Systems*. New York: Kluwer Academic (2005)

[6] P.A. Andersen and L.K. Guerrero. Handbook of Communication and Emotions. Research, theory, applications and contexts. Academic Press, New York, (1998)

[7] Polhemus, L., Shih, L-F and Swan, K., 2001. Virtual interactivity: the representation of social presence in an on line discussion. *Annual Meeting of the American Educational Research Association*.

[8] K. Swan. Immediacy, social presence and asynchronous discussion, in: J. Bourne and J. C. Moore (Eds.): *Elements of quality online education*. Vol. 3, Nedham, MA. Sloan Center For Online Education (2002)

[9] J.N. Bailenson, , K.R. Swinth,, C.L Hoyt, S. Persky, A. Dimov, and J. Blascovich. The independent and interactive effects of embodied agents appearance and behavior on self-report, cognitive and behavioral markers of copresence in Immersive Virtual Environments. *PRESENCE*. 14, 4, 379-393 (2005)

[10] C. Sidner and C. Lee. An architecture for engagement in collaborative conversations between a robot and a human. *MERL Technical Report*, TR2003-12 (2003)

[11] C. Yu, P. M. Aoki and A. Woodruff. Detecting User Engagement in everyday conversations. In *Procs of International Conference on Spoken Language Processing.* 1329-1332 (2004)

[12] A. Pentland. Socially Aware Computation and Communication. *Computer*, 38, 3, 33-40 (2005)

[13] M. G. Core, J. D. Moore, and C. Zinn, The Role of Initiative in Tutorial Dialogue, in: *Procs of 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, April (2003)

[14] F. Shah. Recognizing and Responding to Student Plans in an Intelligent Tutoring System: CIRCSIM-Tutor. Ph.D. thesis, Illinois Institute of Technology (1997)

[15] P. Linell, L. Gustavsson, and P. Juvonen. Interactional dominance in dyadic communication: a presentation of initiative-response analysis. *Linguistics*, 26:415–442 (1988)

[16] A. Woodruff, P. M. Aoki. Conversation analysis and the user experience. *Digital Creativity*, 15 (4): 232-238 (2004)

[17] S. Whittaker. Theories and Methods in Mediated Communication, in *Handbook of Discourse Processes*, LEA, Mahwah, NJ (2003)

[18] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. In: *Procs of the IEEE*, 77,2, 257-286 (1989)

[19] E. Charniak, *Statistical language learning*. The MIT Press (1993)

[20] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26, 3 (2000)

[21] S. E. Levinson, L. R. Rabiner and M. M. Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *B:S:T:J.*, 62, 4, 1035-1074 (1983)

[22] B-H Juang and L R Rabiner. A probabilistic distance measure for Hidden Markov Models. *AT&T Technical Journal*. 64, 2, 391-408, (1985)

[23] D. Walton: Examination dialogue: an argumentation framework for critically questioning an expert opinion. *Journal of Pragmatics,* 38,745-777 (2006)

[24] J. Carletta: Assessing agreement on classification tasks. The Kappa statistics. *Computational Linguistics*, 22 (1996)

[25] I. Mazzotta, F. De Rosis and V. Carofiglio. PORTIA: A user-adapted persuasion system in the healthy eating domain. *IEEE Intelligent Systems*, in press.

[26] E. Levin, R. Pieraccini and W Eckert. Using Markov decision process for learning dialogue strategies. Proceedings of the *IEEE International Conference on Acoustic, speech and signal processing,* 1, 201-204 (1998)

[27] D. P. Twitchell, M. Adkins, J. F. Nunamaker and J. K. Burgoon. Using speech act theory to model conversations for automated classification and retrieval. In *Procs of the 9th International Working Conference on the Language-Action perspective on Communication Modelling* (2004)

[28] A. Soller. Computational modeling and analysis of knowledge sharing in collaborative distance learning. *UMUAI*, 14, 4, 351-381, (2004)

[29] N. Carter, D. Young and J. Ferryman. A Combined Bayesian Markovian Approach for Behaviour Recognition. In Proceedings of the 18th International Conference on Pattern Recognition, (2006)

Towards Affective Natural Language Generation: Empirical Investigations

Ielka van der Sluis and **Chris Mellish** [1]

Abstract. This paper reports on attempts to measure the differing effects on readers' emotions of positively and negatively "slanted" texts with the same basic message. The methods of "slanting" the texts are methods that could be used automatically by a Natural Language Generation (NLG) system. A pilot study and a main experiment are described which use emotion self-reporting methods from Psychology.

Although the main experiment was formulated with the benefit of knowledge obtained from the pilot experiment and a text validation study, nevertheless it was unable to show clear, statistically significant differences between the effects of the different texts. We discuss a number of possible reasons for this, including the possible lack of involvement of the participants, biases in the self-reporting and deficiencies of self-reporting as a way of measuring subtle emotional effects.

1 Introduction: Affective NLG

Much previous research in Natural Language Generation (NLG) has assumed that the purpose of generated texts is simply to communicate factual information to the user [9]. On the other hand, in the real world, texts vary enormously in their *communicative purpose*, for instance they may aim to persuade, amuse, motivate or console. In general, even when a text communicates information, it usually does so in order to affect the reader at a deeper level, and this has an impact on *how* the information should be communicated (the central task of NLG). As a consequence of this, De Rosis and Grasso have defined the notion of "affective NLG" as "NLG that relates to, arises from or deliberately influences emotions or other non-strictly rational aspects of the Hearer" [11]. In practice, however, work on affective NLG mostly emphasises the depiction of emotional states/personalities [8], rather than ways in which texts can induce different effects on readers. To build systems which, from a model of the reader, can intelligently select linguistic forms in order to achieve a particular deep effect, we need a scientific understanding of how the attributes of an individual reader (and the reading process for them) influence the effect that particular linguistic choices have. But in order to evaluate such understanding, we need to have ways of measuring the effects that texts have, beyond simply testing for the recall of facts. The work described in this paper is an initial attempt to find out whether it is possible to measure emotions evoked in the reader of a text. In particular, can we detect the difference between different wordings of a text (that an NLG system might produce) in terms of the emotions evoked in the reader? Although there has been some work on task-based evaluation in NLG cf. STOP [10] and SKILLSUM (Williams and Reiter, In Press), to our knowledge, measurement of emotions invoked in readers is not something that has been investigated before.

This paper is organised as follows: Section 2 introduces our approach to linguistic choice, the composition of affective text and a text validation study. Section 3 discusses potential psychological methods to measure the emotional effect of text and Section 4 a pilot study that was conducted to try these out on text readers. Finally Section 5 brings all together in a full Study in which the texts resulting from our text validation experiments and the most promising affect measurement methods are used to measure the affect of text invoked in readers. The paper closes with a discussion of findings and future work.

2 Linguistic Choice

We decided that a safe way to start would be to aim for large effects in primitive emotions, (e.g. positive versus negative emotions, such as sadness, joy, disappointment, surprise, anger), as opposed to aspects of contextual impact (e.g. trust, persuasion, advice, reassurance). Therefore, although there are many linguistic choices that an NLG system might explicitly control, we focus here on alternatives that relate to simple goals of giving a text a positive or negative "slant". Very often the message to be conveyed by an NLG system has "positive" and "negative" aspects, where "positive" information conjures up scenarios that are pleasant and acceptable to the reader, makes them feel happy and cooperative etc. and "negative" information conjures up unpleasant or threatening situations and so makes them feel more unhappy, confused etc. An NLG system could make itself popular by only mentioning the positive information, but then it could leave itself open to later criticism (or litigation) if by doing so it clearly misrepresents the true situation. For instance, [2] discuss generating instructions on how to take medication which have to both address positive aspects ('this will make you feel better if you do the following') and also negative ones (this may produce side-effects, which i have to tell you about by law). Although it may be inappropriate grossly to misrepresent the provided message, there may be more subtle ways to "colour" or "slant" the presentation of the message in order to emphasise either the positive or the negative aspects.

We assume that the message to be conveyed is a simple set of propositions , each classified in an application-dependent way as having positive, negative or neutral *polarity* in the context of the message.[2] This classification could, for instance, be derived from the information that a planning system could have about which propositions support which goals (e.g. to stay healthy one needs to eat

[1] Computing Science, University of Aberdeen, email: {i.v.d.sluis,c.mellish}@abdn.ac.uk

[2] Note that this polarity is not the same as the one used to describe, for instance, "negative polarity items" in Linguistics

healthy food). We also assume that a possible phrasing for a proposition has a *magnitude*, which indicates the degree of impact it has. This is independent of the polarity. We will not need to actually measure magnitudes, but when we make claims about when one wording of a proposition has a smaller magnitude than another we indicate this with $<$. For instance, we would claim that usually:

"*a few rats died*" $<$ "*many rats died*"

("a few rats died" has less impact than "many rats died", whether or not rats dying is considered a good thing or not). In general, an NLG system can manipulate the magnitude of wordings of the propositions it expresses, to indicate its own (subjective) view of their importance. In order to slant a text positively, it can express positive polarity propositions in ways that have high magnitudes and negative polarity propositions in ways that have low magnitudes. The opposite applies for negative slanting. Thus, for instance, in an application where it is bad for rats to die, expressing a given proposition by "a few rats died" would be giving more of a positive slant, whereas saying "many rats died" would be slanting it more negatively.

Whenever one words a proposition in different ways, it can be claimed that a (perhaps subtle) change of meaning is involved. In an example like this, therefore, there is a question about whether in fact the two different wordings actually correspond to different *messages*, rather than different *wordings* that might be chosen by an NLG system. In this paper, we assume that the choice between these two possibilities would likely be implemented somewhere late in the "pipeline", and so we think of it as being a choice of form, rather than content. This interpretation is supported by our text validation experiments described below.

2.1 Test Texts

We started by composing two messages, a negative and a positive message, within a topic of general interest: food and health issues. The negative message tells the reader that a cancer-causing colouring substance is found in some foods available in the supermarkets. The positive message tells the reader that foods that contain Scottish water contain a mineral which helps to fight and to prevent cancer. The texts are set up in a similar way in that they both contain three paragraphs that address comparable aspects of the two topics. The first paragraph of both texts states that there is a substance found in consumer products that has an effect on people's health and it addresses the way in which this fact is handled by the relevant authorities. The second paragraph of the text extends on the products that contain the substance and the third paragraph explains in what way the substance can affect people's health.

To study the effects of different wordings, for each text a positive and a negative version was produced by slanting propositions in either a positive or a negative way. The slanting was done so that the positive and negative versions of the messages were still reporting on the same event. This resulted in four texts in total, two texts with a negative message one positively and one negatively phrased (NP and NN), and two texts with a positive message one positively and one negatively verbalised (PP and PN). For the negative message, the NP version is assumed to have less negative impact than the NN version. Likewise, the PN version of the positive message is assumed to have less positive impact than the PP version. To maximise the impact aimed for, various slanting techniques were used as often as possible without loss of believability (this was assessed by the intuition of the researchers). The positive and negative texts were slanted in parallel as far as possible, that is in both texts similar sentences were adapted so that they emphasised the positive or the negative aspects of the message. The linguistic variation used in the texts was algorithmically reproducible and can be coarsely classified as, on the one hand, created by the use of quantifiers, adjectives and adverbs to affect the conveyed magnitude of propositions, and, on the other hand, other techniques based on changing the polarity of the proposition (suggested by work on "framing" in Psychology [7];[15]) and changing the rhetorical structure to alter the prominence of propositions. Below, this variation is illustrated with examples taken from the two messages:

SLANTING EXAMPLES FOR THE NEGATIVE MESSAGE

Here it is assumed that recalls of products, risks of danger etc. involve negative polarity propositions. Therefore positive slanting will amongst other things choose low magnitude realisations for these.

Techniques involving adjectives and adverbs:

- "*A recall*" $<$ "*A large-scale recall*" of infected merchandise was triggered
- The substance is linked to "*a risk*" $<$ "*a significant risk*" of cancer

Techniques involving quantification:

- "*Some*" $<$"*Substantial amounts of*" contaminated food was withdrawn
- the substance was used in "*some*" $<$ "*many*" other products
- Since then "*more*" $<$ "*many more*" contaminated food products have been identified
- Sausages, tomato sauce and lentil soup are "*some*" $<$ "*only some*" $<$ of the affected items

Techniques involving a change in polarity

Proposition expressed with positive polarity:

- Tests on monkeys revealed that as many as "*40 percent*" of the animals infected with this substance "*did not develop any tumors*"

Proposition expressed with negative polarity:

- Tests on monkeys revealed that as many as "*60 percent*" of the animals infected with this substance "*developed tumors*".

Techniques manipulating rhetorical prominence

Positive slant:

- "So your health is at risk, but every possible thing is being done to tackle this problem"

Negative slant:

- "So although every possible thing is being done to tackle this problem, your health is at risk"

SLANTING EXAMPLES FOR THE POSITIVE MESSAGE

Here it is assumed that killing cancer, promoting Scottish water etc. involve positive polarity propositions. Therefore positive slanting will amongst other things choose high magnitude realisations for these.

Techniques involving adjectives and adverbs:

- "Scottish Water: "*A cancer-killer*" $<$ "*the Great Cancer-killer*"
- Neolite is a "*detoxifier*" $<$ "*powerful detoxifier*"

- Neolite is "*a possible*" < "*an excellent*" cancer preventative
- Neolite has proven to be "*effective*" < "*highly effective*" at destroying and preventing cancer cells

Techniques involving quantification:

- "*Cancer-killing Neolite*" < "*Substantial amounts of cancer-killing Neolite*" was found in Scottish drinking water
- A campaign for the use of Scottish water in "*consumer products*" < "*many more consumer products*"
- Waterwatch Scotland announced the start of an extensive campaign for the use of Scottish water in "*more*" < "*many more*" consumer products

Techniques involving a change in polarity

Proposition expressed with negative polarity:

- A study on people with mostly stage 4 cancer revealed that as many as "*40 percent*" of the patients that were given Neolite "*still had cancer*" at the end of the study.

Proposition expressed with positive polarity:

- A study on people with mostly stage 4 cancer revealed that as many as "*60 percent*" of the patients that were given Neolite "*were cancer free*" at the end of the study.

Techniques manipulating rhetorical prominence

Negative slant:

- "Neolite is certainly advantageous for your health, but it is not a guaranteed cure for, or defence against cancer"

Positive slant:

- "So Although Neolite is not a guaranteed cure for, or defence against cancer, it is certainly advantageous for your health"

2.2 Text validation

To check our intuitions on the emotional effects of the textual variation between the four texts described above, a text validation experiment was conducted in which 24 colleagues of the Computing Science Department at the University of Aberdeen participated. The participants were randomly assigned to one of two groups (i.e. P and N), group P was asked to validate 23 sentence pairs from the positive message (PN versus PP) and group N was asked to validate 17 sentence pairs from the negative message (NN versus NP). Both the N and the P group sentence pairs included four filler pairs. The participants in group P were asked which of the two sentences in each pair they thought most positive in the context of the message about the positive effects of Scottish water. The participants in group N were asked which of the two sentences in each pair they found most alarming in the context of the message about the contamination of food available for consumption. All participants were asked to indicate if they thought the sentences in each pair could be used to report on the same event. Below, the validations of the N and the P group are discussed separately.

N-Group Results indicated that in 89.75 % of the cases participants agreed with our intuitions about which one of the two sentences was most alarming. On average, per sentence pair 1.08 of the 12 participants judged the sentences differently than what we expected. In 7 of the 13 sentence pairs (17 minus four fillers) participants unanimously agreed with our intuitions. In the other four sentence pairs 1 to, maximally, 4 participants did not share our point of view. In the two cases in which four participants did not agree with or were unsure about the difference we expected, we adapted our texts. One of these cases was the pair:

> "*just 359*" infected products have been withdrawn < "*as many as 359*" infected products have been withdrawn "*already*"

We thought that the latter of the two would be more alarming (and correspond to negative slanting) because it is a bad thing if products have to be withdrawn (negative polarity). However, some participants felt that products being withdrawn was a good thing (positive polarity), because it meant that something was being done to tackle the problem, in which case the latter would be imposing a positive slant. As a consequence of the validation results, it was decided to 'neutralise' this sentence in both the NP and NN versions of the text to "359 infected products have been withdrawn". The second sentence pair on which four participants disagreed was:

> you would be able to notice symptoms resulting from the substance "*just after*" < "*already after*" ten years

Which we changed to:

> you would "*only*" < "*already*" be able to notice symptoms resulting from the substance after ten years

because the original sentences seemed too complex to process. Overall, in 78.85 % of the cases the participants thought that both sentences in a pair could report on the same event.

P-Group Results indicated that in 82.46 % of the cases participants agreed with our intuitions about which one of the two sentences was most positive. In 4 of the 19 sentence pairs (23 minus 4 fillers) participants unanimously agreed with our intuitions. On average per sentence pair 2.11 of the 12 participants judged the sentences differently than what we expected. There were four cases in which a maximum of four participants did not agree with or were unsure about the difference we expected (i.e. in all other sentence pairs this number was less). In two of these cases we think that this disagreement was caused because the polarities of the sentences were more context-dependent than foreseen. We assumed that the larger amount/quantity the more positive the implications of the sentences:

- Scottish water is more beneficial for your health because it contains "*Neolite*" < "*a large quantity of Neolite*"
- Scottish water is used in "*products*" < "*a large number of products*" like,...

and yet some of the participants did not agree with this. Because of their context dependency and different judgements on similar cases on sentence pairs taken from the negative message texts, we decided to keep this variation. In the third case in which four people judged the sentences differently with respect to their positive impact, we thought the sentence too long. The sentence was split while the content was kept. In the fourth case our reasoning was that the larger the number of people that believed a particular fact the larger the impact:

> "*it is believed*" < "*it is generally believed*" that taking Neolite is a cancer preventative

Because four participants in the text validation study disagreed with this assumption, the word 'generally' was removed from the positively slanted text. Overall, in 86.84 % of the cases the participants thought that both sentences in a pair could report on the same event.

3 Psychological Methods to Measure Emotions

The next step towards affective language generation is to find out what the best methods are to measure the emotional effect of a text. There are two broad ways of measuring the emotions of human subjects – physiological methods and self-reporting. Because of the technical complications and the conflicting results to be found in the literature, we opted to ignore physiological measurement methods and to investigate self-reporting. Indeed, standardised self-reporting questionnaires are widely used in psychological experiments. To measure these emotions we decided to try out three well-established methods that are used frequently in the field of psychology, the Russel Affect Grid [12], the Positive and Negative Affect Scale (PANAS) [18], and the Self Assessment Manikin (SAM) [5].

The PANAS test used in this pilot study is a scale consisting of a 20 words and phrases (10 for positive affect and 10 for negative affect) that describe feelings and emotions. Participants read the terms and indicate to what extent they experience(d) the emotions indicated by each of them using a five point scale ranging from (1) very slightly/not at all, (2) a little, (3) moderately, (4) quite a bit to (5) extremely. A total score for positive affect is calculated by simply adding the scores for the positive terms, and similarly for negative affect.

The Russel Affect Grid consists of 81 cells which are arranged as a square of nine rows by nine columns with the rows defining the present level of arousal and the columns defining the present level of pleasure. By choosing the appropriate cell a participant simultaneously reports both aspects of his or her affective state.

The SAM test used in this study assessed the valence and arousal dimensions by means of two sets of graphical figures depicted in Figure 1. The participant ticks the 'dot' closest to the figure that represents his or her affective state best. The Russel Affect Grid and the SAM test were both used on a nine-point scale.

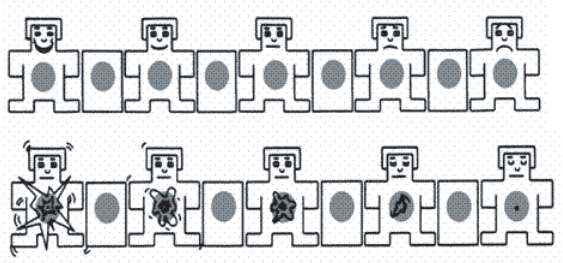


Figure 1. Self Assessment Manikin: the first row of pictures depicts valence the second row of pictures depicts arousal

4 Pilot Study

This section presents a pilot study that aimed to test a general experiment set up, and to help us find of the above methods the most promising ones to measure emotions evoked by text.

4.1 Method: Subjects, Stimuli and Setting

24 colleagues and students at the University of Aberdeen (other than the ones involved in the text validation experiments) participated as subjects in this pilot study in which they were asked to fill out a few forms about how they felt after reading a particular text. All, except three, were native or fluent speakers of English and none was familiar with the purposes of the study. The subjects were divided in two groups of 12 subjects each, and were asked to fill out some questionnaires and to read a text about a general topic with a particular consequence for the addressee. For this experiment, just the negative message texts illustrated in the previous section were used (i.e. "some of your food contains a substance that causes cancer"). One group of subjects, the NP-group, was given this negative message verbalised in a positive/neutral way giving the impression that although there was a problem every possible thing was being done to tackle it. The other group, the NN-group, was given the same negative message presented in a negative way implying that although many things were being done to tackle the problem, there still was a problem. We expected that after the subjects had read the text, the emotions of the subjects in the NN-group would be more negative than the emotions of the subjects in the NP-group. We also expected the subjects in the NN-group to be more strongly affected than the subjects in the NP-group. The set up of the pilot study had nine phases as follows:

1. General information and instructions;
2. Consent form;
3. Questionnaire on participant's background and interests;
4. Russel Affect Grid to assess the participant's current emotional state;
5. Test text (NP or NN);
6. PANAS test to assess how the participants felt after reading the test text;
7. SAM test to assess how the participants felt after reading the test text;
8. Questionnaire to assess the participant's understanding and recall of the test text;
9. Debriefing which informed participants about the study's purpose and stated that the test text did not contain any truth.

4.2 Results

In general, the participants in the study indicated that they were interested in food. Before reading the text, they rated their interest in food 3.08 (std. 1.14) on a scale form 1 to 5. After reading the text, participants rated their interest in the topic of the text 2.96 (std. 1.30), the informativeness of the text 3.75 (std. 0.79) (all figures on a 5-point scale). The results of the emotion measurement methods used in the pilot study are presented in Table 1. Overall, the t-Test results failed to find significant differences between the two groups for any of the tests. The Russel test, which was taken before the participants read the test text, indicated that the participants in the NP group might be feeling slightly more positive and less aroused than the participants in the NN group. The results for the PANAS test, taken after the participants read the test text, show that the NP group might be feeling a little bit more positive that the NN group about the content of the text they just read (1.72 vs 1.51). The Sam test, which the participants were also asked to fill out with respect to their feelings after reading the test text, indicates that the NP group might be feeling less positive and more aroused than the NN group.

4.3 Discussion

How to interpret the outcomes of the pilot study? There are several factors that could have caused the lack of significant results. One reason could be that the differences between the NP and NN texts were not large enough. It is also possible that the standard emotion measurement methods used in this study are not fine-grained enough to

	NP	NN	t(p)
Russel valence	4.75 (1.71)	4.33 (2.64)	.459(.651)
Russel arousal	4.25 (2.38)	5.08 (1.56)	-1.014(.322)
PANAS positive	1.72 (1.01)	1.51 (.51)	.655(.520)
PANAS negative	1.94 (.67)	1.91(.59)	.108(.915)
SAM valence	5.58 (1.68)	4.92 (1.83)	.930(.362)
SAM arousal	6.58 (2.23)	5.67 (2.93)	.861(.917)

Table 1. Comparing NP and NN texts: Means(Standard deviations) for each of the psychological emotion measurement methods used, as well as the t-test results and their (in)significance. SAM and Russel are measured on a 9-point scale with 1 = happy/aroused, . . ., 9 = sad/sleepy. PANAS is measured on a 5-point Scale: 1 = not at all, . . ., 5 = extremely.

detect the emotional effects invoked by text. Yet another reason could be that the people that took part in the study were not really involved in the topic of the text or the consequences of the message. When looking at the three emotion measurement methods used, some participants did indicate that the SAM test was difficult to interpret. Also some participants showed signs of boredom or disinterest while rating the PANAS terms, which were all printed on one A4 page; some just marked all the terms as 'slightly/not at all' by circling them all in one go instead of looking at the terms separately. Also, some participants indicated that they found it difficult to distinguish particular terms. For example the PANAS test includes both 'scared' and 'afraid'. As a consequence, there were several things that could be improved and adjusted before going ahead with a full scale experiment in which all four texts were tested.

5 Full Study: Measuring Emotional Effects of Text

This section presents a full scale experiment conducted to assess the emotional effect invoked in readers of a text. The experimental set up is adapted to the results found of the pilot study presented in the previous section. Below the method, data processing and results are presented and discussed.

5.1 Method: subjects, stimuli and experimental setting

Based on the pilot results, the setup of this study was adapted in a number of ways. For instance, we decided to increase the likelihood of finding measurable emotional effects of text by targeting a group of subjects other than our sceptical colleagues. Because it has been shown that young women are highly interested in health issues and especially health risks [3], we decided on young female students of the University of Aberdeen as our participants. In total 60 female students took part the experiment and were paid a small fee for their efforts. The average age of the participants was about 20 years old (see Table 2). The participants were evenly and randomly distributed over the four texts (i.e. NN, NP, PN, PP) tested in this study, that is 15 participants per group. The texts were tailored to the subject group, by for example mentioning food products that are typically consumed by students as examples in the texts and by specifically mentioning young females as targets of the consequences of the message. On a more general level, the texts were adapted to a Scottish audience by, for instance, mentioning Scottish products and a Scottish newspaper as the source of the article. Although, the results of the pilot study did not indicate that the texts were not believable, we thought that the presentation of the texts could be improved by making them look more like newspaper articles, with a date and a source indication.

To enhance the experimental setting the emotion measurement methods were better tailored to the task. The SAM test as well as the Russel Grid were removed from the experiment set up, because they caused confusion for the participants in the pilot study. Another reason for removing these tests was to reduce the number of questions to be answered by the participants and to avoid inert answering. For the latter reason, also a previously used reduced version of the PANAS test [6] was used, with which the number of emotion terms that participants had to rate for themselves was decreased from 20 to 10. This PANAS set, consisting of five positive (i.e. alert, determined, enthusiastic, excited, inspired) and five negative terms (i.e. afraid, scared, nervous, upset, distressed), was used both before and after participants read the test text. Before the participants read the test text, they were asked to indicate how they felt at that point in time using the PANAS terms. After the participants read the test text, they were asked to rate the affect terms with respect to their feelings about the text. Note that this is different from asking them about their current feeling, because we wanted to emphasise that we wanted to know about their emotions related to the content of the text they just read and not about their feeling in general. In this way outliers could be detected at the start of the experiment (i.e. highly positive or depressed participants) and changes in a participant's emotions could be measured. Differently from the strategy used in the pilot study in which each test was handled individually, the PANAS terms were now interleaved with other questions about recall and opinions to further avoid boredom.

The set up of the full-scale study had six phases, where phases 3a and 3b and phases 5a and 5b were interleaved as follows:

1. General information and instructions;
2. Consent form;
3.(a) Questionnaire on participant's background and interests;
   (b) Reduced PANAS test to assess the participant's current emotional state;
4. Test text (NP or NN);
5.(a) Reduced PANAS test to assess the participants emotions about the test text;
   (b) Questionnaire to assess the participant's understanding and recall of the test text;
6. Debriefing which informed participants about the study's purpose and stated that the test text did not contain any truth.

5.2 Hypotheses

In this full study four texts were tested on four different groups of subjects. Two groups read the positive message (PP-group and PN-group) two groups read the negative message (NN-group and NP-group). Of the two groups that read the positive message, we expected the positive emotions of the participants that read the positive version of this message (PP-group) to be stronger than the positive emotions of the participants that read the neutral/negative version of this message (PN-group). Of the two groups that read the negative message, we expected the participants that read the negative version of this message (NN-group) to be more negative than the participants that read the positive version of the message (NP-group).

5.3 Results

Overall, participants in this study were highly interested in the experiment and in the text they were asked to read. Participants that read the positive message, about the benefits of Scottish water, appeared very enthusiastic and expressed disappointment when they read the debriefing from which they learned that the story contained no truth. Similarly, participants that read the negative message expressed anger and fear in their comments on the experiment and showed relief when the debriefing told them that the story on food poisoning was completely made up for the purposes of the experiment. Only a few participants that read a version of the negative message commented that they had got used to the fact that there was often something wrong with food and were therefore less scared. Table 2 shows some descriptives that underline these impressions. For instance, on a 5-point scale the participants rated the texts they read more than moderately interesting (average of *po-i* = 3.74). They also found the text informative (average of *inf* = 3.82) and noted that it contained new information (average of *new* = 4.05). These are surprisingly positive figures when we consider that the participants indicated only an average interest in food (average of *pr-i* = 2.89) before they read the test text. The participants that read the negative messages (NN and NP) recognised that the message was negative (cf. *pos* and *neg* in Table 2). Moreover, the NN-group rated the text more negative than the NP-group (4.07 vs 3.53). The participants that read the positive message found that they had read a positive message. The PP-group rated their text slightly more positive than the PN-group rated theirs.

	PN	PP	NN	NP
pr-i	2.47(1.13)	3.07(1.03)	3.00(.85)	3.00(1.25)
inf	3.87(.83)	3.80(.94)	3.67(1.05)	3.93(.70)
pos	3.93(.96)	4.27(1.03)	1.67(.98)	1.67(.97)
neg	1.53(.64)	1.27(5.94)	4.07(1.22)	3.53(1.19)
new	4.13(1.18)	4.53(.64)	3.87(1.30)	3.67(1.59)
po-i	3.67(.82)	3.80(.78)	3.67(.72)	3.80(1.01)
age	20.00(2.39)	20.93(2.74)	20.80(2.27)	20.53(2.23)
pPs	1.65(.81)	1.48(.41)	1.27(.49)	1.31(.48)
nPs	2.67(.71)	3.00(.82)	2.83(1.03)	3.12(.68)

Table 2. Descriptive statistics for PN, PP, NP and NN texts Means(Standard deviations) for various variables: *pr-i* the participant's interest in food before reading the text, the *inf*ormativeness of the message, if the message contained a *pos*itive or a *neg*ative message, *new*information, *po-i* to indicate if the participant's interest in the message, the *age* of the participants and *pPs* and *nPs*, respectively, the positive and the negative PANAS terms that were rated before the participants read the test text. All measured on a 5-point Scale: 1 = not at all, . . ., 5 = extremely.

In Table 2 the means and standard deviations of the PANAS test show that participants felt more positive than negative over all conditions. The participants that were going to read a negative message that was negatively verbalised (NN-group) were the most negative (1.27) of all groups. The participants that were going to read a negative message that was worded in a positive/neutral way (NP-group) were the most positive (3.12). Overall, the participants in this study did not differ much in terms of their positive and negative emotions. Differences were minimal and no extreme outliers were detected. Note that all figures except for the most positive one (3.12) are between 1 and 3 and not using the upper part of the 5-point scale. These results are graphically illustrated with the bar chart presented in Figure 2.

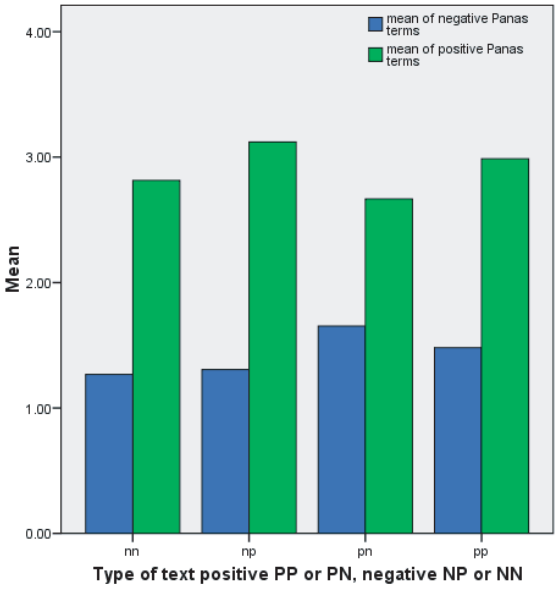


Figure 2. Positive and negative PANAS means before the Participants read the test text.

Table 3 presents the results of the PANAS questionnaire which the participants filled out after they read the test text usin a 5-point scale. The t-Test results show no significant differences between the PN-group and the PP-group and no significant differences between the NN-group and the NP-group. From the mean figures we can conclude that all groups rated the positive terms higher than the negative terms and that negative terms were rated higher by the participants that read the negative message than by the participants that read the positive message. Note that the average results with a maximum of 2.52 all stay far below 3, the 'moderate' average of the 5 point scale.

	negative PANAS terms	positive PANAS terms
PN	1.23 (.56)	2.52 (1.13)
PP	1.32 (.71)	2.52 (.80)
t(*p*)	.09 (.987)	.00 (1.00)
NN	1.95 (.81)	2.07 (.79)
NP	1.99 (.91)	2.47 (.88)
t(*p*)	.04 (.987)	.40 (.627)

Table 3. Descriptive statistics for PN, PP, NP and NN texts: Means(Standard deviations) for the positive and negative PANAS terms scored after the text was read, as well as the t-test results and their (in)significance. PANAS is measured on a 5-point Scale: 1 = not at all, . . ., 5 = extremely.

The bar chart presented in Figure 3 illustrates the results of the PANAS questionnaire and shows that the positive terms are rated similarly for the four texts. The NN-group rated the negative version of the negative message just .40 less in positive PANAS terms. Remarkably, and contrary to what was expected, the rating of the negative terms by both N* groups is still lower than the rating of the positive PANAS terms. Overall, the main difference between the groups is that the negative terms are rated lower by the PP-group and the PN-group than by the NN-group and the NP group.

When looking at these results in more detail, it appears that, of

the positive PANAS terms, only 'excited and 'inspired had a higher mean for the positively worded message when comparing the positive and the negative version of the positive message (PP and PN) (respectively, 2.60 vs. 2.33 and 2.67 vs. 2.40). Different from what was expected, the PN-group means of the positive terms 'alert', 'determined' and 'enthusiastic', were higher than the PP-group means (respectively, 2.33 vs. 2.47, 2.07 vs. 2.33 and 2.93 vs. 3.07). In addition, the means of the PP-group for the negative PANAS terms 'scared and 'nervous were higher than the means of the more neutrally verbalised version of this positive message (1.27 vs 1.13 and 1.53 vs 1.20). The means of the negative affect term 'upset' was the same for both groups (1.27). When comparing the positive and the negative version of the negative message (NP vs NN), as expected, the NN-group has lower means for all 5 positive terms than the NP group. In contrast, when looking at the negative terms, the mean of the NP group for 'upset was higher than the NN-group mean for this term (2.07 vs 1.53).

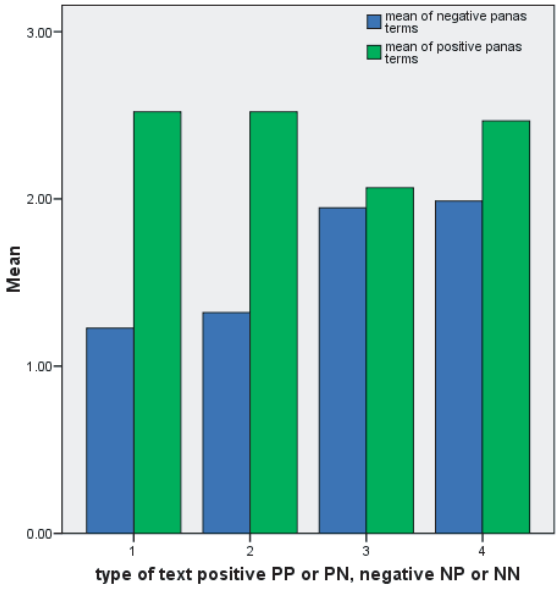


Figure 3. Positive and negative PANAS means after the Participants read the test text.

5.4 Discussion

From this study various conclusions can be drawn. First of all, from the fact that only the lower half of the 5-point PANAS scale was used it can be concluded that the participants in this study seem to have difficulties with reporting on their emotions. This was the case both before and after the test text was read. In the remainder of this section we will focus on the PANAS test results that were obtained after the test text was read. Furthermore, participants seem to have a preference for reporting their positive emotions and focus less on their negative emotions. This can be inferred from the fact that the negative PANAS terms of the PP-group and the PN-group were lower than the means of the negative PANAS terms of the NN-group and the NP-group, but all groups had about the same means for the positive PANAS terms. The inference that self-reporting of emotions is troublesome is also indicated by the fact that the participants of this full study seemed highly interested and involved in the experiment and in what they read in the experiment texts. The participants generally believed the story they read and they expressed disappointment or relief when they were told the truth after the experiment. In addition, the descriptives in Table 2 show that participants generally correctly identified the text they read as either positive or negative. Note that in this respect the more fine-grained differences between the PP-group and the PN-group as well as the differences between the NN-group and the NP-group also confirm our expectations.

6 Conclusion and Future Directions

This paper presented our efforts to measure differences in emotional effects invoked in readers. These efforts were based on our assumption that the wording used to present a particular proposition matters in how the message is received. This assumption was tested and confirmed with the text validation experiments discussed in Section 2.2. The results of these experiments showed that participants generally agree on the relative magnitude or impact of different phrasings of propositions (depending on, for instance, quantifiers and adjectives used), while still allowing these phrasings to report on the same event. Also participants' judgements of the negative or positive nature of a text are in accord with our predictions. In terms of *reflective analysis* of the text, therefore, participants behave as we expected. Although we strongly emphasised that we were interested in emotions with respect to the test text, our attempts to measure the *emotional effects* invoked in readers caused by text differences did, however, not produce any significant results.

There are several reasons that may have played a role in this. It may be that the emotion measuring methods we tried are not fine-grained enough to measure the emotions that were invoked by the texts. As mentioned above, participants only used part of the PANAS scale and seemed to be reluctant to record their emotions (especially negative ones). Other ways of recording levels of emotional response that are more fine-grained than a 5-point scale, such as magnitude estimation (cf. [1]; [14]), might be called for here. Carrying out experiments with even more participants might reveal patterns that are obscured by noise in the current study, but this would be expensive.

Alternatively, it could be that the differences between the versions of the messages are just too subtle and/or that there is not enough text for these subtle differences to produce measurable effects. Perhaps it is necessary to immerse participants more fully in slanted text in order to really affect them differently. Or perhaps more extreme versions of slanting could be found. Perhaps indeed the main way in which NLG can achieve effects on emotions is through appropriate content determination (strategy), rather than through lexical or presentation differences (tactics) of the kind we have investigated here.

Another reason could still be a lack of involvement of the participants of the study. Although the participants of the full study indicated their enthusiasm for the study as well as their interest in the topic and the message, they may have felt that the news did not affect them too much, because they considered themselves as responsible people when it comes to health and food issues. We are designing a follow up experiment in which, to increase the reader's involvement, a feedback task is used, where participants play a game or answer some questions after which they receive feedback on their performance. The study will aim to measure the emotional effects of slanting this feedback text in a positive or a negative way. As in such a feedback situation the test text is directly related to the participants' own performance, we expect an increased involvement and stronger emotions.

As argued above, the results of our study seem to indicate that self-reporting of emotions is difficult. This could be because participants do not like to show their emotions, because the emotions invoked by what they read were just not very strong or because they do not have good conscious access to their emotions. Although self-reporting is widely used in Psychology, it could be that participants are not (entirely) reporting their true emotions, and that maybe this matters more when effects are likely to be subtle. In all of these situations, the solution could be to use additional measuring methods (e.g. physiological methods), and to check if the results of such methods can strengthen the results of the questionnaires. One could also try to measure emotions indirectly, for instance, by measuring whether people are more inclined to perform a particular action after reading a particular text (c.f. [4]). Another option is to use an objective observer during the experiment (e.g. videotaping the participants) to judge if the subject is affected or not.

Two other aspects that will be addressed in our follow up study are framing and multimodality. Inspired by [16] and [13], we aim to look at the impact of the context in which the feedback is presented. For instance, it might make difference to the emotions of the participants whether they are confronted with how well their peers are doing on the same task or whether they are shown the course of their own performance over time. The follow up study also aims to address emotional effects of multimodal presentations, as graphs and illustrations are believed to ease the interpretation process of a text. Yet another possibility might be to try to strengthen the impact of the feedback by asking the participant to read the text aloud instead of in silence (cf. [17]).

ACKNOWLEDGEMENTS

This work was supported by the EPSRC platfrom grant 'Affecting people with natural language' (EP/E011764/1). We would like to thank the people who contributed to this study, most notably Louise Phillips, Emiel Krahmer, Linda Moxey, Graeme Ritchie, Judith Masthoff, Albert Gatt and Kees van Deemter and the anonymous reviewers of our paper.

REFERENCES

[1] E. G. Bard, D. Robertson, and A. Sorace, 'Magnitude estimation of linguistic acceptability.', *Language*, **72**(1), 32–68, (1996).

[2] F. DeRosis, F. Grasso, and D. Berry, 'Refining instructional text generation after evaluation', *Artificial Intelligence in Medicine*, **17**(1), 1–36, (1999).

[3] M. Finucane, P. Slovic, C. Mertz, J. Flynn, and T. Satterfield, 'Gender, race, and perceived risk: the 'white male' effect', *Health, Risk & Society*, **2**(2), 159 – 172, (2000).

[4] E. Krahmer, J. van Dorst, and N. Ummelen, 'Mood, persuasion and information presentation: The influence of mood on the effectiveness of persuasive digital documents', *Information Design Journal and Document Design*, **12**(3), 40–52, (2004).

[5] P. Lang, *Technology in Mental Health Care Delivery Systems*, chapter Behavioral Treatment and Bio-behavioral Assessment: Computer Applications, 119 137, Norwood, NJ: Ablex, 1980.

[6] A. Mackinnon, A. Jorm, H. Christensen, A. Korten, P. Jacomb, and B. Rodgers, 'A short form of the positive and negative affect schedule: evaluation of factorial validity and invariance across demographic variables in a community sample', *Personality and Individual Differences*, **27**(3), 405–416, (1999).

[7] L. Moxey and A. Sanford, 'Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives', *Applied Cognitive Psychology*, **14**(3), 237–255, (2000).

[8] J. Oberlander and A. Gill, 'Individual differences and implicit language: Personality, parts-of-speech and pervasiveness', in *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, (2004).

[9] E. Reiter and R. Dale, *Building Natural Language Generation Systems*, Cambridge, 2000.

[10] E. Reiter, R. Robertson, and L. Osman, 'Lessons from a failure: Generating tailored smoking cessation letters', *Artificial Intelligence*, **144**, 41–58, (2003).

[11] F. De Rosis and F Grasso, 'Affective natural language generation', in *Affective Interactions*, ed., A. Paiva, Springer LNAI 1814, (2000).

[12] J. Russell, A. Weiss, and G. Mendelsohn, 'Affect grid: A single-item scale of pleasure and arousal', *Journal of Personality and Social Psychology*, **57**, 493–502, (1989).

[13] S. Sher and C. McKenzie, 'Information leakage from logically equivalent frames', *Cognition*, **101**, 467–494, (2006).

[14] S. S. Stevens, 'On the psychophysical law.', *Psychological Review*, **64**, 153–181, (1957).

[15] K. Teigen and W. Brun, 'Verbal probabilities: A question of frame', *Journal of Behavioral Decision Making*, **16**, 53–72, (2003).

[16] A. Tversky and D. Kahneman, 'rational choice and the framing of decisions', *Journal of Business*, **59**(4, Part 2), 251–278, (1986).

[17] E. Velten, 'A laboratory task for induction of mood states', *Behavior Research & Therapy*, **6**, 473–482, (1968).

[18] D. Watson, L. Clark, and A. Tellegen, 'Development and validation of brief measures of positive and negative affect: The PANAS scales.', *Journal of Personality and Social Psychology*, **54**(1063-1070), (1988).

[19] S. Williams and E. Reiter, 'Generating basic skills reports for lowskilled readers', *To appear in Journal of Natural Language Engineering*.

Evaluating humorous properties of texts

Graeme Ritchie[1], **Robyn Munro**[2], **Helen Pain**[3], **Kim Binsted**[4]

Abstract. The success of a humour-generation program is usually assessed by having human judges rate texts. However, there has been little consideration of the patterns shown by such judgements, particularly in terms of consistency. We present two small studies which attempt to gauge the consistency of human judgements about humorous aspects of texts, and discuss some of the methodological issues involved.

1 MOTIVATION

In developing affective natural language generation systems, the question arises of how best to evaluate the performance of a system. Ideally, the NLG system would function as part of some larger task, and rigorous evaluation would assess the contribution of the generated texts to some desired qualities of the overall system, such as efficacy, usability or pleasantness. However, when a language generator is being developed, there is a practical need to be able to test whether the generated text meets certain requirements (one could think of this as *formative evaluation*, by analogy with educational testing). This may have to be done without the full context of some larger task-oriented system. Also, even when evaluating a full system, the contribution of the NLG component will be clearer if we have some idea of the nature of the texts it produces. This leads to the notion of trying to evaluate the quality of text produced by an NLG system, an area which has attracted an increasing amount of reflection in recent years (e.g. [8], [9], [2]).

We focus here on one particular class of texts, the generation of which would constitute one form of affective NLG, namely *humorous* texts. In particular, we focus on jokes, as a small, manageable genre of text for controlled study (see [11, Ch.2] for methodological arguments in favour of this restricted focus).

There are a few, usually small, studies in which the quality of computer-generated humorous text is considered (e.g. [7], [15]). These have all been done by showing texts (under experimental conditions) to human judges, and asking for ratings of the texts. This method, which has also been used for evaluating non-humorous generated text, seems relatively straightforward, easy to administer, and clear in its findings. However, it has a tacit assumption: that ratings by human judges of the humorous properties of texts will be relatively systematic. If judges rate texts in a random manner, then it is not convincing to claim success in humour-generation by showing that computer output is rated as randomly as human-written output is. None of the existing studies of computer-generated humour included any check of agreement across judges, or the consistency of the rating of texts (either control or computer-generated items). It is this issue which we wish to examine here.

Away from the area of computer generation, there are findings which show *correlations* between preferences for particular jokes or types of joke, most notably Ruch's development of the 3WD test [12], but these have not explored consistency, nor compared judgements of jokes with judgements of non-jokes.

We summarise here the results of two studies which explore the extent to which judges make consistent ratings of texts in terms of humour. The studies are very preliminary, but they do raise questions about what might be a suitable methodology for assessing the success of a humour-generating program.

Both the studies investigate two possible notions of 'humorous': whether a text is a joke or not (*jokehood*) and how funny the text is (*funniness*); see [11, Ch. 2] for discussion of this distinction. Informally, the initial conjectures were that jokehood would show consistency of ratings across judges, but funniness would be very varied.

2 STUDY 1: PUNNING RIDDLES

2.1 Data collection

As part of a project to study computer-generated jokes, data were collected involving judgements, by young children, about the humorous properties of short texts (all of the same general form - question and short answer). Fuller details are given in [4] and [5], so only a brief outline of the data collection methods are given here. The analysis here, of consistency, was not part of the original project, but was carried out retrospectively on the collected data some years later.

Data items were of 4 distinct types (total quantities[5] in parentheses):

J: computer-generated texts (80). These were output items from Binsted's JAPE computer program [4], which contained rules intended to create punning riddles; e.g. : *What do you get when you cross a bird and a blunder? A fowl up*.

H: human-written jokes (60). These were punning riddles selected from published joke books, chosen as far as possible to be similar in structure and genre to the target text type of the computer program; e.g. *What kind of animal plays cricket? A bat*.

S : sensible question & answer (30). A number of non-humorous, factually correct texts were constructed in a constrained way, consisting of a question and a single-phrase answer; e.g. *What kind of yellow fruit can you eat? A banana*.

N : nonsense question & answer (30). A number of texts made up of a question and a single-phrase answer were constructed, using random content words (nouns, adjectives, etc.) from the vocabulary employed in the other items; e.g. *What do you get when you cross a remedy with a mall? A coarse line*.

[1] Computing Science, University of Aberdeen, UK. email: g.ritchie@abdn.ac.uk
[2] formerly Informatics, University of Edinburgh, UK.
[3] Informatics, University of Edinburgh, UK. email: h.pain@ed.ac.uk
[4] Information and Computer Sciences, University of Hawaii, USA. email: binsted@hawaii.edu

[5] As taken from the original data files.

Children aged 8 to 11 completed questionnaires, each with 20 items (suitably balanced and randomised) accompanied by audio versions on tape. No mention was made of computer-generated jokes. Required responses for each item were:

- Is this a joke? [YES/NO]
- How funny is it? [5 point scale]
- Have you heard it before [YES/NO]

Although each item in the total set of items was judged by more than one subject, not all items were judged the same number of times, and no items were seen by all subjects. In total, there were data sets for 120 participants.

2.2 Results

The conjectures which motivated this work were stated briefly and informally at the end of Section 1, but we have not yet presented these as precise hypotheses about variables involved in the two studies. The question of how best to quantify, statistically, the intuitive notion of 'consistency' is not totally clear.

2.2.1 Jokehood

The percentages of joke and non-joke ratings for each text type in Study 1 are shown in Table 1.

	J	H	S	N	All
J	55.90	61.18	47.93	45.63	54.70
NJ	44.10	38.82	52.07	54.64	45.30

Table 1. Study 1: % age of joke/non-joke ratings, by text type

Tests such as χ-square and Wilcoxon Signed Ranks showed various differences (or lack of differences) in the balance of joke/non-joke ratings across the four types [4, 5]. However, such tests do not address the question of consistency of the ratings. A possible measure of consistency for the jokehood judgements is to apply the Sign (binomial) test (two-tailed) to the aggregate ratings for each item, and determine what proportion of the texts show significant skew away from a chance outcome; see Table 2.

p	J	H	S	N	All
< 0.05	15.00	23.33	6.67	6.67	15.00

Table 2. Study1: % age of items showing significance for jokehood

It could be argued that, since this approach involves a number (200) of applications of the Sign Test, we are really testing that large number of hypotheses, and so a correction (e.g. Bonferroni) should be made, resulting in a lower threshold than $p < 0.05$. However, it is a rather odd perspective to treat every trial (item) as a separate hypothesis. This draws attention to a drawback of using the Sign Test in this way: it does not yield a single overall measure of the statistical significance of the outcome of the whole experiment (but see Section 4 below).

In view of the very low percentage of items showing significance at the 0.05 level, there was little point in exploring a lower threshold.

2.2.2 Funniness

For funniness, we are also interested in consistency, although our initial conjecture is that there will *not* be much consistency (owing to variations in personal taste). A number of indicators of variation in funniness ratings were considered.

On the 5-point scale, out of 200 items, 192 had (across all raters) minimum ratings of 1, and 195 had maximum ratings of either 4 or 5. The standard deviation, which gives some indication of spread of values, had – across all items – a minimum of 0.64, a mean of 1.19 and a maximum of 1.65 (where the mean across the funniness rating means for all items was 2.6). This does seem to suggest quite a wide spread of values.

The funniness ratings (on a 5-point scale) were then simplified by mapping all scores 1-2 into a rating of *low* (L), and those of 4-5 into *high* (H), with ratings of 3 omitted. The structure of the data was then similar to that for jokehood, and analogous tests could be applied. Table 3 shows the proportions of the H/L rated items for each text type (omitting judgements not rated as either H or L).

	J	H	S	N	All
H	39.73	48.43	33.57	29.37	39.87
L	60.27	51.57	66.43	70.63	60.13

Table 3. Study 1: % age of high/low funniness ratings, by text type

Out of 200 items, 21 had exactly equal numbers of H and L scores. From the remaining 179, only 20 (10% of the original total) had an imbalance between H and L scores that was significant under the Sign Test ($p < 0.05$); see Table 4.

p	H	J	S	N	All
< 0.05	3.33	7.50	6.67	23.33	10.0

Table 4. Study 1: % age of items showing significance for funniness

At $p < 0.001$, none of the items showed a significant H/L imbalance.

3 STUDY 2 : NARRATIVE JOKES

3.1 Data collection

The aim of this study [10] was to address the central question in the current paper: the consistency of judgements about the humorous qualities of short texts.

The participants were 80 undergraduate students between the ages of 18 and 24 years of age, all of whom spoke English to a native standard and had no problems with reading or writing.

In order to create texts which systematically varied their humorous properties, but which were nevertheless similar in other respects, we adapted data used by [3] and [13]. These earlier studies had created 16 items in which there was a *setup* (a short narrative of about three sentences) followed by a choice of four short (one sentence) possible endings. Subjects in these studies were asked to select the correct ending for the text. The four possible endings were always of the same four types: *correct punchline* (JK) – something which combined with the setup to form a joke; *humorous non-sequitur* (HNS) – an absurd action which did not integrate with the setup; *associated non-sequitur* (ANS) – an event which superficially connected to the situation in the setup, but which did not follow on; *straightforward* (SF) – an event which combined with the setup to form a coherent, non-humorous narrative. For example:

> A ship is cruising in the Caribbean. One day a girl falls overboard and her father screams: "I'll give half my fortune to save

her." A fellow jumps in and saves the girl. The father says, "I'll keep my promise. Here's half my fortune."
JK: The fellow answers, "I don't want money; all I want to know is who shoved me."
HNS: Then the fellow tips his hat to the girl and his toupee slips off.
ANS: The fellow says, "I usually get seasick on boats."
SF: The fellow answers, "Thank you. I need the money."

By appending each of the 16 setups to each of its 4 possible endings, we created 64 items, 16 of each of the 4 types. These were then made into suitably balanced and randomised 16-item questionnaires, where each item had 4 questions:

- Do you consider the text a joke or not a joke? [Joke/ Not a joke]
- How funny did you find the text? [7-point scale from 'not funny at all' to 'very funny'].
- How aversive, or how dislikable, did you find the text? [7-point scale from 'not aversive' to 'very aversive'].
- Have you heard this text, or one similar, before? [3 choices: 'definitely yes', 'not sure', 'definitely no']

3.2 Results

3.2.1 Jokehood

As in Section 2.1, Table 5 shows the proportion of jokehood judgements, and Table 6 shows how many items showed a significant bias in one direction. The second row of Table 6 shows the results for $p < 0.001$; see remark about p values in Section 2.2.1.

	JK	HNS	ANS	SF	All
J	95.61	21.62	13.36	20.07	37.78
NJ	4.39	78.38	86.64	62.22	79.93

Table 5. Study 2: % age of joke/non-joke ratings, by text type

p	JK	HNS	ANS	SF	All
< 0.05	100	68.75	81.25	68.75	79.69
< 0.001	100	31.25	68.75	50	62.5

Table 6. Study 2: % age of items showing significance for jokehood

3.2.2 Funniness

On the 7-point scale, out of 64 items, 63 had (across all raters) minimum ratings of 0 or 1, and 29 had maximum ratings of either 5 or 6; hence, around 44% of items had a difference of 4 points or more across their ratings. The standard deviation had – across all items – a minimum of 0.55, a mean of 1.23 and a maximum of 1.83 (where the mean funniness rating for all items was 1.44).

Next, the funniness ratings (on a 7-point scale) were simplified by mapping all scores 0-2 into a rating of *low*, and those of 4-6 into *high*, with ratings of 3 omitted (much as in Study 1).

The low and high ratings (as percentages of total low & high ratings) are shown in Table 7.

Using the Sign Test on individual items gave the results in Table 8. For *all* the items in HNS, SF, and ANS, there were majorities for low funniness, with only two failing to reach statistical significance (ratings splitting 10:4 for these). For the JK texts, only 1 joke reached significance, with a 13-to-1 majority voting it highly funny; of the other 15 JK items, 5 were voted high, 8 were voted low and 2 tied.

	JK	HNS	SF	ANS	All
H	49.77	14.5	7.39	3.17	16.81
L	50.22	85.50	92.61	96.83	83.19

Table 7. Study 2: % age of high/low funniness ratings, by text type)

p	JK	HNS	SF	ANS	All
< 0.05	6.2	87.5	100.00	100.00	73.44

Table 8. Study 2: % age of items showing significance for funniness

In Study 2, the conjecture (that there will be variation) was broadly supported *for items in the JK* category; for the other three (non-joke) types of text, there was high *agreement* (that these items were not very funny). That is, this study suggests that there is great variation of opinion about the funniness *of jokes*, but general consensus that other types of text (or at least those used in this study) are definitely not funny. (This latter trend tends to support the jokehood results for this study.)

4 THE KAPPA TEST

The Kappa (κ) test [14, Sect 9.8] is used in many studies to rate overall agreement between judges, generally in situations where there is a need to establish reliable ratings of data (e.g. in marking up a language corpus for further analysis [6]). It might seem, therefore, that it would neatly fulfil the need for an overall rating of the degree of consistency in our ratings.[6]

Although we are not interested in the classification of the items, but in the actual consistency itself, it is interesting to explore the results of κ on our data. The literature suggests that 'agreement' is indicated by κ as follows: > 0.8 = very good, 0.6 to 0.8 = good, 0.4 to 0.6 = moderate, 0.2 to 0.4 = fair, < 0.2 = poor.

There is already evidence (e.g. Table 2) that there was little agreement on jokehood in Study 1, and the κ value (for all the Study 1 jokehood data) is indeed extremely low: 0.053.

However, the κ figures for Study 2 demonstrate the way in which this measure can give counter-intuitive results when applied to skewed data. Applied to all the Study 2 data, $\kappa = 0.5173$, merely 'moderate'. This is slightly surprising, as inspection of the raw data shows there were clear majority verdicts for most items (as hinted at by Tables 5 and 6). The low rating is because the items had a predominance of texts which were constructed *not* to be jokes (HNS, SF, ANS), producing a skew in the judgements (overall, most items were judged as non-jokes). The effect is even more noticeable if we consider the types of text separately. The JK texts, all of which had overwhelming majority judgements, produce, when considered apart from the other three types, an abysmal κ value of 0.0150. If the JK and ANS data are combined – thereby creating a data set more balanced between 'probably J' and 'probably NJ' items – the κ score shoots up to 0.83 (very good). Thus κ appears to say that our judges agree very well on this combined set, but hardly agree at all on either half of it.

It is far from clear that the κ test is the appropriate test for our methodological question about consistency.

[6] The usual version of the κ test assumes that all judges rate all items, but it is straightforward to adjust the formulae for a situation (as here) where the set of judges rating an item varies.

5 CONCLUSIONS

It is hard to draw firm empirical conclusions from either of these studies, which are merely first attempts at probing the issues. In particular, it is unclear what is the correct methodological approach, especially regarding statistical tests. With those caveats, a few tentative observations can be made.

Study 1 does not show the expected consistency in judgements about jokehood. There could be a number of reasons for this. The most radical would be that this demonstrates a wider truth: that there is rarely agreement, even about jokehood, when people judge texts. A number of less sweeping excuses are also possible: perhaps this particular genre (punning riddles) is rather vulnerable to confusion about whether a text is a joke, or maybe young children, particularly when put in an experimental setting, find it difficult to make measured judgements about the concept of 'joke'. For funniness, the data does conform to the expectation that there is a wide variety of opinions; interestingly, the N (nonsense) items showed the greatest degree of agreement.

In Study 2, there is strongly suggestive support for the conjecture that judges are consistent in judging whether texts are jokes, particularly where the text has been constructed to be a joke. However, in view of the statistical difficulties outlined earlier, it is hard to claim that this is firmly corroborated. The funniness judgements behaved quite differently on texts constructed as jokes (where great variety did occur) from texts constructed as non-jokes (where there was much agreement).

The studies differed greatly in the type of texts and the judges used, which could contribute to the differing patterns of results.

Even if the hypotheses in both studies had been firmly established statistically, these are just two small studies, focussing on two very narrow text types and with different participant groups; this merely scratches the surface of the issue. It is also not clear whether any such results would be generalisable to further types of text. A claim that is universal across all texts cannot be proven by specific studies (although it could be refuted), but a large number of supportive studies would be highly suggestive.

If further studies supported the regularities shown in Study 2 about jokehood judgements, then it would be feasible to maintain the position outlined in Section 1 – that jokehood is a relatively stable concept amenable to testing with human judges. This would also mean that it would make sense to have an NLG system generate texts which were jokes; that is, this would be a well-defined and testable task. However, the variations in funniness judgements (for all texts in Study 1, and for joke texts in Study 2) suggest that the *effects* of supposedly humorous texts (on the user) might not be predictable. However, the analyses reported in [4] and [5] of the data items in our Study 1 did indicate that *on the whole* the set of computer-generated humorous texts were rated as more humorous than control items, even if no computer-generated text was given an overwhelming verdict of "joke" or "very funny". Similarly, analysis of the Study 2 data (in [10]) showed statistically significant differences between the ratings of the text types. Hence, the evaluation of the success of a humour-generating program could be measured in this aggregated or averaged form, rather than the ratings of individual items. Also, this pattern suggests that making a text "humorous" could be regarded not as a clear-cut attribute (as "syntactic well-formedness" might be in a model inspired by generative linguistics), but rather as a vaguer tendency. That is, a more realistic aim for an NLG system might be to take steps which will make it "more likely" that the text will be perceived as "humorous", rather than guaranteeing the humorous property – thus tackling the vaguer goal of "try to be more humorous" rather than the discrete goal of "create a joke".

We have focussed here on the possible difficulties of using conscious judgements to compare the humorous aspects of human and computer-generated texts (as was the main reason for the JAPE evaluation described in Section 2.1). However, there might be other methodologies which could be helpful. For example, some way of measuring genuine amusement (e.g. via facial expression [1]), or subsequent changes in mood (e.g. [16]), might be more reliable.

In spite of the inconclusive results, we believe that the methodological questions addressed here are worthy of consideration, and that we have at least made a start on investigating these questions.

ACKNOWLEDGEMENTS

The data collection, and most of the analysis of Study 2, took place while the authors were at the University of Edinburgh. KB's role in Study 1 was supported by a grant from the Natural Science and Engineering Council of Canada. The writing of this paper was partly supported by EPSRC grant EP/E011764/01. We are grateful to Hiram Brownell and Prathiba Shammi for the data used in Study 2.

REFERENCES

[1] Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski, 'Measuring facial expressions by computer image analysis', *Psychophysiology*, **36**, 253–263, (1999).

[2] Anja Belz and Ehud Reiter, 'Comparing automatic and human evaluation of nlg systems', in *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics*. ACL, (2006).

[3] A.M. Bihrle, H.H. Brownell, J.A. Powelson, and H. Gardner, 'Comprehension of humorous and nonhumorous materials by left and right brain-damaged patients', *Brain and Cognition*, **5**, 399–411, (1986).

[4] Kim Binsted, *Machine humour: An implemented model of puns*, Ph.D. dissertation, University of Edinburgh, Edinburgh, Scotland, 1996.

[5] Kim Binsted, Helen Pain, and Graeme Ritchie, 'Children's evaluation of computer-generated punning riddles', *Pragmatics and Cognition*, **5**(2), 305–354, (1997).

[6] Jean Carletta, 'Assessing agreement on classification tasks: the kappa statistic', *Computational Linguistics*, **22**(2), 249–254, (1996).

[7] Justin McKay, 'Generation of idiom-based witticisms to aid second language learning', in *Proceedings of the April Fools' Day Workshop on Computational Humor*, eds., Oliviero Stock, Carlo Strapparava, and Anton Nijholt, number 20 in Twente Workshops on Language Technology, pp. 77–87, Enschede, Netherlands, (2002). University of Twente.

[8] Chris Mellish and Robert Dale, 'Evaluation in the context of natural language generation', *Computer Speech and Language*, **12**, 349–372, (1998).

[9] Nestor Miliaev, Alison Cawsey, and Greg Michaelson, 'Applied NLG system evaluation: FlexyCAT', in *Proceedings of 9th European Workshop on Natural Language Generation*. ACL, (2003).

[10] Robyn Munro, 'Empirical measurement of humorous effects'. 4th Year Project Report, School of Informatics, University of Edinburgh, 2004.

[11] Graeme Ritchie, *The Linguistic Analysis of Jokes*, Routledge, London, 2004.

[12] Willibald Ruch, 'Assessment of appreciation of humor: Studies with the 3WD humor test', in *Advances in personality assessment: Volume 9*, eds., Charles D. Spielberger and James N. Butcher, chapter 2, Lawrence Erlbaum, Hillsdale, NJ, (1992).

[13] P. Shammi and D.T. Stuss, 'Humor appreciation: a role of the right frontal lobe', *Brain*, **122**, 657–666, (1999).

[14] Sidney Siegel and N. J. Castellan, Jr., *Nonparametric Statistics for the Behavioural Sciences*, McGraw-Hill, 2nd edn., 1988.

[15] Oliviero Stock and Carlo Strapparava, 'The act of creating humorous acronyms', *Applied Artificial Intelligence*, **19**(2), 137–151, (2005).

[16] David Watson, Lee Anna Clark, and Auke Tellegen, 'Development and validation of brief measures of positive and negative affect: The PANAS scales', *Journal of Personality and Social Psychology*, **54**(6), 1063–1070, (June 1988).

Affect in Metaphor: Developments with WordNet

Tim Rumbell, John Barnden, Mark Lee and Alan Wallington[1]

Abstract. We discuss an aspect of an affect-detection system used in e-drama by intelligent conversational agents, namely affective interpretation of limited sorts of metaphorical utterance. We discuss how these metaphorical utterances are recognized and how they are analysed and their affective content determined.

1 INTRODUCTION

We present one aspect of a system for extracting affective information from individual utterances, for use in text-based intelligent conversational agents (ICAs). Affect includes emotions/ moods (such as embarrassment, hostility) and evaluations (of goodness, importance, etc.). Our own particular ICA [19] is for use in an e-drama system, where human users behave as actors engaged in unscripted role-play. Actors type in utterances for the on-screen characters they control to utter (via speech bubbles). Our ICA is another actor, controlling a bit-part character. Through extracting affect from other characters' utterances it makes responses that can help keep the conversation flowing. The same algorithms are also used for influencing the characters' gesturing (when a 3D animation mode is used). Our ICA is an addition to an e-drama system produced by an industrial collaborator, Hi8us Midlands Limited, and the 3D animation facility was produced by another industrial collaborator, BT. See [4] for more.

The system aspect demonstrated handles one important way in which affect is expressed in most discourse genres: namely metaphor. Only a relatively small amount of work has been done on computational processing of metaphorical meaning, for any purpose, let alone in ICA research. Major work apart from ours on metaphorical-meaning computation includes ([6], [9], [11], [12], [14], [17]). The e-drama genre exhibits a variety of types of metaphor, with a significant degree of linguistic open-endedness. Also, note that our overarching research aim is to study metaphor as such, not just how it arises in e-drama. This increases our need for systematic, open-ended methods. This paper updates [15]. Since that system-demonstration paper important new developments have taken place in the metaphor processing.

2 METAPHOR AND AFFECT

Conveying affect is one important role for metaphor, and metaphor is one important way of conveying affect. Emotional states and behaviour are often themselves described metaphorically ([10]; [7]), as in 'He was boiling inside' [feelings of anger]. But another important phenomenon is describing something X using metaphorical source terms that are subject to that affect, as in 'My son's room [= X] is a *bomb site*' or '*smelly* attitude' (an e-drama transcript example). Such carry-over of affect in metaphor is well recognized, e.g. in the political domain ([13]). We recently conducted a user study of the system (at four secondary schools in the Birmingham area) and (automatically) recorded the different users' (actors') "speeches". An analysis of the resulting transcripts indicate that this type of affect-laden metaphor is a significant issue in e-drama: at a conservative estimate, at least one in every 16 speech-turns has contained such metaphor (each turn is $\leq$ 100 characters, and rarely more than one sentence; 33K words across all transcripts).

There are other specific, theoretically interesting metaphorical phenomena arising in e-drama that are important also for discourse in general, and plausibly could be handled reasonably successfully in an ICA using current techniques. Some are:

1. Casting someone as an animal. This often conveys affect, from insultingly negative to affectionately positive. Terms for young animals ('piglet', 'wolf cub', etc.) are often used affectionately, even when the adult form is negative. Animal words can have a conventional metaphorical sense, often with specific affect, but in non-conventional cases a system may still be able to discern a particular affective connotation; and even if it cannot, it can still plausibly infer that *some* affect is expressed, of unknown polarity (positivity/negativity) sheerly from the fact of using an animal metaphor.
2. Rather similarly, casting someone as a monster or as a mythical or supernatural being, using words such as 'monster', 'dragon', 'angel', 'devil'.
3. Casting someone as a special type of human, using words such as 'baby' (to an adult), 'freak', 'girl' (to a boy), 'lunatic'. These again can have strong (if context-sensitive) affective connotations.
4. Metaphorical use of size adjectives (cf. Sharoff, 2006). Particularly, using 'a little X' to convey affective qualities of X such as unimportance and contemptibility, but sometimes affection towards X, and 'big X' to convey importance of X ('big event') or intensity of X-ness ('big bully'), and X can itself be metaphorical ('baby', 'ape'). Metaphorical use of size adjectives is often combined with phenomena 1-3. In particular, "little X" where X is an animal can lead to a baby-animal interpretation as one possibility.

Currently, our system partially addresses (1), (2) and (4).

3 METAPHOR RECOGNITION AND ANALYSIS

The approach is split into two parts: recognition of potential metaphors; analysis of recognised elements to determine affect.

3.1 The Recognition Component

The basis here is a subset of a list of metaphoricity signals we have compiled [1], by modifying and expanding a list from [8]. The signals include specific syntactic structures, phraseological items and

[1] University of Birmingham, UK, email: a.m.wallington@cs.bham.ac.uk

morphological elements. We currently focus on two special syntactic structures, *X is/are Y* (in which X could be the pronoun 'you') and *You Y*, and some lexical strings such as '[looks] like', 'a bit of a' and 'such a' (these lexical strings can be interpreted as indicative of similes, but we treat similes and metaphors in the same way). The signals are merely uncertain, heuristic indicators. For instance, in the transcripts mentioned in section 2, we judged *X is/are Y* as actually indicating the presence of metaphor in 38 per cent of cases (18 out of 47). Other success rates are: *you Y* - 61 per cent (22 out of 36); *like* (including *looks like*) - 81 per cent (35 out of 43).

In order to detect signals we use the Grammatical Relations (GR) output from the RASP robust parser [3] This output shows typed wordpair dependencies between the words in the utterance. E.g., the GR output for 'You are a pig' is:

```
|ncsubj| |be+_vbr| |you_ppy| |_|
|xcomp| _ |be+_vbr| |pig_nn1|
|det| |pig_nn1| |a_at1|
```

For an utterance of the type *X is/are Y* the GRs will always give a subject relation (`ncsubj`) between X and the verb 'to be', as well as a complement relation (`xcomp`) between the verb and the noun Y. The structure is detected by finding these relations. As for *you Y*, Rasp also typically delivers an easily analysable structure, but unfortunately the POS tagger in Rasp seems to favour tagging Y as a verb. e.g., 'cow' in 'You cow'. Here the robustness of a parser like Rasp causes problems: the main verb sense of 'cow' (overawe: "subdue, restrain, or overcome by affecting with a feeling of awe") is transitive and so would not normally be relevant to 'You cow', but it is desirable for a robust parser to allow deficient grammar. In such a case, our system looks the word up in a list of tagged words that forms part of the RASP tagger. If the verb can be tagged as a noun, the tag is changed, and the metaphoricity signal is deemed detected.

Once a signal is detected, the word(s) in relevant positions (e.g. the Y position) are pulled out to be analysed. This approach has the advantage that whether or not the noun in, say, the Y position has adjectival modifiers the GR between the verb and Y is the same, so the detection tolerates a large amount of variation. Any such modifiers are found in modifying relations and are available for use in the Analysis Component.

3.2 The Analysis Component

3.2.1 Core Processing

The analysis element of the processing takes the X noun (if any) and Y noun and uses WordNet 2.0 to analyse them. First, we try to determine whether X refers to a person (the only case the system currently deals with), partly by using a specified list of proper names of characters in the drama and partly by WordNet processing (The system also proceeds similarly if X is 'you'). If so, then the Y and remaining elements are analysed using WordNet's taxonomy. This allows us to see if the Y noun in one of its senses is a hyponym of animals or supernatural beings. If this is established, the system sees if another of the senses of the word is a hyponym of the person synset, as many metaphors are already given as senses in WordNet. If the given word contains within its senses different senses that are hyponyms of both animal and person, then we search for evaluative content about the metaphor.

Previously our analysis of evaluative content of a metaphor revolved around finding specific indicative synsets in the hypernym tree of the person sense of the given metaphor word. For example, cow in its metaphorical sense has the 'unpleasant person' synset as a lower hypernym than 'person', which we took as an indicator of negativity). But now, in an important new development, instead of relying on the presence of one of a small set of intermediate nodes for affective evaluation, we have developed a method of automatically detecting the orientation of a given metaphorical word.

Intermediate synsets between the metaphorical sense of the given word and the person synsets contain glosses, which are descriptions of the semantic content of a synset. For example, the gloss of the synset of 'shark' that is a hyponym of 'person' is "a person who is ruthless and greedy and dishonest"; that of 'fox' is "a shifty deceptive person". We search the words and glosses from the intermediate synsets for words that indicate a particular affective evaluation. This search is based on another feature of WordNet, as follows.

WordNet contains a 'quality' synset which has 'attribute' links to four other synsets, 'good', 'bad', 'positive' and 'negative'. We are currently only looking for positive or negative affective evaluations, so this group of synsets provides a core set of affect indicating words to search for in the intermediate nodes. This set is expanded by following WordNet's 'see also' links to related words, to produce lists of positivity and negativity indicators. For example, 'bad' has 'see also' links to five synsets, including 'disobedient' and 'evil'; we then look up the 'see also' links in these five synsets and include these related words in the 'bad' list, and so on, through five iterations, producing a list of over 100 words related to 'bad', and therefore indicating negativity. We search through the words and glosses from the intermediate nodes between the given metaphor synset (arising from the Y component in the sentence) and 'person', tallying the positivity and negativity indicating words found. We can then assign the affective evaluation of the metaphor, so more negativity indicators than positivity indicators suggests that, when the word is used in a metaphor, it will be negative about the target. If the numbers of positivity and negativity indicators are equal, then the metaphor is labeled positive or negative, implying that it has an affective quality but we cannot establish what.

This label is also used in those examples where an animal does not have a metaphorical sense in WordNet as a kind of person (for example, 'You elephant'). See the comment at the end of case 1 in section 2.

3.2.2 Young Animals and Size Adjectives

There is a further complication. Baby animal names can often be used to give a statement a more affectionate quality. Some baby animal names such as 'piglet' do not have a metaphorical sense in WordNet. In these cases, we check the word's gloss to see if it is a young animal and what kind of animal it is (the gloss for piglet, for example, is "a young pig"). We then process the adult animal name to seek a metaphorical meaning but add the quality of affection to the result. A higher degree of confidence is attached to the quality of affection than is attached to the positive/negative result, if any, obtained from the adult name. Other baby animal names such as 'lamb' do have a metaphorical sense in WordNet independently of the adult animal, and are therefore evaluated by means of the Core Processing in section 3.2.1. They are also tagged as potentially expressing affection but with a lesser degree of confidence than that gained from the core processing of the word. However, the youth of an animal is not always encoded in a single word: e.g., 'cub' may be accompanied by specification of an animal type, as in 'wolf cub'. An extension to our processing would be required to handle this and also cases like 'young wolf' or 'baby wolf'.

If any adjectival modifiers of the Y noun were recognized the analyser then goes on to evaluate their contribution to the metaphor's affect. If the analyser finds that 'big' is one of the modifying adjectives of the noun it has analysed the metaphor is marked as being more emphatic. If 'little' is found the following is done. If the metaphor has been tagged as negative and no degree of affection has been added (from a baby animal name, currently) then 'little' is taken to be expressing contempt. If the metaphor has been tagged as positive OR a degree of affection has been added then 'little' is taken to be expressing affection. These additional labels of affection and contempt are used to imply extra positivity and negativity respectively.

4 EXAMPLES OF COURSE OF PROCESSING

In this section we discuss two examples in detail and seven more with brief notes. The examples are mainly from [15] but we have updated several to take account of the gloss-based processing mentioned in section 3.2.1. The first two, the detailed examples, outline the flow of processing and highlight the key analytical decisions made.

4.1 'You piglet'

1. The metaphor detector recognises the *You Y* signal and tags the noun 'piglet' as the *Y* word.
2. The metaphor analyser reads 'piglet' from as *Y* and detects that it is a hyponym of 'animal'.
3. 'Piglet' is not encoded with a specific metaphorical meaning ('person' is not a hypernym). So the analyser retrieves the gloss from WordNet.
4. It finds 'young' in the gloss and retrieves all of the words that follow it. In this example the gloss is 'a young pig' so 'pig' is the only following word. If more than one word is following, then the analysis process is repeated for each of the words following 'young' until an animal word is found.
5. The words and glosses of the intermediate nodes between 'pig' and 'person' contain 0 positivity indicating words and 5 negativity indicating words, so the metaphor is labelled with negative polarity.
6. This example would result in the metaphor being labeled as an animal metaphor which is negative but affectionate with the affection label having a higher numerical confidence weighting than the negative label.

4.2 'Lisa is an angel'

1. The metaphor detector recognises the *X is a Y* signal and tags the noun 'angel' as the metaphor word. 'Lisa' is recognised as a person through a list of names provided with the individual scenarios in e-drama.
2. The metaphor analyser finds that it is a hyponym of 'supernatural being'.
3. It finds that in another of its senses the word is a hyponym of 'person'.
4. The words and glosses of the intermediate nodes between 'angel' and 'person' contain 8 positivity indicating words and 0 negativity indicating words, so the metaphor is labeled with positive polarity.
5. This example would result in the metaphor being labeled as a supernatural being metaphor that is positive.

4.3 Other examples

The following are further examples to show some of the ways in which particular types of utterance are analysed.

1. 'You cow': this is processed as a negative animal metaphor. The synset of 'cow' that is a hyponym of 'person' has the gloss "a large unpleasant woman". Interestingly, 'large' is included in the list of positivity indicators by the current compilation method, but the negativity of the metaphor is confirmed by analysis of the intermediate synsets between 'cow' and 'person', which are 'unpleasant woman', 'unpleasant person' and 'unwelcome person'. These synsets, along with their glosses, contain six negativity indicators, against just the one positivity indicator.
2. 'You little rat': this animal metaphor is determined as negative, having three senses that are hyponyms of 'person', containing three positivity indicators and five negativity indicators. 'Little' provides an added degree of contempt.
3. 'You little piggy': 'piggy' is recognized as a baby animal term and labeled as expressing affection. The evaluation of 'pig' adds a negative label, with no positivity indicators and three negativity indicators, and 'little' adds further affection since the metaphor already has this label from the baby animal recognition. This is therefore recognized as a negative metaphor but meant affectionately.
4. 'You're a lamb': recognized as an animal metaphor and a young animal. It has an 'affectionate' label and is recognized as a positive metaphor, with its two senses that are hyponyms of 'person' contributing two positivity indicators and one negativity indicator. The negative word in this case is 'evil', coming from the gloss of one of the intermediate synsets, 'innocent': "a person who lacks knowledge of evil". This example highlights a failing of using individual words as indicators: negations within sentences are currently not recognized.
5. 'You are a monster': one sense of monster in WordNet is a hyponym of animal. Therefore, this is recognized as an animal metaphor, but affect evaluation reveals three negativity and three positivity indicators, so it is analysed as 'positive or negative'. These indicators are found in two opposed senses of monster: 'monster, fiend, ogre': "a cruel wicked and inhuman person" (analysed as negative); and 'giant, monster, colossus': "someone that is abnormally large and powerful" (analysed as positive, due to the indicators 'large' and 'powerful').
6. 'She's a total angel': a positive supernatural being metaphor, with eight positivity indicators and no negativity indicators from two senses that are hyponyms of 'person', but currently 'total' makes no contribution.
7. 'She is such a big fat cow': a negative animal metaphor made more intense by the presence of big. It has an extra level of confidence attached to its detection as two metaphoricity signals are present but currently 'fat' makes no contribution.

5 FUTURE WORK

Work is ongoing on the four specific metaphorical phenomena listed in section 2 as well as on other phenomena, such as the variation of conventional metaphorical phraseology by synonym substitution and addition of modifying words and phrases, and interpretation of metaphorical descriptions of emotions. We are also looking to broaden metaphor detection, such that, in the case *X is/are Y*, if a hypernym of X is an 'artifact' and of Y is a 'living thing' (or vice-versa) then a metaphor is implied.

Observe that we do not wish simply to 'precompile' information about animal metaphor (etc.) by building a complete list of animals (etc.) in any particular version of WordNet (and also adding the effects of potential modifiers such as 'big' and 'little'). This is because we wish to allow the work to be extended to new versions of WordNet and to generalize as appropriate to thesauri other than WordNet, and because we wish to allow ultimately for more complex modification of the *Y* nouns, in particular by going beyond the adjectives 'big' and 'little'. We recognize that the current counting of positive and negative indicators picked up from glosses is an over-simple approach, and that the nature of the indicators should ideally be examined.

The paper has discussed a relatively 'shallow' type of metaphor processing, although our use of robust parsing and complex processing of a thesaurus take it well beyond simple keyword approaches or bag-of-words approaches. In future work we wish to integrate the processing we have described with the deep semantic/pragmatic reasoning-based approach in our ATT-Meta project [2]. Note also that the carry over of affect in animal (etc.) metaphor as treated above is a special case of a much more general carry-over phenomenon that is central to the ATT-Meta approach (cf. its "view-neutral mapping adjuncts" feature).

Our reason for not reporting full evaluations at this stage is the amount of extensions ongoing or envisioned. We have extra work to do on the system and as such it is premature to engage in a large scale evaluation.

6 RELATED WORK

WordNet glosses have been used elsewhere to extract additional information about metaphor. Veale [18] describes an approach to qualia extraction from WordNet glosses, by attempting to extract relational structure inherent from them. Similarities between this approach to glosses and our own can be found in that glosses are used to determine compatibility between concepts, finding structures not explicitly encoded in WordNet already. Veale also highlights the use of finding relevant analogues of particular concepts to find relations with other concepts, similar to our method of relating words from glosses to the concepts of positivity and negativity. Of course, Veale uses these techniques to understand similarities in relational structure in metaphors and we use them only to determine an affective component. On the other hand, our processing is for extracting information from sentences, whereas Veale's is not directly applied to this.

Other systems of affectively labeling WordNet synsets have been developed. SentiWordNet [5] and WordNet-Affect [16] both annotate synsets with affective labels, and SentiWordNet uses the information available in glosses to do so. Transferring our work to work in these systems could be useful; our processing could potentially fill gaps in these systems. Our processing could still be useful were this not the case, because of our interest, mentioned above, in generalizing the work to non-WordNet resources.

7 CONCLUSIONS

The processing capabilities described make particular but nonetheless valuable and wide-ranging contributions to affect-detection for ICAs. Although designed for an e-drama system, the techniques plausibly have wider applicability. That is, it is both the case that animal, supernatural-being and big/little metaphors appear in many other genres, and that the techniques we have developed for such metaphor can plausibly be generalized to work for a variety of other types of metaphor. The development of the processing in a real-life application is also enriching our basic research on metaphor.

ACKNOWLEDGEMENTS

The research was supported in part by an ESRC/EPSRC/DTI Paccit LINK grant (ESRC RES-328-25-0009) and by EPSRC grant EP/C538943/1. We are indebted to our industrial partners (Hi8us Midlands Ltd, Maverick TV Ltd, and BT), to our colleagues Li (Jane) Zhang and Catherine Smith for intesnive system-development work, and to other colleagues on the e-drama project (Rodrigo Agerri, Sheila Glasbey, Bob Hendley and William Edmondson).

REFERENCES

[1] http://www.cs.bham.ac.uk/ jab/ATT-Meta/metaphoricity-signals.html.
[2] R. Agerri, J. Barnden, M. Lee, and A. Wallington, *Metaphor, Inference and Domain-Independent Mappings*, Proceedings of the Int. Conf. Recent Advances in Natural Language Processing (RANLP'07), Borovets, Bulgaria, 2007.
[3] E. Briscoe, J. Carroll, and R. Watson, *The Second Release of the RASP System*, Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, 2006.
[4] K. Dhaliwal, M. Gillies, J. O'Connor, A. Oldroyd, D. Robertson, and L. Zhang, *Drama: Facilitating Online Role-play Using Emotionally Expressive Characters*, 195–202, Artificial and Ambient Intelligence: Proceedings of the AISB Annual Convention, 2007.
[5] A. Esuli and F. Sebastiani, *SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining*, Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), Genova, Italy, 2006.
[6] D. Fass, *Processing Metaphor and Metonymy*, Ablex, Greenwich, Connecticut, 1997.
[7] S. Fussell and M. Moss, *Figurative Language in Emotional Communication*, 113–135, Social and Cognitive Approaches to Interpersonal Communication, Lawrence Erlbaum Associates, 1998.
[8] A. Goatly, *The Language of Metaphors*, Routledge, London, 1997.
[9] J. Hobbs, *Literature and Cognition*, 21, CSLI Lecture Notes, Centre for the Study of Language and Information, Stanford University, 1990.
[10] Z. Kövecses, *Metaphor and Emotion: Language, Culture and Body in Human Feeling*, Cambridge University Press, 2000.
[11] J. Martin, *A Computational Model of Metaphor Interpretation*, Academic Press, San Diego, CA, 1990.
[12] Z. Mason, 'A computational, corpus-based conventional metaphor extraction system', *Computational Linguistics*, **30(1)**, 23–44, (2004).
[13] A. Mussolf, *Metaphor and political discourse: Analogical reasoning in debates about Europe*, Palgrave Macmillan, Basingstoke, UK, 2004.
[14] S. Narayanan, *Moving Right Along: A Computational Model of Metaphoric Reasoning About Events*, 121–18, Proceedings of the National Conference on Artificial Intelligence, AIII Press, 1999.
[15] C. Smith, T. Rumbell, J. Barnden, B. Hendley, M. Lee, and A. Wallington, *Don't Worry About Metaphor: Affect Extraction for Conversational Agents*, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007), 2007.
[16] C. Strapparava and V. Valitutti, *WordNet-Affect: An Affective Extension of WordNet*, 1083–1086, Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, 2004.
[17] T. Veale, *'Just in Time' Analogical Mapping, An Iterative-Deepening Approach to Structure-Mapping*, Proceedings of the 13th European Conference on Artificial Intelligence (ECAI'98), John Wiley, 1998.
[18] T. Veale, *Systematicity and the Lexicon in Creative Metaphor*, Proceedings of the ACL Workshop on Figurative Language and the Lexicon, the 2003 Applied Computational Linguistics Conference (ACL 2003), Sapporo, Japan, 2003.
[19] L. Zhang, J. Barnden, B. Hendley, and A. Wallington, *Exploitation in Affect Detection in Improvisational E-Drama*, 68–79, Lecture Notes in Computer Science, Springer, Amsterdam, 2006. Also as Report No. 17/92, Computing Laboratory, University of Kent at Canterbury, U.K.

Simulating emotional reactions in medical dramas

Sandra Williams and **Richard Power** and **Paul Piwek** [1]

Abstract.

Presenting information on emotionally charged topics is a delicate task: if bare facts alone are conveyed, there is a risk of boring the audience, or coming across as cold and unfeeling; on the other hand, emotional presentation can be appropriate when carefully handled, but when overdone or mishandled risks being perceived as patronising or in poor taste. When Natural Language Generation (NLG) systems present emotionally charged information linguistically, by generating scripts for embodied agents, emotional/affective aspects cannot be ignored. It is important to ensure that viewers consider the presentation appropriate and sympathetic.

We are investigating the role of affect in communicating medical information in the context of an NLG system that generates short medical dramas enacted by embodied agents. The dramas have both an informational and an educational purpose in that they help patients review their medical histories whilst receiving explanations of less familiar medical terms and demonstrations of their usage. The dramas are also personalised since they are generated from the patients' own medical records. We view generation of natural/appropriate emotional language as a way to engage and maintain the viewers' attention. For our medical setting, we hypothesize that viewers will consider dialogues more natural when they have an enthusiastic and sympathetic emotional tone. Our second hypothesis proposes that such dialogues are also better for engaging the viewers' attention.

As well as describing our NLG system for generating natural emotional language in medical dialogue, we present a pilot study with which we investigate our two hypotheses. Our results were not quite as unequivocal as we had hoped. Firstly, our participants did notice whether a character sympathised with the patient and was enthusiastic. This did not, however, lead them to judge such a character as behaving more naturally or the dialogue as being more engaging. However, when pooling data from our two conditions, dialogues with versus dialogues without emotionally appropriate language use, we discovered, somewhat surprisingly, that participants did consider a dialogue more engaging if they believed that the characters showed sympathy towards the patient, were not cold and unfeeling, and were natural (true for the female agent only).

1 INTRODUCTION

Consider the following three extracts of interactions between a senior nurse and a junior (student) nurse in medical dramas generated by our system:

A **Senior:** Radiotherapy targets cancer cells.
Junior: Cool!

B **Senior:** Anaemia is a condition in which patients feel very tired and may become breathless.
Junior: Right.

C **Junior:** So, let's hope that the packed red cell transfusion took care of the anaemia.
Senior: Yes.

How might viewers perceive the junior nurse's reactions? To the answer in A, the junior responds enthusiastically, perhaps excited by the medical technology, whereas to the one in B, the junior responds more neutrally, perhaps indirectly showing awareness of the patient's discomfort. In C, the junior's summary could be perceived as sympathetic to the patient. Of course, the response in A and summary in C might be perceived as sarcastic and the response in B as unfeeling. If a more direct empathetic response had been attempted in B, e.g., "Oh dear!", or "That's bad!", then it might be perceived as more natural, but it could also be interpreted as patronising or unprofessional. Interestingly, if there were no response at all, the characters might also come across as cold and unfeeling, whilst an inappropriate enthusiastic response such as D might make the characters appear macabre:

D **Senior:** A radical mastectomy is an operation to remove the breast.
Junior: Cool!

We are exploring the simulation of emotions in such responses and their effect on viewers's perceptions of the attitudes of the embodied agents. Our hope is that by generating dialogues in which the characters produce language that is sympathetic to the viewer/patient and enthusiastic about medical technology where appropriate, this will lead to:

- viewers perceiving the dialogue as natural/appropriate;
- engaging the attention of the viewers.

The presentation of emotionally charged information is fraught with difficulties, particularly if the viewer is the patient whose medical record is being discussed (as is our ultimate aim). Our hypotheses connect specific ways of presenting medical information that take emotion into account with perceived naturalness of the resulting dialogues and also the extent to which the dialogues are engaging. The two hypotheses are linked by the underlying idea that appropriate emotional responses will make the dramas more engaging: the viewers' attention will be captured, forcing them to listen more carefully to the interchanges and soak up medical information in the process.

In this preliminary work, we limited our study to enthusiastic responses such as the one in A, neutral responses such as the one in B and sympathetic summaries such as "So, let's hope that the packed red cell transfusion took care of the anaemia". We modified our system to produce such responses in generated dialogue and conducted

[1] The Open University, Walton Hall, Milton Keynes, MK7 6AA, U.K., E-mail: s.h.williams@open.ac.uk, r.power@open.ac.uk, p.piwek@open.ac.uk

a pilot study to elicit viewers' perceptions of two conditions: (a) with emotional responses and summaries and (b) with no responses and neutral summaries (see the Appendix).

2 THE MEDICAL DRAMAS

Our generated medical dramas present a discussion between a senior and junior nurse about a patient's medical record (the system has access to a simulated repository of breast cancer patients' medical records). The senior nurse asks the junior to read the patient's notes for a particular date and, as he reads the notes, the junior nurse also asks questions about medical terms; the senior explains these terms and elaborates on the various medical investigations and interventions that the patient underwent. Consequently, our system generates a type of tutorial dialogue in which the senior nurse is tutor and the junior is student.

The main difference of our approach with other work on tutorial dialogue (e.g., [19]) is that we generate both sides of the conversation as a drama script, just as one might generate a linear text. The differences from generating monologue are that we need to simulate the kinds of questions, answers and explanations that would take place in a dialogue between a tutor and student. One advantage is that we can explore generation of the language of dialogue turns without any necessity for natural language understanding, which would be required in conventional natural language dialogue systems where only half of the conversation is machine-generated.

An obvious consequence is that the user is a viewer, not a participant in the dialogue or the drama. Since the viewer is one step removed she cannot pose her own questions to the system. This might appear a disadvantage on first consideration but it is actually an advantage, for two reasons. First, students rarely have the ability to ask good questions, although they can be taught how ([6]). The viewer can learn from watching the drama unfold, and one important motivation for presenting a tutorial dialogue drama is to demonstrate to viewers how to ask questions. Our aim to provide them with an experience from which they can learn vicariously not only the answers to the questions, but also how to ask questions of their own — a benefit of presentations in dialogue form that has been demonstrated in previous work (e.g., [4, 3]). Second, researchers have found that when people interact with screen characters, they have false expectations of human-like qualities which the characters cannot fulfil, and that sometimes characters can make them feel stupid (see [14]). There is thus a danger that an interactive experience could be frustrating or annoying, so we think our aims are better met by a presentation in which the patient views a video of characters interacting with each other.

Our first pilot experiment was with a version of our system in which medical information was presented as a bare sequence of question and answer dialogue turns with no reactions to the information being presented. Eleven participants listened to a dialogue and a monologue generated from the same underlying electronic health record; they answered some comprehension and preference questions and wrote comments [18]. There was no difference in comprehension or preferences, however; the main comment was that the medical information was too closely packed, so that people had difficulty following it. We came up with a number of solutions for spacing out the medical information and presenting it more slowly. The solution that we will highlight in this paper is that of adding affective reactions to the medical information (other solutions will be reported elsewhere).

Figure 1. Screen shot from an output video.

3 THE SYSTEM

Our NLG system is a data-to-dialogue system — that is, the input is data and the output is a script for a dialogue. It builds a dialogue by querying a simulated relational database of breast cancer patients' medical records; builds concept graphs from the query results (a fragment of a concept graph is shown in Figure 2); adds questions and dictionary definitions to the original concept graph (Figure 3); plans dialogue turns; and realises them as a script for an embodied agent drama. The script is then performed using Loquendo text-to-speech software and Cantoche LivingActor^{TM} character animation (a screen shot of the output is shown in Figure 1). The system is described in more detail in [18].

Figure 2. Part of a concept graph built from data retrieved from a database of medical records.

Figure 2 depicts two concepts, a medical intervention and a medical problem, linked by an arrow representing an INDICATED BY relation between them. The meaning can be paraphrased as "anaemia motivated a packed red cell transfusion". A content planner in the NLG system augments this structure by adding questions and definitions from the system's dictionary of medical terms. Figure 3 illustrates how these would be added to the fragment in Figure 2; the rectangles and arrow from Figure 2 are shown greyed-out and new rectangles representing questions, a definition from the dictionary, and an attribute of the definition, are shown in black.

3.1 Defining medical terms

Our planner adds explanations of medical terms only if they have not been mentioned previously in the dialogue, and only if they are relatively rare in everyday language. Our information on term frequencies was derived from searches of the British National Corpus, a 100 million word corpus of British English (www.natcorp.ox.ac.uk). Our

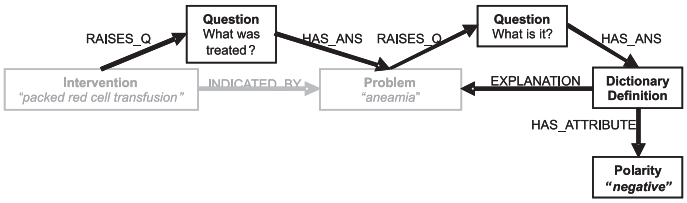


Figure 3. The graph augmented with questions and dictionary definitions.

searches revealed that medical terms such as "anaemia" and "axilla" are infrequent in the BNC with 362 and 3 occurrences, respectively, so these are defined, whereas "breast" was more frequent with 1,615 occurences, therefore it is not defined. However, BNC frequencies did not always coincide with our intuitions about whether people would know a term, for instance, "armpit" only occurs 76 times in the BNC, even though we believe that it is a well-known term. Consequently, we were guided by the BNC, but rather than following a rigid rule to define all terms within a fixed range of BNC frequencies (e.g., 0 to 1,000), we were also guided by our intuitions. In effect we took the medical terms that had a low frequency in the BNC and then selected a subset that we deemed suitable for explanation.

When the content planner adds an explanation of a medical term, it looks it up the definition in its medical term dictionary. The system's dictionary is a text file of definitions that we found on trusted Web sites such as www.cancerresearchuk.org; some fragments of the dictionary follow:

```
TERM anaemia
DEF NPS NEG a condition in which patients feel
 very tired and may become breathless
CAUSE S NEG the blood has fewer red blood
 cells than normal

TERM axilla
DEF NPS NEUTRAL the armpit

TERM CTScan
DEF NPS POS an X-ray scan using a computer
 to construct pictures of the body in cross
 section
```

Here, definitions for the terms "anaemia", "axilla" and "CTScan" are shown. The keyword TERM indicates the beginning of a new term and it is followed by a string containing the term. DEF and CAUSE indicate the beginning of a term's definition and cause (if any), NPS and S are syntactic categories (singular noun phrase and sentence), POS, NEG and NEUTRAL indicate the *polarity* of the definition or cause.

By polarity, we mean whether the definition or cause conveys information that is potentially beneficial, neutral, or detrimental from a patient's point of view. Remember that the medical records input to our system are simulated from patients who have breast cancer. Medical procedures such as radiation therapy or chemotherapy that destroy cancer cells are assigned positive polarity (as is CTScan in the fragment above). Obviously this is a somewhat naive view since although some medical technologies can potentially help patients, some also have unpleasant side effects. Negative polarities are assigned typically to definitions of illnesses, such as anaemia, which describe symptoms that the patient suffers from.

3.2 Adding emotional responses and summaries

When a definition is added to the dialogue, a definition phrase is placed in a template that matches its syntactic category and a response is constructed that accords with its polarity. In a previous version of the system, medical information was presented through sequences of question-answer pairs, that is, questions about entities or relations from the simulator database and answers giving definitions. The new strategy presents the information as question-answer-response triples. These have the effect of slowing down the rate of communication of information, as suggested by our pilot experiment reported in [18]. The NLG system adds a definition question such as "Anaemia?", or "What is anaemia", or "What is that?", along with an answer that includes the negative polarity dictionary definition "anaemia is a condition in which patients feel very tired and may become breathless". Then it adds the emotional response: a neutral response for a neutral or negative polarity definitions is randomly chosen from "right". "okay", and "I see". A positive response for a positive polarity definition is randomly selected from "cool!", "amazing!" "I never knew!", and "just imagine it!". These come with Loquendo text-to-speech software as pre-recorded phrases; we chose these particular ones because their intonation accorded with the emotions that we wanted to convey, i.e., enthusiasm or concern.

The content planner also adds summaries of each medical episode (intervention or investigation) in the patient's record. These clarify and repeat the information. Summaries are of two kinds:

- **Authoratative.** Senior nurse summaries, e.g., "So, a packed red cell transfusion was administered to treat the anaemia."
- **Emotional.** Junior nurse summaries, e.g., "So, let's hope that the packed red cell transfusion took care of the anaemia."

Each embodied agent has a number of built-in gestures that can be associated with textual utterances so that a gesture will play at roughly the same time as a phrase is spoken. However, with Cantoche agents, synchronisation of speech and gestures cannot be fine-tuned to the extent where a gesture can be played to emphasise an individual word or syllable. Three types of gestures are generated by our current system: (a) generated randomly from a small set of fairly neutral speech gestures, e.g., a small raise of the hand, (b) nods or shakes of the head to accompany "yes" or "no" utterances, and (c) the junior nurse takes out a clipboard and reads from it when the senior nurse asks him a question about the patient's medical record.

4 RELATED WORK

The automated generation of dialogue scripts was pioneered by Elisabeth André and collaborators [1]. Extending this work, in the NECA project, script generation was brought together with multimodal NLG [10] and emotive speech synthesis [16] resulting in Fully Generated Scripted Dialogue (FGSD) [17]. The NECA system has a number of important similarities and differences with the current system. First of all, although the NECA platform was domain-independent, the domains to which it was applied, car sales (eShowroom) and social chat (Socialite), put demands on information presentation quite different from those in the medical domain.

Let us illustrate how evaluative comments are dealt with in NECA, following the approach explored first in [1], using the car sales domain. In the domain model, the values of attributes of cars (e.g., horse power, top speed) are given a valence (positive or negative) for each of the dimensions that a potential car buyer might be interested in

(i.e., sportiness, family friendliness, etc.). The system generates dialogues between a virtual car seller and buyer. They might discuss a particular attribute of a car that the user is interested in. Depending on the valence of the attribute and the attribute value, the system can generate evaluative remarks by the buyer character depending on the dimension that interest her (these can be selected by the user). For example, a seller and buyer might discuss the top speed of a particular car with the buyer asking for the top speed and the seller answering 'It has a top speed of 180 mph'. Depending on whether the buyer is interested in environmental friendliness or sportiness of cars, she might then respond with either, for instance, 'Interesting, but isn't that bad for the environment?' or 'Great, that's a very fast car!'.

A difference with the current medical scenario is that whereas in the NECA domains positive/negative valence translates directly to a positive/negative comment (though it is modulated by the personality of the character), in our junior/senior nurse dialogues there is an asymmetry between positive and negative polarity definitions: whereas definitions with a positive polarity attract a positive response, definitions with a negative polarity lead to a neutral response. The rationale is that with the viewer being the patient, emphasing negative information is emotionally insensitive: the aim is to avoid upsetting the viewer and to show sympathy and a positive attitude (enthusiasm) whereever possible and avoid negative emotions.

A further difference is that the ability of the NECA system to generate evaluative remarks was never evaluated; in particular, its relation to naturalness and engagement were not empirically tested. The nearest evaluation of affective natural language in NECA concerned a comparison of two referring expression generation strategies, one for egocentric and one for neutral speakers (see [12]).

More closely related to the current medical domain, the Text-to-Dialogue (T2D) system ([11]) generates dialogue scripts for two computer-animated characters – a pharmacist and a client. T2D, however, generates the scripts from textual input (Patient Information Leaflets) rather than data. Both approaches build on the idea put forward in [13] that (rhetorical) relations between spans of text or data often lend themselves for presentation through dialogue patterns – for example, a causal relation between informational items A and B can be expressed in a dialogue between layman L and expert E of the form $L : Why\ A?\ E : Because\ B$.

In recent years, the topic of affective NLG, in particular for embodied agents, has attracted a lot of interest (see [9] for an overview of work up to 2003; and [2, 15] for collections of papers on embodied agents including a number on generation of affective language). One of the early embodied agents for medical applications, Greta, is described in [8]. Greta is an embodied conversational agent that can play the role of doctor in information-delivering dialogues with patients. It integrates BDI (belief, desire and intention) planning with affective state generation and recognition, and makes use of sophisticated integrated realization of language and gestures that is sensitive to the emotions of the patient. The main difference with our approach is that it aims at direct interaction with the user through dialogue, rather than the use of dialogue between two virtual characters as a means for information delivery. Whereas the Greta agent takes into account whether it is speaking with a patient or a doctor (adjusting its display of emotions accordingly), it does not factor in the possibility of an overhearer who might listen in on a conversation between two doctors, and thus influence their use of language.

The 'Carmen's Bright Ideas' system ([7]) occupies the middle ground between interactive systems, such as Greta, and our system which is aimed purely at *presenting* dramatic dialogue. Carmen's Bright Ideas is intended for parents of children with cancer. It interactively generates dialogues between animated characters using pre-recorded speech. User have some control through clicking on alternative emotional "thought balloons", though the overall storyline is maintained by a director module. This system was subject to a trial in which it replaced a research assistant who was teaching Bright Ideas (a self-help philosophy) to sixteen learners in some of their sessions. Learners responded positively to questions about the helpfulness and clarity of the system.

5 EXPERIMENT

5.1 Materials

We generated a medical drama script from one patient's (simulated) data. The script – see the appendix for the complete script – contained the kinds of emotional reactions to medical information described above. We manually cut out some of the script so that it lasted approximately three minutes (in practice, we cut out repetitions of medical investigations and interventions, e.g., a cancer patient who undergoes chemotherapy often becomes anaemic and consequently has many blood tests and blood transfusions to correct this condition; in such cases we only kept the first occurrence of each type of investigation and intervention). We then recorded a video of the embodied agents "acting" the drama which was shown to participants in the "emotional reactions" group.

A second script was made by manually editing the first one. All emotional reactions to medical information were cut out and emotional summaries made by the junior nurse were replaced with neutral ones, e.g., "So, let's hope that the packed red cell transfusion took care of the anaemia" was replaced with the unemotional "So, a packed red cell transfusion was administered to treat the anaemia". Another video was recorded as before and it was shown to participants in the "no reactions" group.

We designed an on-line questionnaire to elicit judgements about nine statements with an on-line survey tool (www.surveymonkey.com). The statements were arranged into three groups, each on a separate Web page, and a final page where participants could type comments, as follows:

Page 1: The video captured my attention.

Page 2: The woman behaved naturally.
The woman sympathised with the patient.
The woman was cold and unfeeling.
The woman was enthusiastic about medical facts.

Page 3: The man behaved naturally.
The man sympathised with the patient.
The man was cold and unfeeling.
The man was enthusiastic about medical facts.

Page 4: Free text comments.

A set of judgements was associated with each statement ("Strongly disagree", "Disagree", "Disagree a bit", "Don't know", "Agree a bit", "Agree" and "Strongly agree") from which participants were able to select only one. Each judgement was associated with a numerical value on a Likert scale ranging from 1 = "Strongly disagree" to 7 = "Strongly agree".

5.2 Participants

Forty adults, thirty-two females and seven males, who are known by the first author, were invited to participate. They were randomly allocated to one of the two groups, "emotional reactions" or "no reactions", and were sent an e-mail asking them to participate and directing them to a Web page containing the materials relating to their group's condition. Thirty people completed the questionnaire.

5.3 Method

The participants watched a video on a Web site; they were able to view it as many times as they liked. Following successful viewing, they were redirected to another Web site where they were invited to respond to each of the above statements by selecting one judgement. The on-line questionnaire was set up so that participants could not proceed unless they selected a judement for each statement. Their selections were recorded as numerical values on a Likert scale as above. Responses to the questionnaire were collected anonymously by the on-line survey tool (www.surveymonkey.com). The tool records I.P. addresses and does not allow submission of more than one questionnaire from an I.P. address. Since the participants were known to us and because most of them also sent personal e-mails to let us know that they had completed the questionnaire, we are confident that the twenty-eight responses that we received are genuine and valid.

5.4 Results

The main issue is whether the inclusion of emotional reactions influenced viewers' judgements about (a) their interest in the video and (b) the attitudes and behaviour of the embodied characters. Table 1 shows mean judgements for each statement by the two groups (emotional reaction present/absent). As can be seen, the groups gave similar positive judgements on whether the video held their attention (5.13 vs 5.43, n.s.). However, significant differences (independent samples t-test) were found for two judgements (starred): when the man (the junior nurse) gave emotional reactions he was perceived as being more sympathetic towards the patient (4.88 vs 3.57, $p < 0.015$) and more enthusiastic about medical facts (5.06 vs 3.50, $p < 0.003$). Since the woman (the senior nurse) uttered very few emotional responses (apart from agreeing occasionaly with the junior nurse's hope that the treatment worked), we did not expect significant differences between the two conditions in perception of her attitudes.

Table 1. Mean judgements ranging over values from 1 (strongly disagree) to 7 (strongly agree).

Statement	Emotional Reaction n=16	No Reaction n=14
Video captured my attention	5.13	5.43
Woman behaved naturally	4.19	5.14
Woman sympathised with patient	4.31	4.00
Woman cold and unfeeling	2.94	2.71
Woman enthusiastic about medical facts	5.44	4.93
Man behaved naturally	4.06	3.57
Man sympathised with patient*	4.88	3.57
Man cold and unfeeling	2.94	2.93
Man enthusiastic about medical facts*	5.06	3.50

Table 2. Frequencies for Agree, Disagree, Don't know (n=30)

Statement	Agree	Disagree	Don't know
Video captured my attention*	25 (83%)	5 (17%)	0
Woman behaved naturally	20 (67%)	10 (33%)	0
Woman sympathised with patient	10 (33%)	9 (30%)	11 (37%)
Woman cold and unfeeling*	6 (20%)	22 (73%)	2 (7%)
Woman enthusiastic about * medical facts*	24 (80%)	4 (13%)	2 (7%)
Man behaved naturally	12 (40%)	17 (57%)	1 (3%)
Man sympathised with patient	11 (37%)	9 (30%)	10 (33%)
Man cold and unfeeling*	4 (13%)	22 (73%)	4 (13%)
Man enthusiastic about medical facts	18 (60%)	10 (33%)	2 (7%)

The results also show some tendencies that were common to the two groups. Table 2 gives frequencies for positive, negative and neutral responses to the statements, with data pooled so that each row sums to the total number of subjects (30). A judgement is classified as positive (Agree) if it lies in the range 5-7, negative (Disagree) if it lies in the range 1-3, and neutral (Don't know) if it is equal to 4. Overall there is a slight bias (130 vs. 108) for positive responses over negative; taking this into account, an agree-disagree split of 20:10 (or 10:20) has a probability $p < 0.02$ (binomial test) and a 25:5 split a probability of $p < 0.0004$ (binomial test), the starred comparisons are therefore significant. Inspection of the table reveals the following:

- Overall, the video succeeded in holding the viewers' attention, with responses largely positive.
- The characters were not seen as cold and unfeeling. Both for the woman (senior nurse) and the man (junior nurse), this statement was rejected with a significant split.
- The characters were seen as enthusiastic about medical facts, although this tendency was significant only for the woman. This is unsurprising since it was the woman who explained the medical terms. The perceived enthusiasm of the male was dependent on his emotional responses (see Table 1).
- Viewers were divided over whether the characters behaved naturally, with no significant differences, although neutral responses were rare (only one response in the 'Don't know' column).
- Viewers found it hard to make a judgement over whether the characters were sympathetic towards the patient. Overall, only 32 of 270 responses were 'Don't know', and the probability (binomial test) of obtaining as many as 11/30 such responses is significantly low ($p < 0.05$).

Still with data pooled across the two groups, table 3 shows correlations among the subjects' responses to the statements. Here, we think the point of major interest is the first column showing which judgements about the characters are most strongly related to judgements about whether the video was attention-worthy. The results suggest that the video held a subject's attention more when he/she thought the characters showed sympathy towards the patient, were not cold and unfeeling, and were natural (woman only); these correlations are significant ($p < 0.05$, Pearson two-tailed test).

Finally, free text comments were provided by nine participants. The content of these provided valuable clues to their perception of the agents' behaviour. The persistent questions of the male nurse about the meaning of medical terms motivated three people to note that he appeared remarkably ignorant and for one to comment that he seemed to have a poor grasp of English, or worse, poor comprehension which could be dangerous. One respondent thought the

Table 3. Pearson Correlations, n=30, 2-tailed significance in parentheses, attn = the video captured my attention, w = the female embodied agent, m = the male embodied agent, cold = the agent was cold and unfeeling, enth = the agent was enthusiastic about medical facts, nat = the agent behaved naturally, symp = the agent sympathised with the patient.

	attn	w_cold	w_enth	w_nat	w_symp	m_cold	m_enth	m_nat	m_symp
attn	-	-	-	-	-	-	-	-	-
w_cold	-.452*(.012)	-	-	-	-	-	-	-	-
w_enth	-	-.529**(.003)	-	-	-	-	-	-	-
w_nat	.430*(.018)	-.590**(.001)	.510**(.004)	-	-	-	-	-	-
w_symp	.368*(.046)	-.563**(.001)	.557**(.001)	-	-	-	-	-	-
m_cold	-.434*(.017)	.606**(.000)	-	-.516**(.004)	-	-	-	-	-
m_enth	-	-	-	-	-		-	-	-
m_nat	-	-	-	-	-	-	.483**(.007)	-	-
m_symp	.416*(.022)	-.487**(.006)	.414*(.023)	-	-	-.481**(.007)	.764**(.000)	-	-

wording of some of the male nurse's questions made him sound particularly stupid and suggested alternatives, some of which are already part of our system's set – clearly, rather than selecting the form of questions randomly, in future we should derive a better method for choosing appropriate formulations to suit different dialogue situations. Two people liked the female nurse's explicit definitions of technical terms, but whilst one of them liked the repetitions of definitions, the other thought that these should not be repeated verbatim, but reformulated (this is another good candidate for further investigation, but currently it is beyond the scope of our system). Regarding the video interface, one person liked being able to read the text of the speech from the Cantoche agents' speech bubbles, another disliked the background scene showing a desk and plant and two people had problems with slow download and synchronisation of speech and video – these sometimes occur with poor Internet access and different browser versions.

6 CONCLUSIONS

The fundamental purpose of the video is to instruct — to help patients pick up facts and terminology relevant to their condition. At the same time we obviously aim to avoid boring the patient, or giving offence. We have explored in this study the hypothesis that an instructive video will hold the viewer's attention better if the characters display sympathy for the patient and enthusiasm for the medical information given. The outcome does not directly support this hypothesis. By including emotional reactions by the junior nurse, we obtain a significant increase in subjects' ratings of his enthusiasm and sympathy, but no increase in the rating given to the video (i.e., the judgement on whether it held the attention).

Paradoxically, the correlation data (pooling the groups) seem to tell a different story. Here we find a clear indication that subjects who gave higher ratings for sympathy also gave higher ratings to the video. A possible resolution is that the emotional reactions had some effect in increasing attention to the video, but not large enough to override other influences that might vary considerably across small groups of subjects. In this connection, it is also important to note that in our study we had only a single item for each condition. As pointed out by [5], this calls into question any conclusions one might want to draw regarding the influence of the two conditions, because there is no control for random variations in the material that might have influenced the answers of the participants.

Another curious outcome is that sympathy for the patient, the character trait most influenced by the independent variable (presence/absence of emotional reaction), was also the trait that subjects found hardest to assess: out of a total of 60 responses to the sympathy questions, 21 fell into the 'Don't know' category, which was used only 11 times for all the other responses. It seems that subjects are strongly influenced by whether the characters show sympathy towards the patient, but found this hard to judge from the evidence of the video. Perhaps this was because we deliberately avoided any direct expressions of sympathy, for fear that subtle mistakes in tone might give offence. The lesson from our data is that this problem needs to be addressed, tricky though it is, since appropriate displays of sympathy would increase the viewer's attention to the video and its message.

As a final qualification, we should point out that the subjects in this experiment were not cancer patients. They were therefore judging the video, and its characters, in the role of outsiders with (perhaps) some general interest in medicine, rather than people personally affected by the material. However, our results should generalise to instructive videos ('edutainment') for use in education and training, even though special testing would obviously be needed before presentations of this kind could be used as a resource in treatment.

ACKNOWLEDGEMENTS

We would like to thank the two anonymous reviewers for their comments on a draft of this paper. This research was supported by Medical Research Council grant G0100852 under the e-Science GRID Initiative.

REFERENCES

[1] E. André, T. Rist, S. van Mulken, M. Klesen, and S. Baldes, 'The automated design of believable dialogues for animated presentation teams', in *Embodied Conversational Agents*, eds., Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, 220–255, MIT Press, Cambridge, Massachusetts, (2000).

[2] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, *Embodied Conversational Agents*, MIT Press, Cambridge, Massachusetts, 2000.

[3] R. Cox, J. McKendree, R. Tobin, J. Lee, and T. Mayes, 'Vicarious learning from dialogue and discourse: a controlled comparison', *Instructional Science*, **27**, 431–458, (1999).

[4] S D. Craig, B. Gholson, M. Ventura, and A Graesser, 'Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning', *International Journal of Artificial Intelligence in Education*, **11**, 242–225, (2000).

[5] D. Dehn and S. van Mulken, 'The impact of animated interface agents: a review of empirical research', *Int. J. Human-Computer Studies*, **52**, 1–22, (2000).

[6] A. King, 'Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain', *American Educational Research Journal*, **31**, 338–368, (1994).
[7] S. Marsella, W.L. Johnson, and C. LaBore, 'Interactive pedagogical drama for health interventions', in *11th International Conference on Artificial Intelligence in Education, AIED 2003*, (2003).
[8] C. Pelachaud, V. Carofiglio, B. De Carolis, F. de Rosis, and I. Poggi, 'Embodied Contextual Agent in Information Delivering Application', in *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Bologna, (2002).
[9] P. Piwek, 'An annotated bibliography of affective natural language generation', ITRI Technical Report ITRI-02-02, ITRI, University of Brighton, (2002). Version 3 (2003) Available at http://mcs.open.ac.uk/pp2464/affect-bib.pdf.
[10] P. Piwek, 'A flexible pragmatics-driven language generator for animated agents', in *Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL) (Research Notes)*, pp. 151–154, Budapest, Hungary, (2003).
[11] P. Piwek, H. Hernault, H. Prendinger, and M. Ishizuka, 'T2D: Generating Dialogues between Virtual Agents Automatically from Text', in *Intelligent Virtual Agents: Proceedings of IVA07*, LNAI 4722, pp. 161–174. Springer Verlag, (2007).
[12] P. Piwek, J. Masthoff, and M. Bergenstrahle, 'Reference and Gestures in Dialogue Generation: Three Studies with Embodied Conversational Agents', in *Proceedings of AISB05 Virtual Social Agents Symposium*, University of Herfordshire, (2005).
[13] P. Piwek, R. Power, D. Scott, and K. van Deemter, 'Generating Multimedia Presentations from Plain Text to Screen Play', in *Multimodal Intelligent Information Presentation*, volume 27 of *Text, Speech and Language Technology*, 203–225, Springer, Dordrecht, (2005).
[14] J. Preece, Y. Rogers, and H. Sharp, *Interaction Design: Beyond Human-Computer Interaction*, John Wiley & Sons, New York, NY, 2002.
[15] H. Prendinger and M. Ishizuka, *Life-Like Characters: Tools, Affective Functions, and Applications*, Cognitive Technologies Series, Springer, Berlin, 2004.
[16] Marc Schröder, *Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis*, Ph.D. dissertation, Institute of Phonetics, University Saarland, 2004.
[17] Kees van Deemter, Brigitte Krenn, Paul Piwek, Martin Klesen, Marc Schröder, and Stefan Baumann, 'Fully generated scripted dialogue for embodied agents', *Artificial Intelligence Journal*, **to appear**, (2008).
[18] Sandra Williams, Paul Piwek, and Richard Power, 'Generating monologue and dialogue to present personalised medical information to patients.', in *11th European Workshop on Natural Language Generation*, pp. 167–170, (2007).
[19] Claus Zinn, Johanna D. Moore, and Mark G. Core, *Intelligent Information Presentation*, chapter Intelligent Information Presentation for Tutoring Systems, Kluwer, 2005.

7 APPENDIX

Generated dialogue

The dialogue below was used in the experiment. Emotional responses generated by the system are marked (for the condition with no emotional responses, either there was no response, or alternative unemotional responses were generated as shown).

Senior: Could you read the patient's notes for June 21st.
Junior: She had an excision on June 21st.
Junior: excision?
Senior: An excision is the removal of the cancer along with a border of healthy tissue around it.
Junior: Would you repeat it?
Senior: An excision is the removal of the cancer along with a border of healthy tissue around it.

Junior (emotional response): right

Senior: What was treated?
Junior: A tumour.
Senior: Where was the tumour?
Junior: The tumour was in her left breast.
Senior: So, an excision was administered to treat the cancer.

EITHER
Junior (emotional response): And we hope it worked.
Senior: Yes.
OR
Junior: Uh-huh

Senior: Go on.
Junior: From September 27th to October 25th she had a radiotherapy course.
Junior: radiotherapy course?
Senior: Radiotherapy targets cells with radiation.

Junior (emotional response): Cool!

Senior: So, a radiotherapy course was administered to treat the cancer.

EITHER
Junior (emotional response): And we hope it worked.
Senior: ok.
OR
Junior: Uh-huh

Senior: And then?
Junior: Her record says that she had a packed red cell transfusion.
Junior: packed red cell transfusion?
Senior: A packed red cell transfusion is a transfusion of red blood cells.
Senior: Are you following me?
Junior: ok
Senior: Red blood cells contain haemoglobin which carries oxygen around the body.

Junior (emotional response): I never knew!

Senior: And that treatment was for?
Junior: Anaemia.
Junior: anaemia?
Senior: Anaemia is a condition in which patients feel very tired and may become breathless.

EITHER
Junior (emotional response): okay
Junior (emotional summary): So, let's hope that the packed red cell transfusion took care of the anaemia.
OR
Junior: So, the packed red cell transfusion treated anaemia.

Senior: Uh-huh
Senior: Continue please.
Junior: Her record says that she had an examination.
Senior: Where?
Junior: The axillary lymphnodes.
Junior: What are axillary lymphnodes?
Senior: The axillary lymphnodes are the rounded masses of tissue under the arms containing white blood cells.

Junior (emotional response): just imagine it!

Senior: What did the examination reveal?
Junior: lympha, lympha...
Senior: Does it say lymphadenopathy?
Junior: Yes
Junior: What is that?
Senior: Lymphadenopathy is a swelling of the lymph nodes which the doctor can feel when you are examined.
Junior: What did you say?
Senior: Lymphadenopathy is a swelling of the lymph nodes which the doctor can feel when you are examined.

Junior (emotional response): I see

Senior: So, an examination led to detection of the lymphadenopathy.
Junior: Uh-huh.

"*You make me feel…*": Affective Causality in Language Communication

Andrzej Zuczkowski[1] and **Ilaria Riccioni**[2]

Abstract. In this paper we analyse linguistic structures, such as "You make me feel angry", which imply the possibility that a person can cause another person's affect. We illustrate the concept of "affective causality", based on an implicit-naïve theory of interpersonal relations linked to common sense. We present two theoretical alternative viewpoints (deriving from psychology and psychotherapy) which promote people's emotional autonomy. We propose an alternative linguistic model to causal structures**,** in which a listener's feelings are not causally linked to a speaker's utterances. We also show the relation between affective causality in language and in perception. Finally, we outline some possible uses of our model in relation to Affective Computing.

1 AFFECTIVE COMPUTING AND AFFECTIVE CAUSALITY

Affective Computing [13] deals with how a computer can: (1) recognize emotions; (2) express emotions; (3) demonstrate emotional intelligence; and (4) have emotions. As psychologists, whose main research field is language communication, we focuse on how people talk *about* their own feelings and causally relate them to other people's verbal and/or non verbal behaviour. We shall illustrate a naïve viewpoint of "affective causality" (cf. sections 2-5) vs. a critical one (cf. sections 7.1 and 7.4).

Anyway, the problem of affective causality concerns not only language communication, but also perception, as we shall try to show in sections 7.1 and 7.2, where we shall illustrate the difference between expressive qualities and effect qualities.

Furthemore, in section 7.3 we shall try to explain that there is a strong relation between affective causality in language communication and affective causality in perception: the former depends on and is rooted in the latter.

Therefore, we do believe that our paper can be productive in order to find solutions in relation to the theoretical issues in the above mentioned points (1), (2) and (3): these very theoretical issues will throw light on technical aspects concerning Affective Computing (cf. section 8).

2 AN INTRODUCTORY EXAMPLE : "YOU MAKE ME FEEL…"

As an introductory example, let us suppose that after you (= the Speaker S) have said something to me (= the Listener L), I feel angry and I tell you: "You make me feel angry". By this utterance I mean it is you who aroused my anger; because of what you said to me, I got angry. I also mean that if you had not said to me what you did or if you had said something different, I would not have got angry.

In this way, I attribute some effects which are inside me, internal to me (i.e. my feeling), to some causes which are outside me, external to me (i.e. your words and you). Such an attribution implies my firm belief that after your words I had no other way to react except that of feeling angry**.**

In some contexts, besides my thinking of you as the one who caused those effects, I can also attribute to you the intention of and the responsibility for having caused these effects. My "attributive reasoning" can then be paraphrased in the following way: (1) I feel angry (the effect) because (2) *you* made me feel angry by saying what you said (the causal link between your words and my anger); therefore (3) you *intended* to make me feel angry (intention) and you succeeded in doing so (responsibility).

3 AFFECTIVE CAUSALITY

One of the many ways in which affect can be communicated by people in their written and spoken language can take the following forms:

(i) "You *make me feel* angry/sad/glad/happy…";

(ii) "She *amused* me"; "You're *boring* me"; "He'll *astonish* me".

These two linguistic structures are in fact the main ones we normally use in everyday life in order to *describe* perlocutionary acts [1, 17, 18] which produce effects on affects.

Such verbs (to amuse, to bore, to astonish…) and verbal expression (to make somebody feel…) in utterances (i) and (ii) share a *causal* semantic structure which could be schematized in the following way: somebody causes, caused or will cause a certain affect in somebody else: anger, sadness, joy, boredom, astonishment, happiness, amusement…

We are dealing here with somebody's "causation", "production", "generation", "creation" of something new in somebody else: an affect. For that reason, we propose to call this kind of causation "affective causality".

4 FOCUS ON S

Verbs and verbal expressions of type (i) and (ii) solve in a causal way the problem concerning the relation between S, what s/he says and L: it is S who causes, by saying what s/he says, certain affects in L; it is S who amuses, bores, astonishes…L.

[1] University of Macerata, Centre of Research in Psychology of Communication, Italy, email: zuko@unimc.it.

[2] University of Macerata, Centre of Research in Psychology of Communication, Italy, email: i.riccioni@unimc.it

Inside the communicative structure which is made by S, his/her words and L, the whole focus is put on S and on what s/he says: L's effects are caused by S's words and thus by S who utters them. The main role is given to S and to his/her words; s/he does all the job: it is s/he who is the only one who acts, is active, performs an action, while L seems to be passive and dependent on what S says. Utterances (i) and (ii) do not acknowledge any autonomy to L but underline a whole dependence of L on S.

5 AN IMPLICIT THEORY OF INTERPERSONAL RELATIONS (THE COMMON SENSE VIEWPIONT)

Utterances (i) and (ii) *are not simply ways of speaking*, they also *show what people believe* (think, are convinced) happens in language communication, i.e. they convey an implicit theory of interpersonal relations, common sense, confirmed and reinforced every day by the language itself. Such implicit theory is the one illustrated in section 2.

Now the question is: is it proper to use perlocutionary verbs or verbal expressions to describe what happened between S and L and to say "S amused/ bored/astonished L" or "S made L feel angry/sad/glad/happy"? In other words, is it proper to state that S and his/her words *cause* L's affects?

6 PARADOXICAL CONSEQUENCES OF AFFECTIVE CAUSALITY

If I believe that my affects are caused by what others said to me, then I grant myself no power over my feelings and so I do not take upon myself the responsibility; therefore I am not in charge of my feelings.

Thus, I attribute power and responsibility to others: it is you who make me feel the way I feel. Vice versa, I can cause other people's feelings and have power over them; I can control them and take charge of them.

If such were the case, I would be the cause of and the one responsible for your feelings but not my own; you would be the cause of and the one responsible for my feelings but not your own. Each of us would be in the other's thrall. We both would be totally dependent on other people; we could have great power over others' feelings but no power over our own: my feelings depend on other people (it is they who can make me feel bad or good) and other people's feelings depend on me (it is I who can make them feel good or bad).

7 TWO ALTERNATIVE VIEWPOINTS

7.1 FOCUS ON THE RELATION BETWEEN S AND L: GESTALT PSYCHOLOGY

In the field of perception, Gestalt psychologist W. Metzger [9] distinguishes three categories of global qualities of the objects we perceive: *shape* qualities ("round", "linear",…), *material* qualities ("smooth", "transparent",…) and *expressive* qualities ("cheerful", "sad",…). To these he adds a fourth category of qualities ("attractive", "pleasant", "repugnant", "amusing", "boring", interesting",…) which, unlike shape, material and expressive qualities, are not *object* qualities, i.e. they do not belong to objects as objects. These qualities refer to the relation between the perceived object and the perceiving subject, and more precisely to the particular *effect* of the relation on the perceiving subject. We can call them *effect qualities*.

Metzger's viewpoint, applied to language communication [17], is not solely focused on either S or L, but is focused on both, or – more precisely – on the *relation* between S and L. According to Metzger, effect qualities are the global outcome, which is experienced with phenomenal immediacy, of the *interaction* between S and L. Since they are global qualities, i.e. Gestalt qualities, they reflect some phenomenal *conditions* which are *structural* inasmuch as they refer to some features of S, some of L, some of what S says and some of the relation between S and L, i. e. features of the particular and wider Gestalt that they constitute all together.

According to this viewpoint, S and L are on the same level; the term "conditions" recognizes the contribution both of them give to the effect: the effect of a speech act by which S "causes" a feeling in L depends not only on the features of the one who "performs the action" (i.e. S), but also on the features of the one who "undergoes the action" (i.e. L). In contrast, the common sense viewpoint (cf. sections 2-5) tends to consider L's features to be nonexistent and to exalt the features of S and his/her words.

7.2 EXPRESSIVE QUALITIES VS EFFECT QUALITIES

In fact, though effect qualities are properties not of S but of L's feelings and result from L's relation with S, everyday language communication attributes such qualities not to L but to S.

When I say, for example, "That movie is amusing", "This exercise is boring" or "Mary is depressing", I use "amusing", "boring" and "depressing" as if it were object qualities, in particular as if it were expressive qualities of the movie, of the exercise, of Mary, as when I say "That movie is cheerful" or "Mary is sad". Expressive qualities such as "cheerful" and "sad", which are the movie's and Mary's properties, have to be distinguished from effect qualities such as "amusing" and "depressing", which refer to the effect that my relation with the movie or with Mary has on *my* feelings.

A movie may be cheerful and not amuse me, just as my interaction with a sad person may not depress me. Cheerfulness and sadness are the movie's and Mary's properties, they do not depend on my feelings; in contrast, amusement and depression are feelings which *I* experience over the movie and Mary. Another person could experience different feelings. But, instead of saying more correctly "I feel depressed, when I'm with Mary" (because, for example, she is sad), language allows me to say "Mary is depressing me" or even simply "Mary is depressing".

In this way, qualities which are *relative to the perceiving subject* are presented by language as *object's absolute qualities* and the relation between object and subject is presented as a cause-effect relation: it's the "I" who is the experiential or phenomenal subject who is depressed, amused or bored; however, from a linguistic or grammatical point of view, that "I" is presented as an object, i. e. a direct object of the action of something else (movie, exercise) or someone else (Mary) which in its (or her) turn becomes the subject of the sentence: "It's the movie that amuses *me*", "It's the exercise that bores *me*", "It's

Mary who depresses *me*". Yet again a causal structure appears focused on the other-than-I, whether object or person.

Yet, if the feelings I experience over objects and persons depend on me too, how is it possible to maintain that it is that-other-than-I which causes my feelings and to say "That movie is amusing me", "This exercise is boring me", "Mary is depressing me", "You make me feel angry", i.e. how is it possible to use a linguistic structure according to which my feelings depend on objects and on other persons?

7.3 WHY IS CAUSAL LAUNGUAGE FOCUSED ON S?

The common sense answer takes the following line: we talk the way we do, i.e. we use causal expressions (i) and (ii), because affective causality *really exists*; *it is true* that others make us feel a certain way, and just as true that we make others feel a certain way.

According to Gestalt theory, such an answer is to be thought of as "naïve", because it implies that language refers to *transphenomenal reality*: here the existence of affective causality in transphenomenal reality accounts for the existence of affective causality in language.

In contrast, as far as the relations between language and the non-linguistic reality which language refers to are concerned, according to the "critical" or "less naïve" viewpoint of Gestalt theory, language refers to *phenomenal reality,* i.e. not to the world but to our experience of the world. Thus, the answer to our question has to be looked for, first of all, within the scope of the relations between language and phenomenal reality.

Metzger [9] thinks of phenomenal reality as a *continuum* in which it is possible to distinguish *perceived phenomenal reality* (here and now I perceive something) from *represented phenomenal reality* (here and now I think/believe/imagine/remember…something).

A Gestalt specific answer to our question comes from Albert Michotte's [10, 11] experimental phenomenology of the perception of causality. His experiments show that 1) causality is a phenomenal datum, 2) it is a perceived phenomenal datum before becoming a represented phenomenal datum, and 3) it is a perceived phenomenal datum without a transphenomenal correlate.

This means that causality is an immediate perceptual datum which strictly depends on a well-defined system of stimuli, i.e. on well-defined structural conditions of a spatial, temporal and kinetic nature. These conditions make the causal impression coercive: but it is sufficient to change them only a little so that a causal impression disappears. Michotte's experiments show that a causal impression is not a question of "interpretation" due to acquired knowledge or thought: the causal meaning of an event is *intrinsic, immanent* in the event itself and independent of past experience or thought. In other terms, causality – as well as objects' shapes, movements etc. – is a global property, a Gestalt quality which imposes itself in a coercive way on our perception without any mediation of thought or past experience.

Thus, the answer to our question could be found by seeing the perceptual experience which language refers to as governing: then the correlate of the language of affective causality would be seen to lie, first of all, on the perceptual phenomenal level.

If I am convinced (i.e. if I do not doubt it) that other people cause my feelings, it is because I find the highest degree of consistency between what I *experience* daily in the course of my communication with others and the meaning that these particular causal expressions offer me in conceptualizing (or representing) my *experience*. Here the word "experience" has to be understood not as "represented phenomenal reality" but as "perceived phenomenal reality" because the causal link between language and feelings is not a representation, a thought, but a perception, and only afterwards does the link become a representation (thought, belief, conviction or prejudice).

7.4 FOCUS ON L: PSYCHOTHERAPY

In the psychotherapy field, some theories (such as F. Perl's Gestalt Therapy [12], E. Berne's Transactional Analysis [3, 4, 5], R. and M. Goulding's Redecision Therapy [7, 8], R. Bandler and J. Grinder's NeuroLinguistic Programming [2]) maintain a viewpoint which is centred on L's emotional autonomy and independence and which is then antithetical to the one that would take S as its focus. According to them, no one is responsible for other people's actions, thoughts and feelings; each person is responsible not only for his/her own actions but also for his own thoughts and feelings (but not for those of other people); s/he has enough power and capability to be the master of his/her own life**.** Personal responsibility, power and capability are often denied and externalized for different reasons by using, for example, the utterances (i) and (ii), so that people consider out of their control feelings which are their own responsibility.

According to this point of view, in our examples it is not S who, by saying what s/he says, amuses or bores L or makes him/her angry, but, on the contrary, it is L who amuses or bores himself/herself or makes himself/herself angry. In these expressions, perlocutionary verbs are used in a reflexive way: the grammatical subject and direct object are no longer two different persons (as in the case of "*S* makes *me* feel angry"), they are the same and only one person ("*I* make *myself* feel angry").

For that reason, in a psychotherapy session, addressing a client who says "S's talk makes me feel angry", Goulding & Goulding [7, 8] do not ask him/her questions such as "What did S tell you to make you feel angry?", because such questions would confirm the client's belief that his/her anger is caused by S. On the contrary, they ask him/her: "When you are listening to S, what do you tell yourself in your head to make yourself angry?". By this kind of questioning, they shift the focus away from S (where it lays in the client's description) to the client himself/herself and they re-propose to him/her his/her anger as a feeling which is totally his/hers, which is dependent on him/her and not on S, as a feeling, then, which s/he can begin to feel himself/herself the master of and be responsible for.

8 DIFFERENT USES OF OUR MODEL IN RELATION TO AFFECTIVE COMPUTING

How can this model be useful in Affective Computing?

(i) As we have already mentioned in section 1, this model seems useful in relation to our theoretical discussion in point (3): if we aim at creating an *emotionally intelligent* [15, 6] computer, then the adoption of our critical model seems preferable.

(ii) If we want to create a computer able both to perceive other people's emotions and recognize its own, we have to construct a computer able to distinguish expressive qualities from effect qualities.

(iii) Thus, given the strong relation between point (i) and point (ii), this same model can be used both on a linguistic and perceptual level.

9 CONCLUSIONS

In this paper we analysed one of the many ways in which affect can be communicated by people in their written and spoken language. In particular, we referred to linguistic structures, such as (i) and (ii) [cf. section 3], which imply the possibility that a person or an object can cause another person's affects. We illustrated the concept of "affective causality", based on a implicit-naïve theory of interpersonal relations linked to common sense. We presented two theoretical alternative viewpoints: the first one derives from Gestalt theory; the second one refers to relevant theories of psychotherapy which promote people's emotional autonomy.

Throughout the paper, we defend a point of view in favour of reciprocal affective autonomy, within and outside psychotherapy sessions [14]. We conclude that in everyday communication, instead of using sentences with focused-on-S causal expressions such as "You make me feel angry", it would be more proper to use such correlative sentences as "You say what you say *and* I feel angry" (or "I make myself feel angry") or **"** *When* you say what you say, I feel angry" (or "I make myself feel angry")**.** In others words, we can use coordinate or subordinate sentences, in which what you say and what I feel are kept apart and not causally linked [16].

We also believe that the awareness in using alternative options to causal structures in affective communication can increase the communicative skills and effectiveness in order both to produce and to interpret speech acts in verbal communication as well as in order to perceptually recognize our and other people's emotions. That is the reason why we believe that our model can be applied to the main theoretical and technical issues of Affective Computing.

REFERENCES

[1] J.L. Austin, *How to Do Things with Words*, Oxford University Press, Oxford, U.K., (1962).

[2] R. Bandler, J. Grinder, *The Structure of Magic*, Science and Behavior Books, Palo Alto, California, USA, (1975)

[3] E. Berne, *Transactional Analysis in Psychotherapy*, Grove Press, New York, USA, (1961).

[4] E. Berne, *Games People Play*, Grove Press, New York, USA, (1964).

[5] E. Berne, *What do You Say after You Say Hello?*, City National Bank, Beverly Hills, California, USA, (1972).

[6] D. Goleman, *Emotional Intelligence*, Bloomsbury , London, UK ,(1996)

[7] M.M. Goulding, R.L. Goulding, *The Power is in the Patient*, TA Press, San Francisco, USA, (1978).

[8] M.M. Goulding, R.L. Goulding, *Changing Lives through Redecision Therapy*, Brunner e Mazel, New York, USA, (1979)

[9] W. Metzger, *Psychologie*, Steinkopff, Darmstadt, Germany, (1954)

[10] A. Michotte, *La perception de la causalité*, Publications Universitaires, Louvain, Belgium, (1954).

[11] A. Michotte, *Causalité, permanence et réalité phénoménales*, Publications Universitaires, Louvain, Belgium, (1962).

[12] F.S. Perls, R.F. Hefferline, P. Goodman, *Gestalt Therapy*, The Julian Press, New York, USA, (1951)

[13] R. Picard, *Affective Computing*, MIT Press, Cambridge (Mass) (USA), (1995)

[14] I. Riccioni, *La percezione della sintonia dialogica*, Edizioni Junior, Bergamo, Italy (2005).

[15] P. Salovey, J.D. Mayer, Emotional Intelligence, in *Imagination, Cognition and Personality*, 9, 185-211, (1990)

[16] C.M. Steiner, *Achieving Emotional Literacy*, Avon Books, Nee York, USA, (1997)

[17] A. Zuczkowski, *Strutture dell'esperienza e strutture del linguaggio*, CLUEB, Bologna, Italy, (1995).

[18] A. Zuczkowski, *Dialoghi quotidiani: il counselling amicale*, CLUEB, Bologna, Italy, (2004).

Sentiment Analysis: Does Coreference Matter?

Nicolas Nicolov*, **Franco Salvetti**$^{\diamond}$ & **Steliana Ivanova***[1]

Abstract. We investigate the boost in sentiment performance by taking coreference information into account. We mine user generated content for market research. Frequent topic shifts in the data lead us to investigate sentiment algorithms which look at a window of text around topic key phrases. We describe and implement a lightweight sentiment analysis system and show how the system performance can be improved by about 10% when taking into account nominal and pronominal coreference elements.

1 Introduction

User generated content is viewed as a valuable source for mining market intelligence data. We study the unsolicited opinions of millions of users regarding companies, their products and services. As data sources we consider weblogs, message board data, surveys, etc.

In blogs there is substantial topic drift as users are describing different experiences and mentions of companies and products are often periferal. Hence, we determine sentiment around occurrences of topic phrases and not the entire document (blogpost). More specifically, we consider proximity sentiment approaches where the algorithms look at certain left and right window of the topic keyword and in the current sentence. This differs from earlier document-level sentiment work (e.g., movie or product reviews). In error analysis between system output of such algorithms and human annotated data we have noticed cases where richer sentiment contexts are present around nominal and pronominal expressions corefering with topic words. Here are some motivating examples:

A1 *Microsoft retools Zune$_i$ to target Apple's flaws.*

A2 *The upgraded player$_i$ and a new strategy helps Redmond gain ground in its battle to beat the iPod.*

Just looking at the first sentence A1 even humans wouldn't be able to infer sentiment. Some systems might consider *flaws* to influence the sentiment negatively. In the second sentence A2 unigrams like *upgraded*, *helps*, *gain* can influence humans and systems positively. It's the coreference Zune$_i$—player$_i$ that enables us (as humans) to make that inference.

B1 *Well I guess I was one of the ones who were actually able to find a Zune$_i$ 80 as from what I have read they are currently in limited supply.*

B2 *This is actually my first MP3 player purchase, and I had been waiting for the release of these players since I had first heard about the 2nd generation Zunes a few months back.*

B3 *Now I have only had it$_i$ a day, but so far I am completely impressed with the quality of this player$_i$.*

In sentence B1 it is challenging for a system to figure out that *limited supply* licenses a positive context. B2 is even more difficult. It is only in sentence B3 that we are assured of the positivity and even simple systems would be able to get this case (*impressed*, *quality*).

[1] *Umbria Inc., 4888 Pearl East Circle, Boulder, CO 80302, USA $^{\diamond}$University of Colorado & Powerset Inc., emails: {*nicolas,sivanova*}*@umbrialistens.com*, *franco.salvetti@colorado.edu*.

C1 *I can't stop playing with my new Zune$_i$ 80.*

C2 *It$_i$'s lovely to look at and hold, the UI is great, and the sound is super.*

Again, in sentences C1 and C2 it's the context in C2 around the pronoun that reveals the sentiment (*lovely*).

Above we have been agnostic as to what type of sentiment algorithm we assume (even though we have been mentioning unigram and bi-gram features). The approach in this paper applies to sentiment algorithms which do not work at the entire document (document-level sentiment). This paper demonstrates that expanding the context on which the sentiment determination is made based on coreference leads to better sentiment results.

Hurst & Nigam mention the possible benefits of using anaphora resolution in conjunction with sentiment identification [4]. More recently Wu and Oard in the context of the NTCIR-6 Chinese Chinese opinion tasks explore (among other tasks) "a simple approximation to anaphora resolution on the accuracy of opinion holder identification" [6].

Sentiment analysis, in particular volume of positive and negative mentions, is heavily influenced by spam. Banking, telecommunications, electronics and automotive domains are notorious for being spammed. We have developed a number of techniques to eliminate spam ([5]) and for the purposes of this work assume clean data.

The techniques we describe are used in combination with other approaches in a larger deployed system at Umbria for large-scale market analysis.

The structure of the rest of the paper is as follows: In Section 2 we describe a lightweight sentiment approach. In Section 4 we discuss the data we have annotated to research the effects of coreference on sentiment. Statistics on the annotation are presented in Section 5. In Section 6 we show the boost in sentiment from coreference. We discuss further ramifications of this approach in Section 7 and conclude in Section 8.

2 Proximity-based, focused sentiment identification

For our sentiment algorithm we assume we have an English blog posting along with its title (we employ a fast, binary English–non-English language identifier). The text is tokenized (we pay special attention to emoticons as they are sentiment bearing), sentences are identified, tokens are normalized, sentences are part-of-speech tagged, and phrases (noun groups) are identified. We also have the offsets where topic phrases have matched the retrieved document (we will see examples for topic phrases in Section 4). We refer to the matched phrases as *anchors*.

The output of sentiment is a confidence value and a label such as: POSITIVE, NEGATIVE, MIXED, NEUTRAL or UNKNOWN

The sentiment algorithm proceeds by identifying a score for each anchor phrase and then aggregates the scores of all anchors.

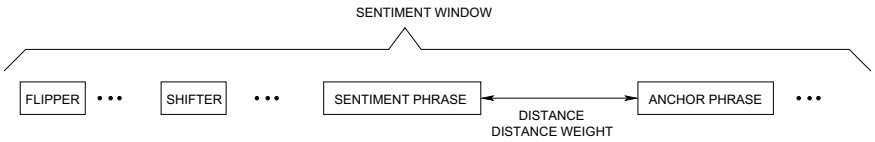


Figure 1. Elements the sentiment algorithm takes into account

2.1 Anchor-level sentiment score

To determine the score for an individual anchor the algorithm considers a sentiment window of 8 tokens before and 6 tokens after the anchor. We dynamically extend the boundaries of the window if the boundaries are inside noun groups. We calculate polarity scores by expanding outward from the anchor and checking if we encounter sentiment phrases. If yes we accumulate the score of the phrase multiplied by the distance weight. The distance weight function is:

$$weight(d) = \begin{cases} -\frac{1}{10}d + 2 & \text{if} \quad d \in \{0,1,2,3\} \\ 1 & \text{if} \quad d \geq 4 \end{cases}$$

If the encountered phrase is a shifter (e.g., adverbs that enhance the sentiment—*horribly wrong*, *really good*) it is allow to influence the score of the word that follows it—we multiply the word's value by the shifter value. The algorithm does separate passes to determine positive (*score*$\oplus$) and negative (*score*$\ominus$) scores. At this point the anchor score is:

$$anchor\_score = \frac{score \oplus - score \ominus}{score \oplus + score \ominus}$$

If flippers we encountered (e.g., negation particle *not*) we multiply the score by $(-1)^{\#flippers}$ (e.g., *not without problems*):

$$anchor\_score = (-1)^{\#flippers} \cdot anchor\_score$$

Figure 1 depicts the elements involved in the anchor sentiment score calculation.

2.2 Aggregating anchor scores

The author opinion of the topic is a combination of the individual sentiment scores for each topic phrase occurrence:

$$\begin{aligned} avg &= average(anchor\_score_1, \ldots, anchor\_score_n) \\ v &= variance(anchor\_score_1, \ldots, anchor\_score_n) \end{aligned}$$

The final decision rule for the sentiment label is:

if avg $< \tau_{\ominus} \rightarrow$ NEGATIVE.
if avg $> \tau_{\oplus} \rightarrow$ POSITIVE.
if $v > \tau_{\text{mixed}} \rightarrow$ MIXED.
else $\rightarrow$ NEUTRAL.

where: $\tau_{\ominus}, \tau_{\oplus}$ and τ_{mixed} are negative, positive and mixed thresholds. We use the values $\tau_{\ominus} = -0.25$, $\tau_{\oplus} = 0.25$ and $\tau_{\text{mixed}} = 0.5$.

The confidence is calculated according to:

$$confidence = \begin{cases} v & \text{if} \quad \text{MIXED} \\ \mid avg \mid & \text{if} \quad \text{POSITIVE, NEGATIVE or NEUTRAL} \end{cases}$$

2.3 Sentiment phrases creation

Our positive and negative phrases (with qualification for strong and weak) augment initially manually created lists by automatic mining through the synonym relation in WordNet [1] using the following starting positive and negative seeds:

Adjectives:

- $\oplus$: {good, nice, excellent, positive, fortunate, correct, superior, beautiful, amazing, successful}
- $\ominus$: {bad, nasty, poor, negative, unfortunate, wrong, inferior}

Nouns:

- $\oplus$: {joy, success}
- $\ominus$: {disaster, failure, pain}

Verbs:

- $\oplus$: {enjoy, like}
- $\ominus$: {suffer, dislike, hate}

For each part-of-speech tag and polarity we do a bounded, breadth-first expansion from each of the seeds, group the extracted items and associated with them the minimum distance to any of the seeds for this polarity. We then consider the positive and negative extracted items; elements for which the difference between the distance to positive and distance to a negative seed is less than a threshold of 2 are removed; the remaining elements in the intersection keep the polarity of the closest seed. Similar techniques are explored by Godbole et al. [3]. Figure 2 illustrates that '*good*' and '*bad*' are not that far from each other in the graph. Subsequently the automatic lists are validated by human annotators. The sentiment resources are continually tuned as part of error analysis.

3 Coreference

The good news: Human languages have evolved to include various shortcuts to refer to previously introduced entities in the discourse. The bad news: Teaching computers the art of figuring which shortcut goes with which entity is not easy.

The latter problem comes under the heading of anaphora/coreference resolution. We consider a scenario where given a text we want to find proper names of entities (e.g., *George W. Bush*) and the other linguistic means to refer to the entities: short names (*George Bush*, *Bush*), nominals (*the U.S. president*) and pronouns (*he*, *his*, *him*). This is essentially the NIST-run automatic content extraction (ACE) task and we have been building systems to perform entity extraction, coreference resolution and relation extraction [2].

In some sense our present task is slightly easier as we can assume a human analyst providing topic terms (names of products) as well as their short forms, possible nominal expressions that can refer to the entity and corresponding pronouns. Still the challenge of referring an occurrence of *man* and *he* to *George W. Bush* vs. *Bill Clinton* who may also be mentioned in the text remains.

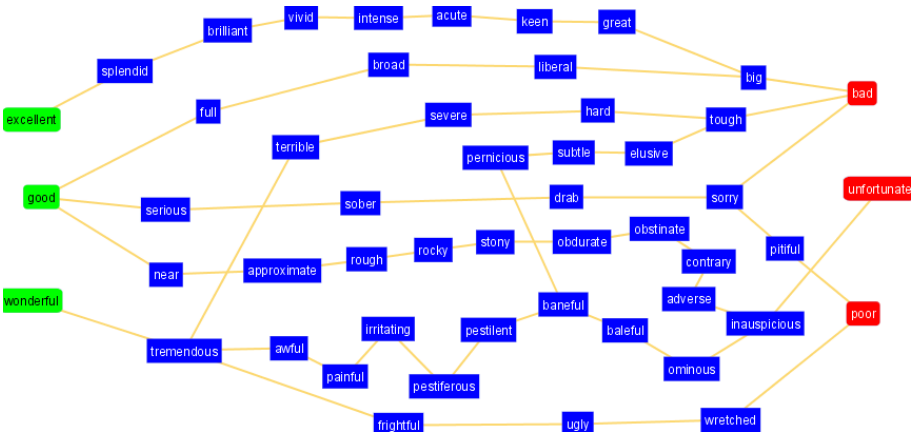


Figure 2. Various paths between positive and negative seeds in WordNet

TOPIC	NOMINALS	COREF EXPR.
R8 -MIDI -Leica	automobile, car, coupe, model, sportscar, supercar, conceptcar, ride, semi-exotic, vehicle, wheels	it, its, that, this
Zune	device, gadget, player	it, its, that, this

Table 1. Topic definitions

It is possible to create reasonably high baselines (in the order of 88% accurate assignments). The investigations in the remainder of this paper are based on perfect coreference from human annotation.

4 Data

We consider blog data for the topics "R8" and "Zune" for the period of October 2007–January 2008. We extract collections of blog postings that mention the topic keywords (cf. Table 1). The system keeps track of the occurrences of the topic terms. We also identify occurrences of the nominal and pronominal expressions for a topic. Each nominal and pronominal expressions is then considered in relation to the occurrences of the topic terms and is marked by human annotators according to the following scheme.

The nominal expressions are marked as referring to:

1. the previous noun group;
2. the first noun group to the left which matches by number and person but not case 1;
3. an anchor phrase but not case 1 nor 2; or
4. another noun group.

Pronominal expressions in addition to the above are marked as:

- modifier: *we didn't think it'd look this good*;
- expletive: *it seems*, *it appears*;
- determiner: *this car*, *that device*;
- wh-pronoun: *the car that captivated the audience*;
- subordinating conjunction/sentential complementizer: *I think that*;
- frozen phrases: *I miss Rowe kids very much. Well, some of them that is.*

If topic words match a different, unintended topic or if the pronoun is in text not part of the posting, we remove these examples:

- exclude: topic words wrongly chosen: *Notes R8*; *Email this*.

In the first example above *R8* refers to release 8; the second example is usually part of a footer of a document—identifying this is part of decomposition of a blog page which is challenging.

5 Annotation

We have annotated the postings extracted for the two topics with coreference information using the scheme described in Section 4. Table 2 show the distribution of different types of the coreference annotation:

Type \ Topic	R8	Zune
previous	11.32%	6.01%
num & pers match	0.93%	1.29%
coreferenced	33.49%	26.18%
modifier	0.78%	0.86%
determiner	14.26%	14.59%
expletive	6.36%	15.88%
complementizer	16.74%	24.03%
frosen phrase	0.93%	1.72%
wh-pronoun	3.57%	6.87%
other	57.36%	36.05%

Table 2. Distribution of coreference annotation types

The inter-annotator agreement between the two annotators participating in this project is 98.91% (overlapping annotations out of all annotations).

We have also annotated the data for sentiment. Figure 3 shows the web-based sentiment annotation tool.

6 Results

To answer the question in the title of the paper—yes, sentiment benefits from coreference information. Here are two examples where originally the sentiment algorithm returned NEUTRAL sentiment. After considering the coreference anchors sentiment changed to POSITIVE:

> *Being first drawn to the Zune last year because of* [it] *'s* [SIC] *style and awesomeness, I decided with it.*

> *My parents got me a Zune.* [It] *has just come out on the market.* [It] *is something i have wanted for a long time and i was very excited about* [it].

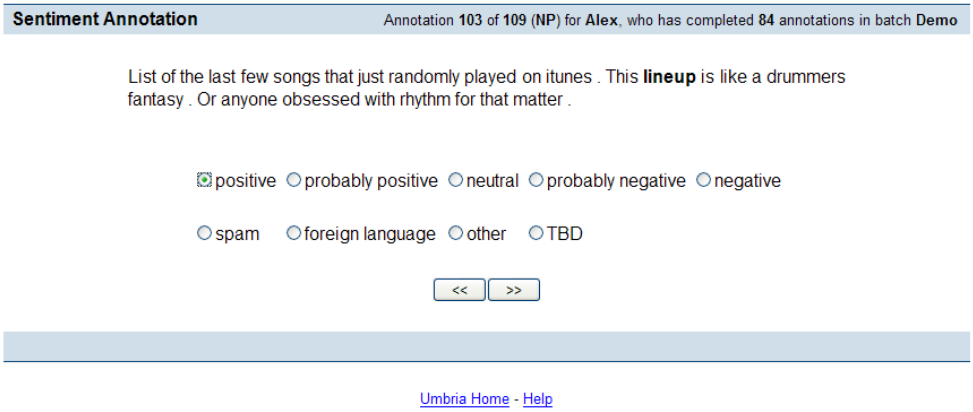


Figure 3. Sentiment annotation tool

Table 3 shows how the sentiment system can improve its performance by taking coreference information into account. The improvement depends on the data and is about 10%. The bottom part of the table shows that in about 14% of the cases the corefencial contexts support the original sentiment determination.

System \ Topic	R8	Zune
nominals	8.20%	5.62%
pronouns	9.13%	6.54%
nominals ∪ pronouns	12.73%	8.31%
nominals support	11.05%	5.77%
pronouns support	11.05%	11.54%
nom. ∪ pro. support	15.24%	13.46%

Table 3. Percent improvement due to coreference over the baseline sentiment analysis. Bottom part: percent of coreference support of original sentiment determined on topic anchors only

7 Discussion

As part of ongoing work we are exploring the coreference effect on sentiment in other domains (in marketing referred to as verticals). We are also investigating different data sources of user generated content—message boards and survey data. In preliminary investigations we have seen a fair use of pronouns in message board data. A major challenge is to properly recognize the quoting structure in such postings. Otherwise we attribute text to an author which is not written by them. For survey data we pay special attention to imperatives (e.g., *have better help and an adaptive user interface for the new release of the software*).

Flat proximity models for sentiment are easy to implement and the performance of such systems hinge upon the quality of the sentiment lists they use. We are investigating the use of dependency parsing in order to consider only those sentiment elements which are related to the anchors. Beyond the adjective modification the nature of that relationship is quite complex and there is additional computational cost associated with the parsing.

Sentiment is expressed not only directly (e.g., *product X is great*) but also indirectly—parts of the product are good; effects of the drug are good, etc. We approach this through automatic clustering. We also benefit from the fact that in our scenario we have human analysts who adjust output suggested by the system in the explorative stage.

We are also looking at validating our techniques across different languages—initially we are looking at German.

Audi R8 has won awards as 'Car of the Year' for 2007 among readership of many magazines. Some have given it this title for 2008. Zune is an MP3 and video player offered by Microsoft.

8 Conclusions

We have considered sentiment analysis on user generated content for market intelligence. Due the frequent topic drift we have argued for focused sentiment analysis which is performed on parts of the document around topic terms. We have described a lightweight, proximity-based sentiment algorithm and have shown that that the system can be improved by about 10% (depending on the topic) by augmenting the focus area of the algorithm using contexts around nominal and pronominal coreference elements.

Acknowledgements

We would like to thank Martha Palmer, Jim Martin and Mike Mozer from the University of Colorado for useful discussions.

REFERENCES

[1] *An WordNet Electronic Lexical Database*, ed., Christiane Fellbaum, The MIT Press, 1998.

[2] Radu Florian, Hany Hassan, Abraham Ittycheriah, Hongyan Jing, Nanda Kambhatla, Xiaoqiang Luo, Nicolas Nicolov, and Salim Roukos, 'A statistical model for multilingual entity detection and tracking', in *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics (HLT-NAACL'2004)*, pp. 1–8, Boston, Mass., (2–7 May 2004).

[3] Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena, 'Large-scale sentiment analysis for news and blogs', in *International Conference on Weblogs and Social Media (ICWSM'2007)*, ed., Nicolov et al., pp. 219–222, Boulder, CO, (26-18 March 2007).

[4] Matthew Hurst and Kamal Nigam, 'Retrieving topical sentiments from online document collections', *In Document Recognition and Retrieval XI*, (2004).

[5] Nicolas Nicolov and Franco Salvetti, 'Efficient spam analysis for weblogs through url segmentation', in *Recent Advances in Natural Language Processing*, volume 292 of *Current Issues in Linguistic Theory*, 125–136, John Benjamins, Amsterdam & Philadelphia, (2007).

[6] Yejun Wu and Douglas W. Oard, 'Chinese opinion analysis pilot task', *In Proceedings of NTCIR-6 Workshop Meeting, May 15-18, 2007, Tokyo, Japan*, (2007).

Towards Semantic Affect Sensing in Sentences

Alexander Osherenko, University of Augsburg, osherenko@informatik.uni-augsburg.de

Abstract. Recently, there has been considerable interest in the recognition of affect in written and spoken language. In this paper, we describe a semantic approach to lexical affect sensing in sentences that uses findings from linguistic literature and from empirical examples. The approach is evaluated using a corpus containing 759 English sentences.

1 INTRODUCTION

Lexical affect sensing is an important field of study whose results can be used in a wide range of applications, e.g. robotics or tutoring systems. Despite its illusory simplicity, the emotional analysis of texts presents a great challenge to computer scientists because of the manifoldness of expressed meanings in texts.

There are two types of approach aimed at solving this problem: statistical and semantic. Statistical approaches make use of data-mining methods, e.g. Support Vector Machines (SVM), and classify emotion in text, for instance by using word counts [8]. However, statistical approaches produce low classification results when classifying short texts.

In contrast, semantic approaches aim to classify affect in texts by using commonsense as well as linguistic information on emotional parts of analyzed texts. For instance, Prendinger and colleagues [9] classify the affective meaning of texts by using emotion words from [13] in word-level analysis, by using lexical modifiers of meaning and negations in phrase-level analysis, or by scrutinizing the grammatical structure of sentences in sentence-level analysis.

2 SYSTEM

We solve the introduced manifoldness problem by analyzing parts of studied texts: the whole text is split into sentences and the sentences into phrases. After the emotional meaning of each part is analyzed, the emotional meaning of the original text is deduced from the emotional meanings of the constituent phrases.

In order to test our idea, we implemented a computer system that uses two functionally complementary parsers: the SPIN parser and the Stanford parser.

The SPIN parser is a semantic parser for spoken dialogue systems, a rule-based framework that parses texts using order-independent word matching [2]. For instance, in text *I like this game* the SPIN parser finds the positive verb *like*. The probabilistic Stanford parser is used for determining parts of speech, lemmatizing words, and splitting text in parts [4]. For example, it takes the text *Finally, I was so angry that I could burst with rage* and splits it into a superordinate subsentence *I was so angry* and a subdominant sentence *that I could burst with rage*.

A text can contain several emotional phrases that have contradictory emotional meaning. We test three strategies for interpreting a text's emotional meaning: as defined by the first or by the last emotional part in the corresponding part (whole text, subsentences, phrases), or by the average meaning (the emotional meaning as defined by the majority of affective votes). For instance, in the sentence *I am happy and sad* the emotional word *happy* (considered a positive word) defines according to the strategy of the first phrase a positive meaning, the emotional word *sad* (considered a negative word) defines according to the strategy of the last phrase a negative meaning, and according to the strategy of an emotional average a neutral meaning (there is no emotional majority).

The affect recognition system classifies the emotional meaning into two stages: in the first stage (division) the system divides the text into parts of particular granularity (analyzes an unchanged text or splits it into subsentences or phrases) and scrutinizes the emotional meaning of each individual part, while in the second stage (consolidation) the system compiles the emotional meaning of the original text by composing it from the emotional meanings of the detected parts.

The proposed algorithm for semantic affect sensing (examined using the example of the emotional sentence *Finally, I was so angry that I could burst with rage*) is therefore as follows (depending on the chosen granularity of the analysis either *Whole text*, *Subsentences*, or *Phrases*):

a. *Whole text.* Apply the chosen classification strategy to the analyzed text (first phrase strategy – emotional meaning of word *angry*, last phrase strategy – emotional meaning of word *rage*, average strategy – average meaning of words, i.e. emotional meaning of words *angry* and *rage*).

b. *Subsentences*. Detect subsentences using the Stanford parser, classify their emotional meaning using the SPIN parser according to the chosen classification strategy (first phrase, last phrase, average), construct an auxiliary text (subsentence combination) out of the emotional meanings of subsentences, and classify the emotional meaning of an original sentence by analyzing the subsentence combination.

   For instance, the system detects the superdominant subsentence *Finally, I was so angry* and the subdominant subsentence *I could burst with rage* and constructs a subsentence combination *superord_high_neg subord_low_neg*, where *superord_high_neg* stands for the high negative meaning of the superordinate sentence and *subord_low_neg* for the low negative meaning of the subordinate sentence. The system classifies the original text as high negative (*high_neg*) by applying patterns for subsentences in Table 2.

c. *Phrases.* In contrast to step *b* above, run an additional intermediate step that facilitates the analysis of the emotional meanings of subsentences, not by using subsentences' text, but rather by using auxiliary texts – phrase combinations. Detect subsentences, then phrases that are contained in the detected subsentences, classify the emotional meaning of phrases according to the chosen

classification strategy (first phrase, last phrase, average), construct an auxiliary text for the emotional structure of the corresponding subsentence (phrase combination), classify the emotional meaning by applying patterns for phrases in Table 3, compile a subsentence combination, and calculate an emotional meaning of the original sentence by using patterns for subsentences in Table 2.

The system detects the dominant subsentence *Finally, I was so angry* and the subdominant subsentence *I could burst with rage*. In the dominant subsentence it extracts four phrases: adverb phrase (*finally*), noun phrase (*I*), verb phrase (*was*), and adjective phrase (*so angry*); and in the subdominant sentence three phrases: noun phrase (*I*), verb phrase (*could burst with*), and noun phrase (*rage*). The system constructs the phrase combination *phrase_null phrase_null phrase_null phrase_high_neg* for the dominant sentence (the phrase *so angry* is classified as *high_neg*), where *phrase_null* corresponds to a neutral meaning and *phrase_high_neg* to the high negative meaning of a phrase, and for the subdominant sentence the phrase combination *phrase_null phrase_null phrase_low_neg*. The system classifies affect in phrase combinations by applying patterns for phrases in Table 3, constructs a subsentence combination *superord_high_neg subord_low_neg that* (cf. step *b*) and classifies it as *high_neg* by applying patterns for subsentences in Table 2.

d. If necessary, calculate the majority vote on the basis of values yielded by the granularities above.

The system calculates the majority vote on the basis of values yielded by the granularities above. It takes from the *Whole text* granularity the *low_neg* value, from the *Subsentences* granularity the *high_neg* value, and from the *Phrases* granularity the *high_neg* value, and calculates the majority vote *high_neg*.

3 CORPUS

We chose in our experiments the Fifty Word Fiction corpus (FWF) containing 759 grammatically correct English sentences that are manually annotated in terms of their sentiment and affect as *positive*, *neutral*, or *negative* [11]. For instance, *We all laughed and ordered beers* is annotated as *positive*. The corpus was collected online and available to the general public for one month, during which some 3,301 annotations were made by 49 annotators. Of the sentences, 82 are annotated as *positive*, 171 as *negative*, and 506 as *unclassifiable*. The inter-coder agreement is 65% (less than 80% – a desirable agreement in line with [1]).

4 SOURCES OF AFFECT INFORMATION

Affect Sensing of Parts

Our system utilizes the following information to classify the affect of parts: information from affect dictionaries, grammatical lexical patterns from linguistic studies, and empirical lexical patterns from our own studies.

Information from Affect Dictionaries

We use emotion words from various affect dictionaries as the basis for our system: Levin verbs [6], GI [12], and WordNet-Affect [13]. We consider 4,527 words from affect dictionaries in our study: 503 words from WordNet-Affect, GI words (1,790 positive and 2,200 negative), and 34 Levin verbs.

Grammatical lexical patterns from linguistic studies

In our system, we use 11 grammatical patterns to scrutinize the emotional meaning of texts [5]:

1. Interjections (299), e.g. *Oh, what a beautiful present!*
2. Exclamations (300a), e.g. *What a wonderful time we've had!*
3. Emphatic *so* and *such* (300b), e.g. *I'm so afraid they'll get lost!*
4. Repetition (300c), e.g. *This house is 'far, 'far too expensive!*
5. Intensifying adverbs and modifiers (301), e.g. *We are utterly powerless.*
6. Emphasis (302), e.g. *How ever did they escape?*
7. Intensifying a negative sentence (303a), e.g. *She didn't speak to us at all.*
8. A negative noun phrase beginning with *not a* (303b), e.g. *We arrived not a moment too soon.*
9. Fronted negation (303c).
10. Exclamatory questions (304), e.g. *Hasn't she grown!*
11. Rhetorical questions (305), e.g. *What difference does it make?*

Empirical lexical patterns from our own studies

We used 25 empirical examples of emotional texts containing negations and intensifiers to build lexical patterns for analyzing emotional meanings. The patterns classify the emotional meanings of texts, facilitating a five-class scheme (*low positive*, *high positive*, *low negative*, *high negative*, *neutral*) using emotion words, negations (*not*, *never*, *any*), and 74 intensifiers of emotional meaning, e.g. *definitely* (Table 1).

Example	*Pattern*
I am so happy.	*<Intensifier> <Emotional word+>→ <Result++>*
I am not happy.	*<Negation> <Emotional word+> → <Result−>*
I am not very happy.	*<Negation> <Intensifier> <Emotional word+>→ <Result−>*

Table 1. Example patterns for modifying affect

Table 1 shows example patterns for modifying affect. The *Pattern* column shows a pattern that matches the example text in the *Example* column. *<Intensifier>* denotes an intensifier word, *<Emotional word+>* a low positive emotional word, *<Result++>* the high positive result of affect sensing, and *<Result−>* the low negative result of affect sensing.

Patterns for Linking Parts

Phrases and subsentences divide the original sentence into parts, with each potentially having its own emotional meaning. For the purpose of compiling the meaning of the original text from constituent parts, the implemented system composes the emotional meaning of the original text out of the emotional meanings of constituent phrases and subsentences.

The proposed system contains 122 empirical patterns for linking subsentences and 19 empirical patterns for linking phrases.

Pattern for linking subsentences	*Example*
<Sup++> <Sup+>→ <Result++>	*It is a very good film and the acting is excellent.*
<Sup++> <Sub−>→ <Result+>	*It is a very good film although the acting seems at first to be not excellent.*

Table 2. Example patterns for linking subsentences

Table 2 shows sample patterns for linking subsentences. The *Pattern for linking subsentences* column shows a pattern that matches the text in the *Example* column. *<Sup++>* represents the high positive emotional meaning of the superdominant subsentence, *<Sup+>* the low positive meaning of the superdominant sentence, *<Sub−>* the low negative emotional meaning of the subdominant subsentence, *<Result++>* the high positive result of affect sensing, *<Result+>* the low positive result of affect sensing, and *<Result−>* the low negative result of affect sensing.

Table 3 shows sample patterns for linking phrases.

Example pattern for linking phrases	*Example*
<Phrase+> <Phrase0> → <Result+>	*exact and accurate*
<Phrase+><Phrase−> → <Result−>	*happy and depressing*

Table 3. Example patterns for linking phrases

Table 3 shows sample patterns for linking phrases. The *Pattern for linking phrases* column shows a pattern that matches the text in the *Example* column. *<Phrase+>* represents the positive emotional meaning of the phrase and <Phrase−> the low negative emotional meaning of the phrase.

5 RESULTS

The baseline for evaluating the proposed approach provides the best recall value, 37.20% averaged over classes, calculated via the statistical approach in [8] using word counts as features and a SVM classifier.

Table 4 shows the results for solving a three-class problem using the proposed approach with and without lexical patterns (using only emotional words). The R^a column represents the recall value averaged over classes and the P^a column the corresponding precision value averaged over classes. The $R^{a\text{-}lp}$ column represents the recall value averaged over classes when classifying texts without lexical patterns and the $P^{a\text{-}lp}$ column signifies the corresponding precision value averaged over classes. The *Gran.* column represents the granularity of the text division (the decision based on the majority vote – no division; the text as a whole; division into subsentences – abbreviated as *Subsent.*; division into phrases), and the *Strategy* column shows the strategy of semantic sensing (first phrase, last phrase, average vote).

Gran.	*Strategy*	R^a	$R^{a\text{-}lp}$	P^a	$P^{a\text{-}lp}$
Majority	*First phrase*	***47.20***	*45.02*	*44.09*	*42.76*
	Last phrase	***47.64***	*46.24*	*44.26*	*43.45*
	Average vote	*45.92*	*45.66*	*43.14*	*43.05*
Whole Text	*First phrase*	*45.41*	***47.30***	*42.90*	*43.90*
	Last phrase	*47.45*	*46.70*	*44.05*	*43.57*
	Average vote	*42.79*	*44.36*	*41.15*	*42.18*
Subsent.	*First phrase*	*47.20*	*45.22*	*44.08*	*42.88*
	Last phrase	*47.24*	*45.84*	*44.03*	*43.22*
	Average vote	*46.04*	*45.66*	*43.22*	*43.05*
Phrase	*First phrase*	*44.79*	*43.71*	*42.90*	*42.13*
	Last phrase	*45.21*	*44.54*	*43.13*	*42.65*
	Average vote	*44.22*	*44.16*	*42.41*	*42.40*

Table 4. Results of affect sensing for three classes

The results corresponding to the word spotting (see the definition in [7]) are shown in the rows of *Whole text* (hereafter referred to as the word-spotting values). Other alternatives, e.g. in the rows of *Phrase*, cannot be considered as word-spotting-processing, since additional patterns for processing combinations (phrase and subsentence combination) are necessary.

The *Majority* rows show the majority vote of the most entities (phrases, subsentences, utterance). If the majority vote cannot be calculated, i.e. classification results are pairwise different, the result of the *Subsentences* classification is taken as the basis.

6 DISCUSSION & FUTURE WORK

The proposed semantic approach is tested on a corpus with English sentences, and the applied patterns improve classification rates compared both with the word-spotting values and with the statistical baseline of 37.20% (Table 4). For instance, the *Majority*, *Last phrase* classification rate, 47.64%, is much higher than the statistical baseline 37.20% and also higher than the word-spotting value of 47.30% for *Whole text*, *First phrase*.

Moreover, the classification rates are higher for the majority evaluation using the full grammar set of applied patterns (47.64% for *Majority*, *Last phrase*). Furthermore, the results are significantly higher compared with the statistical baseline (47.64% vs. 37.20%). In addition, the average vote does not generally bring an enhancement of classification results, e.g. 47.64% for *Majority*, *Last phrase* vs. 45.92% for *Majority*, *Average vote*.

In future, we will revise our approach and collect new corpora containing short emotional texts, for instance through acquiring new data from the Internet.

REFERENCES

1. Craggs, R. 2004. Annotating emotion in dialogue – issues and approaches. 7th Annual CLUK Research Colloquium.
2. Engel, R. 2006. SPIN: A Semantic Parser for Spoken Dialog Systems. In *Proceedings of the Fifth Slovenian And First International Language Technology Conference (IS-LTC 2006)*, 2006.
3. Kollias, S. 2007. ERMIS Project. URL: http://www.image.ntua.gr/ermis/.

4. Klein, D., Manning, C. D. 2003. Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.
5. Leech, G. N., Startvik, J. 2003. A communicative grammar of English. Third edition. Longman.
6. Levin, Beth. 1993. English verb classes and alternations. The University of Chicago Press.
7. Liu, H., Lieberman, H., Selker, T. 2003. A Model of Textual Affect Sensing Using Real-World Knowledge. Pages 125-132 of: Proceedings of IUI-03, the 8th international conference on intelligent user interfaces. Miami, US: ACM Press.
8. Osherenko, A. André, E. 2007. Lexical Affect Sensing: Are Affect Dictionaries Necessary to Analyze Affect? In Proceedings of Affective Computing and Intelligent Interaction (ACII), Springer.
9. Neviarouskaya, A., Prendinger, Ishizuka, M. Textual Affect Sensing for Social and Expressive Online Communication, Affective Computing and Intelligent Interaction (A. Pavia, P. Prada and R. W. Picard (Eds.)), (Proc. Int'l Conf. ACII2007, Lisbon, Portugal), Springer LNCS 4738, pp.218-229 (2007.9).
10. Quirk, R., Greenbaum, S. 1988. A University Grammar of English. Longman Publishing Group; Abridged Ed (January 1, 1973).
11. Read, J. 2004. Recognising affect in text using pointwise-mutual information. Masters thesis, University of Sussex.
12. Stone, P. J., Dunphy, D. C., Smith, M.S., Ogilvie, D. M. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press.
13. Valitutti, A., Strapparava, C., Stock, O. 2004. Developing Affective Lexical Resources. PsychNology Journal. Volume 2, Number 1, 61-83.

Applying a Culture Dependent Emotion Triggers Database for Text Valence and Emotion Classification

Alexandra Balahur[1] **and Andrés Montoyo**[1]

Abstract. This paper presents a method to automatically spot and classify the valence and emotions present in written text, based on a concept we introduced - of emotion triggers. The first step consists of incrementally building a culture dependent lexical database of emotion triggers, emerging from the theory of relevance from pragmatics, Maslow´s theory of human needs from psychology and Neef´s theory of human needs in economics. We start from a core of terms and expand them using lexical resources such as WordNet, completed by NomLex, sense number disambiguated using the Relevant Domains concept. The mapping among languages is accomplished using EuroWordNet and the completion and projection to different cultures is done through language-specific commonsense knowledgebases. Subsequently, we show the manner in which the constructed database can be used to mine texts for valence (polarity) and affective meaning. An evaluation is performed on the Semeval Task No. 14: Affective Text test data and their corresponding translation to Spanish. The results and improvements are presented together with an argument on the strong and weak points of the method and the directions for future work.

1 INTRODUCTION

In recent years, there has been growing interest in studying the methods through which subjectivity is expressed in written text. Whether it is mining for customer opinions, or tracing attitudes towards different topics of interest, tools and applications aiming at discovering subjectivity, spotting and, moreover, interpreting emotion in text are highly applicable to various natural language processing areas. Some important examples include word sense disambiguation [24], multi-document summarization, multi-perspective question answering and speech generation. Moreover, in the context of the interactive web, computers can no longer remain static tools, but must become capable of detecting user-specific characteristics, attitudes and emotions and responding appropriately, thus creating a close to human-human interaction environment. Furthermore, being able to spot specificities of different users can ease the search for appropriate entertainment – beneficial to the user- , and help create personalized advertisement by communicating in a manner persuasive to each user individually – beneficial to the advertised companies. Last, but not least, recent findings have shown that emotional agents, replicating the human, emotion influenced decision process, outperform pure rational agents [7], in that emotion-based decision making better captures the entire process of human reasoning.

Present work in the field has focussed on determining methods to capture emotion and opinion arising from written text, at a word level – identifying positive or negative sentiment of words – in [5], sentence or phrase level- in [9], and document level – in [7]. Lexical resources born from these contributions are WordNet Affect, presented in [22] and SentiWordNet, presented in [4], both for English. Lexical databases were in turn completed in several approaches toward sentiment analysis with lexical and commonsense knowledge databases such as ConceptNet [13], word similarity measures using WordNet [6], rules for determining text polarity using word and part-of-speech composition rules [1], statistical and machine learning methods [26]. To our knowledge, there has been little work done towards obtaining lexical databases of affective terms for languages other than English [19] and no approach including motivational, human-need oriented or relevance directed theories in the study of affect in text. We found that such theories can help spot the emotion inducing terms in text and prioritize their being interpreted, based on their level of importance for human existence. Also, discovering the source that is producing emotion to a reader is directly applicable to real-life, culture dependent situations through projection on commonsense knowledgebases of the considered language.

The observation we set off from is that, on the one hand, there are general terms that trigger emotion in all people – such as war, hunger, freedom, mother, children and, on the other hand, that people are different with respect to social context, values, interests, experience and so on. Therefore, their reaction and interpretations of texts are different. Indeed, when seeing a news headline like ¨Government approves new taxes for car imports¨, one reader will remain neutral, another will be infuriated, a car seller might be glad. Or in the case of a news title such as ¨The Spanish Civil War now a computer game¨, a Spaniard could feel offended, another outraged, another amused, a person of another nationality might feel neutral and a computer games´ addict very happy. On the other hand, a title such as ¨Children killed in bomb attack¨ will most certainly produce a general feeling of sadness and/or fear to the readers.

Finally, relating all these observations, we ask ourselves: What are the causes of these different interpretations? Are there general terms that trigger emotion indifferent of the reader´s culture, social and cultural background? Are there particular terms that trigger emotion only to a group of people; how can we identify those? Is it correct or practically usable to draw a general emotion from a given text? The answer we give is that it depends on what the intention in reading the text is. It is important to notice the clear distinction between the cognitive and emotional aspect of a text. When the text is read as information, the emotion reflected is that of the persons and

[1] Dept. of Software and Computing Systems, University of Alicante, Ap. de Correos 99, 03080 Alicante, Spain. E-mail: {abalahur, montoyo}@dlsi.ua.es.

actions behind the text, the events that it speaks about. However, when interpreting a text from a reader´s point of view, the front-end of the text, is more purposeful to try to spot the reason for the different interpretations, following the general to particular manner, and replicate them in a computer environment.

Thus, we start from the premises that text alone bears no emotional content, at any level, but can contain what we call ¨emotion triggers¨.

We define the notion of ¨emotion trigger¨ as a word or concept expressing an idea, that depending on the reader´s world of interest, cultural, educational and social factors, leads to an emotional interpretation of the text content or not. Examples of emotion triggers are ¨freedom¨, ¨salary¨, ¨employment¨, ¨sale¨, ¨pride¨, ¨esteem¨, ¨family¨ and so on. We will use this defined notion to build a database of such emotion triggers, classify them and integrate them in a system which spots and classifies text valence and emotion.

It is a demonstrated fact that rationality is influenced by emotion [8] and thus, emotion triggers can lead to shifts in the process of inference, based on what motivations and needs they trigger. We call ¨emotional inference¨ the process of conferring an emotional interpretation on a text on the basis of the most relevant emotion triggers of a given context.

Our method was developed in view of a novel perspective of emotion detection and interpretation, based on the defined notions of ¨emotion triggers¨ and ¨emotional inference¨, from a cultural specific perspective. In the end, we try to offer possible answers to the following questions:

1. Does a reader react to all emotional stimuli in a text?
2. From a text with different emotional stimuli, which are the ones that a specific reader will be sensitive to?
3. What makes those emotion triggers be picked for interpretation?
4. To what extent are emotion triggers culture dependent?
5. How can the general/culture-specific emotion triggers be obtained?
6. What feelings are evoked by the different emotion triggers?
7. What influence will the interpretation of the picked emotion triggers have on the further interpretation of the text?

In order to answer these questions, we will further present a method to construct a database of emotion triggers, the proposed model for a text valence and emotion spotting system and the method to integrate the emotion triggers database as part of the system conforming to the proposed model to automatically spot and classify the valence and emotions present in written text.

This paper is organized as follows: we shortly present the theories that lay at the foundation of our proposed method, together with a motivation for their use in Section 2. Further on, in Section 3, we present the model of the system we propose for text valence and emotion spotting. In Section 4, we show the practical steps performed for building the cultural dependent lexical database of emotion triggers and the method used to assign them valence and emotional value. In Section 5, we show the method used for the implementation of the system presented, as well as the resources, cognitive model and computational theory used. We follow by describing the experiment we performed on the system, the evaluation it was subjected to and the results obtained. In Section 6, we compare the obtained results with those obtained by similar systems. Finally, we conclude on the method used and present the lines for future investigation and improvement.

2 THEORIES AND RESOURCES

This section briefly describes the theories lying at the foundation of our text valence and emotion spotting system.

The motivation for introducing the concept of ¨emotion triggers¨ is found in the assumptions and principles of the relevance theory from pragmatics. Abraham Maslow´s theory of human motivation and its corresponding pyramid of human needs offer the method to classify the emotion triggers and create rules of emotion trigger interaction. In parallel, we apply Neef´s matrix of fundamental human needs to create a need-satisfier system of emotion triggers, in view of an alternative set of rules for emotional inference.

2.1 THEORY OF RELEVANCE

There are two ways to conceive of how thoughts can be communicated from one person to another. The first way is through the use of strict coding and decoding, which makes explicit use of symbols, rules, and language. The second way is by making interpretive inferences, which communicate to the hearer information that is left implicit[2]. Relevance theory [21] seeks to explain the second method of communication: implicit inferences. It argues that the human mind will instinctively react to an encoded message by considering information that it conceives to be relevant to the message. By "relevance" it is meant whatever allows the most new information to be transmitted in that context on the basis of the least amount of effort required to convey it. ¨The Theory of Relevance¨, arises from pragmatics, and states in the cognitive principle that ¨human cognition tends to be geared toward the maximization of relevance¨ , that is, from the multiple stimuli present in a communication, be it written or spoken, a reader will choose the one with highest significance to their world of interest. In general, a person will act towards the maximization of the quantity of information it holds, activating the stimuli whose interpretation could bring them a piece of important information and inhibiting those they hold as unimportant. Furthermore, relevance is considered as relative or subjective, as it depends upon the state of knowledge of a hearer when they encounter an utterance. These statements, together with the principles of the relevance theory, can be seen to explain also the process of emotionally interpreting a text, since of all possible terms that could constitute in emotional reactions, the reader will only choose the most important to his/her interest and react to it and the relativity and subjectivity can be considered as cultural and social background specific emotion triggers and their corresponding inferences.

The theory of relevance contains no explicit mentioning or classification of what could constitute stimuli to a person. In that respect, and starting from what we showed in the introductory examples, we considered a good classification the one made by

[2] http://en.wikipedia.org/wiki/Relevance_theory

Abraham Maslow, under the form of a 5-level pyramid of human needs (motivations). We now explain more in depth the basics of this theory:

2.2 MASLOW´S PYRAMID OF NEEDS

Abraham Maslow, in [17], classified the human needs and motivational factors into a 5-level pyramid, from the basic, physiological ones, to the more education and personal level of development dependent ones. Needs such as food, shelter, peace are at the bottom of the pyramid, whereas needs for self achievement, fame, glory are at the top. The basic needs are the general human ones; as we move towards the top, we find the more individual dependent ones.

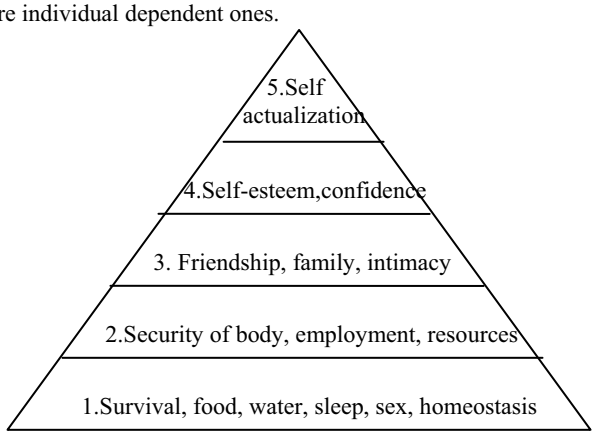


Figure 1. Maslow´s pyramid of human needs and motivations

Maslow also defines a set of rules capturing the change in emotion triggers, based on the basic needs being met or not. Examples of such rules are:

1. If some needs are not fulfilled, a human's physiological needs take the highest priority.
2. Physiological needs can control thoughts and behaviours, and can cause people to feel sickness, pain, and discomfort.
3. When physiological needs are met, the need for safety will emerge.
4. When one stage is fulfilled, a person naturally moves to the next.

We consider the terms in Maslow´s pyramid levels as primary emotion triggers, very general notions that express ideas that are fundamental to all human beings. Their projection to concrete notions, for example – food -> pizza, hot dog etc.; war -> Independence war, civil war is culture dependent.
In order to exploit this classification, we build a lexical database of emotion triggers at the 5 levels. The words found in the five levels are nouns, verbs, adjectives and adverbs.

2.3 NEEF´S MATRIX OF FUNDAMENTAL NEEDS

Among the critics of the Maslow theory of human needs is Manfred Max Neef, in [18]. In turn, Neef´s theory is rooted in the economical perspective of fundamental human needs, complemented by their corresponding satisfiers. In the author´s view[3], human needs are equally important, few, finite and classifiable (and are different from the economical "wants" that are infinite and insatiable) and constant through all human cultures and across historical time periods. Neef states that the only aspect changing over time and between cultures is the way these needs are satisfied. Human needs, according to Neef, are understood as a system - i.e. they are interrelated and interactive. Max-Neef classifies the fundamental human needs as: subsistence, protection, affection, understanding, participation, recreation (in the sense of leisure, time to reflect, or idleness), creation, identity and freedom. Needs are also defined according to the existential categories of being, having, doing and interacting, and from these dimensions, a 36 cell matrix is developed which can be filled with examples of satisfiers for those needs[4].
Therefore, starting in parallel from the matrix of fundamental human needs as primary emotion triggers, we see, on one hand, if a classification of emotion triggers is better than a flat model with rules of inference, and on the other hand, can build a fine-grained taxonomy of terms indicating the precise category in which each type of emotion trigger influences the human affect.

3 EMOTION TRIGGER METHOD

Our emotion trigger method starts from the idea that words in text themselves carry no affectivity, but become emotionally charged depending on the interpretation they are given by each reader´s world of interest and the intention and world of interest of the author. This world of interest is made up of general, personal needs and motivation factors, notions satisfying these needs, knowledge on the historical and social facts, information vehiculated in the media and so on. We call this collection of factors ¨bag of knowledge¨, and believe that by uncovering it, we move an important step forward ¨to try to uncover which are the latent semantic boundaries of human communication¨ [14].

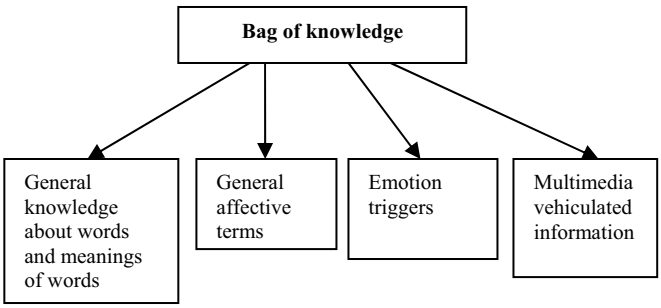


Figure 2. Model for the analysis of emotion in text

The first component of the bag of knowledge is made up of general knowledge about words and meanings of words. It contains what words can mean, the manner in which they are linked, how they change their meaning. The second component is formed of general affective terms, as ¨kind-hearted¨, ¨furious¨, ¨anxious¨, ¨fear¨ and so on. They express emotion, but do not necessarily induce emotion. For example, a title such as ¨Feared

[3] http://en.wikipedia.org/wiki/Fundamental_human_needs
[4] http://www.rainforestinfo.org.au/background/maxneef.htm

opponents, defeated without problems¨ has no connection to the idea of fear. Such classification of words can be found in lexical affective resources such as WordNet Affect or SentiWordNet. The third component is made up of emotion triggers. It contains the terms that carry in themselves an emotion or a conjunct of emotions, each in a certain percentage. Such a resource has not been built so far and constructing it is the starting point of our method. The fourth component is period, culture and place dependent. It consists of the concepts that become emotion triggers due to the degree of importance they are given in the media, in addition to the emotions they are associated with. Also, important events in the history or recent past of an individual, as well as society are considered as being emotion triggers. Examples of such emotion triggers are ¨9/11¨, ¨Second World War¨, etc. It is important to make the observation that these four components are not disjoint sets, neither are they fixed as components or constant among individuals. On the contrary, each can evolve in time, when ordinary words become emotion triggers and when emotion triggers in the fourth component lose impact and become ordinary words. Furthermore, by using the principles of the theory of relevance, we state that the ¨bag of knowledge¨ (BK) consists of different levels of factors, different in importance and by assigning this importance quotient, a system analyzing text will be able to tell the difference between relevant and irrelevant information. We further consider that the interpretation is also dependent on the source of the text and the relation the reader has with it or the a priori knowledge on the degree of trust, reliability of the text source or the attitude of agreement or disagreement of the reader towards the latter. Figure 3 shows the model for the analysis of emotion in text:

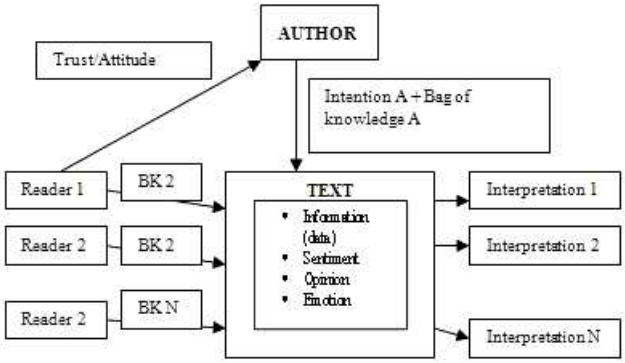


Figure 3. Model for the analysis of emotion in text

The system implemented by following the above model identifies the corresponding ¨bag of knowledge¨ of a reader and uses it to spot and classify text valence and emotion according to it.

In the following subsections, we start by presenting the steps we performed in order to build the lexical databases of emotion triggers for English and Spanish, the process of mapping the concepts found in English to their correspondents in Spanish and the process of projection to culture dependent knowledgebases for both languages. Further, we explain the method used for assigning valence and classifying emotion induced by emotion triggers. We then present the words and rules that influence the basic valence and emotion in a context and finally the rules of emotional inference derived from the theories underlying this method.

3.1 Constructing and Expanding the Core of Emotion Triggers

The core of English emotion triggers is built, at the first stage, of the approximately 37 terms found in Maslow´s pyramid of human needs, structured on 5 levels starting from the terms corresponding to the deficiency needs, found on the four bottom levels and having on top the growth needs terms, of achieving the personal potential, on level 5.

Since most of the words are general notions and their number is relatively small (37), we disambiguate them with the sense numbers they have in WordNet 2.1, in order to ensure that further on, the added words will remain with the intended meaning. For each term, we add all the senses and all grammatical categories that are valid in the context of Maslow´s pyramid levels. We then add to these words the corresponding synonyms and hyponyms from WordNet. For the verbs considered, we also add the entailed actions. We consider as having a negative value the emotion triggers that are antonyms of the nouns found. For each of the nouns and verbs, we further add the corresponding nouns and verbs, respectively, using NomLex [15]. Since NomLex does not assign sense numbers to distinguish between the possible semantics of the nouns and verbs in the collection, we use the Relevant domain concept and corresponding repository [25] to preserve the intended meaning, by taking the top relevant domain of each word sense and assigning the corresponding verb or noun in NomLex the sense number that has the same top relevant domain. If more such senses exist, they are all added.

On the other hand, another core of English words is completed with the terms found in Max Neef´s matrix of fundamental human needs. This matrix is built according to the four main characteristics of the individual: being, having, doing and interacting, for which terms are assigned in order to nine categories of needs: identity, subsistence, affection, creation, protection, freedom, participation, leisure and understanding.

When building the core of words corresponding to the taxonomy proposed by Neef, we apply the same method as the one presented above for the concepts extracted from Maslow´s 5-level pyramid.

3.2 Mapping of Concepts

Using EuroWordNet[5], we map the words in the English lexical database of emotion triggers to their Spanish correspondents, preserving the meaning through the WordNet sense numbers.

3.3 Adding World Knowledge to the Lexical Databases

The final step in building the lexical databases consists of adding real-world situations, cultural-dependent contexts terms to the two lexical databases. For English, we use the ConceptNet to add culture specific actions and terms related to the considered

[5] http://en.wikipedia.org/wiki/EuroWordNet

core of words. For Spanish, we add the cultural context by using the Larousse Ideologic Dictionary of the Spanish Language. Certainly, most words will overlap, since notions like ¨house¨ related to ¨live¨ and ¨sleep¨ related to ¨bed¨ are commonsense in both cultures. However, there are important details that in real-world texts are significantly culturally different, as for example that ¨gazpacho¨ or ¨bizcocho¨ are types of ¨food¨.

3.3.1 ConceptNet

ConceptNet[6] is a freely available commonsense knowledgebase and natural-language-processing toolkit which supports many practical textual-reasoning tasks over real-world documents right out-of-the-box (without additional statistical training) including topic-gisting, affect-sensing, analogy-making, text summarization, contextual expansion, causal projection, cold document classification, and other context-oriented inferences. Commonsense knowledge in ConceptNet encompasses the spatial, physical, social, temporal, and psychological aspects of everyday life. It contains relations such as CapableOf, ConceptuallyRelatedTo, IsA, LocationOf etc.. For the purpose of maintaining the originally intended meaning of the emotional triggers in the lexical database constructed so far, we chose to project the the emotion triggers only based on the relations DefinedAs, LocationOf, CapableOf, PropertyOf and UsedFor.

3.3.2 Larousse Dictionary of the Spanish Language

The Larousse Ideologic Dictionary of the Spanish Language (LIDSL) [2] is made up of four starts: a general classification frame , a synoptic part, an analogic part and an alphabetic index. The Dictionary offers a two-way view on words and ideas they express, thus semantically relating terms pertaining to the same idea and also, given one idea, gathering in frames all concepts related to it. In using this resource, we start from the parallel core of Maslow and Neef concepts representing the levels of needs and motivations, completed as stated before with the synonyms, hyponyms and antonyms found in WordNet, and add the Spanish culture specific terms related to them. For example, from the general concept of ¨comida¨ (¨food¨), we find as subordinated concepts ¨carne¨ (¨meat¨), ¨fruta¨ (¨fruit¨), ¨verdura¨ (¨vegetables¨) etc. These concepts are further refined to specific notions that are types of meat found in the real world: In the case of ¨carne¨, some examples are ¨vaca¨, ¨ternera¨, ¨carnero¨, ¨cordero¨, ¨matanza¨, ¨chicha¨.

3.4 Adding Valence and Classifying Emotion

Having at hand a lexical database of emotion triggers constitutes the first step towards the building of a system conforming to the model described in Figure 3, which spots possible emotional interpretation of texts in a culturally specific way, starting from the general motivational traits applicable to the whole human species.

The next step consists in assigning valence and emotion to the terms in the database. This is done with the following rules, both for the terms in Maslow´s pyramid as well as for those in Neef´s matrix:

1. The primary emotion triggers are assigned the maximum positive valence (100).
2. The terms (also emotion triggers in the final lexical database) synonyms and hyponyms of the primary emotion triggers, as well as the entailed verbs are assigned the maximum positive valence (100).
3. The terms opposed and antonym of those from 1. and 2. are assigned a maximum negative valence(-100).
4. Emotion triggers added further on inherit the valence from the emotion trigger they are related to in case of synonyms, hyponyms and entailment and change their valence from positive to negative or negative to positive in the case of antonyms.
5. Value of all emotion triggers is modified according to the valence shifters they are determined by.

For example, “shelter” from level 1 in Maslow´s pyramid has the valence 100, whereas “homeless”, which is opposed to it, has the valence -100. These values will be weighted according to the pyramid level they are found on (this process is explained in section 4).

Further on, we assign an emotion triggers a value on each of the 6 categories of emotion proposed for classification in the SemEval Task No. 14 – joy, sadness, anger, fear, disgust and surprise, using the following rules:

1. The emotion triggers found in the levels of Maslow´s pyramid of needs and those found in the components of Neef´s matrix of fundamental human needs are manually annotated with scores for each of the 6 categories.
2. The terms (also emotion triggers in the final lexical database) synonym and hyponym of the primary emotion triggers, as well as the entailed verbs are assigned the same scores as the terms they are related to.
3. The terms opposed and antonym of those from 1 and 2.. are assigned initial values.
4. Emotion triggers added further on inherit the valence from the emotion trigger they are related to in case of synonyms, hyponyms and entailment and change their value from positive to negative or negative to positive in the case of antonyms.
5. Value of all emotions of an emotion triggers is modified according to the valence shifters they are determined by.
6. If any of the values calculated in 5. is higher than 100, it is set to 100; if it is lower than -100, it is set to -100.

3.5 Valence Shifters

In order to be able to recognize the change in meaning of emotion triggers due to modifiers, we have defined a set of

[6] http://web.media.mit.edu/~hugo/conceptnet/

valence shifters – words that negate the emotion triggers, intensify or diminish their sense. The set contains:

- Words that introduce negation (no, never, not, doesn´t, don´t and negated modal verbs)
- A set of adjectives that intensify the meaning of the nouns they modify – big, more, better etc.
- A set of adjectives that diminish the meaning of the nouns they modify – small, less, worse etc.
- The set of modal verbs and conditional of modal verbs that introduce uncertainty to the active verb they determine- can, could, might, should, would
- The set of modal verbs that stress on the meaning of the verb they determine - must
- A set of adverbs that stress the overall valence and intensify emotion of the context – surely, definitely etc
- A set of adverbs that shift the valence and diminish emotion of the context – maybe, possibly etc

For each of the valence shifters, we define a weight of 1.5 for the meaning intensifiers and 0.5 for the meaning diminishers. These are coefficients that will be multiplied with the weight assigned to the emotion trigger level and emotions- level association ratio corresponding to the given emotion trigger in the case of emotion triggers built from Maslow´s pyramid. In the case of emotion triggers stemming from Neef`s matrix of fundamental human needs, the weights of the valence shifters are multiplied with the emotion-category association ratio, computed for each emotion trigger and each of the four existential categories. The valence shifters for Spanish are determined by translating the English valence shifters and adding synonymous terms and expressions.

3.6 Emotion Trigger Association Ratio

The association ratio score provides a significance score information of the most relevant and common domain of a word. The formula for calculating it is:

$$AR(w;D) = \Pr(w,D)\log_2 \frac{\Pr(w,D)}{\Pr(w)\Pr(D)}, \text{ where:}$$

- Pr(w,D) is the probability of the word in the given domain
- Pr(w) is the probability of the word
- Pr(D) is the probability of the domain

In our approach, besides quantifying the importance of each emotion trigger in a manner appropriate to the level and emotion it conveys, we propose to use a variant of the association ratio that we call emotion association level. This score will provide the significance information of the most relevant emotion to each level. The corresponding formula is therefore:

$$AR(e;L) = \Pr(e,L)\log_2 \frac{\Pr(e,L)}{\Pr(e)\Pr(L)}, \text{ where:}$$

- Pr(e,L) is the probability of the emotion in the given level
- Pr(e) is the probability of the emotion
- Pr(L) is the probability of the level

3.7 The Construction-Integration Model

The Construction-Integration Model is a psychological model of text comprehension [10], based on the idea that while reading a text, a person will activate the features of words that are appropriate to the context and inhibit those that are not.

The construction-integration model has been so far successfully used in the field of Natural Language Processing for anaphora resolution, generation of representations of word meanings from dictionaries [20] and automatic assessment of summarizations [12]. Also, the author proposed a computational method for metaphor comprehension [11] based on this cognitive model.

The process of text comprehension consists of two phases. The first one – the construction - uses rules in the form of a production system to generate, from the linguistic representation of words, a propositional network of related mental elements. Further, adding the knowledge experience of the reader, a more elaborated propositional network is created.

The second phase – integration- takes as input the crude representation of text in the form of the elaborated propositional network, with nodes linked with positive and negative connections meant to represent the relations between them, and tunes it using connectionist relaxation techniques.

4 SYSTEM FOR VALENCE AND EMOTION

The final system built to classify text at valence and emotion level follows the scheme depicted in Figure 4:

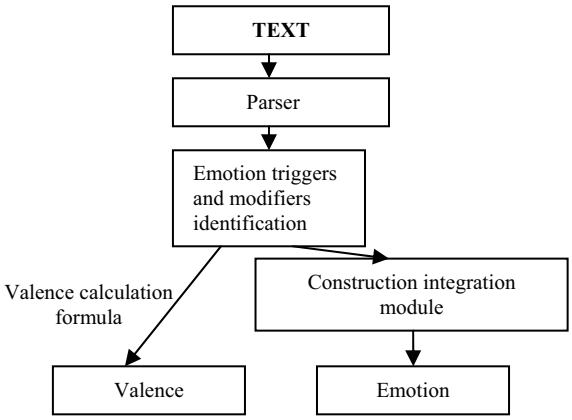


Figure 4. System for valence and emotion classification-components

First, the input text is parsed with Minipar for English and FreeLing[7] for Spanish to obtain for each word the grammatical category, the lemma and its modifiers. Further on, the emotion triggers in the text are identified, together with their corresponding modifiers.

We calculate the valence of the text on the basis of the identified emotion triggers and their modifiers, using the formulas described in what follows.

In the case of emotion triggers obtained from Maslow´s pyramid, we calculate a score called weighted valence of emotion trigger (wv) using the following formula:

[7] http://www.lsi.upc.es/~nlp/freeling/

$wv(et_{ij}) = w(m) * w(l_j) * v(et_i)$, where

- w(m) is the weight of modifier
- $w(l_j)$ is the weight of level
- $v(et_i)$ is the emotion trigger valence
- i is the index of the emotion trigger
- j is the number of the level

The weight of the level is 1 for level 1, 0.8 for level 2, 0.6 for level 3, 0.4 for level 4 and 0.2 for level 5.

In the case of emotion triggers obtained from Neef´s matrix, we calculate a score called weighted valence of emotion trigger(wv) using the following formula:

$wv(et_i) = w(m) * v(et_i)$, where

- w(m) is the weight of modifier
- $v(et_i)$ is the emotion trigger valence
- i is the index of the level

The total valence of the text equals the sum of all weighted valences of all emotion triggers, mapped to -1, if the value is below -50, to 0, if the value is between -50 and 50 and to 1 if the value is above 50. The mapping is done because the final evaluation is done on a coarse (-1,0,1) level.

We will show an example of computation for the headline: "More sleep, healthier slimmer children".

The emotion triggers found in the text are "sleep", "healthier" and "children". The first two terms belong to level 1 in Maslow´s pyramid and are among the core terms; therefore, their values will be 100. The term "sleep" is also determined by the meaning intensifier "more" and thus its final weighted valence is 150. The term "children" is related to the term "family" on level 3, therefore its value will be 60. The sum of these three values is above 50, so the valence of the headline is 1.

Further, we calculate the emotions present in the text, by the following method:

- for each emotion trigger stemming from Maslow´s pyramid, we compute the emotion to level association ratio
- for each emotion trigger stemming from Neef´s matrix, we compute the emotion to category association ratio

We then apply the Construction Integration Model in a manner similar to that described in [12] and construct a spreading activation network. We consider the working memory as being composed of the set of emotion triggers and their emotion association ratio value multiplied with the values for emotions of the emotion trigger, which is considered as activation value. The semantic memory is set up of the modifiers and the top 5 synonyms and antonyms of emotion triggers with their AR value. The number was set empirically in order to prevent a high number of nodes in the network. We set the value of each emotion trigger to 1. We create a link between all concepts in the semantic memory with all the emotion triggers. We consider the strength of link the higher of the two Emotion trigger Association Ratio scores.

The text is processed in the order in which emotion triggers appear and finally we obtain the activation value for each emotion trigger.

The output for the values of the emotions in text is obtained by multiplying the activation values with 100 and adding the scores obtained for the same emotion from different emotion triggers when it is the case. Values of emotions above 50 are set to 1 and values of emotions not exceeding this threshold are considered 0.

5 EXPERIMENTS AND EVALUATION

The evaluation of the system presented was done using the test data provided within the SemEval Task No. 14: Affective Text test set [22] and its Spanish translation. In the task proposed in SemEval, the objective was to assign valence – positive or negative - and classify emotion of 1000 news headlines provided as test set that were previously extracted from news web sites according to 6 given emotions: joy, fear, sadness, anger, surprise and disgust and their translation to Spanish. The results we obtained in the coarse-grained evaluation are presented in Table 1 for valence classification and in Table 2 for one of the 6 emotions- fear :

	Acc	Prec	Rec	F
English	70.15	75.23	65.01	69.74
Spanish	65.02	71.1	66.13	68.52

Table 1. System results for valence annotation

	Acc	Prec	Rec	F
English	95.16	47.21	45.37	46.27
Spanish	95.2	46.01	43.84	44.89

Table 2. System results for emotion annotation for ¨fear¨

For the F measure we considered alpha 0.5.

In the competition, the best results achieved for coarse-grained valence classification were of 55.10 in Accuracy, 61.42 in Precision for one system and Recall 66.38 and an F-measure of 42.43 for another system, whose first two scores were lower. In case of coarse-grained classification of emotion, the best result for Accuracy was 87.9 and Precision 33.33 in a system which however had a very low recall and F measure score. Higher scores for Recall and F-measure were obtained by another system, which had only 75.30 Accuracy and 16.23 Precision. Although the results show relevant improvements over the ones obtained by previously built systems, in using such a complex system, one could and should use a more complex set of emotions. The set of emotions is rather limited and sometimes does not allow for an accurate assignment of the appropriate emotion for the emotion triggers, but according to a conventional classification assigned to remain within the given set of 6 emotions.

6 RELATED WORK

In the field of sentiment analysis, as far as we know, there is no similar approach to the one presented in this paper.

Work in sentiment analysis was done for languages other than English. For example, [24] propose an empirical method of sentiment analysis for Chinese and [2] presents an approach for multilingual sentiment analysis based on SentiWordNet and performs an evaluation on German movie reviews.

However, the difference in our approach is not given by the fact that we study sentiment in Spanish. The novelty consists firstly in the fact that we determine words that until now have not been

thought of as “sentiment-bearing” words, and show that their emotional value is given by their close link to human needs and motivations – the factors that determine to a high extent emotion. Secondly, our approach signals the distinction between the cognitive and the emotional aspects of sentiment analysis. Whereas others concentrated on seeking emotion in text in order to discover the sentiment present in the events described, we concentrated on the emotion raised by the text to the reader.

7 CONCLUSIONS & FUTURE WORK

In this paper we presented a method to assign valence and classify emotion in text starting with a database of cultural dependent emotion triggers derived from a theory in pragmatics and two motivational and need-based theories. The final classification of texts was done using the cognitive model of construction and integration, the emotion to level and emotion to category association ratio and taking into account valence shifters. It was observed that the system outperformed previously obtained results. In order for the system to be complete, we should also build the fourth component of the system, by applying the system on large corpora of news and of world and culture specific data. Part of the future work is thus applying the described method to classify valence and emotions from recent news texts and also world and culture dependent historical facts describing texts. Another direction for future work consists in applying a larger set of emotions for classification.

REFERENCES

[1] Al Masum Shaikh, M., Prendinger, H., Mitsuru, I.. Assessing Sentiment of Text by Semantic Dependency and Contextual Valence Analysis. In Lecture Notes in Computer Science. Volume 4738/2007. Pages 191-202

[2] Denecke, K.: Using SentiWordNet for Multilingual Sentiment Analysis. In Proceedings of the International Conference on Data Engineering (ICDE 2008), Workshop on Data Engineering for Blogs, Social Media, and Web 2.0, Cancun, 2008

[3] Diccionario Ideológico de la Lengua Española, Larousse Editorial, RBA Promociones Editoriales, S.L., ISBN 84-8016-640-1

[4] Esuli, A., Sebastiani, F. SentiWordNet: A Publicly Available Resource for Opinion Mining. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.

[5] Esuli, A., Sebastiani, F.: Determining the semantic orientation of terms through gloss analysis. In Proceedings of CIKM, pp. 617-624 (2005)

[6] Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database, MIT Press, Cambridge, Massachusetts (1999)

[7] Hu., M. Liu, B.: Mining and summarizing customer reviews. In Proceedings of KDD (2004)

[8] Jiang, H., Vidal, J. M. From Rational to Emotional Agents. In Proceedings of the AAAI Workshop on Cognitive Modeling and Agent-based Social Simulation, 2006.

[9] Kim, S.M., Hovy, E.H.: Identifying and Analyzing Judgement Opinions. In Proceedings of HLT-NAACL 2006, ACL, p. 200-207 (2006)

[10] Kintsch, W. Comprehension: A Paradigm for Cognition. Cambridge Press (1999)

[11] Kintsch, W. Metaphor Comprehension: A computational theory. Psychonomic Bulletin & Review, 2000

[12] Lemaire, B., Mandin,S., Dessus, Ph., Denhière, G. Computational cognitive models of summarizatiuon assessment skilla. In Proceedings of the 27th Annual Conference of the Cognitive Science Society (CogSci' 2005), B. G. Bara, L. Barsalou and M. Bucciarelli, Ed.Mahwah: Erlbaum, 2005, pp. 1266–1271.

[13] Liu, H. & Singh, P. (2004) ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal, To Appear. Volume 22, forthcoming issue. Kluwer Academic Publishers.

[14] Liu, H., Maes, P. (2007): Introduction to the Semantics of People & Culture (Editorial Preface), International Journal on Semantic Web and Information Systems, Special Issue on Semantics of People and Culture (Eds. H. Liu & P. Maes), 3(1), Hersey,PA: Idea Publishing Group.

[15] Macleod,C., Grishman,R., Meyers,A., Barrett,L., Reeves, R. (1998) NOMLEX: A Lexicon of Nominalizations. Proceedings of EURALEX'98, Liege, Belgium, August 1998.

[16] Magnini, B., Cavaglià, R.. "Integrating Subject Field Codes into WordNet". In Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S. and Stainhaouer G. (Eds.) Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece, 31 May - 2 June, 2000, pp. 1413-1418.

[17] Maslow, A.H. A Theory of Human Motivation, Psychological Review 50 (1943):370-96.

[18] Max-Neef, M. A. (1991): Human scale development: conception, application and further reflections. The Apex Press. New York

[19] Mihalcea, R., Banea, C., Wiebe, J. Learning Multilingual Subjective Language via Cross-Lingual Projections, in Proceedings of the Association for Computational Linguistics (ACL 2007), Prague, June 2007.

[20] Powell, C., Zajicek, M., Duce, D. (2000): "The generation of representations of word meanings from dictionaries", In ICSLP-2000, vol.3, 482-485.

[21] Sperber, D., Wilson, D. (2004) "Relevance Theory" in G. Ward and L. Horn (eds) Handbook of Pragmatics. Oxford: Blackwell, 607-632.

[22] Strapparava, C. Valitutti, A. "WordNet-Affect: an affective extension of WordNet". In Proceedings ofthe 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, May 2004, pp. 1083-1086.

[23] Strapparava, C., Mihalcea, R. (2007) SemEval-2007 Task 14: Affective Text. In Proceedings of the th International Workshop on Semantic Evaluations(SemEval 2007), Pages 70-74. Prague, June 2007.

[24] Tan, S. , Zhang, J., An empirical study of sentiment analysis for chinese documents, Expert Systems with Applications (2007), doi:10.1016/j.eswa.2007.05.028

[25] Vázquez,S., Montoyo, A., Rigau, G. Using relevant domains resource for word sense disambiguation. In ICAI, pages 784-789, 2004.

[26] Wiebe, J., Mihalcea,R. Word Sense And Subjectivity, in Proceedings of the Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06), Sydney, Australia, July 2006.

[27] Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation 39(2-3), pp. 165-210 (2005)

Feeler: Emotion Classification of Text Using Vector Space Model

Taner Danisman[1] and **Adil Alpkocak**[1]

Abstract. Over the last quarter-century, there is increasing body of research on understanding the human emotions. In this study, automatic classification of anger, disgust, fear, joy and sad emotions in text have been studied on the ISEAR (International Survey on Emotion Antecedents and Reactions) dataset. For the classification we have used Vector Space Model with a total of 801 news headlines provided by "Affective Task" in SemEval 2007 workshop which focuses on classification of emotions and valences in text. We have compared our results with ConceptNet and powerful text based classifiers including Naive Bayes and Support Vector Machines. Our experiments showed that VSM classification gives better performance than ConceptNet, Naive Bayes and SVM based classifiers for emotion detection in sentences. We achieved an overall F-measure value of 32.22% and kappa value of 0.18 for five class emotional text classification on SemEval dataset which is better than Navie Bayes (28.52%), SVM (28.6%). We have tested and discussed the results of classification using cross-validation technique for emotion classification and sentiment analyses on both the ISEAR and SemEval datasets. In addition to the classification experiments we have developed an emotion enabled video player which automatically detects the emotion from subtitle text of video and displays corresponding emoticon.

Keywords: Emotion Detection in Text, Vector Space Model, Emotion Perception, Human Computer Interaction.

1 INTRODUCTION

Current state-of-art in computer human interaction largely ignores emotion whereas it has a biasing role in human-to-human communication in our everyday life. In the mean time, a successful computer human interaction system should be able to recognize, interpret and process human emotions. Affective computing could offer benefits in an almost limitless range of applications. However, the first step is Human Emotion Recognition (HER), and it is getting more attention recently. In HER, the data gathered to recognize human emotion is often analogous to the cues that humans use to perceive emotions in others. Hence, human emotion recognition is multimodal in nature, and includes textual, visual and acoustic features. Text seems to be the most studied modality since the text is relatively easier to process than others.

HER from text can be simply envisioned to be a classification problem of a given text according to predefined emotional classes. In this case, it first requires a preparation of proper training set for each emotional class and selection of good features. One of the solutions for this issue is Bag of Word (BoW). It's very similar to keyword spotting [5] and lexical affinity [6]. BoW approach that is widely used in information retrieval, and tries to generate a good lexicon for each emotional class and feature extraction. However, creation of emotional lexicon is both time consuming and labor-intensive task since usually requires manual annotations. On the other hand, the number of words in lexicons is very limited, and it is not desired for most classifiers using the BoW approach. Moreover, user's vocabulary may differ from the document vocabulary. In literature, an alternate approach for this issue has been proposed by [1]. They use blog based emotion datasets, where every blog document is already labeled by authors. It seems that it is good for generating large scale lexicon for a better representation for a given language. Blogs have more than 200-300 words per document on average. However, assigning a single emotional label to a document having many words is not very meaningful. Therefore, a better training set for each emotional class must consider sentences and words, not paragraphs. After preparing a proper training set and selecting good features, the next task is to classify a given text.

To date, many approaches have been proposed for HER from text. These approaches can be grouped into three main groups: keyword spotting, statistical NLP, and ontology based approaches. Each approach has its own advantages and disadvantages. In addition, there is no rigid line between these approaches. Keyword spotting is easy to implement, and based on predetermined set of terms to classify the text into emotion categories. Despite its simplicity, creation of an effective lexicon is difficult too since only 4% of words used in texts have emotional value [14]. For these reasons it is not suitable for wide range of domains. The second group is based on statistical NLP approaches. This approach is similar to lexical affinity where affinities of words are still used but as a feed for a machine learning algorithm. In case of lexical affinity, words have some probabilistic value representing the affinity for a particular emotion class. However, it requires high quality, large-scale training dataset for a better classification. The third groups is based on ontologies, heavily uses semantic networks like WordNet-Affect [4] and ConceptNet [15] are linguistic resources for lexical representation of affective information using commonsense knowledge. ConceptNet is an integrated commonsense knowledgebase with a natural language processing toolkit MontyLingua which supports many practical textual reasoning tasks over real world documents without additional statistical training.

In this paper, we propose a VSM approach for HER from text.

[1] Computer Engineering Department, Dokuz Eylul University, Tinaztepe Campus, 35160 Izmir/TURKEY.
Email: {taner,alpkocak}@cs.deu.edu.tr

- We have used sentences from ISEAR [2] dataset, emotional words from Wordnet-Affect and polarity of words from WPARD datasets.
- Our approach uses Vector Space Model for HER.
- We measured the effect of stemming and emotional intensity on emotion classification in text.
- Third, we have developed an emotion enabled video player which automatically detects the emotions from subtitle text and displays emotions as emoticons during video play.

The rest of the paper is organized as follows. In section two, we have explained the emotion classification problem in text. In section three, Vector Space Model and methodology is presented. Section four shows the experimental results performed on the SemEval test set. Finally, section five concludes the study and provides a general discussion.

2 PROBLEM DEFINITION and RELATED WORK

One side of the problem is the selection of a qualified dataset for machine learning methods. In order to cover most of the words in a given language, a large-scale dataset is needed. In addition this dataset should have variation of emotional content, independent emotional responses from different cultures to eliminate cultural affects of emotion.

Manual creation of large-scale datasets is difficult and time consuming task. Blog based datasets provides large-scale lexicons as presented in [1]. They worked on large collection of blog posts (122,624 distinct web pages) for classifying blog text according to the mood reported by its author during the writing. According to their results, increasing the amount of training data leads an additional increase in classification performance. On the other hand, the quality of the dataset is important for better classification.

All these requirements lead us to use ISEAR (International Survey on Emotion Antecedents and Reactions) dataset in our experiments. ISEAR consists of 7,666 sentences and snippets in which 1096 participants from fields of psychology, social sciences, languages, fine arts, law, natural sciences, engineering and medical in 16 countries across five continents completed a questionnaire about the experiences and reactions to seven emotions in everyday life including joy, fear, anger, sadness, disgust, shame, and guilt. Surprisingly, ISEAR dataset is not studied yet for text based emotion classification. Previous studies using the ISEAR dataset try to find relationships among emotions and different cultures, genders, ages, and religions. On the other hand this corpus is well suited to use for emotional text classification purposes. Table 1 shows samples from this dataset for the anger emotion.

"A close person lied to me".
"A colleague asked me for some advice and as he did not have enough confidence in me he asked a third person".
"A colleague asked me to study with her. I could not explain things as perfectly as she had expected. So she reacted in an aggressive manner."
....

Table 1. ISEAR anger samples

2.1. Related Work

Achievements in this domain can be used in next generation intelligent robotics, artificial intelligence, psychology, blogs, product reviews, and finally development of emotion-ware applications such as emotion-ware Text to Speech (TTS) engines for emotional reading of text. CRM and service oriented companies like Right Now Technologies and NICE Systems produces customer service software SmartSense™ and NICE Perform™ respectively which recognizes customer emotions using keyword spotting technique and prosodic features of speech then performs flagging, prioritizing and routing inquiries and customers based on emotional content.

[7] developed a new aggregator to fetch news from different news resources and categorize the themes of the news into eight emotion types using semantic parsers and SenseNet [8]. [9] studied the natural language and affective information using cognitive structure of affective information. They developed ALICE chat-bot based on Artificial Intelligence markup language (AIML) script to improve interaction in a text based instant messaging system that uses emoticons or avatar that represents the sensed emotion to express the emotional state.

According to [10] emotion annotation for text is a hard problem and inter-annotator agreement value k=.24-.51. [11] employed a commonsense knowledgebase OMCS (Open Mind Common Sense) having 400,000 facts about everyday world to classify sentences into basic emotions (happy, sad, angry, fearful, disgusted, and surprised) categories. [5] developed an emotion extraction engine that can analyze the input text in a chat dialogue, extract the emotion and displays the expressive image on the communicating users display. Their parser only considers sentences in present continuous tense, sentences without starting auxiliary verbs (No question sentences allowed), positive sentences, etc. [12] considered the emotional expressions for text-to-speech engines and emotional reading. They partitioned the text into nouns adjectives and adverbs and used the frequency of words to determine the emotional class. [13] tried to detect emotion from both speech and textual data. They manually defined the emotional keywords and emotion modification words. They have used "very" and "not" as a modification word where the only difference between "very happy", "happy", and "not happy" is the emotional intensity. As they are using keyword-spotting technique (they have 500 words labeled as emotion words), they reported that textual recognition rate is lower than speech based recognition. According to their work, emotion recognition performance of multimodal system is better than performance of individual modalities.

3 VECTOR SPACE MODEL

Vector Space Model (VSM) is widely used in information retrieval where each document is represented as a vector, and each dimension corresponds to a separate term. If a term occurs in the document then its value in the vector is non-zero. Let us assume that we have *n* distinct terms in our lexicon. Then, lexicon, ℓ, is represented as a set of ordered terms, and more formally, it is defined as follows:

$$\ell = \{t_1, t_2, t_3, ..., t_n\}$$

Then, an arbitrary document vector, $\vec{d}_i$, is defined as follows:

$$\vec{d}_i = \langle w_{1i}, w_{2i}, ..., w_{ni} \rangle$$

where w_{ki} represents the weight of k^{th} term in document i. In literature, there several different ways of computing these weight values have been developed. One of the best known schemes is *tf-idf* weighting. In this scheme, an arbitrary normalized w_{ki} is defined as follows;

$$w_{ki} = c(t_k, d_i) = \frac{tf_{ik} \log(N/n_k)}{\sqrt{\sum_{k=1}^{n} (tf_{ik})^2 [\log(N/n_k)]^2}} \text{ where;}$$

$t_k = k^{th}$ term in document d_i

tf_{ik} = frequency of the word t_k in document d_i

$idf_k = \log\left(\frac{N}{n_k}\right)$ inverse document frequency of word t_k in entire dataset

n_k = number of documents containing the word t_k,
N = total number of document in the dataset.

Each emotion class, M_j, is represented by a set of documents, $M_j = \{d_1, d_2, \ldots, d_c\}$. Then, we have created a model vector for an arbitrary emotion, $\vec{E}_j$, by taking the mean of $\vec{d}_j$ vectors for an arbitrary emotion class. More formally, each $\vec{E}_j$ is computed as follows:

$$\vec{E}_j = \frac{1}{|M_j|} \sum_{\vec{d}_i \in M_j}^{|M_j|} \vec{d}_i$$

where $|M_j|$ represents the number of documents in M_j. After preparing model vectors for each emotion class, the whole system is represented with a set of model vectors, $D = \{E_1, E_2, ..., E_s\}$ where s represents the number of distinct emotional classes to be recognized.

In VSM, documents and queries are represented as vectors, and cosine angle between the two vectors used as similarity of them. Then normalized similarity between a given query text, Q, and emotional class, E_j, is defined as follows:

$$\text{sim}(Q, E_j) = \sum_{k=1}^{n} w_{kq} * E_{kj}$$

In order to measure the similarity between a query text and the D matrix of size $s \times n$, first we convert the query text into another matrix $n \times 1$ similar to D where n is the size of the lexicon and s is the number of emotions. Then for each emotion (each row of D matrix), we make multiplication between the query matrix Q and one row of D matrix. After these multiplications we have m scalar values representing the cosine similarity. The index of the maximum of these values is selected as the final emotional class. More formally:

The classification result is then,

$$\text{VSM}(Q) = \arg\max_j (\text{sim}(Q, E_j))$$

The basic hypothesis in using the VSM for classification is the contiguity hypothesis where documents in the same class form a contiguous region, and regions of different classes do not overlap.

4 EXPERIMENTATION

Before starting on a research on emotion classification, the first question is "Which emotions should be addressed?" There are many different emotion sets exists in the literature including basic emotions, universal emotions, primary and secondary emotions, neutral vs. emotional, and for some cases the problem is reduced to a two class classification problem (Sentiment Analysis) using the Positive and Negative values as class labels. Simple classification sets give better performance than expanded sets of emotions which require cognitive information and deeper understanding of the subject. In our research study, we have used five emotion classes (anger, disgust, fear, sad, and joy) that form the intersection between the ISEAR dataset and the SemEval test set. Therefore, the number of emotion classes s=5.

For the classification, we have used Naïve Bayes, Support Vector machines and Vector Space Model classifiers. We have considered the effect of the stemming, negation and intensity of emotions on classification performance. We have used WEKA tool [16] for the Naïve Bayes and SVM classification. In order to compare the performance of VSM and other classifiers, we have considered the mean F1-measure value and the kappa statistics which considers the inter-class agreements.

First, we have used set theory, which deals with collections of abstract objects to find the intersections and set differences of objects in a given set. For the graphical simplicity, we only show three emotional classes (anger, disgust, and fear) with a few words in Figure 1 where each circle represents an emotional class and entries represent the words.

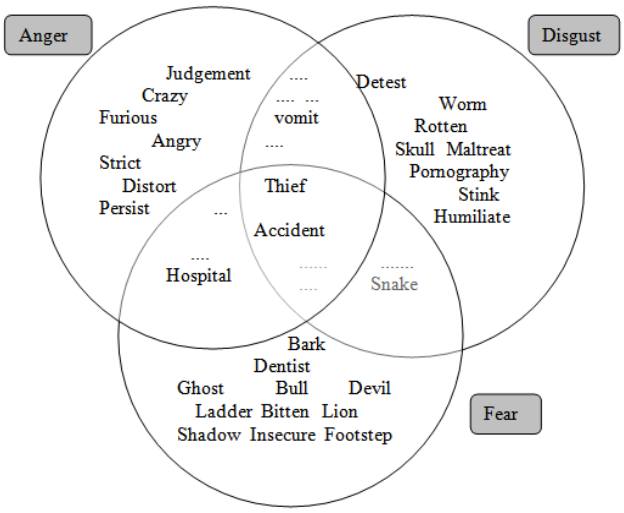


Figure 1. Words belonging to specific emotions in ISEAR dataset after finding set differences

Using the set difference words, the word "ghost, bark, dentist and lion" is appeared only in sentences representing the fear emotion whereas "successful, rejoice, sunshine, tremendous,

accredit, and carnival" appeared only in joy sentences. The following table represents the automatically extracted emotion words, which are non-intersected words in ISEAR database.

Anger	Disgust	Fear	Joy	Sad	Shame	Guilt
rules	detest	ghost	happiest	weep	aloud	pinch
strict	worm	bark	joyous	sadden	zip	decline
disregard	rotten	dentist	envelop	coma	infidelity	feign
judgment	skull	devil	success	farewell	underwear	counselor
crazy	hypocrite	ladder	rejoice	drought	smear	harsh
hatred	stink	bitten	ecstatic	tragic	dialect	caretaker
persist	humiliate	lion	sunshine	saddest	spell	insinuate
…	…	…	…	…	…	…

Table 2. Emotion related words from ISEAR dataset using set difference for each emotion set.

These results give hope us to use tf-idf (Term Frequency-Inverse Document Frequency) values for emotion classification because these non-intersected words have very low term frequency values (among 1-10) but have high inverse document frequency values. In addition, by using the tf-idf values we are also able to use the words in intersected areas. Therefore, we selected to use VSM (Vector Space Model) for emotion classification in text.

4.1. Stop Word removal strategy

This study also showed that some words are only appeared in specific sentences belonging to a single emotion, so stop-word removal based on minimum term frequencies is not suitable for emotion detection. Stop words are usually the most frequent words including articles (a, an, the), auxiliary verbs (be, am, is, are), prepositions (in, on, of, at), conjunctions (and, or, nor, when, while) that do not provide additional improvement for search engines but increase the computational complexity by increasing the size of the dictionary. The important aspect of stop-word removal in emotion detection is the words, not their frequencies. There are several publically available stop-word lists available where these lists consist of approximately 400-500 most frequent words in a given language. However, public stop-word lists consider the information retrieval and they do not consider words carrying emotional content. Therefore we first need to remove some of the emotional words from the stop-word list including negative verbs (not, is not, does not, do not, should not, etc.). In addition, we replaced the word "very" with blank and the word "blank not blank" is replaced by "blank not". In addition, Words in Table 2 are removed from the stop-word list to improve the classification rate. We ignored the part of speech tagging on input text because of its effect of reducing the classification accuracy as described in [17].

Since non-alpha tokens are automatically removed by TMG [18], the exclamation marks and question marks are replaced by descriptive new words "XXEXCLMARK" and "XXQUESMARK" respectively. Negative short forms are also replaced by negative long forms such that "doesn't" is replaced by "does not". After these replacements, the following sentences are changed as follows:

"I don't love you!" => "I do not love you XXEXCLMARK" => "I do NOTlove you XXEXCLMARK"

"I am not very happy." => "I am not happy." => "I am NOThappy."

As seen in the above examples, the word "happy" and "love" is used to create new words "NOTlove" and "NOThappy". In this way, we can discriminate the word "love" having positive meaning and "NOTlove". In the same way, the new word "NOThappy" has a negative meaning.

Initially we have used stemming for finding morphological root of a given word. Stemmers in linguistic are widely used in search engines and query based systems to improve the efficiency of these systems. For emotion classification, stemming also removes the emotional meaning from the words. We found that tense information also affects the emotional meaning of the words. For example the words "marry" and "love" is frequently shown in joy sentences while the words "married" and "loved" are appeared in sad sentences.

4.2. Training and Test sets

For training, we have used combination of ISEAR, Wordnet-Affect and WPARD datasets. Testing is performed on SemEval Task 14 "Affective Text" test set.

Our main training dataset, ISEAR, is further expanded by adding emotional words from Wordnet-Affect [4] and WPARD (Wisconsin Perceptual Attribute Rating Database) [3] to improve the emotional classification of sentences. Each word in Wordnet-Affect and WPARD is replicated up to average number of terms per document which is 16 (as seen on Table 4) in our experiment to make ISEAR like sentences. In this case, the sentences are constructed using the same words.

WPARD is like a polarity dataset were collected from 342 undergraduate students using online form to rate how negative or positive were the emotions they associated with each word, using a scale from -6 (very negative feeling) to +6 (very positive feeling), with 0 being a neutral feeling. Table 3 shows samples from this dataset.

Word	Value	Word	Value
rape	-5.60	hope	+4.43
killer	-5.55	honeymoon	+4.48
funeral	-5.47	home	+4.50
slavery	-5.41	sunset	+4.53
cancer	-5.38	beach	+4.58
corpse	-4.95	family	+4.58
slave	-4.84	friend	+4.60
war	-4.78	peace	+4.62
coffin	-4.73	kiss	+4.64
morgue	-4.72	holiday	+4.73
cigarette	-4.49	fun	+4.91

Table 3. Sample cross-section from WPARD [3] dataset

Before extracting the features, we have preprocessed the ISEAR dataset and manually eliminated some of the inconsistent and incomplete entries (such as "[No response]" lines). Normalization is performed using the TMG toolbox [18] and get the following distribution as seen in Table 4.

Emotion	Number of sentences	# of words before stop word removal	Average # of terms before normalization	Average # of terms after normalization
Angry	1,072	26,3	24.8	17.7
Disgust	1,066	22,8	21.6	15.8
Fear	1,080	25,6	23.9	17.1
Joy	1,077	21,1	19.8	14.2
Sad	1,067	21,3	20.2	14.6
Shame	1,052	24,9	23.9	16.9
Surprise	1,053	23,5	22.6	15.9
Average	1,066	23,7	22.4	16.0

Table 4. Number of sentences per emotion in ISEAR Dataset

SemEval Task 14 "Affective text" test set is used for testing. Table 5 shows the sample cross-section in XML format and Table 6 shows corresponding ground truth for this test set.

```
<corpus task="affective text">
<instance id="500">Test to predict breast
cancer relapse is approved</instance>
<instance id="501">Two Hussein allies are
hanged, Iraqi official says</instance>
<instance id="502">Sights and sounds from
CES</instance>
<instance id="503">Schuey sees Ferrari unveil
new car</instance>
…
```

Table 5. Sample cross-section from SemEval test set

Instance Id	Anger	Disgust	Fear	Joy	Sadness	Surprise
500	0	0	15	38	9	11
501	24	26	16	13	38	5
502	0	0	0	17	0	4
503	0	0	0	46	0	31
…	…	…	…	…	…	…

Table 6. Corresponding ground truth data for SemEval test set

4.3. Experiments

In order to build up the D matrix, first we made normalizations including limited stop-word elimination, term-length thresholds, which is 3 in our case. We did not consider global and local thresholds. Average number of terms per document before the normalization is 22.43 and after the normalization number of index terms per document is 16 and the dictionary size is 5,966 terms. This result leads us to a D matrix of size 7,466×5,966. As the size of average number of index term elements per document is 16, the D matrix is very sparse. After computing E_j vectors, the new size is 5×5,966.

After the normalization step, we have computed the term frequency and inverse document frequency (tf-idf) values that provide a level of information about the importance of words within the documents. The tf–idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document and how important the word is to all the documents in the collection.

Experiment 1: We studied the effect of emotional intensity to classification performance on the SemEval test set. In our experiment we have selected emotions having either positive or negative valence value greater than a threshold T where T is between 0-70. According to Figure 2, F1-Measure value increases proportionally with the T when T is between 30 to 70. It shows that increased emotional intensity also increases the classification performance.

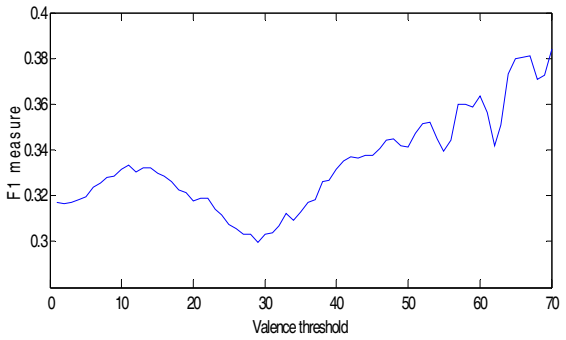


Figure 2. Valence threshold versus F1-measure on VSM classifier

Experiment 2: We have studied the effect of stemmers on emotion classification in text using 10 fold cross-validation on the ISEAR dataset and on unseen test set SemEval. A stemmer is an easy to implement algorithm which determines a stem (or morphological root) form of a given inflected (or, sometimes, derived) word form. In some cases it is known as suffix remover. We found that some words having the same morphological root can have different emotional meaning. For example, the word "marry" classified as joy and "married" is classified as sad emotions. In spite of this, those samples are very limited. Our experiments showed that use of the stemming algorithms still gives additional increase in classification accuracy as seen in Table 7, Table 8, and Table 9 for Naïve Bayes, SVM and VSM classifiers. Bold values represent the best scores considering three classifiers.

		Naïve Bayes Classifier 5 Class Emotional Classification					
		10 Fold cross validation on the ISEAR dataset			Test on SemEval test set		
Training Set	Stemming	Kappa	Mean F1	Accuracy	Kappa	Mean F1	Accuracy
ISEAR	Yes	.59	67.0	67.2	.14	27.1	31.3
	No	.59	67.2	**67.4**	.09	23.3	26.8
ISEAR+WPARD+ WORDNET_AFFECT	Yes	.51	61.2	60.8	.16	**29.0**	35.0
	No	.46	57.8	57.0	.12	25.3	30.8

Table 7. Naive Bayes results for five class emotion classification

Training Set	Stemming	Support Vector Machine 5 Class Emotional Classification					
		10 Fold cross validation on the ISEAR dataset			Test on SemEval test set		
		Kappa	Mean F1	Accuracy	Kappa	Mean F1	Accuracy
ISEAR	Yes	.59	67.5	**67.4**	.11	24.5	27.2
	No	.58	67.0	66.9	.09	23.4	26.4
ISEAR+WPARD+ WORDNET_AFFECT	Yes	**.61**	**68.3**	70.2	.12	24.9	27.0
	No	.56	65.0	67.1	.09	23.7	28.0

Table 8. Support Vector Machine results for five class emotion classification

Training Set	Stemming	Vector Space Model Classifier		
		SemEval test set		
		Kappa	Mean F1	Accuracy
ISEAR	Yes	0.16	28.7	**36.0**
	No	0.11	26.1	32.0
ISEAR+WPARD+ WORDNET_AFFECT	Yes	**0.17**	28.5	34.8
	No	0.11	25.5	32.0

Table 9. Vector Space Model results for five class emotion classification

In addition to stemming experiment, we have considered the effect of adding emotional words from Wordnet-Affect and WPARD dataset into our training set. Results showed that, classification performance increased for Naïve bayes and SVM classifiers but in case of VSM the performance is reduced and there is only a small increase in kappa. This is because; we only added the word itself not sentences in our training set. Therefore during the normalization step, words come from Wordnet-Affect and WPARD behaved like a document which results a decrease in accuracy as seen in Table 8.

Experiment 3: In this experiment, we only considered positive and negative classes. Therefore, we combined the anger, disgust, fear, and sad emotions in Negative class while joy is the only member of the Positive class. Table 10 shows the results of this classification for different classifiers where the best performance for cross-validation comes from SVM classifier with 79.5% F-Measure value and 59.2% with VSM classifier.

Previous studies achieve up to 42.4% F1-measuse using coarse-grained evaluation for polarity detection on this dataset as reported in [19] while VSM approach achieves 59.2% F1-measure.

For emotion classification, previous studies on this dataset achieves up to 30.3% F1-measure for single class and 11% on average for six-class emotion classification using coarse-grained evaluation. Evaluation criteria of these studies can be found in [19].

Our results achieve up to 49.6% F1-measure for single classes and 32.2% on average for five-class emotion classification as seen on Table 11.

Classifier	Test method/set	Positive	Negative	Overall F1
Naïve Bayes	10Fold Cross Validation /ISEAR	64.1	89.9	74.8
Naïve Bayes	SemEval	55.3	60.6	57.8
libSVM	10Fold Cross Validation /ISEAR	**69.0**	**93.8**	**79.5**
libSVM	SemEval	49.9	**66.3**	56.9
VSM	SemEval	**59.1**	59.4	**59.2**

Table 10. Experimental results for polarity in terms of F-Measure using cross-validation on the ISEAR dataset

For the stop word experiment , "English.stop" file from Porter stemmer and "common_words" file from TMG are used. As seen on Table 11, almost all best F-Measure (mean of precision and recall) scores come from our classifier with 32.22% value.

In case of ConceptNet, we have used XML-RPC based client to communicate with ConceptNet server. For the evaluation, ConceptNet outputs a prediction vector $P(S) = \langle p_1(S), p_2(S),..., p_m(S) \rangle$ of size m where S represents a sentence or a snippet, $p_i(S)$ represents prediction value of i^{th} emotion class for the sentence S. Final classification result selects the maximum of $p_i(U)$ and assigns the corresponding class label using

$$P(S) = \arg\max_i \; p_i(S)$$

Classifier	Stop Word	Anger	Disgust	Fear	Joy	Sad	Overall F1
Naïve Bayes	Porter	20.2	5.2	41.9	39.6	32.6	27.9
Naïve Bayes	Tmg	21.5	5.4	**42.7**	40.5	32.5	28.5
libSVM	Porter	17.7	**9.5**	39.0	42.7	34.1	28.6
libSVM	Tmg	14.5	8.8	40.0	42.0	33.9	27.8
VSM	Porter	22.1	9.1	40.1	49.2	**37.1**	**31.5**
VSM	Tmg	**24.2**	9.3	41.1	**49.6**	36.7	**32.2**
ConceptNet	N/A	7.8	**9.8**	16.8	**49.6**	26.3	22.1

Table 11. Experimental results (in terms of F1-Measure) for emotions trained from ISEAR and tested on SemEval Test set

We have also create a video player which detects the emotion of subtitle texts and speech signal using the VSM and SVM classifiers trained on the ISEAR and displays the corresponding

emoticon as seen in Figure 3. Emotion detection in speech signal is performed using ensemble of support vector machines.

Figure 3. Emotion-ware video player screenshot from Finding Nemo[2]

5 CONCLUSION and FUTURE WORK

In this paper, we proposed a VSM approach for HER from text. We measured the effect of stemming and emotional intensity on emotion classification in text. We showed that Vector Space model based classification on short sentences can be as good as other well-known classifiers including Naïve Bayes and SVM and ConceptNet.

We also studied the effect of stemming to emotion classification problem. According to our experiments, use of stemming removes and decreases the emotional meaning from words. But these examples are very rare in our test set therefore use of stemming still increases the classification performance for all classifiers.

Finally, we have developed an emotion enabled video player, which shows video, emotional states and valence information at the same time.

As future work, we are planning to combine multiple modalities in video (audio, visual and text) to improve the classification performance.

Acknowledgments. This study was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) Project No: 107E002.

6 REFERENCES

[1] G. Mishne, Experiments with mood classification in blog posts. In: *Style2005 – 1st Workshop on Stylistic Analysis of Text for Information Access*, at SIGIR (2005).

[2] K. R. Scherer and H.G. Wallbott, Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66:310-328 (1994).

[3] D. A. Medler, A. Arnoldussen, J.R. Binder, and M.S. Seidenberg, The Wisconsin Perceptual Attribute Ratings Database. http://www.neuro.mcw.edu/ratings/ (2005).

[4] C. Strapparava and A. Valitutti., WordNet-Affect: an affective extension of WordNet. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC 2004), Lisbon, pp. 1083-1086, May (2004).

[5] A. C. Boucouvalas and X. Zhe, Text-to-Emotion. Engine for Real Time Internet Communication. In: *Proceedings of the 3rd International Symposium on CSNDSP*, Staffordshire University, UK, pp. 164-168, (2002).

[6] A. Valitutti, C. Strapparava, and O. Stock, Developing affective lexical resources. *PsychNology Journal*, 2(1):61-83 (2004).

[7] M. Shaikh, H. Prendinger, and M. Ishizuka, Emotion sensitive news agent: An approach towards user centric emotion sensing from the news. *The 2007 IEEE/WIC/ACM International Conference on Web Intelligence* (WI-07), Silicon Valley, USA, pp. 614-620 Nov. (2007).

[8] M. Shaikh, H. Prendinger, and M. Ishizuka, SenseNet: A linguistic tool to visualize numerical-valence based sentiment of textual data (Poster). *Proceedings 5th International Conference on Natural Language Processing* (ICON-07), Hyderabad, India, pp. 147-152, (2007).

[9] M. Shaikh, H. Prendinger, and M. Ishizuka, A cognitively based approach to affect sensing from text, In: *Proceedings of 10th Int'l Conf. on Intelligent User Interface* (IUI 2006), Sydney, Australia, pp.349-351 (2006).

[10] C. O. Alm, D. Roth, and R. Sproat, Emotions from text: Machine learning for text-based emotion prediction. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579-586, Vancouver, British Columbia, Canada, (2005).

[11] H. Liu, H. Lieberman, and T. Selker, A model of textual affect sensing using real-world Knowledge. In: *Proceedings of International Conference on Intelligent User Interfaces* (IUI-03) pp. 125–132, (2003).

[12] F. Sugimoto and M. Yoneyama, A Method for Classifying Emotion of Text based on Emotional Dictionaries for Emotional Reading, In: *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications*, Austria, pp. 91-96 (2006).

[13] Z.J. Chuang and H. Wu, Multi-Modal Emotion Recognition from Speech and Text. *International Journal of Computational Linguistics and Chinese Language Processing*, 9(2):1-18 (2004).

[14] J. W. Pennebaker, M. E. Francis, and R. J. Booth, Linguistic Inquiry and Word Count: LIWC 2001. Mahwah, NJ: Lawrence Erlbaum, (2001).

[15] H. Liu and P. Singh, ConceptNet: A Practical Commonsense Reasoning Toolkit. *BT Technology Journal* 22(4):211-226. Kluwer Academic Publishers, (2004).

[16] I. H. Witten and E. Frank (2005) *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, (2005).

[17] E. Boiy, P. Hens, K. Deschacht, and M.-F. Moens, (2007) Automatic Sentiment Analysis in On-line Text, ELPUB2007. Openness in Digital Publishing: Awareness, Discovery and Access In : *Proceedings of the 11th International Conference on Electronic Publishing*,Vienna, Austria 13-15, June (2007).

[18] D. Zeimpekis and E. Gallopoulos, Tmg: *A matlab toolbox for generating term-document matrices from text collections*. Technical report, University of Patras, Greece, (2005).

[19] C. Strappava and R. Mihalcea, SemEval-2007 Task 14: Affective Text, In: *Proceedings of the 4th International Workshop on Semantic Evaluations* (SemEval-2007), pages 70–74, Prague, June (2007).

[2] © Finding Nemo is a Copyright owned by Walt Disney and Pixar Entertainment

Old Wine or Warm Beer: Target-Specific Sentiment Analysis of Adjectives

Angela Fahrni & Manfred Klenner[1]

Abstract. In this paper, we focus on the target-specific polarity determination of adjectives. A domain-specific noun, the target noun, is modified by a qualifying adjective. Rather than having a prior polarity, adjectives are often bearing a *target-specific* polarity. In some cases, a single adjective even switches polarity depending on the accompanying noun. In order to realise such a 'sentiment disambiguation', a two stage model is proposed: Identification of domain-specific targets and the construction of a target-specific polarity adjective lexicon. We use Wikipedia for automatic target detection, and a bootstrapping approach to determine the target-specific adjective polarity. It can be shown that our approach outperforms a baseline system that is based on a prior adjective lexicon derived from SentiWordNet.

1 INTRODUCTION

Approaches to sentiment analysis range from counting the (prior) polarity of words [9] to systems that do a full compositional semantics analysis of sentence affect [5]. Specific resources have been developed, e.g. adjective lists [4], SentiWordNet [2] or WordNet-Affect [8], that compile the prior polarity of words. It has been noted, however, that the polarity of words is not in any case domain-independent [9]. An 'unpredictable plot' in the movie domain might be a good thing, but an 'unpredictable boss' surely is not. Moreover, as we would argue, even within a domain, the polarity of adjectives can vary. Take the adjective 'cold'. While a 'cold coke' is positive, a 'cold pizza' is not. 'Coke' and 'pizza' are domain-specific targets. Note that the adjective 'cold' has the same WordNet sense in both contexts (i.e. temperature reading), but the polarities are inverse. A kind of target-specific sentiment disambiguation seems to be necessary.

We propose a two stage model. First, the targets of a domain are identified. We use Wikipedia's and Wikionary's category system to get a comprehensive and moreover dynamic (since both resources are growing and growing) target list. In a second step, the target-specific polarity of adjectives is determined in a corpus-driven manner by searching for combinations of a target-specific adjective with adjectives that have a known prior polarity (e.g. good, excellent etc.). In order to evaluate our approach, we have derived an adjective lexicon with prior polarities from SentiWordNet. It serves as a baseline in our experiments carried out with 3891 automatically extracted and – by two independently working annotators[2] – manually classified (positive, negative, neutral) noun phrases.

[1] Institute of Computational Linguistics, University of Zurich, Switzerland, email: angela.fahrni@swissonline.ch, klenner@cl.uzh.ch

[2] Annotation mismatches have been resolved afterwards.

The domain of fast food restaurants was chosen as a test bed for our approach. We have downloaded about 1600 (manually classified) texts from epinions.com, a website with a huge amount of customer opinions concerning a broad spectrum of topics (holiday resorts, cars, credit cards, lawyers ..). Although these texts are manually classified along five categories: from very bad (one star) to really good (five stars), they do not establish a gold standard for our task, which is NP polarity detection. However, the ultimative goal of our work is to do sentiment detection on the sentence and eventually on the text level.

2 WIKIPEDIA-BASED TARGET DETECTION

Wikipedia's category system[3] is used to organise the stock of Wikipedia articles. It is hierarchical, but it does not constitute a genuine taxonomy, since it is based on pragmatic rather than ontological considerations. Although Wikipedia's hierarchy might be questionable, it actually does identify crucial domain-specific concepts. Moreover, the category tree also specifies named entities such as product names, proper names and brand names. This is a big advantage, since these items most often are the targets we are interested in. Adapting to a new domain boils down to identify the crucial Wikipedia categories (on an appropriate hierarchical level).

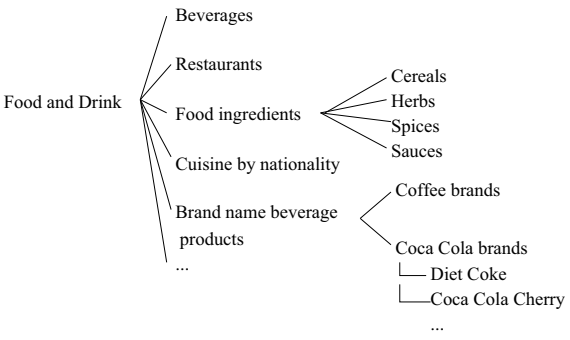


Figure 1. Wikipedia category 'Food and Drink'

In the fast food domain, /Food and drink/ is the most interesting category[4], it identifies 46807 targets. See Fig. 1 for a fragment of the category tree. Rather than using a flattened list of these targets, we keep the hierarchy in order to propagate polarities. For example, if it is known that 'cold coca cola' is positive then 'cold coca cola cherry'

[3] http://en.wikipedia.org/wiki/Portal:Contents/Categorical_index

[4] /Furniture/ and /Service/ might as well provide additional targets.

also is. Note that in the literature, e.g. [7], what we call targets, is sometimes called *features* and *attributes*.

3 TARGET-SPECIFIC OR PRIOR POLARITY

We argue that only a few adjectives do have a prior positive or negative polarity. Especially vague adjectives such as 'big', 'young', 'large', 'deep' are best understood as bearing neutral prior polarity. They adapt, however, to the context by either acting as intensifiers of the intrinsic positive or negative polarity of a target noun (e.g. 'deep insight', 'deep disappointment') or they combine with a neutral noun to form a non-neutral polarity (e.g. 'old bread'). Extreme examples such as 'cold pizza' and 'cold coke' where a single neutral adjective yields positive or negative polarity depending on the (neutral) noun are best explained as the violation ('cold pizza') and affirmation ('cold coke') of intrinsic or common sense properties of target objects (pizza, coke). In the absence of common sense reasoning, we propose a corpus-driven approach to determine such target-specific polarities.

Nevertheless, we sometimes need prior polarities, since they help us to explain certain compositional effects. Take the adjective 'lost' which has a (prior) negative polarity. A 'lost virtue' (virtue=positive) is negative, 'lost glasses' (glasses=neutral) is negative, but 'lost anger' (anger=negative) is positive. If 'lost' had no prior (negative) polarity, it could not combine with a negative word to form a positive noun phrase ('lost anger'). It is also not a simple valency shifter, otherwise we could not explain how the combination with a neutral noun forms a negatively qualified noun phrase ('lost glasses').

In the literature, adjectives with a clear prior polarity have been used as a seed list in order to identify the polarity of additional adjectives, e.g. [9]. The assumption of these approaches was that the augmented list again establishes a set of adjectives having a prior polarity. Contradicting polarities of an adjective encountered in a corpus were interpreted as a kind of noise and are resolved to one (predominant) polarity using statistical measures. We argue that often it is not noise what is encountered but target-specific sentiment ambiguity.

While we are relying on the same methods to identify the polarity of non-seed adjectives discussed in the literature, namely contextual pattern such as coordination, we aim at building a target-specific adjective lexicon instead of a domain-independent lexicon. Our seed adjective lexicon consists of 120 negative and 80 positive adjectives, where the polarity is supposed to be domain- and target-independent[5]. A few examples of positive polarity adjectives are: wonderful, tasteful, superb, positive, perfect, nice, ideal, great, excellent, delightful, delicious, good, beautiful.

4 TARGET-SPECIFIC POLARITY LEXICON

To get a target-specific polarity lexicon, two different corpora are being used. The one previously described (1600 texts from epinions.com, henceforth corpus I) and the world wide web (corpus II). Corpus I is tagged and all targets are identified. The most frequent targets from corpus I are used to find new texts in corpus II. Corpus I acts as a kind of reference corpus: we know that these texts are fast food ratings and thus we know that the adjectives used there and the targets from our (Wikipedia derived) target list (which might be noisy) that occur in these texts actually are relevant for the task at hand: the construction of a target-specific adjective lexicon. Corpus II is the pool used to identify the polarity of the non-seed adjectives from corpus I with respect to specific targets.

After we have identified the adjectives and targets we are interested in, we proceed as follows: We search both corpora for tag sequences that relate a target and at least two adjectives. It must hold that:

- the noun or noun sequence is a target
- at least one of the adjectives is from the seed list
- at least one of the adjectives comes from the stock of target-relevant adjectives

Currently, two sequence patterns are considered:

- adjective coordination (incl. modifying adverbs) e.g. 'good and tasteful burger'
- copula constructions, e.g. NP BE Adj Adj+ e.g. 'the french fries are soggy and rather tasteless'

It is assumed that adjectives in such constructions share the polarity[6]. As already mentioned, this kind of pattern directed polarity determination of adjectives is not new. However, in contrast to previous approaches, we require a *target* to be present relative to which the sentiment disambiguation is done. Moreover, the adjective and target must be of *interest* according to a reference corpus (corpus I).

Table 1. Examples of polarity tagged noun phrases

ADJ	Target	OWN-Pol.	SWN-Pol.
hot	burger	1	0
cheap	burger	0.949	-1
fresh	fruit	0.75	0
mouth-watering	burger	0.975	1
sized	sandwich	1	0
supersonic	burger	0.95	0

OWN System SWN SentiWordNet

Table 1 gives some examples of polarity tagged pairs generated by our systems. All noun phrases receive positive polarity from our system (OWN-Pol.) but quite different polarities from our baseline system that relies on an adjective list derived from SentiWordNet (last column, SWN-Pol., see section 5). The *polarity values* are gradual, ranging from -1 (very bad) to 1 (very good); 0 means neutral.

The polarity values of the seed adjectives are manually set, in some cases we took information from SentiWordNet into account. Polarity values of non-seed adjectives are given as the mean of the polarity values of their peers (where a peer is e.g. a seed adjective that occurs together with it in a coordination)[7].

All those adjectives that have a single polarity with all of its targets receive a *domain-specific* polarity. These and only these adjectives are combined with the seed list to form an augmented seed list. If the polarity of an adjective depends on the target, an adjective-target pair is added to the polarity-specific lexicon. Those adjectives from corpus I that have not received a polarity in the first cycle (since they never occurred e.g. in a coordination with a seed adjective) get a second (third and fourth) change. They might get a polarity in another round, on the basis of the incrementally augmented seed list. Currently, we run four such incremental cycles.

[5] Of course, figurative language readily produces counterexamples.

[6] There are, of course, exceptions, e.g. 'rich and poor people'.

[7] We have also experimented with a *confidence value* of a polarity classification, which is meant to tell us how strong a decision was.

5 A PRIOR-POLARITY LEXICON DERIVED FROM SentiWordNet

[2] introduce a semi-automatic approach to derive a version of WordNet where word senses are bearing polarities. The resource is called SentiWordNet and is freely available for research purposes. The developers rely on the same idea as described above, namely a seed of paradigm words with a clear polarity.

Table 2. SentiWordNet: 'unpredictable'

POS	synset	pos.	neg.	word	sense
a	1781371	0.0	0.625	unpredictable#a	#1
a	708935	0.0	0.0	unpredictable#a	#2
a	566807	0.0	0.25	unpredictable#a	#3

Table 2 shows the entry of 'unpredictable'. The numbers below the polarity tags (pos., neg.) indicate the polarity strength (1 indicates maximal strength). Word sense 1 and 3 of 'unpredictable' have negative polarity, while word sense 2 is neutral.

In SentiWordNet, the adjective 'hot' has 22 senses, 7 of them have neutral, 5 have negative and 10 have positive polarity. Since we are only interested in the polarities, we merge the positive, negative and neutral senses into one polarity entry, respectively. Each entry receives as its polarity weight the weighted sum of its SentiWordNet scores, e.g.

$$weight('hot' = pos) = \frac{\sum_{i \in swn\_pol('hot'=pos)} swn\_score(i)}{\mid synsets('hot') \mid}$$

where $swn\_pol('hot' = pos)$ denotes the set of synsets of 'hot' bearing positive polarity and $swn\_score(i)$ is the value of the SentiWordNet entry of word sense i of 'hot'.

This way, the adjective 'hot' in its neutral reading gets a weight of 0.6, while positively interpreted it receives 0.28, leaving a 0.12 score to the remaining negative case. Applying this strategy to SentiWordNet, we have generated an adjective lexicon with prior polarities that blends the numerical weights of an adjective's SentiWordNet entry into three discrete polarity classes. Altogether, 21194 adjective entries has been derived. Note that some of them has received three ('e.g. 'hot'), some two (e.g. 'unpredictable') and other only one polarity entry (e.g. 'good').

6 EVALUATION

We have carried out an evaluation of our system on the basis of 3891 manually classified noun phrases[8]. The resulting gold standard comprises 1832 positive, 415 negative and 1644 neutral instances.

Three different experimental settings are distinguished. First, we compared the polarity decisions of SentiWordNet (our baseline system) and our system for the whole data set (all). Second, we took only those classifications that received different polarities from the two systems ($conflict$). Third, only the instances where both systems agreed in their polarity assignment are taken ($agree$).

Table 3 shows the accuracy under these conditions. Given the whole data set (3891 NPs), our system outperforms SentiWordNet by 6.8%. This setting is the 'realistic' one, so, given domain-specific texts, a substantial improvement can be achieved with the methodology we propose. If we (only) evaluate the conflicting classifications (1937 NPs), our system shows its strength. Here an improvement of 13.7% was achieved. The evaluation of those cases where both systems assign the same polarity (1954 NPs shows that we can design a high-accuracy system by combining both resources, SentiWordNet and our system.

[8] which corresponds to 2426 NP types

Table 3. Accuracy under 3 experimental settings

	SWN	OWN
all	63.4 %	70.2%
conflict	39.2 %	52.9%
agree	87.4 %	87.4%

SWN SentiWordNet OWN System

Table 4. Evaluation of (all) 3891 noun phrases

	SWN			OWN		
	prec	rec	f-meas	prec	rec	f-meas
pos	97.5 %	39.5%	55.9%	66.0%	91.2 %	76.5%
neg	89.8 %	34.2%	49.5%	82.3%	37.2 %	51.1%
neut	53.7 %	97.4%	69.3%	77.3%	55.2 %	64.4%
⊘			58%			64%

SWN SentiWordNet OWN System ⊘ arithmetic mean

Table 4 shows the results (whole data set) for each single class. We can see that our approach clearly outperforms SentiWordNet with respect to the positive NPs (76.5% F-measure compared to 55.9%), but only slightly given the negative NPs (1.6%). Given neutral NPs, SentiWordNet wins (4.9%). A closer look at the data shows that SentiWordNet has a strong bias towards neutral classifications.

Table 5. Evaluation of (conflict) 1937 classification conflicts

	SWN			OWN		
	prec	rec	f-meas	prec	rec	f-meas
pos	37.5 %	1.2%	2.4%	53.1%	98.4 %	68.9%
neg	58.9 %	11.3%	19.3%	51.5%	17.2 %	25.7%
neut	38.9 %	95.7%	55.3%	50.8%	4.2 %	7.8%
⊘			25.5%			34.2%

SWN SentiWordNet OWN System ⊘ arithmetic mean

From Table 5 we can see that our system has a bias towards positive classifications, but precision is still reasonable, so a F-measure of 68.9% was achieved. Note that it is the class of conflicting classifications where our system (a target-specific approach) has to prove is advantages over a system with prior polarities. The overall difference in performance is 8.7%. But since normally one is interested in positive and negative polarities rather than neutral, our system improves at the right place. If we look at these two classes, our approach is 70.9% superior to SentiWordNet (to be more precise: our adjective list derived from SentiWordNet). However, we clearly have to come to a more balanced performance.

Finally, from Table 6 ($agree$) we can see that a combination of the two approaches can act as a high-precision system correctly identifying 98.3% of the positive and 99.9% of the negative NPs. Note however that those NPs receiving the same vote from both systems most often include a seed adjective. So they won't have a target-specific polarity.

Table 6. Evaluation of 1953 (agree) decisions.

	SWN + OWN		
	prec	rec	f-meas
pos	98.3 %	83.2%	90.1%
neg	99.9 %	56.4%	72.1%
neut	78.9 %	98.9%	87.8%
⊘			83.3%

⊘ arithmetic mean

7 RELATED WORK

Our approach to the identification of polarity of adjectives is based on the ideas of [3] (among others). However, [3] only identify prior polarity, not contextual.

Work on contextual polarity detection is described in [11]. Here, a (supervised) machine learning approach is used to find the contextual polarity of words. In our (semi-supervised) approach, the notion of a domain-specific target is stressed, while in their approach this is left implicit as a problem to be solved by the machine learning component. Note that [11] are striving to cope with a more challenging domain, namely news texts. Accordingly, the empirical performance reported there is worse than the one reported here. But we can not seriously compare both.

Our approach to target detection is based on Wikipedia's category systems. Others, e.g. [7] have used contextual, e.g. meronymy discriminators such as 'the X of Y' where X is identified as an *attribute* of the *feature* Y. We plan to improve our Wikipedia based approach by also taking Wikipedia articles into account. Then, contextual discriminators but also available tools such as those described in [6] might prove helpful.

There are several approaches to derive polarity tagged adjective lists form WordNet, e.g. [1], [4]. Since we plan to use SentiWordNet [2] also as a source for noun and verb polarity, we have already worked with it to derive a baseline system for adjective polarity detection.

Finally, the interaction between word sense disambiguation and subjectivity has been discussed by [10]. However, in their system sentiment detection helps word sense disambiguation while in our approach a single word sense might even give rise to two inverse target-specific polarities. Our solution to that problem has the side effect that word sense disambiguation becomes superfluous. If an adjective changes polarities depending on the target, both, adjective and target are added to the target-specific lexicon[9].

8 CONCLUSION AND FUTURE WORK

We have introduced a semi-supervised approach to NP polarity detection that is based on a target-specific polarity lexicon induced from a seed lexicon and two corpora. This enables our system to assign different polarities to NPs with the same adjective but a different target noun (e.g. 'cold french fries' and 'cold buttermilk'). Our system outperforms a baseline system derived from SentiWordNet. The SentiWordNet baseline establishes a kind of upper bound for approaches that rely on *prior* polarity information only.

Although we have started to experiment with both a measure of polarity strength and a confidence metric, the results have not been sufficiently evaluated and thus are not presented here. Polarity strength tells us how strong a positive or negative evaluation (here NP polarity) is, confidence indicates how reliable a polarity decision is.

Currently, NP polarity depends exclusively on adjective polarity. This is an artefact of the chosen domain where nouns mostly are neutral (food, furniture, employees etc.). But NP polarity often is compositional (as is sentence polarity). For example, a positive adjective and a negative noun ('excellent forgery', 'perfect spy') combine to a negative polarity. Therefore, but also to prove the domain independence of our model, we plan to switch to another domain.

We are also working on a model of sentence-level sentiment analysis. Currently, our main focus lies on the identification of basic dependency structure that reliably indicate 'subject verb object' constellations ('I love this little book'). We then will focus on negation ('never'), intra-sentential valency shifters ('but') and complex compositional phenomena ('this could not fail to get nasty').

[9] However, neither approach is fully satisfying, since there are cases where different senses of an adjective do have inverse polarities when combined with a single noun, e.g. the 'inexpensive' versus 'poor quality' reading of 'cheap food'.

REFERENCES

[1] Alina Andreevskaia and Sabine Bergler, 'Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses', in *Proceedings EACL-06, the 11rd Conference of the European Chapter of the Association for Computational Linguistics*, pp. 209–216, Trento, IT, (2006).

[2] Andrea Esuli and Fabrizio Sebastiani, 'SentiWordNet: A publicly available lexical resource for opinion mining', in *Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation*, Genova, IT, (2006).

[3] Vasileios Hatzivassiloglou and Kathleen R. McKeown, 'Predicting the semantic orientation of adjectives', in *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics*, pp. 174–181, Madrid, ES, (1997). Association for Computational Linguistics.

[4] Jaap Kamps, Robert J. Mokken, Maarten Marx, and Maarten de Rijke, 'Using WordNet to measure semantic orientation of adjectives', in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 1115–1118. European Language Resources Association, Paris, (2004).

[5] Karo Moilanen and Stephen Pulman, 'Sentiment composition', in *Proceedings of the Recent Advances in Natural Language Processing International Conference (RANLP-2007)*, pp. 378–382, Borovets, Bulgaria, (September 27-29 2007).

[6] Simone Paolo Ponzetto and Michael Strube, 'An API for measuring the relatedness of words in Wikipedia', in *Companion Volume to the Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp. 49–52, Prague, Czech Republic, (June 2007).

[7] Ana-Maria Popescu and Oren Etzioni, 'Extracting product features and opinions from reviews', in *Proceedings of HLT-EMNLP-05, the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing*, pp. 339–346, Vancouver, CA, (2005).

[8] C. Strapparava and A. Valitutti, 'Wordnet-affect: an affective extension of wordnet', in *Proceedings of LREC-04, the 5th Conference on Language Resources*, (2004).

[9] Peter D. Turney and Michael L. Littman, 'Measuring praise and criticism: Inference of semantic orientation from association', *ACM Transactions on Information Systems*, **21**(4), 315–346, (2003).

[10] Janyce Wiebe and Rada Mihalcea, 'Word sense and subjectivity', in *Proceedings of COLING/ACL-06, the 21st Conference on Computational Linguistics / Association for Computational Linguistics*, pp. 1065–1072, Sydney, AUS, (2006). Association for Computational Linguistics.

[11] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, 'Recognizing contextual polarity in phrase-level sentiment analysis', in *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, CA, (2005).

Detecting and Adapting to Student Uncertainty in a Spoken Tutorial Dialogue System

Diane Litman [1]

Abstract. We are currently building an adaptive tutorial spoken dialogue system, with the goal of using spoken and natural language processing to monitor and respond to the affective state of student uncertainty. First, I will discuss the empirical approach used to design and implement our system. To detect student uncertainty, we use machine learning to develop a predictive model based on lexical and prosodic features of student utterances. To develop system responses to student uncertainty, we use a bigram analysis of a human tutoring corpus to identify dependencies between uncertain student answers and subsequent tutor responses. I will conclude the talk by presenting initial evaluation results from a first controlled experiment, comparing wizard versions of both our new adaptive and our original non-adaptive tutorial dialogue systems.

[1] Department of Computer Science & Learning Research and Development Center University of Pittsburgh Pittsburgh, PA USA

Adjectives and Adverbs as Indicators of Affective Language for Automatic Genre Detection

Robert Rittman and **Nina Wacholder**

School of Communication, Information and Library Studies
Rutgers, The State University of New Jersey
New Brunswick, New Jersey

Abstract. We report the results of a systematic study of the feasibility of automatically classifying documents by genre using adjectives and adverbs as indicators of affective language. In addition to the class of adjectives and adverbs, we focus on two specific subsets of adjectives and adverbs: (1) trait adjectives, used by psychologists to assess human personality traits, and (2) speaker-oriented adverbs, studied by linguists as markers of narrator attitude. We report the results of our machine learning experiments using Accuracy Gain, a measure more rigorous than the standard measure of Accuracy. We find that it is possible to classify documents automatically by genre using only these subsets of adjectives and adverbs as discriminating features. In many cases results are superior to using the count of (a) nouns, verbs, or punctuation, or (b) adjectives and adverbs in general. In addition, we find that relatively few speaker-oriented adverbs are needed in the discriminant models. We conclude that at least in these two cases, the psychological and linguistic literature leads to identification of features that are quite useful for genre detection and for other applications in which identification of style and other non-topical characteristics of documents is important.

1 INTRODUCTION

This paper reports on the use of adjectives and adverbs to discriminate text genres characterized by affective expressions (e.g., fiction) from genres in which affective expressions are typically inappropriate (e.g., academic writing).[1] We adopt the definition of genre given by Lee [2].

> [G]enre is a document-level category assigned on the basis of **external** criteria such as intended audience, purpose, and activity type, that is, it refers to a conventional, culturally recognised grouping of texts based on properties other than lexical or grammatical (co-)occurrence features, which are, instead, the **internal** (linguistic) criteria forming the basis of text type categories.

Thus, a news report is intended to inform, an editorial or opinion piece is intended to persuade, and a novel is intended to entertain.

The paper is organized as follows. First, we review discriminating features selected in automatic genre classification research. In Section 3, we summarize how adjectives and adverbs are generally indicative of affective language, and describe the characteristics of two small subsets of adjectives (trait adjectives) and adverbs (speaker-oriented adverbs). In Section 4, we describe our methodology for discriminating documents by genre using these features. In Section 5, we present our results. In Section 6, we discuss our conclusions and provide direction for future work.

2 FEATURE SELECTION IN GENRE CLASSIFICATION RESEARCH

In previous research in genre discrimination, researchers have focused on identifying any features that are useful in discriminating genres. Toward this end, they have identified discriminating features of four basic types: (a) syntactic (parts of speech, e.g., adverbs, nouns, verbs, and prepositions), (b) lexical (terms of address, e.g., *Mr.*, *Mrs.*, *Ms.*; content words; most frequent words in a corpus, e.g., *the*, *of*, *and, a*); (c) character-level (e.g., punctuation, character count, sentence count, word length in characters); and (d) derivative (ratio measures, e.g., average words per sentence, average characters per word, type/token ratio).

They have applied these features to discriminate different sets of documents and different genres. Using a set of features that were relatively easy to identify automatically in combination with a machine learning method, and working with 500 documents from the Brown Corpus, Karlgren and Cutting [3] selected a set of 20 features, such as first person pronouns, adverbs, prepositions, and nouns; characters per document; average words per sentence; and type/token ratio. Similarly, Kessler at al. [4] classified 500 documents from the Brown Corpus with 55 features (lexical, character-level, and derivative features). Using 500 documents from the LIMAS German corpus, Wolters and Kirsten [5] took a hybrid approach, combining the traditional IR "bag of words" method with a natural language processing method they called, "bag of 'tagged' words." They represented documents as vectors of the frequency of content word lemmas and function words, and combined it with part of speech information. Inspired by research in author attribution, Stamatatos et al. [6] selected the 50 most common words in the British National Corpus (e.g., *the*, *of*, *a*, and *in*), as well as eight of the most frequent punctuation symbols (period, comma, colon, semicolon, quotes, parenthesis, question mark, and hyphen). Using a subset of these features Ng et al. [7] selected four punctuation

[1] These results are taken from a much larger study by Rittman [1] of automated classification of documents by genre using adjectives and adverbs as discriminating features.

marks (comma, period, colon, and semicolon) to classify a collection of *Wall Street Journal* and *Federal Register* documents in an investigation of features independent of syntax and semantics. Subject-classified and genre-classified training data was used by Lee and Myaeng [8] to select features based on three criteria: (a) find terms that occur in many documents belonging to one genre which are distributed evenly among all subject classes; (b) eliminate terms that are specific to a particular subject; and (c) downgrade terms that are common to many genres. Web-based technology features based on HTML tags and URL information were selected by Lim et al. [9] in addition to features used by other researchers (e.g., part of speech, punctuation, average words per phrase, and frequency of content words). More than one-hundred features (including syntactic, lexical, and character-level features) were selected by Santini et al. [10] and Santini [11] to address the problem of emerging genres in the Web. The problem of Web genre detection was also addressed by zu Eissen and Stein [12] using thirty-five derivative features, such as average number of mail links, average number of help symbols, and average number of various parts of speech (e.g., nouns, verbs, prepositions, and adverbs). Finally, Finn and Kushmerick [13] classified two sets of Web-generated corpora representing documents as (a) a bag-of-words (vector indicating the presence or absence of a word), (b) part-of-speech statistics (vector of 36 parts of speech features); and (c) text statistics (e.g., average sentence length, average word length, and frequency of punctuation).

The approach to genre identification that characterizes these studies might be called a 'bag of features' approach: researchers applied machine learning techniques to any features that could be identified automatically. Since the focus of their research was genre identification, this approach was completely appropriate. But the use of bags of features, along with different sets of documents and genres, has made it difficult to systematically study the contribution of affective language to genre identification. Only one of the studies described above (Wolters and Kirsten [5]) specifically mentioned that adjectives and adverbs are useful for distinguishing genre, as opposed to a mix of many kinds of features they tested. Furthermore, the authors report their results using the standard measure of Accuracy; this measure does not take into consideration the impact of the percentage of documents that belong to each class on the outcome; this too makes it hard to compare results.

In what follows, we discuss the characteristics of adjectives and adverbs that make them particularly useful for identifying expressions of affect and assess their contribution to automatic genre classification.

3 ADJECTIVES AND ADVERBS AS FEATURES OF AFFECTIVE LANGUAGE

As a grammatical category, adjectives modulate the meaning of nouns by emphasizing important or surprising properties of the noun being modified (e.g., a *safe* / *historical* / *unusual* building). The properties that are highlighted frequently represent a judgment or opinion. The statement, *She wrote a poem*, is a statement of (presumed) fact. The statement, *She wrote a beautiful* / *horrendous* poem, mixes a statement of fact with human judgment. Research indicates a correlation between human perceptions of subjectivity and the occurrence of adjectives in (a) sentences (Bruce and Wiebe [14], Wiebe [15], and Wiebe et al. [16]) and (b) documents (Rittman et al. [17]). This relationship is expected because of the nature of adjectives themselves. Subjective expressions necessarily involve judgments and opinions about people and things, and we frequently use adjectives to express our judgments.

In a similar way, adverbs modulate the meaning of verbs, adjectives, other adverbs, and noun phrases. This is especially true of the many adverbs derived from adjectives by adding the suffix *–ly* (*beautiful => beautifully*; *horrendous => horrendously*); adverbs typically inherit the subjective connotation of the adjectives from which they have been derived.

Within the larger set of adjectives or adverbs in the context of a sentence, researchers in psychology and linguistics have each indicated a subset of words that appear to be particularly expressive of affect. Psychologists have identified trait adjectives and linguists have identified speaker-oriented adverbs.

3.1 Trait Adjectives

The significance of adjectives in description and judgment has long been noted in psychology. Psychologists use trait adjectives to describe human personality traits (e.g., *nervous, energetic, accommodating*, and *careful*). Trait adjectives are classified by the type of personality they indicate, based on theories of psychology. Using factor analysis on various lists of adjectives Goldberg [18] proposed five dimensions of personality that are generally accepted as the "Big Five": I. Extraversion (*active, assertive, bold*), II. Agreeableness (*agreeable, considerate, cooperative*), III. Conscientiousness (*careful, conscientious, efficient*), IV. Emotional Stability (*imperturbable, relaxed, undemanding*), and V. Intellect (*artistic, bright, complex*). Some researches (e.g., Nowson et al. [19], Argamon et al. [20], and Mairesse et al. [21] have studied the relationship between personality traits of experimental subjects and their use of language features in different genres.

We turn the Big Five on its side and select adjectives that are used by psychologists as indicators of personality as features for genre detection. Although psychologists use these adjectives to scientifically characterize human personality in the context of written and spoken text, when these adjectives are used in non-scientific language, they represent expressions of judgment. What is virtuous to one person may be sinful to another. Furthermore, trait adjectives frequently have an affective connotation; for example, the adjectives *perfidious* and *vulgar* almost always represent a negative judgment while the adjectives *loyal* and *intriguing* almost always represent a positive one. These connotations are pertinent whether the adjectives are used to describe people or some other kind of entity. The trait adjectives in Appendix A (a subset of 44 trait adjectives which we derive from the full list reported by Peabody and De Raad [22], along with the adverbs derived from them (Appendix B), are therefore particularly likely to express affect.

3.2 Speaker Oriented Adverbs

Adverbs that express sentiment typically serve three grammatical functions: disjuncts, adjuncts and subjuncts (Quirk et al. [23]). Disjuncts (1.3) are peripheral to the

sentence and "express an evaluation of what is being said either with respect to the form of the communication or to its meaning … [And they express] the speaker's authority for, or comment on, the accompanying clause" (Quirk et al. [23]) For example, in (1.3), *frankly* is a description of the speaker's attitude about the statement, *I am tired*. In contrast, adjuncts (1.1) and subjuncts (1.2) are integrated within the structure of the clause. For example, in (1.1) and (1.2), *slowly* and *kindly* focus internally on the grammatical subject or verb phrase (i.e., they walked *slowly*, you wait *kindly*). This is quite different than the disjunctive use of *frankly*, which focuses externally on the speaker's behavior. As Mittwoch [24] explains, disjuncts are a way to "refer to one's own words." For this reason, disjuncts are referred to as *speaker-oriented adverbs* (SOAs).

(1.1) *Slowly* they walked back home. (Adjunct)
(1.2) Would you *kindly* wait for me? (Subjunct)
(1.3) *Frankly,* I'm tired. (Disjunct)

The potential usefulness of SOAs in identifying expressions of affect is supported by Jackendoff [25] and Ernst [26], who indicate that (a) adverbs can refer to the speaker (narrator), the grammatical subject, or the manner in which an event occurs, (b) sentence position of adverbs affects meaning, (c) adverbs can occur in some positions and not in others, and that (d) adverb phrases can frequently be paraphrased using corresponding adjective phrases. SOAs refer to the speaker of the sentence, subject-oriented adverbs refer to the grammatical subject of the sentence, and manner adverbs refer to the main verb of the sentence. In summary, SOAs provide a grammatical mechanism by which a speaker can insert an indication of mood or attitude at the periphery of the sentence. We use a set of 30 SOAs derived from Ernst [26] (Appendix C).

3.3 Adjectives and Adverbs in Relation to Genre Identification

Since adjectives and adverbs frequently perform some degree of evaluation, it follows that the occurrence of a relatively high number of adjectives and adverbs should indicate the presence of expressions of judgment in a document. This characteristic makes the frequency of adjectives and adverbs in text a likely feature for discriminating genres that include expressions of sentiment and judgment.

Trait adjectives and the adverbs inherited from them frequently have evaluative connotations, at least in context; we expect that they will be most frequent in genres that describe people's behavior, such as fiction. SOAs characterize the narrator's perspective, and are indicative of intent and behavior.

Since genre is indicative of the author's purpose, intended audience, and type of activity (Lee [2]), we explore the contribution of adjectives and adverbs in general, and trait adjectives and SOAs in particular, to the identification of genre.

4 METHODOLOGY

In the first part of this section, we describe the materials used in our study; these include the collection of documents, the genre labels and the features, and the classification problems that we used machine learning methods to solve. In the second part, we describe Accuracy Gain, a measure of the contribution of features to a classification task that is more rigorous than the standard measure of Accuracy used in most genre identification tasks.

4.1 Experimental Materials

To systematically study the relationship of adjectives and adverbs to genre, we needed a set of documents that had been classified by genre and tagged by part-of-speech. Fortunately, the freely available British National Corpus, World Edition (BNC2 [27]) satisfied these requirements. Lee [28, 29] originally assigned each of the 4,054 [2] documents in BNC2 to one of 70 genres; 46 were written and 24 were spoken. However, the large number of genres meant that relatively few documents were assigned to each genre. Davies [30] therefore organized these 70 genres into six supergenres which he labeled *academic*, *fiction*, *news*, *non-fiction*, *other*, and *spoken*. Our experimental task is to assess the contribution of adjectives and adverbs to automatic classification of these six genres.

We consider two basic kinds of genre classification problems. The easier problem is one-against-one; the harder problem is one-against-many. One-against-one discriminates one genre from another (e.g., academic vs. fiction or fiction vs. news). One-against-many discriminates one genre from all other genres in a corpus (e.g., academic vs. fiction, news, non-fiction, other, and spoken, or news vs. academic, fiction, non-fiction, other, and spoken). One-against-one is easier because it considers one document subset of a corpus against another subset, and because only two genre categories are considered. One-against-many is harder because it treats one genre against the entire corpus; the latter consists of documents from multiple genres.

For one-against-one, we chose three of Davies [30] supergenres that are mutually exclusive: academic, fiction and news. This presents three one-against-one classification problems: academic vs. fiction; academic vs. news; and news vs. fiction. For one-against-many, we used the same three supergenres, comparing the performance of classifiers on distinguishing the supergenres from documents consisting of all of the other genres in BNC2. Our one-against-many problems are (1) academic vs. not-academic (fiction, news, non-fiction, other, and spoken), (2) fiction vs. not-fiction (academic, news, non-fiction, other, spoken); and (3) news vs. not-news (academic, fiction, non-fiction, other, and spoken).

We chose the supergenres of academic, fiction, and news for several reasons. First, they are based on Lee's [2] criteria for genre (intended audience, purpose, and activity type). Second, Davies' [30] organization is exclusive. For instance, the set of documents labeled *fiction* excludes academic, news, non-fiction, spoken and other (although non-fiction can be similar in content to other classes, such as academic for instance except it is intended for a different audience). What he calls *spoken* includes everything spoken, regardless of

[2] BNC2 includes 4,054 documents. We exclude one document because it is a duplicate (see Lee [2]).

whether it is academic, fiction, news, non-fiction, or other. We do not use *other* because it is not a conceptually distinct class. Third, selecting only three supergenres directly (academic, fiction, and news) limits the number of discriminant tests to six problems (three one-against-one and three one-against-many), as opposed to 21 problems if we selected all six supergenres (15 one-against-one and six one-against-many). Finally, the supergenres of non-fiction, other, and spoken are treated indirectly in our one-against-many classification problems.

From the complete corpus (N=4,053), we randomly selected 50% of the documents for training (n=2,029) and 50% for testing (n=2,024). Based on this selection, we broke out six document sets. Table 1 shows the number of documents in each set.

Problem	Doc Sets	50% Training	50% Testing	Total
		2,029	**2,024**	**4,053**
Acad vs. Fiction	Acad	276	229	505
	Fict	218	246	464
	Total	494	475	969
Acad vs. News	Acad	276	229	505
	News	249	269	518
	Total	525	498	1,023
Fict vs. News	Fict	218	246	464
	News	249	269	518
	Total	467	515	982
Acad vs. NotAcad	Acad	276	229	505
	NotAcad	1,753	1,795	3,548
	Total	2,029	2,024	4,053
Fict vs. NotFict	Fict	218	246	464
	NotFict	1,811	1,778	3,589
	Total	2,029	2,024	4,053
News vs. NotNews	News	249	269	518
	NotNews	1,780	1,755	3,535
	Total	2,029	2,024	4,053

Table 1: Training-Testing Document Sets

In the larger study, Rittman [1] systematically tested the impact of a variety of adjective and adverb features on genre identification on these six problems. First, features were represented as types and tokens. Type indicates whether a word occurs in a document at least once; token indicates the frequency of a word in a document. These two measures give rise to two related representations of features: count and vector. Count is the aggregate of all members of a class in a document. For example, suppose that word_1, word_2 and word_3 all belong to the target class under discussion. And suppose that word_1 occurs three times in a document, word_2 does not occur, and word_3 occurs 7 times. The count of types is 2 (1+0+1). The count of tokens is 10 (3+0+7). The vector approach also identifies types and tokens but represents them a different way. For example, if a word occurs in the document, then type=1; else type=0. If type=1, then token=frequency of the word in the document; else token=0. Thus, in the example above for the three hypothetical words, the vector for the document is (1, 3, 0, 0, 1, 7).

In some cases, the sentence position of a word was marked, such as *sentence-initial position* and *not sentence-initial position* which conceptually is the union of *any sentence position*. Finally, each feature was represented as a *normalized* and a *non-normalized* variable. The normalized variable is calculated by dividing frequency by the total count of words in a document.

We call the various ways of representing features a *method*. Each method includes a unique set of features (such as speaker-oriented adverbs) and different ways of representing the features (such as sentence position, vector or count, or as normalized or non-normalized). The intersection of a method and a classification problem represents a model. Since there are 54 methods for six problems, we analyzed a total of 324 models. In what follows, we report only on a portion of the 324 models identified in the larger study, using only the results of the normalized variables for the adjective and adverb features that are most interesting. Table 2 lists these 17 sets of features

Feature / Method	Word Class	Types in BNC2 *	Vector or Count
SOA1	30 Speaker-Oriented Adverbs (Appendix C) in any sentence position using a vector length of 60	30	V
SOA2	SOA1 in sentence-initial+not-sentence-initial position using a vector length of 120	30	V
SOA3	Count of SOA1 in any sentence position	30	C
RB	All words tagged in BNC2 as adverbs	9,324	C
RB-ly	All RB ending in *ly*	6,372	C
JJ	All words tagged in BNC2 as adjectives	147,038	C
JJ1	Subjective adjectives identified by [31] (in BNC2)	1,322	C
JJ2	Trait Adjectives full list derived from [22] (in BNC2)	732	C
JJ3	Subset of JJ2 (Appendix A) identified in pilot study as indicative of subjectivity	44	V / C
RB1	Trait Adverbs derived from JJ2 by adding *ly*	539	C
RB2	Subset of RB1 (and JJ3) (Appendix B) using vector length of 72, also as a count	36	V / C
JJ3+RB2	Union of JJ3+RB2 using a vector length of 160	80	V

SOA1+JJ3 +RB2	Union of SOA1+JJ3+RB2 (Appendix A, B, C) using a vector length of 220	110	V
All-JJ&RB	SOA3, JJ, JJ1, JJ2, JJ3, RB, RB-ly, RB1, RB2	165,437	C
NN	Nouns	430,415	C
VB	Verbs	60,083	C
Punc	Punctuation	71	C
* For each feature we consider both types and tokens. Thus the number of variables for each feature is doubled.			

Table 2: Selected Features

We include the count of all adjectives and adverbs that were tagged as such in BNC2, including all adverbs ending in *–ly*, and all adjectives identified by Wiebe [31] as *subjective adjectives*. For our vector experiments, we select the 30 speaker-oriented adverbs derived from [26] (Appendix C), the subset of 44 trait adjectives derived from Peabody and De Raad [22] (Appendix A), and a list of 36 *trait adverbs* derived from Appendix A by adding *-ly* (Appendix B). We also combine these three lists of words in a union set of 110 adjectives and adverbs (Appendix A, B, and C). We also included nouns, verbs and punctuation as a benchmark to compare the performance of models using our adjective and adverb features.

Each set of features in Table 2 was used to build different models using tools for discriminant analysis provided by SPSS (version 11.0). The machine learning method of discriminant analysis is a widely used classification method of multivariate statistics used in genre classification work by Karlgren and Cutting [3], Stamatatos et al. [6], and Ng et al. [7]. For each classification problem, we test the performance of various sets of features and methods.

4.2 Accuracy Gain [3]

The standard measure of performance for classification problems is Accuracy [32]. Simply put, Accuracy is the fraction of correctly predicted cases. However, Accuracy does not consider the proportion of members of a particular class. As such, it does not take into account the most rigorous and expected baseline that a hypothetical classifier can achieve; this baseline is equal to proportionate size of the majority group. For example, if 88 of 100 documents are academic and 12 are news, the best strategy for a classifier is to guess *academic* every time; this hypothetical classifier will have an accuracy performance of 88%. This is what we call a *best guess* classifier.

The standard accuracy measure therefore under-rates the performance of a classifier that does better than the best guess and over-rates the performance of a classifier that does less well than best-guess. Suppose that one classifier identifies 88% of 100 documents correctly and another identifiers 90% of 100 correctly. There is only a 2% difference in performance by the two classifiers.

Accuracy gain (AG) achieves a more realistic measure of the performance of the classifiers by treating the baseline (or best-guess) performance as zero; the performance of a classifier that does better than best guess is represented with a positive number; the performance of a classifier that does less well than the baseline is represented as a negative number. Table 3 compares the measures of Accuracy and Accuracy Gain for a case where 88 documents belong to one class and 12 belong to another.

Classifier Performance (Baseline = 0.88)	
Accuracy (%)	Accuracy Gain (%)
86	-17
88	0
90	17
91	25
100	100

Table 3: Accuracy Computed as Accuracy Gain

With AG, a classifier that achieves only baseline Accuracy (88%) performs at 0%, i.e., no better or worse than it would have achieved with a best-guess strategy. But the classifier that achieves a two-point improvement (90%) over the baseline (88%) has a 17% Accuracy Gain. This more accurately reflects the performance of the classifier with the apparently small (2%) improvement.

Figure 1 shows how we calculate AG by re-scaling (or normalizing) the rigorous baseline to zero. The resulting fraction (AG) represents the improvement over the best guess procedure compared to the maximum possible improvement (100%).

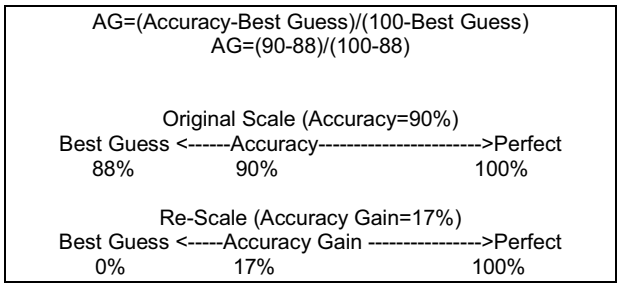


Figure 1: Calculation of Accuracy Gain (AG)

Another advantage of the AG measure is that it allows consistent comparison of results for studies that use classes with different populations. All of the studies cited in Section 2 use the standard Accuracy measure; only half report the proportionate size of genre categories (Karlgren and Cutting [3], Kessler et al. [4], Ng et al. [7], Lee and Myaeng [8], Santini et al [10] and Santini [11], and zu Eissen and Stein [12]).

We therefore report our results using AG as the most rigorous performance measure and as a method that promotes comparability of future research.

5 RESULTS

Table 4 shows the impact of the different features (methods) for discriminating the three genres. As expected, the results for the one-against-one classification problems are better than one-against-many. Still, for both sets of problems, models

[3] See Rittman [1].

using adjective and adverb features outperform models containing nouns, verbs, or punctuation. For example, for academic vs. fiction, NN achieves an AG of 81.4%, higher than VB or Punc. But four features (SOA1; SOA2; SOA1+JJ3+RB2; All-JJ&RB) do better than these standard categories. This shows that adjectives and adverbs should be used as discriminating features for genre identification.

We also find that the highest AG for all six problems is achieved by the combination of many kinds of adjective and adverb features (All-JJ&RB); this row is highlighted in Table 4. For example, distinguishing academic writing from fiction, using this feature achieves an astonishing 98.8% of the possible gain in accuracy (AG). The same method applied to fiction vs. news and academic vs. news scores the second and third highest AG of 93.0% and 90.8%, respectively. The same is true for the harder problems: news vs. not-news (AG=13.5%), academic vs. not-academic (AG=10.6%) and an impressive AG of 52.5% for fiction vs. not-fiction.

Feature / Method	Problem					
	One-Against-One			One-Against-Many		
	Acad vs. Fict	Acad vs. News	Fict vs. News	Acad vs. Not-Acad	Fict vs. Not-Fict	News vs. Not-News
SOA1	91.6	75.6	69.8	-1.8	16.4	-1.5
SOA2	89.4	73.0	72.0	1.8	13.9	0.8
JJ	72.2	65.0	61.6	9.7	0.0	6.8
RB	77.6	41.8	88.4	0.0	-4.9	-0.8
SOA1+JJ3 +RB2	88.6	79.6	70.4	-2.7	14.8	2.3
All-JJ&RB	98.8	90.8	93.0	10.6	52.5	13.5
NN	81.4	68.6	83.0	-2.7	0.0	3.8
VB	72.2	69.0	76.0	-10.6	-8.2	0.0
Punc	76.8	20.4	83.4	0.0	28.7	-1.5

Table 4: Best Performing Models Using Adjectives and Adverbs Compared to Other Features

Table 4 also shows that for five of the six problems, the next best performance is achieved by vector models using the 30 SOAs (SOA1) (academic vs. fiction, AG=91.6%), the 110 adjectives and adverbs (SOA1+JJ3+RB2) (academic vs. news, AG=79.6%), or models using the simple count of all adjectives (JJ) (academic vs. not-academic, AG=9.7%; news vs. not-news, AG=6.8%) or all adverbs (RB) (fiction vs. news, AG=88.4%). For fiction vs. not-fiction, a model using the simple count of all punctuation achieves the second best result (AG=28.7%), compared to using all adjective and adverb features (All-JJ&RB) (AG=52.5%).

Another way to assess the performance of our methods is to see which choices of features produce an accuracy gain for all six classification problems. Table 4 shows that these methods include the count of all adjective and adverb features (All-JJ&RB), and the vector of the 30 SOAs (SOA2), although we see that AG for the hard problems of academic vs. not academic and news vs. not-news is only 1.8% and 0.8% respectively. Nevertheless, benchmark models using nouns, verbs, and punctuation do not achieve a positive AG for all six problems.

Furthermore, the model representing the 30 SOAs (SOA2) that yields a positive AG for all six problems contains only 11 to 19 unique words in the final discriminant model (Table 5). Fewer than 20 unique SOAs (SOA2) can do the work of thousands of words (All-JJ&RB).

Speaker-Oriented Adverbs (SOA2)		
Problem	**AG**	**Unique Words in Model ***
Acad vs. Fict	89.4	17
Acad vs. News	73.0	11
Fict vs. News	72.0	13
Acad vs. Not-Acad	1.8	13
Fict vs. Not-Fict	13.9	19
News vs. Not-News	0.8	16
* A unique word in a discriminant model can be represented as both a type and a token variable, or only as a type or a token		

Table 5: Number of Unique SOAs in Models Yielding Accuracy Gain for All Problems

We also assess classifier performance of vector models using combinations of the three sets of the 110 words (Appendix A, B, C). We find that models representing only the 30 SOAs are most effective for academic vs. fiction, fiction vs. news, and fiction vs. not-fiction (Table 6). When combined with trait adjectives and trait adverbs, performance improves slightly for academic vs. news and remains stable for fiction vs. not-fiction.

Feature / Method	Problem					
	One-Against-One			One-Against-Many		
	Acad vs. Fict	Acad vs. News	Fict vs. News	Acad vs. Not-Acad	Fict vs. Not-Fict	News vs. Not-News
SOA1	91.6	73.8	71.2	-3.5	14.8	-0.8
JJ3	68.4	48.6	46.0	-3.5	-5.7	1.5
RB2	47.8	21.6	38.2	-2.7	-1.6	-2.3
JJ3+RB2	71.8	51.8	48.8	-4.4	-0.8	2.3
SOA1+JJ3 +RB2	88.6	79.6	70.4	-2.7	14.8	2.3

Table 6: Contribution of Speaker-Oriented Adverbs, Trait Adjectives, and Trait Adverbs to Models

Classifier performance is slightly better than the best guess for news vs. not-news using trait adjectives and trait adverbs alone. However, performance is not effective for academic vs. not academic using any combination of the three sets of the

110 words. This suggests that, as a class, SOAs contribute more to vector models than do trait adjectives and trait adverbs, and none of the 110 words are effective for distinguishing academic from not-academic documents.

Finally, we assess the contribution of the 30 SOAs (Appendix C) as compared to the 44 trait adjectives (Appendix A) and the 36 trait adverbs (Appendix B) by ordering the relative contribution of these 110 words to our vector models. We rank the 110 words that we entered into the various models by assigning scores according to the contribution (weight) each word made to a model, and by giving credit for the number of models each word contributed to. This method evaluates the contribution of words to all possible models (though it does not show which words are best for discriminating between particular genres). We find that only 95 of the 110 words contributed to a vector model. These 95 words include the 30 SOAs, 40 of the 44 trait adjectives, and only 25 of the 36 trait adverbs (we indicate these words in Appendix A, B, and C). On average, SOAs made the greatest contribution to vector models, generally ranking higher than trait adjectives and trait adverbs. For example, half of the 30 SOAs (e.g., *maybe, generally, surely, necessarily, clearly, specifically, strangely, and seriously*) rank in the top 25% of most effective words, whereas only small numbers of the other classes occur above the same cutpoint (9 of 40 trait adjectives; e.g., *bad, moral, natural, characterless*, and *honest*), and only 3 of 25 trait adverbs: *fairly, badly*, and *naturally*). Clearly, SOAs contributed most significantly to our genre classification models.

This may be indicative of a relationship between narrator behavior (marked by the use of SOAs in text) and author intent (one of several distinguishing criteria of genre). It also shows that the use of a linguistically defined construct guides us directly to the essential feature of the statistical models. Indeed, a model representing only 30 SOAs (SOA1) is comparable to the best-performing model (All-JJ&RB) for academic vs. fiction (AG=91.6%) (Table 4). It is most difficult to distinguish the three genres from all other genres in the corpus as expected, although fiction vs. not-fiction is relatively distinct using this feature (SOA1) (AG=16.4%).

6 CONCLUSION AND FUTURE WORK

Motivated by research in psychology and linguistics, we demonstrate that using adjective and adverb features in discriminant models is generally superior to benchmark models containing nouns, verbs, or punctuation features. In particular, vector models representing only 110 words (SOAs, trait adjectives and trait adverbs) are comparable to models using the count of thousands of words. As a class, the 30 SOAs are generally more effective than the class of 44 trait adjectives and 36 trait adverbs for the classification experiments we conducted.

But in the long term, our specific results are less important than the evidence that our approach to systematically studying the contribution of adjectives and adverbs to genre identification provides useful clues about how expressions of affect can be recognized by computer systems and how this information can be used for any application that depends on accurate identification of characteristics of affective language. Accuracy Gain rigorously measures the contribution of features to classification problems.

We recommend that our principles and methods be applied to (a) solving problems in other applications, (b) using other corpora, and (c) finding other features. Other applications include author identification, detection of subjectivity versus objectivity, classification of texts for question-answering, natural language generation, detection of customer review opinions in business environments, detection of personality traits in text, and detection of people who might be susceptible to certain beliefs. Other corpora include weblogs, email logs, and chat room logs. Possible sources of features include the domains of stylistics, communication, journalism, content analysis, and political discourse.

REFERENCES

[1] R. Rittman. *Automatic Discrimination of Genres: The Role of Adjectives and Adverbs as Suggested by Linguistics and Psychology*. Ph.D. Dissertation, Rutgers, The State University of New Jersey, New Brunswick, NJ. (2007). https://www.scils.rutgers.edu/~rritt/Rittman_dissertation_20070427.pdf

[2] D. Lee. Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path Through the BNC Jungle. *Language, Learning, and Technology*, 5(3):37-72 (2001).

[3] J. Karlgren and D. Cutting. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. *Proceedings of the 15th. International Conference on Computational Linguistics*, Coling 94, Kyoto. 1071-1075. (1994).

[4] B. Kessler, G. Nunberg, and H. Schutze. Automatic Detection of Text Genre. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 32 - 38. (1997).

[5] M. Wolters and M. Kirsten. Exploring the Use of Linguistic Features in Domain and Genre Classification. *9th Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway, 142-149. (1999).

[6] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Text Genre Detection Using Common Word Frequencies. *18th International Conference on Computational Linguistics* (COLING 2000), Luxembourg, Vol. 2, 808-814. (2000).

[7] K.B. Ng, S. Rieh, and P. Kantor. Signal Detection Methods and Discriminant Analysis Applied to Categorization of Newspaper and Government Documents: A Preliminary Study. In *Proceedings of Annual Conference of American Society for Information Science*, pp. 227-236. (2000).

[8] Y. Lee and S. Myaeng. Text Genre Classification with Genre-Revealing and Subject-Revealing Features. *Proceedings of the 25th Annual International ACM SIGIR*, 145-150. (2002).

[9] C. Lim, K. Lee and G. Kim. Multiple Sets of Features for Automatic Genre Classification of Web Documents. *Information Processing and Management*, 41, 1263–1276. (2005).

[10] M. Santini, R. Power and R. Evans. Implementing a Characterization of Genre for Automatic Identification of Web Pages. *Proceedings of 21st International Conference on Computational Linguistics* and *44th Annual Meeting of the Association for Computational Linguistics*, 699-706. (2006).

[11] M. Santini. Some Issues in Automatic Genre Classification of Web Pages. *Proceedings of JADT 2006: 8es Journées Internationales d'Analyse statistique des Données Textuelles*. (2006).

[12] S. M. zu Eissen and B. Stein. Genre Classification of Web Pages: User Study and Feasibility Analysis. *Proceedings of the 27th Annual German Conference on Artificial Intelligence (KI 04)*, 256-269. (2004).

[13] A. Finn and N. Kushmerick. Learning to Classify Documents According to Genre. *Journal of the American Society for Information Science and Technology*, 57 (11), 1506-1518. (2006).
[14] R. Bruce and J. Wiebe. Recognizing Subjectivity: A Case Study in Manual Tagging. *Natural Language Engineering* 5(2):187-205. (1999).
[15] J. Wiebe. Learning Subjective Adjectives from Corpora. *Proceedings of the 17th National Conference on Artificial Intelligence* (AAAI-2000), Austin. (2000).
[16] J. Wiebe, R. Bruce, and T. O'Hara. Development and Use of a Gold Standard Data Set for Subjectivity Classifications. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park: University of Maryland, 246-253. (1999).
[17] R. Rittman, N. Wacholder, P. Kantor, K.B. Ng, T. Strzalkowski, B. Bai, et al. Adjectives as Indicators of Subjectivity in Documents. *Proceedings of the 67th Annual Meeting of the American Society for Information Science and Technology*, 349-359. (2004).
[18] L. Goldberg. The Development of Markers for the Big-Five Factor Structure. *Psychological Assessment*, 4:26-42. (1992).
[19] S. Nowson, J. Oberlander, and A. Gill. Weblogs, Genres and Individual Differences. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, Stresa, Italy. (2005).
[20] S. Argamon, S. Dhawle, M. Koppel, and J.W. Pennebaker. Lexical Predictors of Personality Type. *Proceedings of the Classification Society of North America*, St. Louis. (2005).
[21] F. Mairesse, M.A. Walker, M.R. Mehl, R.K. Moore. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30, 457-500. (2007).
[22] D. Peabody and B. De Raad. The Substantive Nature of Psycholexical Personality Factors: A Comparison Across Languages. *Journal of Personality and Social Psychology*, 83(4):983-997. (2002).
[23] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. *A Comprehensive Grammar of the English Language*. London: Longman. (1985).
[24] A. Mittwoch. How to Refer to One's Own Words: Speech Act Modifying Adverbials and the Performative Analysis. *Journal of Linguistics*, 13:177-189. (1977).
[25] R. Jackendoff. Adverbs. In R. Jackendoff. *Semantic Interpretation in Generative Grammar* (pp. 47-107). Cambridge, MA: MIT Press. (1972).
[26] T. Ernst. The Semantics of Predicational Adverbs, and The Scopal Basis of Adverb Licensing. In T. Ernst. *The Syntax of Adjuncts* (pp. 41-91, 92-148). New York: Cambridge University Press. (2002).
[27] BNC2. British National Corpus, World Edition. http://www.natcorp.ox.ac.uk/ (2001).
[28] D. Lee. BNC World Index. http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/home/BNC_WORLD_INDEX.ZIP (2003).
[29] D. Lee. David Lee's Genre Classification Scheme. http://homepage.mac.com/bncweb/manual/genres.html (n.d.).
[30] M. Davies. VIEW: Variation in English Words and Phrases. http://view.byu.edu/ (n.d.).
[31] J. Wiebe. [Learning Subjective Adjectives from Corpora]. Unpublished list of subjective adjectives. http://www.cs.pitt.edu/~wiebe/pubs/aaai00/adjsMPQA (2000).
[32] I. Witten and E. Frank. *Data Mining*. San Diego: Academic Press. (2000).

APPENDIX A: 44 Trait Adjectives (JJ3)

avaricious *	bad *	biased *	calculating *	characterless *
decent *	deceptive *	dishonest *	disloyal *	ethical *
fair *	faithful *	frank *	honest *	hypocritical *
insincere *	intriguing *	just *	loyal *	lustful *
lying *	malicious *	materialistic *	mercenary *	moral *
natural *	noble *	perfidious	pharisaical	principled *
rapacious *	righteous *	sincere *	trustworthy *	truthful *
underhanded	unfaithful *	unreliable *	unscrupulous *	untruthful *
upright *	venal	virtuous *	vulgar *	
* Indicates 40 words that contributed to a vector model				

APPENDIX B: 36 Trait Adverbs (RB2)

avariciously	badly *	calculatingly *	decently *	deceptively *
dishonestly *	disloyally *	ethically *	fairly *	faithfully
hypocritically *	insincerely *	intriguingly *	justly *	loyally *
lustfully *	maliciously *	materialistically *	morally *	naturally *
nobly *	perfidiously *	pharisaically	rapaciously	righteously *
sincerely *	truthfully *	underhandedly	unfaithfully	unreliably *
unscrupulously	untruthfully	uprightly	valiantly	virtuously *
vulgarly				
* Indicates 25 words that contributed to a vector model				

APPENDIX C: 30Speaker-Oriented Adverbs (SOAs)

amazingly	briefly	candidly	certainly	clearly
confidently	curiously	definitely	frankly	generally
honestly	ideally	luckily	maybe	necessarily
normally	obviously	oddly	possibly	predictably
preferably	probably	roughly	seriously	simply
specifically	strangely	surely	surprisingly	unfortunately
Note: All SOAs contributed to a vector model				

Verbs as the most "affective" words

Marina Sokolova and **Guy Lapalme** [1]

Abstract. We present a work in progress on machine learning of affect in human verbal communications. We identify semantic verb categories that capture essential properties when human communication combines spoken and written language properties. Information Extraction methods then are used to construct verb-based features that represent texts in machine learning experiments. Our empirical results show that verbs can provide a reliable accuracy in learning affect.

1 Introduction

In some social situations, there is a tendency to avoid adjectives and adverbs with explicit negative connotations. Their absence, or near absence, can create additional problems to Text Data Mining for automated and statistical learning of affect and emotions. We attribute this to the fact that negative adjective and adverbs discriminate more between positive and negative opinions than those with a positive affect [13]. In the absence of negative words in texts, the accuracy of affect and emotion classification usually declines. To overcome this problem, we have looked for other sets of features to represent texts in machine learning experiments not involving positive and negative words.

In this paper, we show that, under certain conditions, people's actions, expressed by verbs, allow an accurate machine learning of the conscious subjective aspect of feeling or emotion, i.e., affect. We apply Communication Theory to build semantic verb categories, then formalize their use by language patterns and apply Information Extraction to construct text features from them. Debates from the US Congress and consumer-written forum messages provide appropriate data for empirical support, because in both cases data contributors consciously state their feelings towards the discussed matters.

In the empirical part of the paper, we apply machine learning technique to the texts represented by the verb-based features by running regression and classification experiments. Regression problems of sentiment and emotion analysis have not been widely studied before as previous studies mainly focused on binary classification [10], sometimes solving a three-class classification problem [16]. Joined regression and classification learning allows a more detailed analysis of the applicability of our approach. In the absence of a direct affect labelling, we use given opinion labels as their estimates.

Our method does not rely on domain-specific and content words thus, it is applicable to study affect on data belonging to different domains. Our results can also be used to *recover* affect in situations in which participants do not give an explicit negative evaluation of the discussed matter.

[1] RALI, Département d'informatique et de recherche opérationnelle, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, Québec, Canada, H3C 3J7, {sokolovm,lapalme}@iro.umontreal.ca

Category	Refers to	Examples
cognition	mental state	consider, hope, think, know
perception	activity of the senses	see,feel,hear
attitude	volition and feeling	enjoy,hate,love
activity	a continuing action	read, work, explain
event	happening or transition to another state	become,reply,pay,lose
process	continuing or eventual change of state	change,increase,grow

Table 1. The list of non-modal verb categories, their main semantic references, and examples of corresponding verbs.

2 Verb categories in human communication

Learning from records of human communication is one of the fastest growing areas of language and machine learning technologies. Such problems are more subjective and difficult to solve than traditional text classification and mining tasks [11], especially when the learning goal is the analysis of a communicated affect. They also require the development of methods to capture the relevant characteristics from a vast amount of data.

Stimulated by the English saying *"Actions speak louder than words"*, we looked at how verbs, which express actions in language, reveal a person's affect towards the discussed matter either emotionally or rationally. The emotional part may be expressed by attitude (enjoy, hate) and, partially, by the perception of the situation (smell, feel). The rational part may require the person to list facts such as events (meet, send) or the state of affairs (depend, have). To increase or diminish the communicative effect, people can use logic and politeness or imply possibility or necessity, that can be shown through the use of primary modals (can, will) or more conditional secondary modals (could, should) as was shown by the studies of Leech [8, 9].

We also consider that, under certain conditions, human communication combines characteristics of spoken and written communication. This happens when humans communicate through the Web or speak according to a prepared scenario such as in political debates. When such a situation occurs, we want to represent texts with the verbs that are most likely used in both spoken and written language.

For example, verbs denoting activity (play, write, send) and cognition verbs (think,believe) are the two most frequent categories when opinions are expressed in spoken-like language. Activity, the largest among verb categories, is the most frequent in all types of texts. Verbs denoting process (live, look, stay) often appear in written-like language, sometimes as often as activity verbs [2]. The high frequency of mental verbs is specific for spoken language [9, 14]. They are separated in three categories: attitude, perception and cognition. We defined the semantic categories, shown in Table 1, from verbs given in [8] to which we added their synonyms found the Roget's Interactive Thesaurus [1].

In Figure 1, we outline some involvement implications for patterns containing verbs. At the highest level, we consider whether the

closeness	→	*firstPerson* (*logic* | *physicalAction* | *mentalAction* | *state*)
distancing	→	`you` (*logic* | *physicalAction* | *mentalAction* | *state*)
logic	→	*primaryModal* | *secondaryModal*
physicalAction	→	[*modifier*] (*activity* | *event* | *process*)
mentalAction	→	[*modifier*] (*cognition* | *perception* | *attitude*)
state	→	[*modifier*] *havingBeing*
firstPerson	→	`I` | `we`
primaryModal	→	`can` | `may` | `will` | | `must` | ...
secondaryModal	→	`could` | `might` | `should` | `would`
activity	→	`read` | `work` | `explain` | ...
event	→	`become` | `reply` | `pay` | `send` | ...
process	→	`change` | `increase` | `stay` | ...
cognition	→	`believe` | `consider` | `hope` | ...
perception	→	`feel` | `hear` | `see` | `smell` | ...
attitude	→	`enjoy` | `fear` | `like` | `love` | ...
havingBeing	→	`have` | `be` | `depend` | `consist` | ...
modifier	→	*negation* | *adverb*

Figure 1. Rules (*non-terminal* → alternative$_1$ | alternative$_2$ | ...) generalizing the use of verb categories. | separate alternatives, [] indicate optional parts and parenthesis are used for grouping. *non-terminal* must be replaced by one of the alternatives. Alternatives are composed of other non-terminals and `terminals` which are the pieces of the final string.

person involves herself in the statement (*firstPerson*) or projects on interlocutors (*you*):

closeness uses `I` or `we` to indicate a direct involvement of the author; sub-rules indicate different degrees of the author's involvement:

- ***logic*** expresses permission, possibility, and necessity as the representation of logic, and superiority, politeness, tact, and irony as the representation of practice:
  - ***primaryModal*** such as `can` and `may` express direct possibility, permission or necessity of an action;
  - ***secondaryModal*** uses a more polite, indirect and conditional pattern than a primary modal and indicates more hypothetically and tentatively the author's intentions.
- ***physicalAction*** denotes an author's goal-oriented actions (*activity*), actions that have a beginning and an end (*event*) and a series of steps towards a defined end (*process*). This pattern corresponds to a direct and active involvement of the author;
- ***mentalAction*** uses mental action verbs, being more polite and tentative, that are a common face-saving technique and that mark openness for feedback;
- ***state*** indicates personal characteristics and corresponds to actions without definite limits and strong differentiations.

distancing uses second person pronouns and shows how an author establishes distance from the matter.

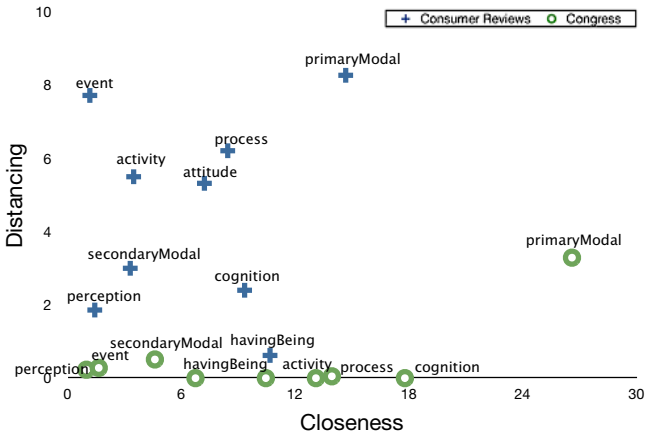


Figure 2. Distribution of verb categories in Congress debates and Consumer data.The horizontal axis estimates *closeness* (in per cent), the vertical axis – *distancing* (in per cent). Crosses denote Consumer reviews categories, circles – those in Congress debates. Labels indicate the verb categories of Figure 1.

3 Empirical support

In our empirical studies, we assume that positive and negative affect is revealed through the stated positive and negative opinion about the discussed matter. This is why we consider that opinion scores given in the corpus approximate the affect scores. We experimented with two kinds of data both combining spoken and written language properties: the first are consumer-written product reviews posted on the web which are loosely-edited, free structured texts, presumably written by the general population; the second are records of US Congress debates; they are more structured, edited and professionally written.

Consumer reviews introduced by Hu and Liu [4] in which text segments are manually tagged according to positive or negative opinions expressed by the reviewers, such as the following which is labelled +3 which means highly positive:

> this `is` *my* first digital camera , and what a ' toy ' *it* `is`! *i* `am` a software engineer and `am` very keen into technical details of everything *i* `buy`, *i* `spend` around 3 months before `buying` the digital camera; and *i* `must say`, g3 worth every single cent ...

To learn the strength of opinions, for the regression problem, we computed three numerical labels for each text: the number of positive tags, the number of negative tags, a signed sum of the two numbers. To solve classification problems, we applied unsupervised equal-frequency discretization to each of the numerical labels [3].

Congress debates We also used 1117 Congress debates [15] that either support or oppose a proposed legislature. Thomas et al. labeled texts by numerical polarity scores, computed by SUPPORT VECTOR MACHINE. SVM builds a decision surface that separates positive and negative texts. A score is the distance from a text to the surface. It can be positive or negative. The following excerpt has a positive score of 0.712:

> *we* `have known` that small businesses and working families `need` tax relief, and *we* `have fought` hard to `make` that happen so that *we* `see` the opportunity right there ...

For regression problems, we keep the scores as the data labels. For classification purposes, we use score signs as the data labels.

Verb distribution To illustrate and compare data's verb distributions, we calculated their frequencies and projected them with respect to *closeness* vs *distancing* axes as shown in Figure 2. As the resulting sets of points do not overlap, we conclude that the category distributions differ across these dimensions. The points of each set form a near-convex cluster with only one outlier: *havingBeing*, for consumer reviews and *primaryModal*, for Congress debates.

Information Extraction We constructed three feature sets based on the terminals of Figure 1:

1. The first feature set presents density and diversity of the words in each category. For a text T, for each verb category, we computed the number of word tokens and the number of word types for present, past and continuous forms. As a result, for each non-modal verb category we built six features. To represent modal verbs, we built four features, making 40 features in total.
2. The next set uses the 301 individual terminals as its features. Each terminal is represented by its occurrences in the text.
3. The third feature set expands the terminals with words occurring more than 5 times after or before a terminal.

See [12] for more information on the extraction process of verb information.

4 Learning experiments

We ran some learning algorithms available on Weka [17]. As mentioned in Section 3, we assumed that positive or negative subjective feeling is consciously revealed by the stated opinion. This assumption allowed us to use opinion labels as substitutes for the data affect labels.

Our first goal was to tackle regression (*quantitative*) learning problems. So far, machine learning experiments of sentiment analysis concentrated on classification (*qualitative*) tasks. Because of the novelty of this application, we wanted to try different types of algorithms to see what paradigms better learn the strength of the revealed affect. We chose KNN, a prototype-based algorithm, an optimization algorithm, SVM, and M5 TREES, a decision-based one. We applied BAGGING (bootstrap aggregating) to assess the influence of training data. In our experiments, BAGGING improved performance of M5 TREES, but not KNN nor SVM. We normalized each representation to eliminate the bias introduced by the text length.

Table 2 reports smallest relative absolute error RAE and corresponding root relative squared error $RRSQ$ obtained by the algorithms. The best performance, with the smallest error, was obtained on the Congress data. Positive consumer opinions were learned better than negative and overall opinions. An interesting phenomenon emerges when comparing algorithm performance – in terms of the learned correlation coefficients. The best performing algorithm in terms of accuracy is BAGGED M5 TREES. Since better accuracy implies that the algorithm learns dependencies between opinions and expressed actions better than other algorithms, we conclude that the output decision trees provide a reliable model of the data.

For Congressional debates, all output tree models agree that `demand`, `has` and `have` are the most important features, followed by `should` and `would`. Recall that we only report here the results of the best performing algorithms. Since this implies that the algorithms model better dependencies than other algorithms, we conclude that the strong language verbs have a positive correlation with attitude toward proposed legislations. On consumer review data, bagged trees placed `can`, `are` and `find` as the most important features for learning the overall opinions. Somewhat expectedly, `like` was among most decisive features for learning positive opinions. Learning negative opinions relied more on `be`, `am`, `would` and `should` than on other verbs.

To better display abilities of our approach, we performed a more traditional task of opinion classification. Again, we normalized each representation to eliminate the bias introduced by the text length. We chose SUPPORT VECTOR MACHINE (SVM) which is well-known for its high accuracy in text classification problems. Its use enabled us to compare our results with those of [15] obtained on the Congress debate data. They reported a test accuracy of 66.05 for positive/negative classification on the same data that we used for this work. The accuracy increased to 76.16 when they linked each data entry with previous speeches of the same speaker.

Our Congress results (78.14) have a better accuracy, even though we did not use previous records of speakers or other data reinforcements; the results are reported in the right part of Table 3. These results thus show that the *expressed actions do speak loud*. Under certain conditions, they reveal more than the previous history of the same speaker. For consumer reviews, learning positive opinions was easier than learning negative and overall opinions. Our method's accuracy is close to human-human agreement on positive and negative sentiments, when it is based on verbs [5]. More details on learning with verb-based features are provided in [12].

5 Related work

Sentiment analysis focuses on whether a text, or a term is subjective, bears positive or negative opinion or expresses the strength of opinion. Application of learning algorithms - through classification - has been pioneered by Lee et all [10]. However, Lee and many authors that followed her, used machine learning algorithms on reviews written by only four professional critics. This means that the algorithms were trained and tested on overly specific undiversified data. To achieve a comparable accuracy on the Congress data, they had to enhance data with previous speeches of speakers. Our goal is to seek general enough methods that can work with an unrestricted number of data contributors.

For automating recognition and evaluation of the expressed opinion, texts are represented through N-grams or patterns and then classified as opinion/non-opinion, positive/negative, etc. [6]. Syntactic and semantic features that express the intensity of terms are used to classify opinion intensity [16]. These works do not consider a hierarchy of opinion disclosure. We, however, built a pragmatic-lexical hierarchy of the use of semantic categories that allows us to interpret machine learning models formulated in lexical items and in terms of the pragmatics.

Various verb semantic classification schemes have been suggested and used for different purposes. Biber et al [2] examine word distribution, lexico-grammatical patterns and grammatical/discourse factors of four text genres: conversation records, fiction, news and academic writing. The authors suggest seven verb categories: activity, mental, communication, existence, occurrence, causative, aspectual. We think that these verb categories are not specific enough to distinguish between the verb's use in communicating personal opinions and other texts. We opted to build verb categories that reflect peculiarities of expressing personal opinions.

VerbNet [7] assigns verbs to 57 lexical-semantic categories such as *urge* (ask,persuade), *order* (command,require), *wish* (hope, expect), *approve* (accept,object). Since this work do not consider whether texts exhibit communication characteristics, the verb categories the

Algorithms	Consumer reviews						Congress	
	positive		negative		overall		debates	
	RAE	$RRSE$	RAE	$RRSE$	RAE	$RRSE$	RAE	$RRSE$
KNN	91.19	87.97	90.77	88.70	93.56	96.50	78.74	86.60
SVM	80.98	84.15	89.33	96.71	91.38	94.38	90.89	94.80
BM5P	*80.26*	82.21	*87.21*	85.81	*89.82*	96.61	*73.73*	78.84

Table 2. Smallest $RelativeAbsoluteError$ and $RootRelativeSquaredError$ obtained by the algorithms. Rows report results for each algorithm. Columns report results for each problem. For each problem, the smallest RAE is in $italic$.

Features	Consumer reviews						Congress	
	positive		negative		overall		debates	
	Acc	*Recall*	*Acc*	*Recall*	*Acc*	*Recall*	*Acc*	*Recall*
Categories	74.52	74.50	63.64	61.50	66.24	67.30	65.70	67.90
Terminals	76.12	75.80	66.56	67.20	70.06	74.50	69.63	72.00
Terminals-B	76.43	75.70	67.83	73.20	73.60	75.20	70.61	73.40
Collocations	77.75	79.00	68.33	69.50	73.82	78.90	75.18	77.60
Collocations-B	**78.87**	80.10	70.95	71.40	75.21	79.70	**78.14**	81.10

Table 3. Accuracy and corresponding true positive rates obtained by SVM. Rows report results for each feature set. Columns report results for each problem. For each problem, the largest accuracy is reported in **bold**. Baselines are the majority class accuracy: for the consumer data – 52.22, for Congress – 59.76.

authors suggest do not capture specifics of communication. We focused on verb categories in communicative texts, in which speakers communicate their opinions about the discussed matters.

6 Conclusion and future work

In this study, we have shown the importance of relations between expressed actions and affect. We formalized expressed actions by building language patterns of modal, event, activity, process, cognition, perception, state verbs and personal pronouns. We applied machine learning methods to establish quantitative relations between the use of verb categories and affect.

Our use of regression and classification methods allows to perform a more detailed learning than previous studies that usually defined their problems either as binary classification or multi-class classification problems. On two data sets, consumer reviews [4] and the US Congress debates [15], we showed that regression problems were successfully learned by BAGGED M5 TREES, whereas SVM obtained a reliable accuracy in classification problems. Our method extracts all its information from only the given data. Other methods could only achieve a similar accuracy by adding personal information about speakers, such as the history of previous comments [15]. However, such type of additional information is not often easily available.

Learning affect from the used verbs becomes practically justified and, indeed, desirable when a social context dictates avoidance of negative adjectives and adverbs, because empirical results showed that negative adjective and adverbs discriminate better between positive and negative emotions than positive ones. In the future, we intend to analyze the use of different types of verb modifiers (always, never). We are also interested in learning the correspondence between a revealed affect and pragmatics of communication, e.g. intensity and immediacy. Another venue for future work is to investigate the phenomenon of impression building, i.e. how texts allow inference of an author's abilities or intentions.

Acknowledgements

This work has been supported by *Natural Sciences and Engineering Research Council* of Canada. The authors thank Fabrizio Gotti for technical support of experiments.

REFERENCES

[1] Roget's interactive thesaurus, 2006. http://thesaurus.reference.com/.
[2] D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, *Longman Grammar of Spoken and Written English*, Longman, 1999.
[3] M. Boulle, 'Optimal bin number for equal frequency discretizations in supervized learning', *Intelligent Data Analysis*, **9**(2), 175–188, (2005).
[4] M. Hu and B. Liu, 'Mining opinion features in customer reviews', in *Proceedings of Nineteeth National Conference on Artificial Intelligence (AAAI-2004)*. AAAI Press, (2004).
[5] S.-M. Kim and E. Hovy, 'Determining the sentiment of opinions', in *Proceedings of the of the 20th international conference on Computational Linguistics (COLING - 2004)*, pp. 1367–1373, (2004).
[6] S.-M. Kim and E. Hovy, 'Crystal: Analyzing predictive opinions on the web', in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 1056–1064, (2007).
[7] K. Kipper, A. Korhonen, N. Ryant, and M. Palmer, 'Extending verbnet with novel verb classes', in *Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 25–32, (2006).
[8] G. Leech, *Meaning and the English Verb*, Longman, 2004.
[9] G. Leech and J. Svartvik, *A Communicative Grammar of English*, Longman, third edn., 2002.
[10] B. Pang, L. Lee, and S. Vaithyanathan, 'Thumbs up? sentiment classification using machine learning techniques', in *Proc Empirical Methods of Natural Language Processing EMNLP'02*, pp. 79–86, (2002).
[11] M. Sokolova and G. Lapalme, 'Performance measures in classification of human communication', in *Proceedings of the 20th Canadian Conference on Artificial Intelligence (AI'2007)*, pp. 159 – 170. Springer, (2007).
[12] M. Sokolova and G. Lapalme. Do actions speak loud? semantic verb categories in expression of opinions, 2008. submitted elsewhere.
[13] M. Sokolova and G. Lapalme. A simple and effective information extraction method for opinion analysis, 2008. submitted elsewhere.
[14] M. Sokolova and S. Szpakowicz, 'Language patterns in the learning of strategies from negotiation texts', in *Proceedings of the 19th Canadian Conference on Artificial Intelligence (AI'2006)*, pp. 288 – 299. Springer, (2006).
[15] M. Thomas, B. Pang, and L. Lee, 'Get out the vote: Determining support or opposition from congressional floor-debate transcripts', in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 327–335, (2006).
[16] T. Wilson, J. Wiebe, and R. Hwa, 'Recognizing strong and weak opinion clauses', *Computational Intelligence*, **22**(2), 7399, (2006).
[17] I. Witten and E. Frank, *Data Mining*, Morgan Kaufmann, 2005.

eXTRA: A Culturally Enriched Malay Text to Speech System

Syaheerah L. Lutfi[1], **Juan M. Montero**[1], **Raja N. Ainon**[2] **and Zuraida M.Don**[3]

Abstract. This paper concerns the incorporation of naturalness into Malay Text-to-Speech (TTS) systems through the addition of a culturally-localized affective component. Previous studies on emotion theories were examined to draw up assumptions about emotions. These studies also include the findings from observations by anthropologists and researchers on cultural-specific emotions, particularly, the Malay culture. These findings were used to elicit the requirements for modeling affect in the TTS that conforms to the people of the Malay culture in Malaysia. The goal is to introduce a novel method for generating Malay expressive speech by embedding a localized 'emotion layer' called eXpressive Text Reader Automation Layer, abbreviated as eXTRA. In a pilot project, the prototype is used with Fasih, the first Malay Text-to-Speech system developed by MIMOS Berhad, which can read unrestricted Malay text in four emotions: anger, sadness, happiness and fear. In this paper however, concentration is given to the first two emotions. eXTRA is evaluated through open perception tests by both native and non-native listeners. The results show more than sixty percent of recognition rate, which confirmed the satisfactory performance of the approaches.

1 INTRODUCTION

Intelligent systems have been shown to increase their effectiveness by adapting to their individual users. In recent years, it has been recognized that affective factors can play an important role in this adaptation. One of the core conveyers of affect is speech. However, most of the research within the field of intelligent systems uses synthetic characters that are usually based on a full-blown simulated individual personalities whose behaviours are psychologically [1]}or biologically-driven (e.g., Blumberg, 1994 in [2]. This includes having conversational styles that are limited to a certain content and manner of speech, which reduces the believability and trustworthiness of such characters. It is believed that the reason for this is because such character's design ignores the pivotal mask of affect, which is the *socio-cultural grounding* [3];[4];[2];[5];[6]. Little research within the emotion-oriented technology field aims at understanding cultural differences which influence vocal affect.

While some aspects of emotion are universal to all cultures, other aspects may differ across cultures [4];[2]. For example, Americans and Asians have slightly different conceptions of self. American culture promotes a view of the self as independent. On the other hand, most Asian cultures, such as those of Japan and China, promote a view of the self as interdependent (collectivist culture). People from these cultures tend to describe themselves in terms of which groups they belong to. They learn to rely on others, to be modest about achievements, and to fit into groups. Maldonado and Hayes Roth [2] pointed out that biological or psychological model of synthetic characters do not express "the essence of humanity of their constructions", for example, particularly referring to speech, the intonation in the utterances in a particular topic addressed may not conform to a certain culture in terms of *what*, *how* and *when* it is said. This is more obvious when the topic being addressed is within the realm of persuasion. Taking into considerations the findings of these studies, we therefore proposed that the modeling of affective speech for a TTS to be expanded to include pursuits of cultural variability, producing a culturally-specific synthetic speech

2 BACKGROUND STUDIES

2.1 Culturally-dependent Vocal Affect

The expression and perception of emotions may vary from one culture to another [7]. Recent studies [8];[9] reveal that localized synthetic speech of software agents, for example, from the same ethnic background as the interactor are perceived to be more socially attractive and trustworthy than those from different backgrounds. The participants in Nass's [9] experiments conformed more to the decisions of the ethnically matched characters and perceived the characters' arguments to be better than those of the ethnically divergent agents. Just as with real people, users prefer expressive characters that "sound" like them; that they can better relate with, because it's easier to understand and predict. Therefore, cultural localization is critical even in the effort of actively matching the user's ethnicity, and perhaps also central psychological tendency.

Particularly in speech, the acoustic characteristics such as intonation, pronunciation, timbre, and range of vocal expressions that are localized to certain extent of variability, are constantly used in everyday activities to differentiate individuals across cultures [2]. These conversational aids can be used to determine not only the geographical origin of a particular person or character, but even their cultural influences and places of residences.

Based on these studies, we realized that it is crucial to infuse a more familiarized set of emotions to a TTS system whereby the users are natives. This is because; a TTS system that produces affective output that is better 'recognized' would have a reduced artificiality and increased spontaneousness, hence offering users more comfort when interacting with the TTS system

Additionally, by concentrating on the culturally-specific manner of speaking and choices of words when in a certain

[1] Language Technology Group, Technical University of Madrid, email: {syaheerah,juancho}@die.upm.es

[2] Language Engineering Lab, University Malaya, email: ainon@um.edu.my

[3] Faculty of Language and Linguistics, University Malaya, email: zuraida@um.edu.my

emotional state, the risk of evoking confusions or negative emotions such as annoyance, offense or aggravation from the user is minimized, other than establishing a localized TTS.

2.2 Vocal Affect in Malay Culture In Relation To Anger and Sadness

In an attempt to understand emotions from the Malay perspective especially with regard to anger and sadness, we refer quite substantially to the work by Wazir Jahan [10]. According to her the description of the emotions in Malay was not based on empirical research but based on passing observations and intuitive reasoning. She concedes that many studies have been carried out on *latah* (for women) and *amuk* (for men, English a*mok*), since these two expressions of emotion are closely related to the understanding of the 'Malay mind' then brought about by rebellious reactions against colonization. Wazir Jahan examined the observations of the Malay mind by several western anthropologists who believe that the Malay people look 'externally impassive' but are actually very sensitive even to something as normal as 'the accidents of every day life'. Evidence gathered from past observations seem to show that the Malays are inclined to keep their emotions in check until the time when they cannot contain them anymore and that is when they explode. These observations seem to be in line with what is expressed by the former Prime Minister, Tun Dr. Mahathir in his book *The Malay Dilemma*, "*the transition from the self-effacing courteous Malay to the amok is always a slow process. It is so slow that it may never come about at all. He may go to his grave before the turmoil in him explodes*" [11]In this article we are not interested in the phenomenon of amok in itself but in its expression since it bears elements of a culturally specific form of anger.

A study carried out by Silzer [7] illustrates that the expression of human emotions are cultural specific, e.g. how anger is expressed in English is different from how 'marah' (anger) is expressed in Malay. He explains that the causal component of *marah* is more specific such that marah "is the result of intentional personal offence, where the offender knowingly committed the "bad" act, while realizing it would cause the other person unpleasant feeling". This causes the offended party to inform the offender in a certain tone of voice that he or she has done something wrong, and should rectify the problem. It is also observed that when expressing anger, Malays are inclined to shout. This way of expressing anger could probably be caused by the accumulation of negative feelings which when released manifest in the form of shouting or yelling.

Preliminary studies show that Malay utterances when uttered in anger tend to have a slightly higher overall pitch while sadness is accompanied by lower overall pitch when compared to English utterances [12]

2.3 Issues in Affective Speech Modelling

In recent years, there have been an emerging number of studies focusing on Malay text-to-speech conversion [12-16]. These are concatenative speech conversion systems, which mostly apply phonological rule-based approach for prosody modification in order to invoke imitation of humans' pronunciation. Nonetheless, though these prosodic models were introduced in the hope of providing a high degree of naturalness, it is still insufficient to localize the output (to make it culture-dependent), hence, limiting its naturalness.

Three major issues that contribute to this problem have been identified; firstly, there are various linguistic features that interactively affect the phonological characteristics, making it difficult to gather complete rules to describe the prosody diversity [17]. The second challenge in modeling an affective component is the variability in speech. A speaker may not repeat what he says in the same way; he may not use the same words to convey the same message twice knowingly or not (even in read speech) [18]. One can also say the same word in many different ways depending on the context. Therefore, the instances of the same word will not be acoustically identical. This is quite difficult to map in a TTS system, especially when using qualitative rules, which causes the repetition of the same set of prosody when reading long sentences.

The usual practise is that, the linguistic features and speaking styles will be translated into prosodic patterns, which are repeatedly applied to speech. While this may be good for a small amount of sentences, repeated tones become boring and tedious for reading whole paragraphs of text. Apart from that, the same sets of tones do not fit different types of sentences with varying contents and lengths [14]. Therefore, applying fixed qualitative rules to prosodic variation patterns or ranges comes with great limitations. Lastly, there is a dearth of prerequisite studies on various human emotions. Consequently, to find a solution for these issues, a novel approach using emotion templates to apply expressiveness to the output of TTS system was investigated. This paper presents the completed work of the prototype.

3 THE MALAY SPEECH DATABASE

The findings from the studies above lead us into building a database that consists of speech data with emotions that are more 'agreeable' to the Malay people. This is done by directing the speaker to record her speech by speaking them in the two emotional states *suitable* with the Malay identity. There are two sets of utterances: one with neutral contents, and the other with emotionally-inherent contents. Each set contains thirty two utterances. For each of the utterances with emotionally-inherent contents, an accompanying scenario that elicits the corresponding emotion is given. For example "Kamu sungguh kurang ajar" (You are so rude) and "Reaksi terhadap anak murid yang menendang kerusi guru dengan sengaja" (Reaction towards a student of yours who kicked your chair on purpose) were sentence and scenario, respectively. Having such elicitation scenario helps to reduce the interpretation variations.

To ensure that the intended emotions elicited in the speech samples are recognized by listeners, a series of open perceptual tests was conducted. Results show that the speech samples with neutral content set have a low recognition rate while the samples with emotionally-inherent content are highly recognized with minimum effort, in other words, these samples are perceived as intended by the native participants. The results are shown in section 6.

4 THE MALAY LANGUAGE SYLLABLE STRUCTURE

It is observed that in Malay language, the structure of syllables is straightforward. In addition, the intonational or prosodic relationship between syllables within a word is more obvious than between two words. The simple syllable structure that the Malay language is based on allows for the use of an algorithm that focuses on the number of syllables rather than other linguistic features [15]. In Malay, the syllable structure units are as follows:

- CVC (Consonant-Vowel-Consonant)
- CV (Consonant-Vowel)
- VC (Vowel-Consonant)

5 IMPLEMENTATION

A prototype by Wu and Chen [17] on template-driven generation of prosodic information for their concatenative Chinese TTS system has inspired the use of templates to generate emotions for Malay synthesized speech. Additionally, the findings in the interdisciplinary studies discussed in previous sections shaped the idea to propose a hybrid technique for building an effective emotion component to be used with concatenative Malay TTS system. The justifications are listed below:

i. Since the hosting system uses diphone concatenative synthesis (Multi-Band Resynthesis OverLap Add or MBROLA), the employment of this technique is compulsory.
ii. The facts about Malay language syllable structure discussed section 6, added with the restrictions of phonological rule-based systems mentioned in section 4, shaped the idea to create a *syllable-sensitive* rule-based system.
iii. The effectiveness of the template-driven method proposed by Wu and Chen [17] has brought the idea to adapt this method and combine it with the techniques in (i) and (ii).

Table 1: Detailed Information on the Child Components

Child Components	Responsibility
of Template Selector: SyllableReader TemplateMatcher	 This component reads and analyses syllables This component matches the results of the analysis, with data from the database in order to select the correct emotion template.
of Merger: Composer Emotionizer	 This component provides for navigable structures of phonemes (input and template are separately handled). This component applies a rule-based algorithm to Composer data in order to merge input-derived data with template data.

5.1 Template-driven Emotion Generation

The combination of the techniques in (i), (ii) and (iii) above derives the eXpressive Text Reader Automation Layer, or eXTRA. Figure 1 exposes eXTRA module's detailed internal architecture, while Table 1 explains the responsibilities of each of the child component.

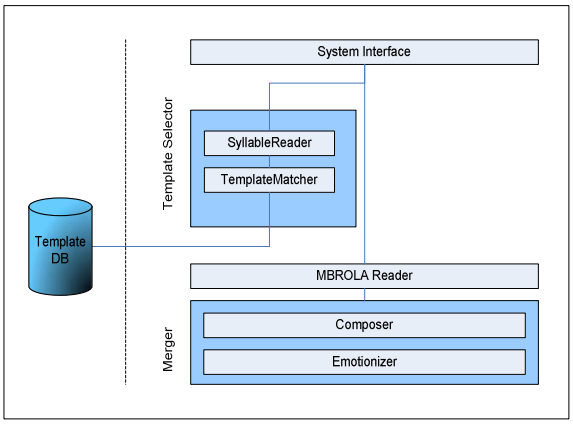


Figure 1: Low-level Architecture of eXTRA

For the generation of highly natural synthetic speech, the control of prosodic parameters is of primary importance. Diphone synthesis allows maximum control of prosodic parameters. Therefore, attempts to model the emotions in eXTRA took advantage of model-base mapping or "copy synthesis" to build the emotion templates. In other words, the emotional prosodic parameters from the actor's speech samples are 'copied' into the templates. First, the actor's speech data is annotated on phoneme level using speech analysis software, Praat [19]. Then, the exact pitch and duration information from each phoneme is extracted and transferred into acoustical data in templates, which ensures more natural emotional-blended speech when the target template is applied to the speech. The next section explains how the prototype works in more detail.

5.1 How eXTRA Works

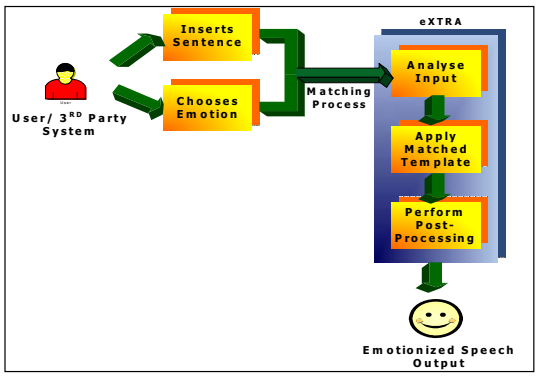


Figure 2: A simplified framework of eXTRA

Figure 2 provides the visual illustrations of eXTRA's framework Using the syllable-sensitive algorithm, each word from user input is analyzed and chunked into syllables in reverse order (stack) to determine syllable count; the input sentence is processed from the last word to the first. The result is then matched against the emotion template that contains the sentence with the same syllable-count and sequence. In other words, the template selection is done by identifying the integers that represent the syllable sequence of the template-sentence – "2222", "2332" etc. This is done by using a template selector module. When matched, the prosodic information from the template will be transferred to input at the level of phoneme. To ensure a more natural tune, the post-processing is done. It involves assigning silence and default parameters to additional phonemes correlating to each word wherever necessary. Figure 2 shows the framework of eXTRA while Figure 3 below presents a screenshot of the Fasih extended with eXTRA.

Consider the input sentence "Awak tidak tahu malu" (you have no shame) is to be spoken in anger. This sentence has a syllable sequence set of "2222". Therefore, the anger template that will be selected from the database also comprises the syllable sequence set "2222". The sentence in this template is "Kamu sungguh kurang ajar" (You are so rude). Consequently, the anger template is applied to the input sentence to produce an emotionized output. This is done by matching the emotional prosodic parameters from the template-sentence to the input-sentence at the level of phonemes. The matching process is explained in the next section. To ensure a more natural tune, the post-processing is done. It involves assigning silence and default parameters to additional phonemes correlating to each word wherever necessary.

In other words, the eXTRA module then enhances this speech data to become emotional speech data by applying an emotion template. Thus, the generating of emotional speech output requires three essential components: *input data* and *template data* (both representing speech data) and a *rule-based algorithm* that renders the data into output (Figure 3).

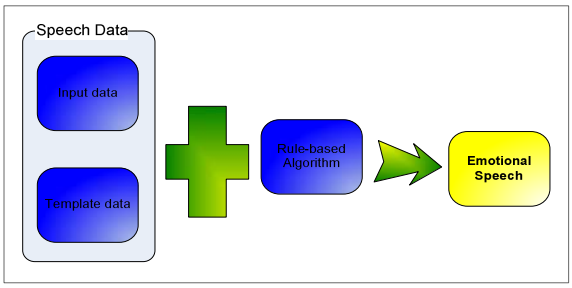


Figure 3: Major Components That Produce Emotional Speech

5.1.1 Matching Process

Consider the input sentence is "Awak tidak tahu malu" (you have no shame) and the selected emotional state is anger. This sentence has a syllable sequence set of 2222, therefore the matched anger template would be the template that has the same syllable sequence as well. From the template database, the particular matched template consists of the sentence "Kamu sungguh kurang ajar". Appendix A shows how the input phonemes are matched against the template phonemes. Vowels usually have longer duration than consonants, thus, contributing to more pitch points. However, vowel pitch points are not suitable to be transfered to consonants, since this may produce longer sound than expected. To solve this issue, syllabic and categorical matching are applied. Syllabic matching refers to the matching of phonemes between the input and template according to syllables. In other words, a pattern of syllables from the sentence is first identified in order to establish a match against another sentence's syllable pattern. Categorical matching refers to the matching of phonemes of the same type; vowels are matched against vowels while consonants are matched against consonants. This is illustrated in Table 2, where the vowels from the input sentence are matched against the vowels from the template sentence according to syllables. This also applies for consonants.

In the case where a phoneme is left without a match, a default duration value or silencing is assigned. A default duration value is assigned to the unmatched phonemes in the input sentence while the unmatched phonemes in the template are put to silence.

Table 2: The Organization of Matching Between the Template and the Input Contents

Line No.	Contents of Template Sentence			Contents of Input Sentence		
	SAMPA symbol	Parameters		SAMPA symbol	Parameters	
		Duration(ms)	Pitch Points pairs		Duration(ms)	Pitch Points pairs
1	k	silenced				
2	V	105	0 287 31 253 69 281 100 296	V	105	0 287 31 253 69 281 100 296
3	m	59 (<92)	50 309 100 323	w	92	50 309 100 323
4	U	65	18 321 49 309 100 278	V	65	18 321 49 309 100 278
				k	92(default)	

Legend:
Unmatched phoneme

Table 2 shows that the relevant prosodic parameters from the phonemes in the template are transferred to the matched phonemes in the input. A post-processing is also done for the purpose of assigning silence and default values to the 'left-over', unmatched phonemes. The example shows that consonant /k/ in the template is put to silence while consonant /k/ in input is given a default value of 92 for the opposite reason. Such value is given so that the consonant produces a basic sound when concatenated. This value is copied from Fasih, which assigns *only* duration parameter to its consonants.

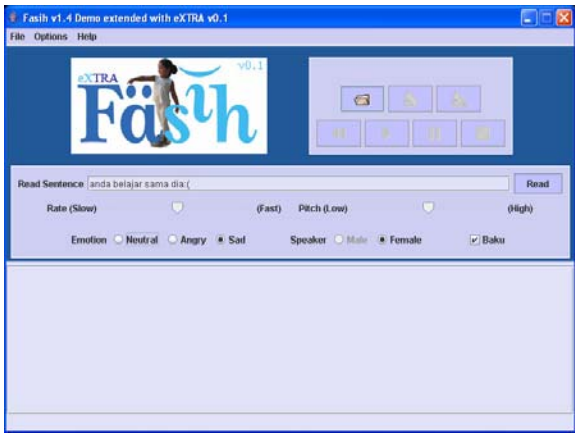
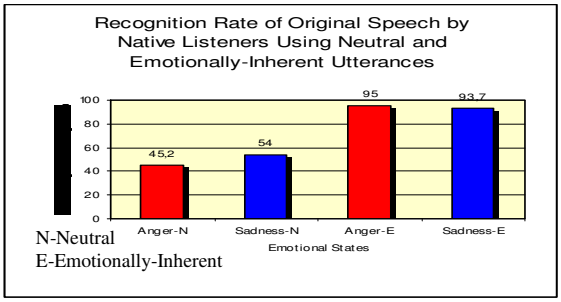


Figure 4: A screenshot of Fasih extended with eXTRA

6 EVALUATIONS

The prototype is evaluated in an open perceptual test participated by 20 native and 20 non-native listeners who were not aware of the test stimuli. They are Malaysian and international students of University Malaya, Kuala Lumpur. Native listeners were asked to listen to both sets of neutral and emotionally-inherent utterances (64 utterances) while non-native listeners only listened to emotionally-inherent utterances (32 sentences). A week earlier, they were asked to listen to the same set of utterances of the original samples (actress' speech). The results obtained with native listeners are presented in Figure 5 and Figure 6 respectively, while with non-native listeners, it is in Figure 7.



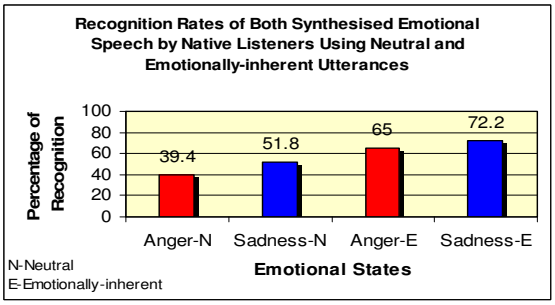


Figure 5 and 6: Recognition results for both original and synthesized speech samples using neutral and emotionally inherent content

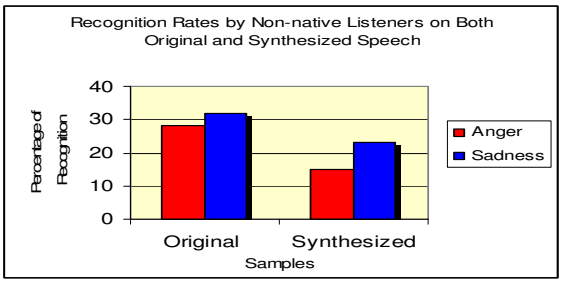


Figure 7: Results with non-native listeners

7 SUMMARIES OF RESULTS AND CONCLUSION

Comparing Figure 4 and Figure 5, the differences in recognition for samples of neutral contents between both charts are very small for both the emotions, which suggests that possibly the actress was relatively less successful in expressing the intended emotions using neutral contents. On the other hand, the difference of recognition for samples with emotionally-inherent contents shows that there is still room for improvement by approximately 25%, with regards to the modeling of the synthesized speech. The intonation and duration model for the emotional speech is used to generate the emotional utterances from neutral speech. In the original samples (Figure 4), the difference of recognition between the semantically-neutral and semantically-emotional speech bars is 45% on average while for the same comparison; the difference is 23.5% for synthesized speech. We suggest that these differences in recognition shown are due to content of the speech. It is observed that participants tend to focus on the contents rather than the tones elicited despite repeated reminders. Nevertheless, this kind of response is expected, because in real life situations, meaning and context are a bigger clue to the emotional state of the speaker. Lastly, the significant differences shown in the results from the experiments between neutral and emotionally-inherent contents proved that utterances that have no conflicting content and localized emotions are more suitable for use in building templates.

As for non-native listeners, there is a high difference of recognition between original and synthesized speech in anger compared to sadness. More participants are able to accurately detect sadness in synthesized speech compared to anger. This is possibly because of most sad speech samples exhibits lower F0 and longer duration and therefore it is easy to point out that the speaker is sad. Our data showed that in this open test, most participants from Middle East tend to perceive anger as "neutral", while participants from Western countries tend to presume sadness as "anger". This discovery is interesting to us as in some cases; even samples that produce clear angry expression (that can be easily detected by Malay listeners) are deemed as neutral. The low recognition rates clearly show that the non-native participants may have different perceptions from native participants. The findings based on data also proved Silzer's statement that "expression and perception of emotions vary from one culture to another"

Overall, the recognition rates by native listeners show higher figures compared to previous research work([20]; [9];[21];[22].. Basically, these results indicated over sixty percent recognition rates for both intended emotions expressed in the synthesized

utterances, which are encouraging, considering that people recognize only sixty percent emotion in *human* voice (Shrerer, 1981 in Nass *et al.*, 2000). This is possibly because due to the effort in localizing the emotion for better perception.

8 ACKNOWLEDGEMENTS

We are deeply grateful to Dr. Normaziah Nordin and Mr. Kow Weng Onn from the Pervasive Computing Lab, MIMOS for their fruitful suggestions and advise on improving this prototype. We are also greatly indebted to Mr. Imran Amin H. de Roode from Fullcontact, for providing professional guidance and assistance in technical effort. This work has also been partially supported by ROBINT (DPI2004-07908-C02-02) and ROBONAUTA (DPI2007-66846-C02-02).

9 REFERENCES

[1] Nass, C., Isbister, K., & Lee, E. , *Truth Is Beauty: Researching Embodied Conversational Agents*, in *Embodied Conversational Agents*, J. Cassell, et al., Editors. 2000, MIT Press: Cambridge, MA. p. 374-402.

[2] Maldonado, H. and B. Hayes-Roth, *Toward Cross-Cultural Believability in Character Design*, in *Agent Culture: Human-Agent Interaction in a Multicultural World* S. Payr and R. Trapple, Editors. 2004, Lawrence Erlbaum Associates: Mahwah, NJ. p. 143-175.

[3] Physorg.com (2007) *Research Finds that Culture is key to interpreting facial expressions*. DOI: Article 96297525

[4] Krenn, B., et al., *Life-Like Agents for the Internet: A Cross-Cultural Case Study*, in *Agent Culture: Human-Agent Interaction in a Multicultural World*, S. Payr and R. Trappl, Editors. 2004, Lawrence Erlbaum Associates: Mahwah,NJ.

[5] Hayes-Roth, B. and P. Doyle, *Animate Characters.* Autonomous Agents and multi-agent systems, 1998. **1**(2): p. 195-230.

[6] Reeves, B., & Nass, C, *The Media Equation: How People Treat Computers, Televisions, and New Media Like Real People and Places.* 1996, NY: Cambridge University Press.

[7] Silzer, P.J. *Miffed, upset, angry or furious? Translating emotion words*. in *ATA 42nd Annual Conference*. 2001. Lost Angeles, CA.

[8] Brave, S. and C. Nass, *Emotion in Human-Computer Interaction*, in *The Human-Computer Interaction Handbook: FundamentALS, Evolving Technologies and Emerging Applications*, J.A. Jacko and A. Sears, Editors. 2003, Laurence Elbaum Associates (LEA): Mahwah, NJ. p. 81-93.

[9] Nass, C., et al., *The effects of emotion of voice in synthesized and recorded speech*. 2000: Standford, CA.

[10] Wazir-Jahan, K., ed. *Emotions of Culture: A Malay Perspective*. ed. W.J. Karim. 1990, Oxford University Press: NY.

[11] Mohamad, M., *The Malay Dilemma*. 1981, Kuala Lumpur: Federal Publications.

[12] Razak, A.A., M.I.Z. Abidin, and R. Komiya. *Emotion pitch variation analysis in malay and english voice samples*. in *The 9th Asia-Pacific Conference on Communications 2003*. 2003.

[13] El-Imam, Y.A. and Z.M. Don, *Text-to-speech conversion of standard Malay.* International Journal of Speech Technology, 2000. **3**(2): p. 129-146.

[14] Syaheerah, L.L., et al. *Template-driven Emotions Generation in Malay Text-to-Speech: A Preliminary Experiment*. in *4th International Conference of Information Technology in Asia (CITA 05)*. 2005a. Kuching, Sarawak.

[15] Syaheerah, L.L., et al. *Adding Emotions to Malay Synthesized Speech Using Diphone-based templates*. in *7th International Conference on Information and Web-based Applications & Services (iiWAS 05)*. 2005. Kuala Lumpur, Malaysia: University Malaya.

[16] Tiun, S. and T.E. Kong, *Building a Speech Corpus for Malay TTS System*, in *National Computer Science Postgraduate Colloquium 2005 (NaCPS'05)*. 2005.

[17] Wu, C.-H. and J.-H. Chen. *Template-driven generation of prosodic information for chinese concatenate synthesis*. in *IEEE International Conference on Acoustics, Speech and Signal Processing*. 1999. Phoenix, Arizona.

[18] Murray, I.R. and J.L. Arnott. *Synthesizing emotions in speech: is it time to get excited?* in *4th International Conference on Spoken Language Processing 1996*. 1996.

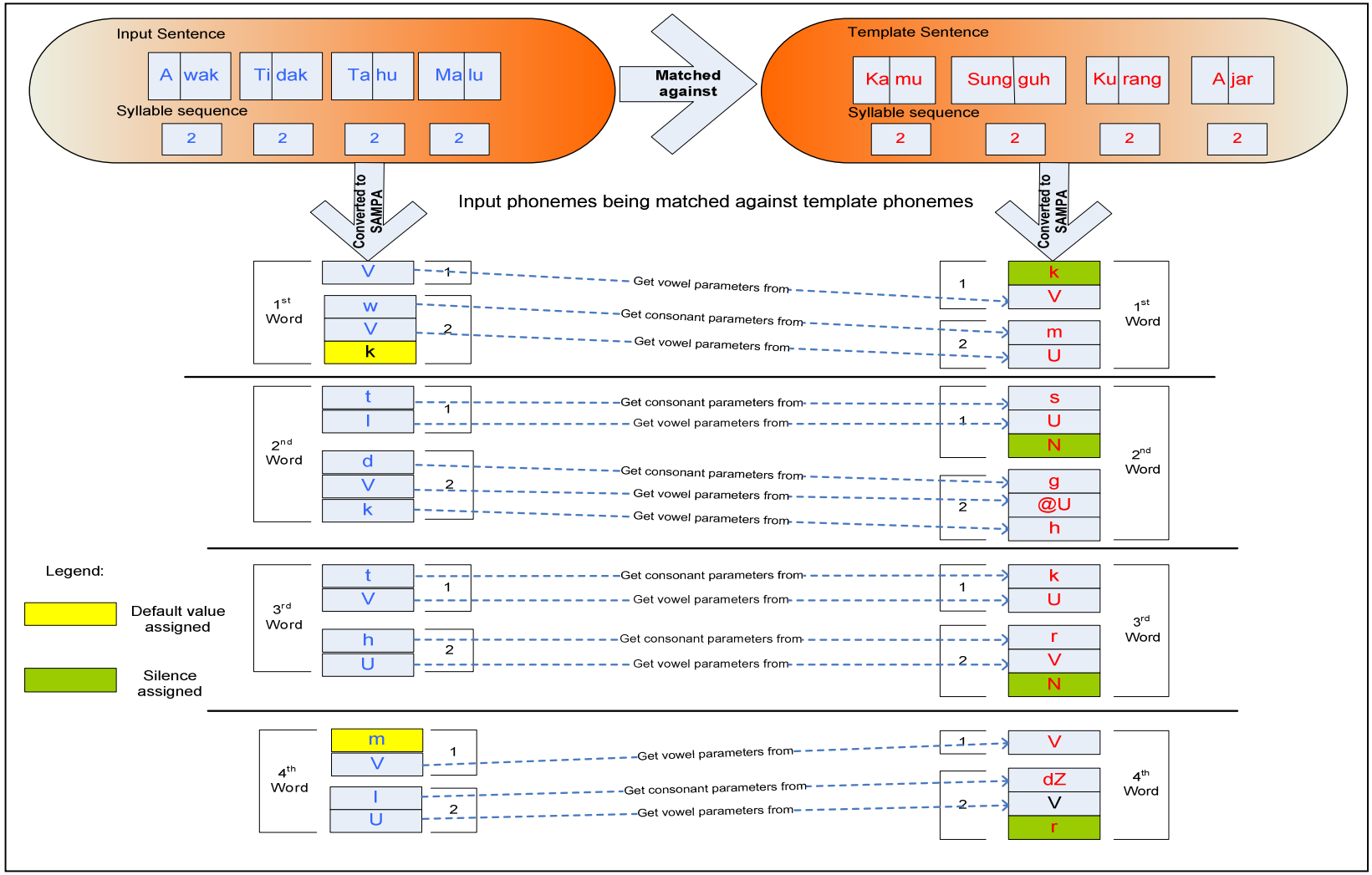
[19] Boersma, P. and D. Weenink, *Praat*. 2005: Amsterdam, NL.

[20] Bulut, M., S. Narayanan, and A.K. Syrdal. *Expressive speech synthesis using a concatenative synthesizer*. in *ICSLP*. 2002. Denver, CO.

[21] Murray, I.R. and J.L. Arnott, *Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion.* Journal Acoustical Society of America, 1993. **93**(2): p. 1097-1108.

[22] Murray, I.R., J.L. Arnott, and E.A. Rohwer, *Emotional Stress in synthetic speech: progress and future directions.* Speech Communication, 1996. **20**(1-2): p. 85-91.

APPENDIX A: A Visualization of the Phonemes Matching Process



Single Speaker Acoustic Analysis of Czech Speech for Purposes of Emotional Speech Synthesis

Martin Grůber [1] and **Milan Legát** [2]

Abstract. This paper deals with an acoustic analysis of the sets of Czech sentences uttered by single speaker. The data used in this analysis consists of both emotional and neutral sentences. We have been especially interested in some features which are supposed to influence the perception of speech, such as F0, phoneme duration, formant frequencies or energy. The analyzed sets of sentences were composed of utterances expressing various speaker's attitude. We have tried to reveal some acoustically measurable differences among various speaker's attitudes so that we could incorporate this knowledge into our speech synthesis system [8] to obtain emotional synthetic speech.

1 Introduction

Without question, contemporary speech synthesis techniques produce high quality and intelligible speech. However, the synthetic speech cannot sound completely natural until it expresses a speaker's attitude. Thus, emotional speech synthesis is a frequently discussed topic and has become a concern of many scientists. In spite of the fact that some results have already been presented, this issue has not been satisfactorily solved yet. Some papers which deal with this problem include, but are not limited to [1] [2] [3] [4] [5] [9] [11] [13].

To incorporate some expressivity into the synthetic speech, we firstly need to find out which factors are important for listeners to perceive spoken speech as expressive speech. In the first phase of this research, we have focused on acoustic characteristics of the expressive speech. The results of our analysis cannot be generalised as we have analysed sentences uttered by single speaker and due to this reason they are not statistically representative. Nevertheless, we can utilize the revealed acoustic characteristics for incorporation of emotions into our speech synthesis system.

This paper is organised as follows. Section 2 deals with the description of the data used in the analysis. In section 3 the acoustic analysis as such is described. In this section we list the features that were measured on the data and the techniques which were used for their acquisition. Section 4 is dedicated to an overview of the attained results. Some conclusions and future work are also presented in this section.

2 Speech material used in analysis

Just for experimental purposes, we have recorded a database of utterances containing various speaker's attitudes. The speech data were uttered in an anechoic room by a semi-professional female speaker with some radio-broadcasting experience. Before the recording of emotional sentences, the speaker was instructed to try her best to portray the emotions.

The database is composed of four sets of sentences uttered in neutral speaking style - 100 wh questions (referred to as *whQuest*), 97 yes-no questions (*ynQuest*), 91 imperative sentences (*imperSen*) and 100 indicative sentences (*indicSen*). We consider these four sets of sentences to be emotionally neutral and we have used them as referential ones in our analysis.

In addition, the database contains six sets of sentences in which two emotions are expressed by the speaker - happiness and sadness. These two contrasting emotions have been chosen because they are supposed to be well distinguishable, according to [4], [10] and [12]. Another reason for the selection of these two emotions is that the emotional speech synthesis is a very complex task and our short-term plan is to enable our synthesis system to use sad, neural and happy speaking style. For each emotion, three sets of utterances were analysed.

The first pair of sets (*happyHC* / *sadSC*) contains sentences with emotionally dependent content corresponding with the particular emotion, in each of these sets there are 100 sentences. The second pair of sets (*happySel* / *sadSel*) is a selection from the first one. These sets contain the amount of 30 and 20 items, respectively. The selection was made by a few listeners, whose task was to mark sentences which seemed to them to correspond perfectly with the given emotion. The last pair of sets (*happyNC* / *sadNC*) is similar to the first one but the content is emotionally neutral and identical for both emotions. Again, both of these sets contain 100 sentences. The speaker was instructed to utter the same sentences using both happy and sad speaking style.

This division of emotionally uttered sentences has been intended to show whether the content of a sentence affects the speaker when portraying the emotion. Further, a comparison between two given emotions was required. We have decided to compare the sets of sentences with neutral content for exclusion of an influence of the content. The reason for making a selection of some emotional sentences by listeners was to find out whether these perceptively slightly different sentences differ also from other emotional utterances in terms of their acoustic characteristics.

3 Acoustic analysis

In the following subsections, there are presented the results of the acoustic analysis of expressive speech. Each subsection corresponds with one feature which is supposed to influence the perception of speech significantly. These are the fundamental frequency F0 (in sec-

[1] University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Cybernetics, Czech Republic, email: gruber@kky.zcu.cz
[2] University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Cybernetics, Czech Republic, email: legatm@kky.zcu.cz

tion 3.1), duration of phonemes (in section 3.2), values of the formant frequencies F1, F2 and F3 (in section 3.3) and values of the RMS energy (in section 3.4).

The part of the database containing emotional utterances was recorded at another time and with slightly different settings of recording equipment (assembling/disassembling of the recording devices because of sharing the anechoic room with other projects) than the part containing questions, indicative and imperative sentences. This is due to the fact that emotional utterances were recorded only for experimental reasons whereas the other sentences were selected from the huge speech corpus which was recorded in neutral speaking style and which is currently used by our speech synthesizer. Unfortunately, the different settings of recording equipment resulted in a slight difference in the intensity level of these two sets of utterances. Because of this fact, we have not performed the analysis of RMS energies of these two groups of sentences.

3.1 F0 analysis

To determine F0 contours, we took advantage of having corresponding glottal signals recorded along with speech signals. We have used Robust Multi-Phase Pitch-Mark Detection Algorithm [7] for marking of pitch pulses in speech and derived the F0 contour from this sequence. First, we obtained local F0 estimates calculated as median of inverse values of distances between four consecutive pitch marks. Then, the sequence of these local F0 estimates was smoothed by median filter of order 3 (see Fig.1).

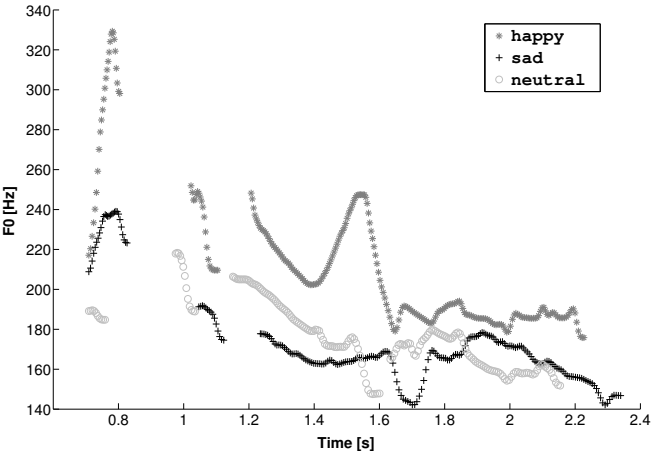


Figure 1. F0 contour of neutral sentence and sentences expressing emotions (selected from *sadNC* and *happyNC* sets).

Table 1. Mean values and standard deviations of the F0.

set of sentences	mean value [Hz]	standard deviation [Hz]
sadSC	184.82	28.55
sadSel	181.27	28.31
sadNC	**181.32**	**29.01**
happyHC	202.40	44.73
happySel	209.57	49.34
happyNC	**203.62**	**46.28**
indicSen	**193.76**	**36.63**
whQuest	188.96	43.67
ynQuest	197.72	32.94
imperSen	198.78	39.15

The results summarized in Tab. 1 show that all the sentences representing happiness have higher F0 than the sentences representing sadness. The F0 mean value of the neutral utterances is in the middle of the values for two emotional sets.

A major difference between *happySel* and the other sets for happiness could be also noticed. It could suggest, that the listeners' selection may express the given emotion more than the sets containing all sentences. However, the same conclusion cannot be drawn for the sentences representing sadness.

Some differences were also found among the sets expressing various speaker's attitude, i.e. *indicSen*, *whQuest* and *ynQuest*. Note that these sets of sentences were all uttered in a neutral speaking style.

In Fig. 1, there is shown the F0 contour for a neutral sentence and for sentences representing sad and happy emotion. All three sentences have the same content - *"A připíjí vínem"* [a pQ\ipi:ji: vi:nem] - according to the Czech version of SAMPA phonetic alphabet. The difference in mean values and variances of F0 is visible as is the different duration of the whole sentence, as described in 3.2.

3.2 Duration analysis

For the determination of durations of phonemes, an automatic segmentation technique using HTK Tools improved by a statistic approach [6] was utilized. To calculate the duration of a phoneme, the time of its end and its beginning were simply subtracted. The results of this analysis are shown in Tab. 2.

Table 2. Mean values and standard deviations of the phonemes duration.

set of sentences	mean value [ms]	standard deviation [ms]
sadSC	96.5	60.3
sadSel	98.7	63.6
sadNC	**97.6**	**62.0**
happyHC	91.6	48.6
happySel	92.7	50.8
happyNC	**85.5**	**43.2**
indicSen	**84.2**	**47.1**
whQuest	77.0	47.7
ynQuest	79.4	44.9
imperSen	81.3	44.6

The average duration of the phonemes appearing in neutral sentences is 84.2 ms. For sadness, the duration is longer, for *sadSC* set the mean value is 96.5 ms (by 15% above the neutral set). The differences among three sets with sentences expressing sadness are not statistically significant, according to the performed t-test.

For happiness, quite surprising results were obtained. The average duration for *happyHC* and *happySel* was about 92 ms, that is longer than the average duration in neutral sentences (by 9%), but the mean value in the *happyNC* set is almost at the same level as the mean value in the neutral one, no statistically significant difference was detected between these two sets. It suggests that the speaker was not able to portray the happy emotion in the sentences with neutral content, in terms of phone duration.

As mentioned above, the listeners' selections have almost the same average phonemes duration as the sets containing all emotional sentences with emotionally dependent content. The average phone durations for sentences expressing various speaker's attitudes were also different. The mean value for *whQuest* set was the lowest across all the sets.

3.3 Formant analysis

To obtain formant frequency estimates, we used Speech Filing System[3]. We employed the `formanal` program which is referred to as currently the best one in SFS to perform fixed-frame formant analysis. This program was originally implemented in the Entropic Signal Processing System and it is used under licence from Microsoft. The formant estimates which are presented in Tab. 3 come from the middle parts of vowels which were found by cutting the initial and the final quarter of the vowel length.

Table 3. Mean values and standard deviations of the formant frequencies for Czech short vowels.

a	**F1 [Hz]**		**F2 [Hz]**		**F3 [Hz]**	
sets	mean	std	mean	std	mean	std
indicSen	640	124	1377	223	2623	380
happyNC	695	126	1405	270	2657	441
sadNC	649	140	1320	235	2598	375
e	**F1**		**F2**		**F3**	
sets	mean	std	mean	std	mean	std
indicSen	487	96	1859	352	2697	224
happyNC	513	116	1949	313	2754	209
sadNC	492	137	1803	350	2634	250
i	**F1**		**F2**		**F3**	
sets	mean	std	mean	std	mean	std
indicSen	387	61	2054	445	2739	167
happyNC	371	59	2180	355	2782	226
sadNC	375	77	1975	381	2657	214
o	**F1**		**F2**		**F3**	
sets	mean	std	mean	std	mean	std
indicSen	444	57	1050	197	2719	222
happyNC	443	74	1026	211	2741	387
sadNC	421	74	1063	183	2683	275
u	**F1**		**F2**		**F3**	
sets	mean	std	mean	std	mean	std
indicSen	377	109	985	267	2698	184
happyNC	344	76	933	257	2634	391
sadNC	339	71	969	280	2544	396

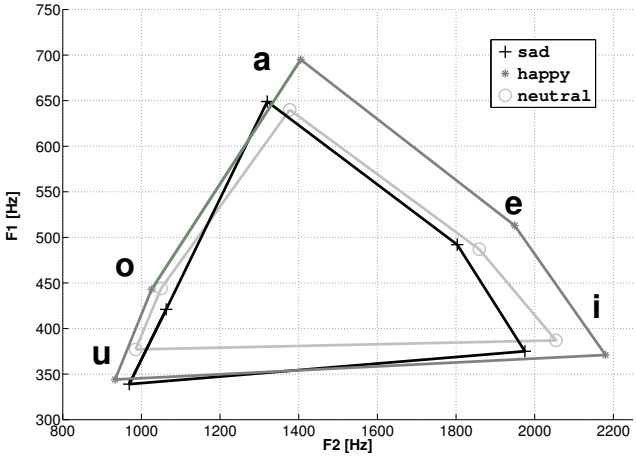


Figure 2. F1-F2 diagram for 5 Czech short vowels. A comparison of vocalic triangles for sentences uttered in neutral speaking style and happy and sad emotion (values measured for *happyNC* and *sadNC* sets).

[3] Speech Filing System – http://www.phon.ucl.ac.uk/resource/sfs

Regarding the analysis of formant frequencies, various differences were detected across all the sets in terms of various vowels. Results for *happyNC*, *sadNC* and *indicSen* (as a referential set) are shown in Tab. 3. Since tendencies of formant frequencies shifts are not clear from this table, the vocalic triangle is depicted in Fig. 2. It represents the distribution of Czech short vowels in the F1-F2 space for various speaker's attitudes. Unfortunately, the influence of different emotions on F3 is not visible from this figure.

It seems that the vocalic triangle is expanded for values measured in sentences portraying happiness, it means that low formant frequencies are lowered and high ones are increased, in comparison with neutral speech. This phenomenon applies to both F1 and F2. The formant frequencies obtained for sentences conveying sadness cause a counter-clockwise rotation of the vocalic triangle. Again, these results apply only to our speaker and further analysis of more speakers would be necessary to generalise this phenomenon.

There were detected also differences among the mean values for various sets representing the same emotion. The differences were not tested by any test in order to find out whether they are statistically significant or not, but it could be said that the emotions are well represented by sets *happyNC* and *sadNC*. The differences between particular emotions were greater than the differences among the sets representing the same emotion.

Regarding the neutral sets of our database, no considerable differences were found.

3.4 RMS analysis

RMS[4] energy is a value that characterizes the intensity of a speech signal. Using this feature, the differences of intensity level in different sets of sentences can be measured. For this analysis, we had to divide our speech material into two parts and analyze them separately. One group contains the emotional sentences and the other one contains the neutral sentences. This separation was necessary due to slightly different settings of technical equipment for recording, as explained in the second paragraph of the section 3.

For the calculation of the RMS energy (1) of a sentence, initial and final pauses were cut off.

$$RMS = \sqrt{\frac{\sum_{i=1}^{n} s(i)^2}{n}}, \tag{1}$$

where $s(i)$ is $i-th$ sample of the signal and n is the length of the signal.

The results obtained for the emotional part of the corpus are shown in Tab. 4. It is obvious that there is a difference between given emotions. The RMS energy of the sentences portraying happiness is higher than for the sentences portraying sadness. It means that the happy sentences are spoken louder. The difference is statistically significant which was proved by t-test. On the other hand, the differences between sets representing the same emotion are not statistically significant, except the sets *happyNC* and *happyHC*. In this case, the p-value reached the value 0.0407, which means that these two sets can be regarded as equal in terms of the mean value of the RMS energy considering lower significance level, e.g. $\alpha = 0.01$.

The results for the neutral part of our database are presented in Tab. 5. Comparing *indicSen* vs. *whQuest* and *ynQuest* vs. *imperSen*,

[4] RMS = Root Mean Square, also known as the quadratic mean; a statistical measure of the magnitude of a varying quantity. It is especially useful when variates are positive and negative, e.g. waves.

Table 4. Mean values and standard deviations of the RMS energy (signal range $\langle -1,\ 1 \rangle$).

set of sentences	mean value	standard deviation
sadSC	0.0232	0.0039
sadSel	0.0232	0.0031
sadNC	**0.0224**	**0.0038**
happyHC	0.0307	0.0050
happySel	0.0299	0.0055
happyNC	**0.0294**	**0.0039**

Table 5. Mean values and standard deviations of the RMS energy (signal range $\langle -1,\ 1 \rangle$).

set of sentences	mean value	standard deviation
indicSen	0.1333	0.0179
whQuest	0.1282	0.0217
ynQuest	0.1480	0.0393
imperSen	0.1518	0.0373

no statistically significant differences can be observed, the other cases seem to be different.

4 Conclusions & future work

In this study, we have compared and contrasted some emotional and neutral utterances in terms of F0, phoneme duration, formant frequencies and RMS energies. Some results are briefly summed up in Tab. 6, where the analysed emotional speech is compared with the referential one in terms of F0, duration and RMS. Initially, this paper was intended to cover single speaker analysis of both emotional sentences and sentences expressing various speaker's attitude in spite of being uttered in neutral speaking style. However, we are currently more concerned with the synthesis of emotions which is why we decided to prefer analysis of emotional utterances to obtain some results useful for speech synthesis.

Table 6. Brief overview of the acoustic analysis.

set of sentences	**F0**		**duration**		**RMS**
indicSen	194	•	84.2	•	—
happyNC	204	⇑ 5%	85.5	⇑ 2%	0.0294
sadNC	181	⇓ 7%	97.6	⇑ 16%	0.0224

The discussion of the results of the formant analysis seems to be too complex and it is out of scope of this paper. Moreover, at the present time, incorporation of the results into our speech synthesis system would require more modifications of the current approach in comparison with incorporation of F0, duration and RMS energy results. However, in Fig. 2 there is depicted an influence of emotion being present in the spoken speech on the formant frequencies.

The results reached confirmed that all the features measured on the speech signal are important acoustic correlates of various speaker's attitudes. Nevertheless, it cannot be concluded that these features are sufficient for the distinction of all emotions from speech signal. In the future a similar analysis should be performed on more extensive database containing more emotions, e.g. anger, boredom and contentment.

The results found in this analysis could be confirmed by classification task using F0, duration, RMS energy and formant frequencies as predictors for determining emotion from speech signal. The reference data would be obtained by means of more complex listening tests. In the case that any classification model were able to give good results using these predictors, we could conclude that it would be sufficient to modify these characteristics of neutral speech to obtain emotional output.

Our future work will be focused on the incorporation of the obtained results into our speech synthesis system. It includes the modelling of prosodic features based on emotionally recorded data for single instance concatenation approach and the extension of a feature set for the unit selection approach.

ACKNOWLEDGEMENTS

This work was funded by the Ministry of Education of the Czech Republic, project No. 2C06020, and in part by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

The access to the METACentrum computing facilities provided under the research intent MSM6383917201 is highly appreciated.

REFERENCES

[1] A.W. Black, 'Unit selection and emotional speech', *Proc. of Eurospeech 2003*, 1649–1652, (2003).

[2] M. Bulut, S.S. Narayanan, and A.K. Syrdal, 'Expressive speech synthesis using a concatenative synthesiser', *Proc. of the 7th International Conference on Spoken Language Processing*, 1265–1268, (2002).

[3] W. Hamza, R. Bakis, E.M. Eide, M.A. Picheny, and J.F. Pitrelli, 'The ibm expressive speech synthesis system', *Proc. of the 8th International Conference on Spoken Language Processing*, 2577–2580, (2004).

[4] G. Hofer, K. Richmond, and R. Clark, 'Informed blending of databases for emotional speech synthesis', *Proc. of Interspeech 2005*, 501–504, (2005).

[5] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, 'A corpus-based speech synthesis system with emotion', *Speech Communication*, **40 n. 1-2**, 161–187, (2003).

[6] Matoušek J., Tihelka D., and Psutka J., 'Automatic segmentation for czech concatenative speech synthesis using statistical approach with boundary-specific correction', *Proc. of Eurospeech 2003*, 301–304, (2003).

[7] M. Legát, J. Matoušek, and D. Tihelka, 'A robust multi-phase pitch-mark detection algorithm', *Proc. of Interspeech 2007*, 1641–1644, (2007).

[8] J. Matoušek, J. Romportl, D. Tihelka, and Z. Tychtl, 'Recent improvements on artic: Czech text-to-speech system', *Proc. of Interspeech 2004 - ICSLP, 8th International Conference on Spoken Language Processing*, **III**, 1933–1936, (2004).

[9] J.M. Montero, J. Gutiérrez-Ariola, S. Palazuelos, E. Enríquez, S. Aguilera, and J.M. Pardo, 'Emotional speech synthesis: From speech database to tts', *Proc. of the 5th International Conference of Spoken Language Processing*, 923–926, (1998).

[10] J.A. Russell, 'A circumplex model of affect', *Journal of Personality and Social Psychology*, **39**, 1161–1178, (1980).

[11] M. Schroder, 'Emotional speech synthesis: A review', *Proc. of Eurospeech 2001*, 561–564, (2001).

[12] C.M. Whissell, *The Dictionary of Affect in Language*, 113–131, Robert Plutchik and Henry Kellerman (Ed.), Emotion: Theory, Research, and Experience, Academic Press, New York, 1989.

[13] E. Zovato, A. Pacchiotti, S. Quazza, and S. Sandri, 'Towards emotional speech synthesis: A rule based approach', *Proc. 5th ISCA Speech Synthesis Workshop*, 219–220, (2004).

Interplay between pragmatic and acoustic level to embody expressive cues in a Text to Speech system

Enrico Zovato[1], Francesca Tini-Brunozzi[2] and Morena Danieli[1]

[1]LOQUENDO S.p.A. – [2] EIKON INFORMATICA

Torino, Italy

Abstract. This paper[1] deals with the problem of generating emotional speech within the Unit Selection approach to text to speech synthesis. By taking into account state-of-the-art research in different fields, from psychology to linguistics, we claim that a complex interplay between the phonetic level and the pragmatic level of language constitutes the basis of voice expression of emotions, and that the phonetic-pragmatics interplay can be accounted for in a text-to-speech system by providing accurate representations of contextually relevant discourse markers. The availability of an inventory of expressive cues implementing discourse markers, can improve the naturalness and expressivity of generated speech, moving toward the ambitious goal of emotional speech generation.

1 INTRODUCTION

In the last few years the demand of more expressive and natural human-computer interfaces has increased along with the range of applications of computer mediated communication. In particular, when natural language is the medium of the interaction between humans and computer, the complexity of the linguistic interface requires that the problem of naturalness and emotional attitude be approached at several different levels. For example, the linguistic competence of the interface should be sophisticated enough to capture the emotional state of the human conversational partner, and it should express context relevant contents in a natural-sounding, expressive, and emotional believable voice.

The fact that rhythm and intonation of the human voice are elective *loci* where emotional experiences can be represented, is considered almost an uncontroversial datum by scholars working on theoretical and experimental models of emotions within different fields, including linguistics (see [3], [7], [8], and [20]), psychology (see [9], [12] and [21]), automatic speech processing (see [14]), neurosciences and psychoanalysis (see [18]). Of course, despite of that *consensum*, the study of how emotional states can affect voice characteristics has been approached in very different ways within those very different research areas. In the past few years, experimental psychologists focused on some individual aspects of emotional speech: In particular, their works were based on the partition between basic and complex emotions, aiming to discover their distinctive features in the acoustic signal. Most of those studies hypothesized the psychological and neurophysiological onset of a set of basic emotions, independently from other relevant aspects of the emotional experiences. For example, a huge amount of research has been devoted to the identification of the prototypical *intonation profiles* (F_0) associated with "anger", "joy", "pain", "depression", and so on (see [12], and [19] for a review). Cognitive psychologists claim that affects are the prime movers of human motivation: In this view *affects* are neurophysiologically generated, sensationlike signals. The activation of the neurobiologic affect generators triggers tone of voice as a motor stereotype, and this in turn transmits the physical data underlying empathy and emotional communication.

Despite of the tremendous advancements in the field of emotional studies, state-of-the-art research on acoustic and prosodic properties of voice expression of emotions does not provide yet an in-depth understanding of how the different features of the emotional experience show themselves in speech in terms of acoustic correlates and lexical choices. However, converging evidence from neuroscience, psychology, and linguistics can provide a promising framework for investigation. In particular, linguists take seriously into consideration the fact that speaking is a motory activity occurring on an affective basis. They showed that the analysis of speech segments can hardly show pure, prototypical intonation profiles of single emotional states if they are kept apart from the speaker's conscious or unconscious intentions.

In our work we exploited the intention-based linguistic analysis proposed by Cresti ([7] and [8]). Cresti's work is based on the

[1] This work has been partially supported by the EU's 6th framework project "COMPANIONS", contract IST 034434.

analysis of the intonation profiles of utterances from a large corpus of Italian spontaneous speech, collected in several real situations of language uses. Her results show that acoustical characteristics of emotions are influenced not only by the internal state of the speakers, but also by the external *context of occurrence* of their utterances, where the *context* includes the complex interplay between subjects' affects, and the network of interpersonal relationships where emotional experiences occur.
In our view some of the difficulties in implementing emotional behaviour in artificial agents, often reported by artificial intelligence and voice scientists, reflect the importance of taking into account the way in which context affect different parameters such as lexical choices, occurrence of extra-linguistic phenomena, and the onset of voice parameters.

In this paper, we will show that useful insights in the desired direction can come from combining linguistic, acoustic and pragmatic evidences. In particular, for improving the expressiveness of the Loquendo text-to-speech system we created an inventory of discourse markers and speech acts that constitute contextually relevant part-of-speech that can be combined with neutral speech in order to generate emotionally connoted speech.

The plan of the paper is as follows: Paragraph 2 sets the problem from the point of view of text-to-speech technology, paragraph 3 proposes the linguistic analysis underlying the selection of discourse markers and speech acts, and paragraph 4 offers details related with implementation issues.

2 EXPRESSIVE SPEECH SYNTHESIS

Speech synthesis systems, also called Text to Speech systems exploit different technologies providing very different degrees of quality and naturalness. The most effective systems are the so called *corpus based synthesizers*. They are based on the concatenation of variable length speech units, selected from a large database containing speech data from a single speaker. The core technology of these systems is the search algorithm that, given the input text, has to detect the best fitting units in the database depending on the phonetic and prosodic representation of the same input.

The naturalness and intelligibility of these systems is mainly due to the fact that "exact" replicas of human speech are concatenated, avoiding any kind of signal processing that could introduce artefacts. The longer the average length of the selected units is, the better is the naturalness and acoustic quality, since fewer concatenations are necessary (every concatenation is a sort of discontinuity point). Consequently, for a given language, the goal is to design a database providing an adequate statistical coverage of the most frequent phonetic contexts. Several hours of recording sessions are therefore necessary to collect the audio data and an important point is that talents have to maintain their reading style uniform throughout the various sessions. Generally, this is a neutral "standard" reading style, i.e. no expressive attitude has to be adopted as well as no emphasis has to be introduced in the read sentences.

In this way corpus based synthesis systems, despite their intelligibility and quality, are extremely static in terms of expressive capabilities, since only one, and generally "flat" style is adopted. Adding expressivity to synthetic speech is a matter of research and investigation whose results have led to two main approaches, even if not completely satisfactory.

The first approach is based on the acquisition of speech data not only providing good phonetic coverage, but also providing a sort of expressive coverage. This is obtained by adopting different expressive styles beyond the neutral one when speech data is recorded [10]. Of course, only a limited number of styles is affordable and a preliminary choice has to be done also depending on the domain of the application. This solution is particularly effective in specific contexts, since the output quality is comparable to the one of the neutral database, but it is not flexible.

The second approach is based on the signal manipulation of the output signal obtained through the selection of speech units from a mono-stylistic (neutral) database. This kind of operation mainly aims at modifying the prosody and voice quality of the concatenated speech. In practice, the intonation path, the speech rate, intensity and spectral shape are jointly manipulated according to models that indicate how these parameters change depending on the context and the target expressive style [13,15,25]. In order to get effective models, statistical analysis of significant amounts of annotated data is necessary. The critical aspect of this approach is that, despite its flexibility, the algorithms exploited to impose the target contours often introduce distortions and compromise the naturalness of the original waveforms [23].

The paradigm here proposed is different from the two approaches previously described and less ambitious in terms of general purposes expressive synthesis. The key idea is in fact to start from a linguistic point of view that, considering the most common application domains, takes into account expressive cues that have a pragmatic function, like for example greetings, apologies, recalls, etc. These prompts are recorded and inserted into the voice database and used only in certain contexts, providing expressivity to the synthesised speech.

3 PRAGMATIC FEATURES OF THE EXPRESSIVE CUES

As discussed in the previous paragraph the goal of reaching naturalness and emotional expressivity of the synthesized speech can hardly be met by modifying only the acoustic and spectral features of the speech signal, given current state of the art of speech synthesis technologies. However, the investigation of linguistic phenomena lying at the interface between phonology and pragmatics, has showed helpful for selecting the lexical structures carrying out expressive and emotional contents.

Our goal was the creation of a rich acoustic inventory of expressive cues [5], in order to be able to integrate them in the synthesized message [11], without impairing the naturalness and the fluidity of the speech provided by the *Unit Selection* technique.

The expressive cue inventory is language specific. It includes phrases classified into different categories of speech acts [8,17] and discourse markers [3,22], and some extra-linguistic elements such as interjections, onomatopoeia and human sounds. It is worth noticing that the acoustic-prosodic and lexical structures of these phrases contribute to increase the pragmatic values of the sentences that include them. This is particularly important in a range of applications, such as human-machine interaction, e-learning, human-human computer-mediated communication, among others.

Underlying this approach is the hypothesis that the expression of emotions in human voice, can seldom, if any, be separated from the linguistic contexts [1,17] where the speech acts occur. In its turn the context affects a number of parameters of the speech act, including acoustic modifications, lexical choices, and extra-linguistic voice mediated phenomena, such as deep sigh, winks, back-channelling sounds, and so on [2,16].

Recent research in pragmatic linguistics has pointed out that the structure of human speech is modulated pragmatically [4] thanks to the competent use of discourse markers by the speakers. Bazzanella [3] offers an in-depth analysis of discourse markers, showing their potential inter-relational function. In particular, this scholar classifies discourse markers on the basis of two points of view, both of them necessary for the success of the conversation, that is the point of view of the speaker and the point of view of the co-conversant. From both point of views discourse markers are linguistics expressions that derive their meaning and pragmatic values mainly from the utterance context. For example, from the speaker's point of view, the author proposes a large set of inter-relational functions such as:

1. taking and leaving turn (i.e., *well*, *but,* …)
2. fillers (i.e. *you know*, *see,* …)
3. requests of attention and agreement (c*an you understand this? do you agree? don't you, …?*)
4. phatisms (*in my view*, …)
5. request of agreement (*do you agree?...*)

From the point of view of the co-conversant, she identifies the following functions, among others:

1. interruption (*but*, *yes but*, …)
2. back-channels (*aha, mhm*, *I see*, ...)
3. confirmation of attention (*sure*, *OK*)
4. phatisms (*you are welcome)*
5. reinforcement (*true*, *of course*, …)

On the basis of this analysis we have identified a set of expressive cues also reported in Table 1 [24].

SPEECH ACT	EXAMPLE
Refuse	*Absolutely not! …*
Approval	*Exact! …*
Disapproval	*Absurd! …*
Recall	*Let's keep in touch! …*
Announce	*Here I am! …*
Request of Confirmation	*Isn't it? …*
Request of Information	*Why? …*
Request of Action	*Help! …*
Prohibition	*This is forbidden! …*
Contrast	*I don't think so! …*
Disbelief	*That's unbelievable! …*
Surprise	*What a surprise! …*
Regret	*I'm so sorry! …*
Thanks	*Thanks a lot!*
Greetings	*Welcome! …*
Apologies	*I'm sorry! …*
Compliments	*Congratulations! …*

Table 1. Speech acts categories with examples

The items and phrases we selected are representative of speech acts that reflect the speaker's attitude with respect to her/his conversational partner in different contexts of uses. For doing this, the linguistic analyses have been done on a corpus basis.

Also the communicative potential of human sounds is relevant. For example, a throat could communicate embarrassment, distancing, request of attention, or it could play the role of *back-channelling*. The inventory also includes human sounds as throats, bitter and hearty laughs, sobbings.

On the basis of this inventory, we could implement the acoustic counterpart of a limited, but rich, set of speech acts, including: refuse, approval/disapproval, recall in proximity, announce, request of information, request of confirmation, request of action/behaviour, prohibition, contrast, disbelief, surprise/astonishment, regret, thanks, greetings, apologies, and compliments.

4 IMPLEMENTATION OF THE EXPRESSIVE FEATURES

The design of the speech acts corpus, as previously explained, is based on linguistic rather than phonetic criteria. There is a substantial difference in the way these data are recorded with respect to the acquisition of the baseline voice data, where emphasis and too marked intonation movements are avoided. In fact, in this case, we asked our speaker to adopt the more suitable voice registry according to the semantic and pragmatic function of the scripts. Nevertheless, during the recording sessions, talents had to be accurately directed, particularly concerning the level of activation. This hasn't to be too high because the stylistic difference with the base synthetic voice would be too marked and consequently judged as unnatural. The main goal is adding expressivity without compromising the continuity of prosodic patterns. As concerns the acquisition of the speech data, for each linguistic category a set of samples is recorded. Some sets are bigger than others depending on the variety of the speech acts and on their frequency in the considered spoken language. Generally, for each voice, the database is composed of about 500 expressive utterances. The speaker is free to interpret the scripts and adopt the suitable attitude while the director only controls his/her level of activation and the pronunciation accuracy. At the end of the acquisition the best samples are selected in terms of acoustic quality, effectiveness and reliability. These data are then normalised to better match the acoustic characteristics of the base voice data and the same coding is also applied. The expressive speech data corresponding to the illocutionary acts is also automatically labelled like the neutral speech data. In fact phonetic and prosodic labels are assigned to each elementary unit (phoneme). One more label identifies the stylistic class of the utterance which the unit was extracted from. These classes are, for example, declarative, interrogative, marked interrogative, exclamation, etc.

In the selection phase the TTS avoids mixing units belonging to different categories and, in particular, will choose the expressive utterances only when an exact matching with the phonetic counterpart of the graphemic input string and the target stylistic class occur. Regarding the latter this is simply obtained through the analysis of the final punctuation of the sentence. On the contrary, if the marked input text has no correspondence in

the expressive set of data, then the concatenation of neutral speech segments is exploited.

Beyond the expressive utterances, the paralinguistic events are also recorded (laughs, coughs, hesitations, etc.). Of course, these data are not analysed at segmental and phonetic level. Only one identification label is assigned to each of them as a whole. The "synthesis" of these events is realised by inserting in the input text these labels preceded by a special control tag.

In order to make the expressive features effective, they have to be a priori known by the user. To this end we have developed a client application suitable for producing vocal prompts. One important feature of this application is the possibility to show all the available linguistic and paralinguistic events, having them classified according to the previously described categories. In this way the user can easily choose and insert the expressive cues into the synthesised speech, obtaining a more colourful synthetic speech, in terms of intonation movements and voice quality.

5 CONCLUSIONS

In this paper we approached the problem of generating emotional speech within the Unit Selection approach to text to speech synthesis by taking advantage of state-of-the-art research in different fields, from psychology to linguistics. A common evidence of research in those areas concerns the relevance of the utterance contexts, and the identification of different levels from where each human speech act is marked in terms of intentionality, expressivity, content, and emotion expressions. These results are in some sense disruptive for the traditional organization of levels of linguistic models. Actually, in the near past it was unusual to hypothesize interfaces between acoustic models and pragmatic modules when describing the implementation of computational models of language analysis and generation. On the contrary, we claim that the complex interplay between the acoustic level and the pragmatic level of language constitutes an important aspect of voice expression of emotions. We also claim that the phonetic-pragmatics interplay can be accounted for in a text-to-speech system by providing accurate representations of contextually relevant discourse markers. The availability of an inventory of expressive cues implementing discourse markers, can improve the naturalness and expressivity of generated speech, moving toward the ambitious goal of emotional speech generation.

REFERENCES

[1] Akman V., Bazzanella C. (eds.) 2003 *Context*, special issue 35, *Journal of Pragmatics*, 321-504.

[2] Bazzanella C. 2004 Emotions, Language and Context,. In Weigand E. (ed.) 2004 *Emotion in dialogic interaction. Advances in the complex*. Amsterdam/Philadelphia, Benjamins., 59-76.

[3] Bazzanella, C. 2006 Discourse Markers in Italian: towards a 'compositional' meaning. In Fischer K. (ed.) 2006. *Approaches to discourse particles,* Amsterdam, Elsevier, 504-524.

[4] Brown P., Levinson S. 1978, 1987: *Universals in language usage: Politeness phenomena,* in E. N. Goody (ed.): *Questions and Politeness. Strategies in Social Interaction*. Cambridge, Cambridge University Press., 56-248; ed. 1987: *Politeness*. Cambridge UP, Cambridge.

[5] Campbell, N. (2002), Towards a grammar of spoken language: incorporating paralinguistic information. In: 7th ISCA International Conference on Spoken Language Processing, Denver, Colorado, USA, September 16-20, 2002.

[6] Cresti, E. (2000), *Corpus di Italiano Parlato*. Volume I: Introduzione. Firenze, Accademia della Crusca.

[7] Cresti, E. (2003), L'intonation des illocutions naturelles représentatives: analyse et validation perceptive, (http://lablita.dit.unifi.it/publications/)

[8] Cresti, E. (2005), Per una nuova classificazione dell'illocuzione a partire da un corpus di parlato (LABLITA). In: Burr E. (Ed.), Atti del VI Convegno internazionale SILFI (giugno 2000, Duisburg), Cesati, Pisa.

[9] Danieli, M., Emotional speech and emotional experience. *VIII International Conference of Neuro-Psychoanalysis (N-PSA): Neuro-Psychoanalytical Perspectives on Depression,* July 18-22, 2007, Wien.

[10] A. Iida Akemi, N. Campbell, F. Higuchi & M. Yasumura, A corpus-based based speech synthesis system with emotion. In: Speech Communication, Vol. 40, 2003: 161-187.

[11] Hamza W., Bakis, R., Eide, E.M., Picheny, M. A., & Pitrelli, J. F. (2004), The IBM Expressive Speech Synthesis System. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP), Jeju, South Korea, October, 2004.

[12] Johnstone, T. & Scherer, K.R., (1999). The effects of emotions on voice quality. In *Proceedings of XIV Int. Congress on Phonetic Science.*

[13] Montero, J.M., Gutiérrez-Arriola, J.M., Palazuelos, S., Enríquez, E., Aguilera, S. & Pardo, J.M., Emotional Speech Synthesis: from Speech Database to TTS, In Proceedings of ICSLP 1998, Sidney.

[14] Moore, E.I.I. Clements, M. Peifer, J *et al.*. Comparing objective feature statistics of speech for classifying clinical depression. In *Engineering in Medicine and Biology Society*, p17-20.

[15] Murray, J.R. & Arnott, J.L., Synthesising emotions in speech: is it time to get excited?, In Proceedings of ICSLP 96, Philadelphia, pp.1816-1819.

[16] Ochs E., Schieffelin B. 1989: *Language has a heart*, in E. Ochs (ed.): *The Pragmatics of Affect*, numero speciale di *Text* 9.1, 7-25.

[17] Ochs E. Schegloff E.A., Thompson S.A. (eds.) 1996 *Grammar and Interaction*. Cambridge, Cambridge University Press.

[18] Pally, R. (2001). A Primary Role for Nonverbal Communication in Psychoanalysis. In *Psychoanalytical Inquiry*, v21 p71-93.

[19] Panksepp, J. (1999), Emotions as Viewed by Psychoanalysis and Neuroscience: An Exercise in Consilience, *Neuro-Psychoanalysis*, 1:15-38

[20] Poggi, I. & Magno Caldognetto, E. (2004). Il parlato emotivo. Aspetti cognitivi, linguistici e fonetici. In F. Albano Leoni, F. Cutugno, M. Pettorino & R. Savy. (Eds.), *Atti del Convegno "Italiano parlato"* (Napoli 14-15 febbraio 2003). Napoli: D'Auria Editore, CD-Rom.

[21] Scherer, K.R. (2003), Vocal communication of emotion: A review of research paradigms, *Speech Communication*,

[22] Schiffrin, D. 1987: *Discourse markers*. Cambridge, Cambridge University Press.

[23] Schröder, M. (2001). Emotional Speech Synthesis: A Review, In Proceedings of EUROSPEECH 2001, pp. 561 – 564, Scandinavia, 2001.

[24] Tini Brunozzi F., Quazza S. & Zovato, E., (to appear), Atti illocutivi e segnali discorsivi. Un contributo linguistico a un sistema TTS verso la sintesi vocale espressiva, atti del XL Congresso Internazionale di Studi della SLI "Linguistica e modelli tecnologici di ricerca", Vercelli 21 - 23 settembre 2006, Roma, Bulzoni.

[25] Zovato, E., Pacchiotti, A., Quazza, S. & Sandri, S., Towards emotional speech synthesis: a rule based approach. In: 5th ISCA Speech Synthesis Workshop, Pittsburgh USA, 2004.