

An emotional humanoid partner

Anzalone S.M.¹, Cinquegrani F.¹, Sorbello R.² and Chella A.²

Abstract. In this paper we propose an emotional humanoid robot based on Latent Semantic Analysis, that exhibits an emotional behaviour in the interaction with human. Latent Semantic Analysis (LSA) paradigm is capable to encode the semantics of words using a statistical computation of a large corpus of text. We illustrate how the creation and the use of this emotional conceptual space allows the building of “Latent Semantic Behaviour” because it simulates the emotional associative capabilities of human beings. The presented approach integrates traditional knowledge representation and intuitive capabilities provided by geometric and sub-symbolic information modelling. To validate the effectiveness of the approach we implemented the system using the humanoid robot from Aldebaran, Nao.

1 INTRODUCTION

The interaction between humans, robots and environment is a relevant topic for many entertainment applications. Classical rule based control models and totally symbolic knowledge representation systems result often too restrictive. To overcome these limitations much research involved hybrid symbolic or sub-symbolic approaches. Many attempts have been made to merge connectionist and symbolic approaches, and, often in robotics, faces the problem of linking a conceptual level to a symbolic one. On the other hand, interacting with humans means the ability to recognize and understand emotions. The interaction of the entertainment robot with the humans and the environment should be not mechanical and not deterministic in order to avoid a predictable behaviour. In order to reach this goal many approaches tried to use ethological or emotional models.

In the last years many studies have been conducted on the knowledge representation in order to provide intelligent interaction with the environment. Many of these models try to incorporate ethological or emotional models to be used in entertainment robots based on behavioural system approach.

From our point of view a robot is made for entertainment use when essentially is an emphatic robot.

The Latent Semantic Analysis (LSA) is a paradigm aimed on the extracting and representing the meaning of words by statistical computations applied to a large corpus of texts usually taken through specific queries from internet. LSA is a method based on vector spaces: words are coded in a large semantic space and the semantic similarity between them can be computed

using a geometric distance between their representative vectors. The interesting feature of LSA is the capability to simulate several human cognitive phenomena as well as word categorization, sentence-word semantic priming, discourse understanding, judgments of essay quality. It has been shown that it is possible to label the basis vectors of the semantic space generated by the LSA paradigm, in order to see this space as a data-driven conceptual space where words documents and concepts, represented using a verbal description, can be mapped. Using this space it is possible to simulate some human psycholinguistic cognitive phenomena that involve association capabilities and semantic similarity measurements.

Objects in the real world, as words in sentences, are connected between them by complex relations: these can be partially captured using complex data structures and functions. A more simple way could be to let automatically emerge these latent associations from the data by the creation of an high dimensional vector space where objects can be mapped and relations among them will emerge as geometric distances in this space. Using this particular interpretation of Latent Semantic Analysis paradigm it is possible to design humanoid robots equipped with a conceptual space capable to modify the behaviours coherently with the information that the robot perceives.

In this paper is presented the use of an LSA-based conceptual space to build the emotional behaviours of an humanoid robot because the space simulates the emotional-associative capabilities of human beings. A subset of words and sentences, coded in the data driven conceptual space, is linked to specific behavioural categories.

The humanoid robot will interact with the environment by coding stimuli using an hidden semantic layer. In this layer the LSA paradigm will be used to represent the logical relations that classical rule based paradigms fail to represent.

The aim of this project is to create an emotional robotic partner for entertainment purposes. This robot should be capable to perceive emotions that emerge from environmental situation or from people behaviours and will be able, according to an empathy mechanism, to activate behaviours that are related to the inferred emotional state.

As an example of this system, let's consider a scenario in which a man talking to his robotic partner says “I love to play soccer”. In this case, the humanoid, using the hidden semantic mechanism, associates the sentence with a particular behaviour, that will consist of searching for a ball and carry it to the human partner. According to the Thagard theory about empathy analogies, during the training phase the robot encodes achievable situations and their potential correlated emotions. This sub-symbolic encoding will allow the building of a map, of a set of analogies capable to bind the current situation with some aspects of the knowledge gained during previous emotional situations.

Studies from Ortony, Clore and Collins show that emotions are descriptors of the people reactions about some events that

¹ Dipartimento di Ingegneria Informatica, DINFO.
Università degli Studi di Palermo, Palermo, Italy.
Email: {anzalone, cinquegrani, sorbello}@dinfo.unipa.it.

² Dipartimento di Ingegneria Informatica, DINFO.
Università degli Studi di Palermo, Palermo, Italy.
Email: {sorbello, chella}@unipa.it.

happens in the environment, so, if the environment can be described using words it is possible to use them to in the emotional space, to infer emotions not only about a verbal communication but also from environmental stimuli. In order to do so, a set of words and sentences related to the environments knowable by the emotional robot is chosen.

The conceptual space used by the humanoid robot will be created using a training set capable to represent emotions. Each time an input stimulus is perceived, it is mapped on the conceptual space. From this mapping will emerge an emotion closely related to the stimulus, then this will activate the most suitable behaviour capable to represent the inferred emotion. In order to implement a non deterministic mechanism, each emotion will have to a set of behaviours coherent with it, so the executed action will be randomly chosen from them.

2 THE EMOTIONAL SPACE CREATION

The LSA paradigm is based on a vector space method: given a text corpus of N documents and M words, LSA defines a mapping between the M words and the N documents into a continuous vector space S , where each word, as well as each document is associated to a vector in S .

Let A be the $M \times N$ matrix whose (i, j) -th entry is the square root of the sample probability of the i -th word belonging to the j -th document. The Singular Value Decomposition of the matrix A is performed, so that A is decomposed in the product of three matrices: a column-orthonormal $M \times N$ matrix U , a column-orthonormal $N \times N$ matrix V and a $N \times N$ diagonal matrix Σ , whose elements are called singular values of A .

$$A = U\Sigma V^T$$

Let us suppose that A 's singular values are ranked in decreasing order. Let R be a positive integer with $R < N$, and let $\tilde{U}$ be the $M \times R$ matrix obtained from U by suppressing the last $N-R$ columns, $\tilde{\Sigma}$ the matrix obtained from Σ by suppressing the last $N-R$ rows and the last $N-R$ columns and $\tilde{V}$ be the $N \times R$ matrix obtained from V by suppressing the last $N-R$ columns. Then

$$\tilde{A} = \tilde{U}\tilde{\Sigma}\tilde{V}^T$$

is a $M \times N$ matrix of rank R . It can be shown that, according to the illustrated procedure, $\tilde{A}$ is the best rank R approximation of A (among the $M \times N$ matrices) with respect to the Hellinger distance, defined by

$$d_h(X, Y) = \sqrt{\sum_{i=1}^M \sum_{j=1}^N (\sqrt{x_{ij}} - \sqrt{y_{ij}})^2}$$

which allows to interpret the *TSVD* technique as a statistical estimator.

The two matrices $\tilde{U}$ and $\tilde{V}$ obtained after truncated decomposition reflect a breakdown of the original relationships identified by A into linearly independent vectors. The R columns of $\tilde{U}$ constitute a basis of a semantic space S , which can be interpreted as a "conceptual space". The term "conceptual space" may be confusing, since it recalls the well known Gardenfors conceptual spaces. Gardenfors spaces have to be "manually" built by extracting from the knowledge base the quality dimensions, so they could be not the optimal choice, for practical reasons, in order to describe the knowledge base of a robot. The conceptual interpretation of the semantic space allows the automatic building of "conceptual spaces" by sub-symbolic processing of the raw sample data, so that the fundamental "latent semantic concepts" rise spontaneously. These fundamental coordinates will be represented by the orthonormal basis generated by the application of the LSA methodology, using the Hellinger distance instead of the Frobenius distance. More formally, the independent R dimensions of the R^R space can be labelled, in order to characterize the "fundamental" concepts residing in the data driven space. The technique is to identify, for each column of the matrix $\tilde{U}$ the words associated to the components with the highest value these words will constitute a label for each axis. Emotional states and external situations can be verbally described and mapped into this semantic space. Besides, an associative capability is simulated between a generic verbal-coded stimuli s_i and a generic emotion e_j with a closeness geometric distance given by the cosine between s_i and e_j .

3 THE ARCHITECTURE

The architecture of the presented system is organized in three main areas:

- The *Sub-conceptual Area* processes the perceptual data that comes from the sensors.
- The *Emotional Area* contains the "conceptual space of emotional states" which constitutes the sub-symbolic representation of emotions, acquired during the training phase. This area gives to the robot emotional capabilities.
- The *Behavioural Area* activates a non-deterministic behaviour related to the humanoid emotional state.

The Sub-Conceptual Area.

The sub-conceptual area is allowed to receive perceptions from the different modalities of the robot and carry on also the control of its actuators. This area is composed by two main modules:

- The *MotionModule* controls the actuators of the robot in order to perform the movement requested by the Behavioural Area.
- The *PerceptualModule* processes the raw data coming from robot sensors in order to obtain high level information about the environment.

A model as similar as possible to human beings has been prepared by considering five input channels for the Perceptual Module associable to the five human senses: sighting, touching, hearing, tasting and smelling. This framework will constitute a

base for future development of real humanoid robots equipped with ad-hoc sensors in order to completely simulate human senses.

The perceptual information sensed by the processing of the raw sensor data is associated to their English description: as an example, if the system recognizes teddy bear, it will be associated to the sentence “I can see a teddy bear”. The use of a verbal description of the information retrieved from the environment allows their mapping into the emotional area in an easy way. These natural language descriptions associated to each modality will be the input of the conceptual area.

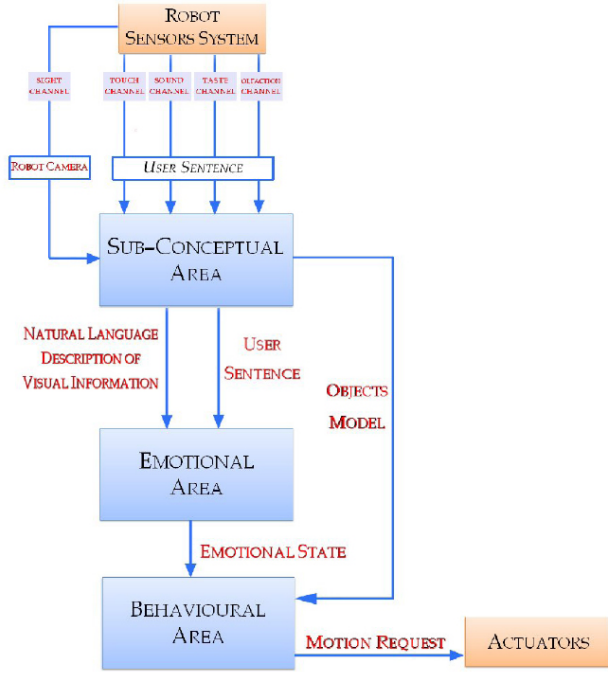


Figure 1: A sketch of the emotional robotic partner architecture.

The Emotional Area.

The emotional area makes the robot able to find emotional analogies between the current status and the previous knowledge stored on the robot using the semantic space of emotional states.

A corpus of documents dealing with emotions has been chosen in order to infer the emotional semantic space. Emotional states have been coded in this space using verbal description of situations capable to evoke emotions. Environmental incoming stimuli are encoded using the natural language and then are mapped in the space, to find analogies with the emotions encoded in it.

The set of documents used has been obtained through an accurate selection of excerpts associated to feelings. We have selected the following emotional expressions: sadness, fear, anger, joy, surprise, love and a neutral state have been considered. These classification takes inspiration from Parrot [25].

A large amount of documents has been selected from several publicly available on-line sources. The excerpts have been organized in homogeneous paragraphs both for text length and emotion. A matrix has been organized where the 6 emotional states and the neutral state have also been coded according to the LSA paradigm. A corpus of about 1000 documents, equally distributed among the seven states, has been built. This set of documents represents the affective knowledge base of the system.

We did not use a parsing elaboration of the documents. Each document has been only processed in order to remove all words in literature named “stopwords” that do not carry semantic information like articles, prepositions and so on. The documents classification based on the 6+1 emotions was conducted from only one human expert annotator. According to the technique outlined in section two, a 87×1000 terms documents matrix (A) has been created where $M=80+7$ is the number of words plus the emotional states and $N=1000$ is the number of excerpts. The generic entry a_{ij} of the matrix is the square root of the sample probability of the i -th word belonging to the j -th document. The TSVD technique, with $K=150$, has been applied to A in order to obtain its best approximation according to the Hellinger distance. This process leads to the construction of a $K=150$ dimensional conceptual space of emotions S . The axes of S represent the “fundamental” emotional concepts automatically induced by TSVD procedure arising from the data. In the obtained space S , a subset of n_i documents for each emotional state corresponding to one of the six “basic emotion” E_i has been projected in S using the folding-in technique.

According to this technique each excerpt is coded as the sum of the vectors representing the terms composing it. As a result, the j -th excerpt belonging to the subset corresponding to the emotional state E_i is coded as a vector em_j and each emotional state E_i is represented in S by the associated subset of n_i vectors.

The inputs from the five sense channels are coded in natural language words or sentences describing them and projected in the conceptual space using the folding-in technique. At time t , the input from the visual channel is coded as a vector $v(t)$, the input from hearing channel is coded as a vector $h(t)$ and the input from touch channel is coded as a vector $t(t)$. The tasting and smelling channel are out of the scope of the present research and will be implemented in the future version of the system. These vectors, representative of the inputs from the channels, are merged together as a weighted sum in a single vector $f(t)$ that synthesizes the inputs stimuli from environment at instant t :

$$f(t) = \alpha \cdot vis(t) + \beta \cdot hea(t) + \gamma \cdot tou(t) + \delta \cdot sme(t) + \eta \cdot tas(t)$$

where α , β , γ , δ and η are weights that allow assigning different relevance to the specific inputs coming from each one of the three sense channels.

The emotional semantic similarity between the vector $f(t)$ and the vectors $em_1, em_2, \dots, em_7$, that code the six emotions in S , plus the “neutral” state, can be evaluated using the cosine similarity measure between each em_i and $f(t)$:

$$sim(f(t), em_j) = \frac{f(t) \cdot em_j}{\|f(t)\| \cdot \|em_j\|}$$

A higher value of $sim(f(t), em_j)$ corresponds to an higher value of similarity between the emotion evoked from the input and the emotion E_i associated with the vector em_j .

The semantic similarity measure is calculated between $f(t)$ and each em_j . The vector em_j which maximizes the quantity expressed in formula will be the inferred emotional state with an associated weight, or intensity. This emotional state will activate a behavioural stimulus with a given intensity given by $sim(f(t), em_j)$.

The Behavioural Area.

The purpose of the Behavioural Area is to execute and control the behaviours coherent to the emotional state inferred by the Emotional Area. A behaviour is described as a sequence of primitive actions sent directly to the robot actuators. Each emotional state is related with different behaviours in order to give to the robot a non-monotonous, non-deterministic, and non-boring response. A behaviour is selected by the evaluation of a score associated to each one of them:

$$w_{b_i} = \alpha_w r + \beta_w dt$$

where r is a random value ranging from 0 to 1, dt is the time elapsed by the instant at which the behaviour b_i has been executed and the instant at which this valuation is effected; α_w and β_w are respectively the weight assigned to the random value and to the time. The response with the highest weight is chosen and executed. The emotional stimulus is weighted, so the reaction will be executed with the same intensity: movements of the parts of the body will be faster or slower.

If the emotional state is classified as “neutral” a standard behaviour (to lie down, to sleep, and so on) is randomly selected.

4 THE HUMANOID ROBOT NAO

The humanoid robot Nao is the robotic platform used. It is from Aldebaran Robotics corporation from Paris. It is 60cm tall and its weight is 4,3 Kg. It has 25 degrees of freedom distributed along the body: 4 joints for each arm; 2 for each hand; 5 for each leg; 2 for the head; one used to control the hip.

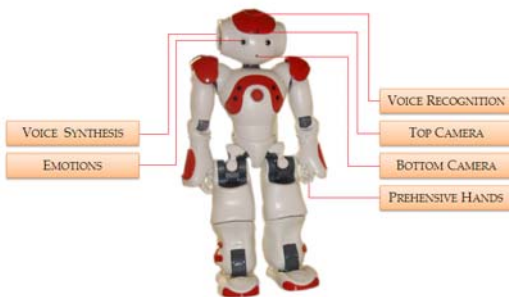


Figure 2: The humanoid robot Nao, from Aldebaran robotics.

Nao uses some force sensors called “Force Sensitive Resistors” to measure the resistance versus the quantity of pressure applied. In other words, it calculates the distribution of the force along the foot area in order to measure the inertial moment. Each time the robot makes a step, the gravity acceleration is calculated and summed to the acceleration that comes from the robot step itself, calculated in order to balancing the reaction force from the contact surface.

The robot has four ultrasonic sensors posed on the chest in order to calculate distances with objects inside the range between 25cm and 100cm.

Nao is capable to calculate its position respect the floor using the Inertial Measurement Unity (IMU), through an accelerometer and a gyroscope. In this way it is possible to check if the robot falls during the accomplishment of a behaviour.

Two cameras integrated in the head of the robot are present: with a maximum resolution of 640x480 and with a maximum frame-rate of 30fps it is possible to capture the environment and to recognize objects.

In order to improve the emphatic capabilities of this robot, different coloured led are present along the body, from the ears, to the eyes and also on the chest and on the feet. Colours can be used to express emotion. For example red as rage, blue or green as joy, and so on.

Also some buttons are present on the robotic platform, on the head and on the chest. In particular the big button on the chest is used to turn on and turn off the robot, and also to modify standard behaviours of the robot itself.

Nao carries on a x86 AMD Geode CPU capable to work with 500 MHz, an internal memory SDRAM of 256 MB, 1 GB of flash memory, two speakers, a vocal synthesizer, WiFi, Ethernet and uses an embedded 32 bit Linux. It is possible to control it using C/C++ libraries, but also Python and the Gostai’s framework Urbi.

The robot Nao uses a lot of sensors and an high number of degrees of freedom and this is justified because the wide capabilities of using it in different missions, for many purposes and in several environments.

5 THE EMOTIONAL ROBOT

The emotional humanoid robot presented in this paper is an Aldebaran’s Nao supplied with empathy capabilities. The emotional robot follows the empathic model based on the emotional conceptual space, yet described. Components running on the robot or on a support computer have been implemented in order to distribute the computational weight of the entire system. Three modules, running on a computer, are the *Perceptual module*, that implements the Sub-conceptual area, the *Emogen module*, that realizes the Emotional Area, and the *Controller*, that carry on the Behavioral Area.

The humanoid behaviours are implemented inside the robot Nao, in the *NaoMotion module*. This module is capable to accept request of behaviours and to execute them by controlling the joints of the robot. Two kinds of behaviours have been implemented: walks and actions. The walking algorithm used has been supplied by Aldebaran instead of the action behaviours that are implemented as interpolation of different poses of the robot. The actions implemented are greeting, clapping, looking

around and kicking. Other actions are used to express the six emotions: fear, angry, sadness, joy, love and surprise.

Inside the Perceptual module a vision system has been implemented in order to sense objects and translate them in symbolic instances. It uses the data captured from the camera posed on the head of the robot to recognize objects of the world using colours and geometric features. This information is used by the Controller module and, as verbal description, by the EmoGen module.

A human user can also realize an interaction with the robot through an user-friendly interface. It accepts sentences from the user by a computer keyboard, sentences that are sent to EmoGen module. In the future development the keyboard will be replaced with voice recognition as functionality of the robot.

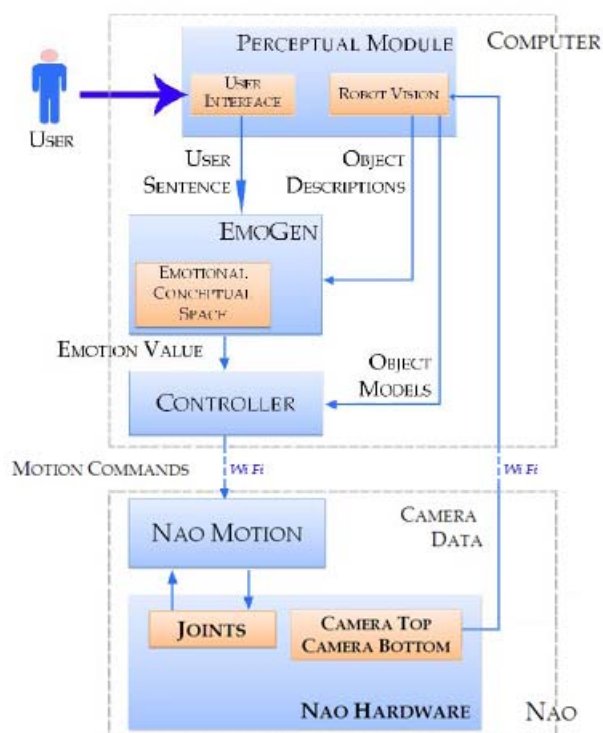


Figure 3: The different modules of the system.

The EmoGen module elaborates sentences from user and from visual information and infers, according with the emotional conceptual space, the correct emotional state.

At last, the emotional state and the information about the objects recognized are used by the Controller module to generate the most coherent action. The controller will send through TCP-IP connection this information to the robot. The NaoMotion module will respond to this command by the activation of one of the correspondent emotional behaviours.

6 EXPERIMENTAL RESULTS

Different experiments have been conducted using the robot Nao in order to test the emotional capabilities of the robot by reproducing various emotive situations.

In a first experiment the robot is posed in front of a Robovie-M from VStone corporation, a yellow robot simply to recognize as a puppet, and an orange ball. Visual stimulus is converted in a small and simple sentence as *“The puppet throws the ball”*. The dominant emotional state activated in the emotional conceptual space of the robot is “Surprise” and the robot executes a “Surprise” behaviour.

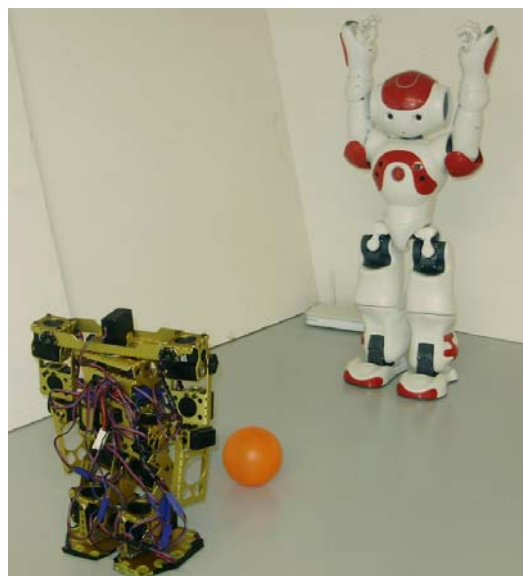


Figure 4: In the first trial the robot express a Joy behaviour.

In a second experiment the input phrase provided by an naïve human user is *“brother and sister play with the dog”*. The dominant emotional state activated in the emotional conceptual space of the robot is “Love” and the robot executes a “Love” behaviour.

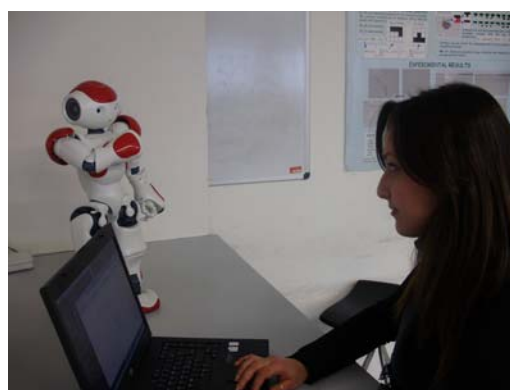


Figure 5: The robot express a Love behaviour.

The Table 1 shows four experiments results. Different sentences are elaborated by the system in order to achieve the highest emotion value.

<i>My sister and my dog are running together</i>			
Sad -0.0098	Angry -0.2417	Fear -0.1077	
Joy 0.0842	Surprise -0.1219	Love 0.0343	Neutral -0.0377
<i>Brother and sister play with the dog</i>			
Sad 0.1509	Angry -0.0032	Fear -0.0371	
Joy 0.1794	Surprise 0.0624	Love 0.2548	Neutral -0.0962
<i>The child is sad while he look at the window</i>			
Sad 0.4566	Angry 0.0821	Fear 0.2377	
Joy -0.2512	Surprise 0.0481	Love 0.2366	Neutral -0.0654
<i>The puppet throws the ball</i>			
Sad 0.0132	Angry -0.0413	Fear 0.0056	
Joy -0.0423	Surprise 0.0139	Love -0.0259	Neutral -0.0241

Table 1: The Numerical results of different experiments.

7 CONCLUSIONS & FUTURE WORK

The results shown in this paper confirm that it is possible to generate spontaneous and non trivial behaviour for entertainment humanoid robot by the creation of a high dimensional sub-symbolic representation of concepts and emotions. The robot is capable to perceive utterances and visual stimulus and to respond in an empathic way.

ACKNOWLEDGMENT

The authors would like to thank Giovanni Pilato and Giorgio Vassallo for the valuable and important help during the development of the robot architecture system.

REFERENCES

- [1] Takahashi M., Suzuki T., Cinquegrani F., Sorbello R., Pagello E.: *A mobile robot for transport applications in hospital domain with safe human detection algorithm*. In: 2009 IEEE International Conference on Robotics and Biomimetics (ROBIO 2009). Guilin, Guangxi, Cina, 18-22 December (2009).
- [2] Chella A., Pilato G., Sorbello R., Vassallo G., Cinquegrani F., Anzalone S. M.: *An Emphatic Humanoid Robot With Emotional Latent Semantic Behavior*. On: First International Conference, SIMPAR 2008, Simulation, Modeling, and Programming for Autonomous Robots, 3-6 November, Venice, Italy (2008).
- [3] Menegatti, E., Silvestri, G., Pagello, E., Greggio, N., Cisternino, N., Mazzanti, F., Sorbello, R., Chella, A.: *3D models of Humanoid Soccer Robot in USARSim and Robotics Studio simulators*. In: IJHR 2008, International Journal of Humanoid Robotics (2008).
- [4] Chella A, Barone R.E, Pilato G, Sorbello R.: *An Emotional Storyteller Robot*. In: 2008 AAAI Spring Symposium on. Stanford University, April (2008).

- [5] Chella A, Pilato G, Sorbello R., Vassallo G, Cinquegrani F, Anzalone S.M.: *An Emphatic Humanoid Robot With Emotional Latent Semantic Behavior in Microsoft Robotic Studio*. In: . Cittadella dei Musei, Cagliari, Italia, 11-13 September (2008).
- [6] Chella A, Pilato G, Sorbello R., Vassallo G, Barone R.E.: *The Latent Semantic Analysis as a Framework for Behaviours of EmphyBot Robot in Entertainment Application*. In: 13th IEEE/IFAC International Conference on Methods and Models in Automation and Robotics. Szczecin, Poland, 27-30 August (2007).
- [7] Chella A., Sorbello R., Anzalone, S.M., Cinquegrani, F., Caccia, D.: *A New Architecture Based on a Simulation Environment for Four Legged and Humanoid Robots*. In: 13th IEEE IFAC International Conference on Methods and Models in Automation and Robotics, MMAR 2007, August (2007).
- [8] Ortony A., Clore G., Collins A.: *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge (1988).
- [9] Chella A., Frixione M. and Gaglio S.: *An Architecture for Autonomous Agents Exploiting Conceptual Representations*. In: Robotics and Autonomous Systems 25, pp. 231–240 (1998).
- [10] Agostaro F., Augello A., Pilato G., Vassallo G., Gaglio S.: *A Conversational Agent Based on a Conceptual Interpretation of a Data Driven Semantic Space*. In: Lecture Notes in Artificial Intelligence, Springer-Verlag GmbH, vol. 3673, pp 381-392 (2005).
- [11] Landauer, T. K., & Dumais, S. T.: *A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge*. In: Psychological Review, 104, 211-240 (1997).
- [12] Landauer, T.K., Foltz, P.W., Laham, D.: *Introduction to Latent Semantic Analysis*. In: Discourse Processes, 25, 259-284 (1998).
- [13] Thagard, P., Shelley, C.P.: *Emotional analogies and analogical inference*. In: D. Gentner & K. H. Holyoak & B. K. Kokinov (Eds.), The analogical mind: Perspectives from cognitive science, pp. 335-362. Cambridge, MA: MIT Press (2001).
- [14] Berry, M.W., Dumais, S.T., Gavin W. O'Brien: *Using Linear Algebra for Intelligent Information Retrieval*. In: SIAM Review, Vol. 37, No. 4, pp. 573-595 Dec., (1995).
- [15] Peters, S., Widdows, D.: *Word vectors and quantum logic experiments with negation and disjunction*. In: Mathematics of Language, 8, Bloomington, Indiana, June, (2003).
- [16] P. Gardenfors. *Conceptual Spaces*. MIT Press, Bradford Books, Cambridge, MA, (2000).
- [17] Ultsch, A.: *The Integration of Connectionist Models with Knowledge based Systems: Hybrid Systems*. Proc. of the IEEE SMC 98 International Conference, 11-14 October, San Diego, pp. 1530-1535. (1998).
- [18] Zhu, H.: *Bayesian geometric theory of learning algorithms*. In: Proceedings of the International Conference on Neural Networks (ICNN'97), vol. 2: pp. 1041–1044. (1997).
- [19] Landauer, T. K., Laham, D., & Foltz, P. W., *Learning human-like knowledge by Singular Value Decomposition: A progress report*. In: M. I. Jordan, M. J. Kearns & S. A. Solla (Eds.) Advances in Neural Information Processing Systems 10, pp. 45-51. Cambridge: MIT Press (1998).
- [20] Dautenhahn, K., Billard, A.: *Bringing up robots or psychology of socially intelligent robots: From theory to implementation*. In: Proc. 3rd Int. Conf. on Autonomous Agents. Seattle WA My (1999).
- [21] Fujita, M.: *Digital Creatures for Future Entertainment Robotics*. In: IEEE International Conference on Robotics and Automation, pages. 801-806. San Francisco, California (2000).
- [22] Arkin R., Fujita M., Takagi T., Hasegawa R.: *Ethological Modeling and Architecture for an Entertainment Robot*. In: IEEE Int. Conf. on Robotics & Automation, p.p. 453-458. Seoul. (2001).
- [23] Lakemayer, G.: *On Sensing and off-line interpreting in GOLOG*. In: Logical Foundations for Cognitive Agents, Contributions in Honor of Ray Reiter, pages 173-187, Springer, Berlin, (1999).
- [24] Colon, E., Sahli, H., Baudoin, Y.: *CoRoBa, a Multi Mobile Robot Control and Simulation Framework*. In: Int. Journal of Advanced Robotic Systems, (2006).
- [25] Parrott, W. (2001), *Emotions in Social Psychology*, Psychology Press, Philadelphia.

Machine Learning and Affect Analysis Against Cyber-Bullying

Michal Ptaszynski¹ Pawel Dybala¹ Tatsuaki Matsuba² Fumito Masui³ Rafal Rzepka¹ Kenji Araki¹

Abstract. Online security has been an important issue for several years. One of the burning online security problems lately in Japan has been online slandering and bullying, which appear on unofficial Web sites. The problem has been becoming especially urgent on unofficial Web sites of Japanese schools. School personnel and members of Parent-Teacher Association (PTA) have started Online Patrol to spot Web sites and blogs containing such inappropriate contents. However, countless number of such data makes the job an uphill task. This paper presents a research aiming to develop a systematic approach to Online Patrol by automatically spotting suspicious entries and reporting them to PTA members and therefore help them do their job. We present some of the first results of analysis of the inappropriate data collected from unofficial school Web sites. The analysis is performed firstly with an SVM based machine learning method to detect the inappropriate entries. After analysis of the results we perform another analysis of the data, using an affect analysis system to find out how the machine learning model could be improved.

1 INTRODUCTION

Online security has been an urgent problem ever since the creation of the Internet. Some of the most well known issues of online security include hacking, cracking, data theft and online espionage. However, for several years, a problem that has become much more visible and therefore influential and socially harmful, is the problem of exploitation of online open communication means, such as BBS forum boards, or social networks to convey harmful and disturbing information. In USA, a great focus on this issue began in 2001 after the 9.11 terrorist attack. However, similar cases have been noticed before in other countries not on such a scale. In Japan, on which this research is focused, a great social disturbance was caused by cases of sending alarming messages by criminals or hijackers on the Internet just before a committing a crime. One famous case of this kind happened in May 2000, a year before the 9.11 in USA, when a frustrated young man sent an message on, then fairly popular, Japanese BBS forum *2channel*, informing readers he was going to hijack a bus, just before he proceeded with his plan. Growing number of similar cases around the world opened a public debate on whether such suspicious messages could not be spotted early enough to prevent the crimes from happening [1] and on the freedom of speech on the Internet in general [2]. Some of the famous research in this matter was performed

by the team of Hsinchun Chen, who started the project aiming to analyze Dark Web Forums in search of alarming entries about planned terrorist attacks [3, 4]. Another research of this kind was performed by Gerstenfeld [5], who focused on extremist groups.

However, there have been little research performed on a problem less lethal, although equally serious, namely online slandering and bullying of natural persons, known generally as 'cyber-bullying'. In Japan the problem has become serious enough to be noticed by the Ministry of Education, Culture, Sports, Science and Technology (later: MEXT) [6]. At present school personnel and members of Parent-Teacher Association (PTA) have started Online Patrol to spot Web sites and blogs containing such inappropriate contents. However, countless number of such data makes the job an uphill task. Moreover, the Online Patrol is performed manually and as a volunteer work.

Therefore we started this research to help the Online Patrol members. The final goal of this research is to create a machine Online Patrol crawler automatically spotting the cyber-bullying cases on the Web and reporting them to the Police.

In this paper we present some of the first results of this research. We first focused on developing a systematic approach to spotting online cyber-bullying entries automatically to ease the burden of the Online Patrol volunteers. In our approach we use machine learning to train a system for spotting the undesirable contents and then perform affect analysis of these contents to find out how the system could be further improved.

The paper outline is as follows. In Section 2 we describe the problem of cyber-bullying in more details. In Section 3 we present the prototype SVM-based method for detecting the cyber-bullying entries and evaluate it. In Section 5 we present the results of affect analysis of these data and propose some ideas about how the system described in Section 3 could be improved. Finally, in Section 5 we conclude the paper and provide some hints on further work in this area.

2 WHAT IS CYBER-BULLYING?

Although the problem of sending harmful messages in the Internet has appeared for several yaers, it has been officially defined only lately and named as cyber-bullying⁴. The National Crime Prevention Council in America state that cyber-bullying happens "when the Internet, cell phones or other devices are used to send or post text or images intended to hurt or embarrass another person." ⁵ Other definitions, such as the one by Bill Belsey, a teacher and an anti-cyber-bullying activist, say that cyber-bullying "involves the use of

¹ Graduate School of Information Science and Technology, Hokkaido University, {ptaszynski,paweldybala,kabura,araki}@media.eng.hokudai.ac.jp

² Graduate School of Engineering, Mie University, matsuba@ai.info.mie-u.ac.jp

³ Department of Computer Science, Kitami Institute of Technology, f-masui@mail.kitami-it.ac.jp

⁴ other terms used are cyber-harassment, and cyber-stalking

⁵ <http://www.ncpc.org/cyberbullying>

information and communication technologies to support deliberate, repeated, and hostile behavior by an individual or group, that is intended to harm others.” [7]

Some of the first robust research on cyber-bullying was done by Hinduja and Patchin, who performed numerous surveys about the subject in the USA [8, 9]. They found out that the harmful information may include threats, sexual remarks, pejorative labels, false statements aimed at humiliation. When posted on a BBS forum or a social network, such as Facebook, it may disclose personal data of the victim which, published with a humiliating material about them or in their name, aims to defame or ridicule them personally.

2.1 Cyber-bullying and Online Patrol in Japan

In Japan, after a several cases of suicides of cyber-bullying victims who could not bare the humiliation, MEXT has considered the problem serious enough to start a movement against the problem. In a manual for spotting and handling the cases of cyber-bullying [6], the Ministry puts a great importance on early spotting of the suspicious entries and messages, and distinguishes several types of cyber-bullying noticed in Japan. These are:

1. Cyber-bullying appearing on BBS forums, blogs and on private profile web-sites;
 - (a) Entries containing libelous, slanderous or abusive contents;
 - (b) Disclosing personal data of natural persons without their authorization;
 - (c) Entries and humiliating online activities performed in the name of another person;
2. Cyber-bullying appearing in electronic mail;
 - (a) E-mails directed to a certain person/child, containing libelous, slanderous or abusive contents;
 - (b) E-mails in the form of chain letters containing libelous, slanderous or abusive contents;
 - (c) E-mails send in the name of another person, containing humiliating contents;

In this research we focused mostly on the cases of cyber-bullying appearing on informal web sites of Japanese secondary schools. Informal school web sites are web sites where school pupils gather to exchange information about school subjects or contents of tests, etc. However, as was noticed by Watanabe and Sunayama [10], on such pages there have been a rapid increase of entries containing insulting or slandering information about other pupils or even the teachers. Cases like that make other users uncomfortable using the Web sites and cause undesirable misunderstandings.

To deal with such malicious entries, a movement of Online Patrol have started. In this movement usually participate teachers and PTA members, who, based on the MEXT definition of cyber-bullying, read through all available entries, decide whether an entry is dangerous or not and, if necessary, send a deletion request to the web page administrator or the Internet provider and send a report to the police. The typical Online Patrol route is presented on Figure 1

Unfortunately, at present state of affairs, the school personnel and PTA members taking part in Online Patrol perform all tasks manually as voluntary work, beginning from reading the countless numbers of entries and deciding about the appropriateness of their contents,

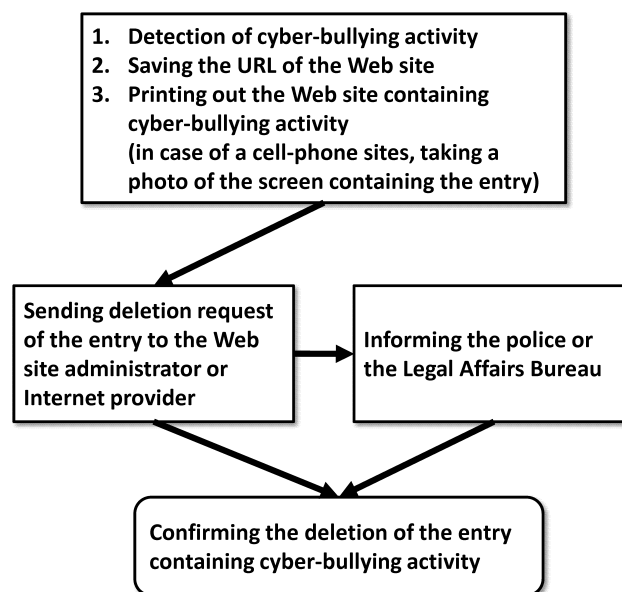


Figure 1. Route of Online Patrol.

through printing out or taking photos of the pages containing cyber-bullying entries, and finally sending the reports and deletion requests to the appropriate organs. Moreover, since the number of entries rises by the day, surveillance of all of the Web becomes an uphill task for the small number of patrol members.

With this research we aim to create a Web crawler capable to perform this difficult task instead of humans, or at least to ease the burden of the Online Patrol volunteers.

3 MACHINE LEARNING METHOD FOR CYBER-BULLYING DETECTION

In this section we describe a machine learning method developed to handle cyber-bullying activities. The method consists of several stages, including creation of a lexicon of vulgar, slanderous and abusive words, slanderous information detection module, ranking of the information according to the level of their harmfulness, and visualization of the harmful information. The system flow chart is represented on Figure 2. The creation of the system has separated into two general phases: training phase and processing (test) phase. Below we present the details of each phase.

1. Training phase;

- (a) Crawling the school Web sites;
- (b) Detecting manually the cyber-bullying entries;
- (c) Extraction of vulgar words and adding them to the lexicon;
- (d) Estimating word similarity with Levenshtein distance;
- (e) Part of speech analysis;
- (f) Training with SVM;

2. Processing (test) phase;

- (a) Crawling the school Web sites;
- (b) Detecting the cyber-bullying entries with SVM model;

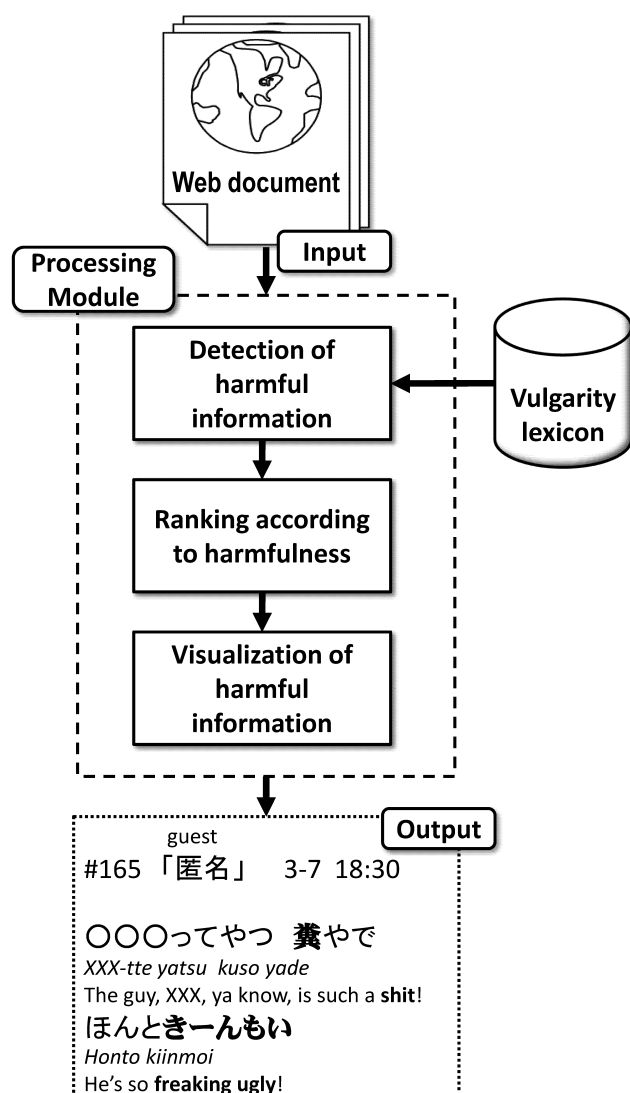


Figure 2. Flow chart of the cyber-bullying activity detecting system.

- (c) Part of speech analysis of the detected harmful entry;
- (d) Estimating word similarity with Levenshtein distance;
- (e) Marking and visualization of the key sentence;

3.1 Construction of Vulgarity Lexicon

We defined the keywords distinguishable for cyber-bullying as vulgarities. Vulgarities are defined as "obscene or vulgar words which connote offenses against particular persons or society." Examples of such words in English are *shit*, *fuck*, *bitch*. Examples in Japanese include, e.g., *uzai* (freaking annoying), or *kimoi* (freaking ugly).

Since vulgar words are often not detected by part of speech (POS) taggers, or marked as "unknown word", we decided to create list of vulgarities and add it to a basic lexicon of POS tagger. This was performed according to the procedure described below. At first we performed a study on the vulgar keywords constituting the harmful information. We obtained a set of informal school Web pages, read them and performed a manual categorization into "harmful" and

"non-harmful" entries based on the MEXT classification of cyber-bullying. From these Web pages we obtained 750 harmful entries including 216 unique vulgar keywords distinguishable for cyber-bullying activity. The extracted keywords were finally added to the list of vulgarities. Finally, we added the grammatical information to the extracted keywords and added them to the POS tagger. An example of addition of grammatical information is presented in Table 1.

Table 1. Example of a newly registered word.

<i>kimoi</i> (freaking ugly)
POS: Adjective;
Headword: <i>kimoi</i> (hit-rate: 294);
Reading: <i>kimoi</i> ;
Pronunciation: <i>kimoi</i> ;
Conjugated form: uninflected;

3.2 Estimation of Word Similarity with Levenshtein Distance

Users would often change spelling of words and write them in a unnormalized way. It is a part of jargonization of the language used online. Examples of this phenomenon in English, would be writing phrases like, e.g., "C.U." in the meaning of "See you [later]!" (usually on the end of e-mails or online messages), or "brah" in the meaning of "bro[ther], friend"⁶, etc. Some examples of such colloquial transformations in the Japanese online language are shown in Table 2.

Table 2. Three examples of colloquial transformations of Japanese words in online jargon.

original word	colloquial transformation
<i>kimoi</i> (freaking ugly, gross)	<i>kimosu</i> , <i>kishoi</i> , <i>kisho</i> , ...
<i>uzai</i> (freaking annoying)	<i>uzee</i> , <i>UZAI</i> , <i>uzakkoi</i> , ...
<i>busaiku</i> (ugly bitch)	<i>buchaiku</i> , <i>bussaiku</i> , ...

With this variation, words having the same meaning would be classified as separate samples, which would cause hit-rate dispersion. Therefore to unify the same words written with slightly different spelling we calculated the similarity of the extracted words. The similarity was calculated using Levenshtein Distance [11] in a way similar to [12, 13]. The Levenshtein Distance between two strings is calculated as the minimum number of operations required to transform one string into another, where the available operations are only deletion, insertion or substitution of a single character.

However, as Japanese is transcribed using three character types, Chinese characters (*kanji*) encapsulating from one to several syllables and two additional syllabary (*katakana* and *hiragana*), calculating the distance would become very imprecise. To solve this problem, every word was automatically transformed in their alphabetical transcription. An example of distance calculation is presented on the back-transformation of the word *kimosu* to its original spelling *kimoi* in Table 3. The distance between both words is calculated as 2.

3.3 SVM Based Classification of Cyber-bullying

To classify the entries into either harmful (cyber-bullying) or non-harmful, we used Support Vector Machines. Support Vector Ma-

⁶ <http://www.internetslang.com/>

Table 3. Three examples of colloquial transformations of Japanese words in online jargon.

transformed word	performed operation
<i>kimosu</i>	
→ <i>kimoiu</i>	substitution of 's' to 'i'; distance = 1;
→ <i>kimoi</i>	deletion of final 'u'; distance = 2;

chines (SVMs) are a method of supervised machine learning developed by Vapnik [14] and used for classification of data. They are defined as follows. With a set of training samples, divided into two categories A and B, SVM training algorithm generates a model for prediction of whether test samples belong to either category A or B. In the traditional description of SVM model, samples are represented as points in space (vectors). SVM constructs a hyperplane, in a space of a higher dimension than the base one, with the largest distance to the nearest training data points (support vectors). The larger the margin the lower the generalization error of the classifier. Since SVM has been successfully used for text classification [15] we decide to use them in this research as well. In our research the category A contains cyber-bullying cases and the category B contains all other cases, which do not consist of socially harmful information. As the software for building SVM models we used SVM_light (ver6.02)⁷.

3.4 Extraction of Key Sentences

In the process of automation of Online Patrol, apart from the classification of cyber-bullying entries, there is a need to appropriately determine how harmful is a certain entry. A ranking according to the harmfulness of entries is important to detect the most dangerous cases. In our approach an entry is considered as the more harmful, the more vulgar keywords appear in the entry.

The harmfulness of an entry is calculated using T-score. T-score is a measure answering the question of how confident can one be that the association measured between two words is an actual collocation and not a matter of chance. The higher occurrence frequency a word has in a corpus, the higher is the value of T-score. A T-score of a word associating with words A and B is calculated according to the equation below:

$$T_{score} = \frac{a}{b} \quad (1)$$

where,

$$a = [\text{word co - occurrence frequency}] - \frac{([\text{occurrence of word A}] * [\text{occurrence of word B}])}{[\text{all words in the corpus}]}$$

and,

$$b = \sqrt{[\text{word co - occurrence frequency}]}$$

We calculate the harmfulness of the whole entry as a sum of T-scores calculated for all vulgar words. This way the more frequently occurring words there are in the entry, the higher rank the entry achieves in the ranking of harmfulness.

3.5 Evaluation of the Method

To verify the performance of the method we evaluated three procedures:

1. Classification of Cyber-bullying entries with SVM;
2. Word similarity calculation with Levenshtein distance;
3. Extraction of key sentences;

3.5.1 Evaluation of SVM Model

To apply SVM to the detection of harmful information from unofficial school BBS sites, we needed to prepare the data for training the SVM model. At first we performed morphological analysis of the BBS entries to be used as the training data. For every part of speech from the analyzed and parsed data we used the POS label names and strings of characters to identify the strings. As the features in the part of speech identification we used parts of speech like nouns (person's name), nouns (other than name), verbs and adjectives. In the identification of the whole entries we used feature sets consisting of the features of each part of speech and the strings of characters containing the whole entry. Based on the amount of identified features, SVM model calculates the probability of affiliation of a character string to a certain class.

As the training data we used 966 entries gathered during an actual Online Patrol, from which human annotators (Online Patrol members) classified 750 entries as harmful and 216 as non-harmful. This time however, we did not apply the string similarity calculation to the SVM model, to evaluate both techniques separately.

The above conditions are applied to test the model trained with SVM_light. In the evaluation we calculated the system's result as balanced F-score, using 10-fold cross validation for Precision and Recall. In 10-fold cross validation data is first broken into 10 sets of size $n/10$. Then, 9 datasets are used to train on and 1 as a test. This procedure is repeated 10 times and the overall score is the mean accuracy from all 10 tests. The balanced F-score, Precision (P) and Recall (R) are calculated as follows,

$$F_{score} = 2 * \frac{P * R}{P + R} \quad (2)$$

where,

$$Precision = \frac{s}{n} \quad Recall = \frac{n}{c}$$

and,

s = cases correctly classified by the system as harmful
 n = all cases classified by the system as harmful
 c = all harmful cases

The result was Precision = 79.9%, Recall = 98.3% and the F-score = 88.2%.

3.5.2 Evaluation of Similarity Calculation

When preparing the conditions for evaluation of word similarity calculation method with Levenshtein distance, we noticed that the larger was the threshold, the higher was the probability of matching a word with completely different meaning. Therefore we performed an optimization of the similarity calculation algorithm. In the optimization we applied two heuristic rules shown in Table 4.

On BBS sites, where informal language is widely used, such as unofficial school Web sites, prolonging of syllables is used mostly to express changes in user attitude, or mood, usually highlighting the unofficial character of the entry. However, such operation has no influence on the semantics of a word. Therefore before the phase of similarity calculation we added a rule deleting all syllable prolongations.

⁷ <http://svmlight.joachims.org>

Table 4. Two heuristic rules applied in optimization of similarity calculation algorithm.

Rule	Example
1. deletion syllable prolongations	<i>kimoooi</i> → <i>kimoi</i>
2. unification of word first letter	In case of <i>uzai</i> we will consider only the words beginning with <i>u</i>

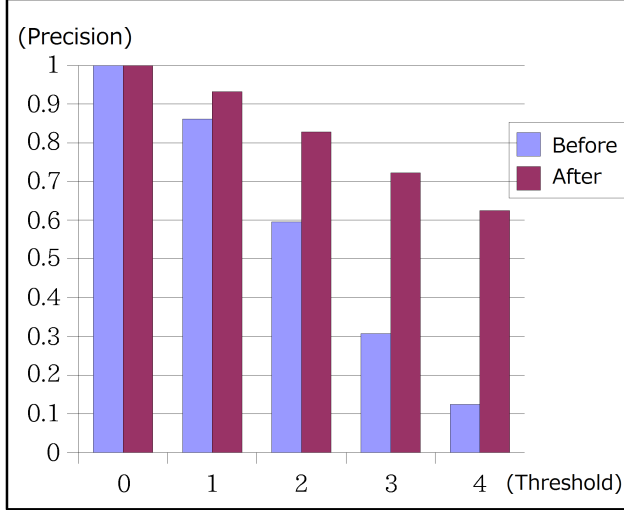


Figure 3. Precision of similarity calculation before and after applying the heuristic rules.

The second rule was unification of the first letters of matched words. This was done because in the Japanese language only the final letters change during conjugation. Therefore we could assume that it is irrelevant to calculate similarity for words differing with first letters.

The results for Precision of similarity calculation before and after applying the heuristic rules are presented on Figure 3. The Precision was greatly improved after applying the rules. With the threshold set on 2, the Precision before applying the rules was 58.9% and was improved to 85.0%.

3.5.3 Evaluation of Key Sentence Extraction

To set the ranking of cyber-bullying entries according to how harmful they are, we calculated the T-score of vulgar vocabulary they contain. The results are represented in Table 5. As one can see in Table 5, the first three scores are much higher than the other ones. This is caused by numerous repetitions of the same vulgar words in one entry, which artificially increases the occurrence rate. Moreover, in the cases, appearing in the table on places 4-6, one of the words was used repeatedly, which also caused artificial increase of occurrence and eventually the result of T-score, although smaller than for the first three cases. In such cases, although the occurrence of a certain pair is not high, T-score is artificially increased and therefore biased.

To solve this problem we considered the same vulgar words appearing in one entry as a one word. However, some vulgar words may appear as collocations. In such cases the identical numerous collocations are considered as one. The change in the results is represented in Table 6. The results indicate the bias have disappeared. However, the T-score became too small causing many word pairs appear on the same place in the ranking making the ranks difficult to set. This is

caused by the fact that the words with the same meaning appear on the Web sites in different transcription and therefore there is a large number of word sets with a small occurrence.

We solved this problem by calculating word similarity before calculating T-score. The results are represented in Table 7. Thanks to word similarity calculation a rise in T-score was observed and the number of diversified word pairs increased, which made rank setting much easier. There was 202 vulgar word pairs, from which 40 pairs occurred more than twice.

Table 5. The results of T-score calculation for sets of vulgar words. As mentioned in section 3.2 words in Japanese can be usually transcribed in three different systems: *hiragana*, *katakana* and *kanji*. The differences in transcription are represented in the table as markers after the words, with [h] for *hiragana*, [k] *katakana* and [K] for *kanji*.

word A	word B	co-occurrence	T-score
<i>baka</i> [h] (stupid)	<i>baka</i> [h]	861	29.34
<i>shine</i> [K] (fuck you)	<i>shine</i> [K]	552	23.49
<i>shine</i> [h]	<i>shine</i> [h]	58	7.62
<i>tarashi</i> [k] (pimp)	<i>shine</i> [K]	17	4.12
<i>busu</i> [k] (ugly bitch)	<i>shine</i> [K]	16	4.00
<i>kimosu</i> [h]	<i>shine</i> [h]	16	4.00
<i>shine</i> [h]	<i>busu</i> [h]	6	2.45

Table 6. Change in the results of T-score calculation for sets of vulgar words, when two or more words were considered as one.

word A	word B	co-occurrence	T-score
<i>shine</i> [K]	<i>shine</i> [K]	7	2.65
<i>shine</i> [h]	<i>shine</i> [h]	4	2.00
<i>pashiri</i> [k] (looser)	<i>shine</i> [K]	3	1.73
<i>debu</i> [k]	<i>kiero</i> [K]	3	1.73
<i>kiero</i> [K] (get lost)	<i>kiero</i> [K]	2	1.41
<i>shine</i> [K]	<i>kiero</i> [K]	2	1.41
<i>uzai</i> [h]	<i>kimoi</i> [k]	2	1.41

Table 7. Change in the results of T-score calculation for sets of vulgar words, with calculated word similarity.

word A	word B	co-occurrence	T-score
<i>shine</i> [K]	<i>shine</i> [K]	11	3.32
<i>kimoi</i> [h]	<i>shine</i> [K]	11	3.32
<i>kimoi</i> [h]	<i>busaiku</i> [h]	8	2.83
<i>uzai</i> [h]	<i>kimoi</i> [h]	7	2.65
<i>panko</i> [k] (slut)	<i>panko</i> [k]	6	2.45
<i>kimoi</i> [h]	<i>kimoi</i> [h]	6	2.45
<i>busaiku</i> [h]	<i>busaiku</i> [h]	6	2.45

3.6 Discussion

This time in morphological analysis of vulgar words we used only a small set of manually found words, which we added to the lexicon. It is difficult to include all existing vulgar words and more such vocabulary will appear in the future as well. Therefore as a further work we need to develop a method for automatic extraction of vulgar vocabulary from the Internet.

The results of SVM model used to distinguish between harmful and non-harmful information were 79.9% of Precision and 98.3% of Recall. However, on the unofficial school Web pages used as the data in this research there were numerous entries consisting of only one sentence, or even one word. Therefore the feature set for training was not sufficient and the overall result was not ideal (F-score = 88.2%).

As for word similarity calculation, many vulgar words are short and a change of even one letter might cause a change of meaning. Therefore, two different words would be matched as similar by Levenshtein distance, when the threshold is too wide. This problem might be solved by automatically setting the threshold according to the word length.

As for the extraction of key sentences, although we were able to calculate non biased T-score for vulgar expressions and set the ranking, over 80% of vulgar words appeared only once. This caused over half of the cases to be attached with similar ranks. This problem could be solved by increasing the number of training data or applying a different method of rank setting.

As was shown by Chen and colleagues [3, 4], analysis of affect intensity of Dark Web Forums often helps specifying the character of the forum. Aiming to find any dependencies between expressing emotions and cyber-bullying activities we performed an additional study and analyzed affective level of the cyber-bullying data.

4 AFFECT ANALYSIS OF CYBER-BULLYING DATA

For the comparative affect analysis we obtained additional Online Patrol data containing both cyber-bullying activities and normal entries. The data used in affect analysis contained 1,495 harmful and 1,504 non-harmful entries. The affect analysis was performed with ML-Ask system developed by Ptaszynski and colleagues [16].

4.1 ML-Ask - Affect Analysis System

The affect analysis system employed in the analysis of the cyber-bullying data described in this paper is ML-Ask developed by Ptaszynski et al. [16]. ML-Ask (eMotive eLements / Emotive Expressions Analysis System) was developed for analyzing the emotive contents of utterances. The system uses a two-step procedure: 1) Analyzing the general emotiveness of an utterance by detecting emotive elements, or emotemes, expressed by the speaker and classifying the utterance as emotive or non-emotive; 2) Recognizing the particular emotion types by extracting expressions of particular emotions from the utterance. This analysis is based on Ptaszynski's [21] idea of two-part classification of realizations of emotions in language into:

1) Emotive elements or Emotemes. Elements conveyed in an utterance indicating that the speaker was emotionally involved in the utterance, but not detailing the specific emotions. The same emotive element can express different emotions depending on context. This group is linguistically realized by subgroups such as interjections, exclamations, mimetic expressions, or vulgar language. Examples are: *sugee* (great!), *wakuwaku* (heart pounding), *-yagaru* (a vulgarization of a verb);

2) Emotive expressions. Parts of speech used to describe emotional states. However, they function as expressions of the speaker's emotions only in utterances where the speaker is emotionally involved. In non-emotive sentences they fulfill the function of simple descriptive expressions. The group is realized by various parts of speech, like nouns, verbs, adjectives, etc. Examples are: *aijou* (love), *kanashimu* (feel sad), *ureshii* (happy), respectively.

The hand-selected emotive element database consists of interjections, mimetic expressions, endearments, vulgarities, and representations of non-verbal emotive elements, such as exclamation marks

or ellipses. The emoteme database collected and divided in this way contains 907 elements in total.

A system for analysis of emoticons was also added, as they are symbols commonly used in everyday text-based communication to convey emotions.

The database of emotive expressions is based on Nakamura's collection [22] and contains 2100 emotive expressions, each classified into the emotion type they express.

4.1.1 Affect Analysis Procedure

On textual input provided by the user, two features are computed in order: the emotiveness of an utterance and the specific type of emotion. To determine the first feature, the system searches for emotive elements in the input to determine whether the input is emotive or non-emotive. In order to do this, the system uses McCab [17] for morphological analysis and separates every part of speech. McCab recognizes some parts of speech which belong to the group of emotemes, such as interjections, exclamations or emphatic sentence-final particles, like *-zo*, *-yo*, or *-ne*. If these appear, they are extracted from the utterance as emotemes. Next, the system searches and extracts every emoteme based on the system's emoteme databases (907 items). The input is then processed by CAO [18], a system for analysis of emoticons. Emoticons belong to both emotemes, and emotive expressions, as their appearance in an input always indicates emotional attitude of the user and for most emoticons it is possible to specify the emotion type they convey. All of the extracted elements mentioned above (exclamations from McCab, emotemes and emoticons) indicate the emotional level of the utterance. The emotional value of an input is calculated quantitatively, the more emotemes there were, the higher was the emotional value, however with the upper limit of 5.

Secondly, in utterances classified as emotive, the system uses a database of emotive expressions to search for all expressions describing emotional states. Here, emotions specified by CAO are added to the result of ML-Ask basic procedure. This determines the specific emotion type (or types) conveyed in the utterance. Examples of analysis performed by ML-Ask (system output) are represented below. In these examples, from top line there are: an example in Japanese, emotive information annotation (emotemes-underlined, emotive expressions-bold type font) and English translation.

4.1.2 Contextual Valence Shifters in ML-Ask

One of the problems in the procedure described above was confusion of the valence polarity of emotive expressions. The cause of this problem was extracting from the utterance only the emotive expression keywords without its grammatical context. One case of such an input is presented below in example (3). In this sentence the emotive expression is the verb *akirameru* (to give up [verb]) but the phrase *-cha ikenai* (Don't- [particle+verb]) suggest that the speaker is in fact negating and forbidding the emotion expressed literally. Such phrases are called Contextual Valence Shifters (CVS).

The idea of Contextual Valence Shifters (CVS) application in Sentiment Analysis was first proposed by Polanyi and Zaenen [23]. They distinguished two kinds of CVS: negations and intensifiers. The group of negations contains words and phrases like "not", "never", and "not quite", which change the valence polarity of the semantic orientation of an evaluative word they are attached to. The group of intensifiers contains words like "very", "very much", and "deeply",

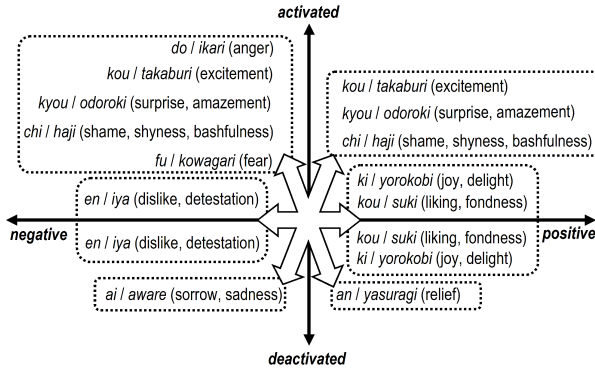


Figure 4. Grouping Nakamura's classification of emotions on Russell's space.

which intensify the semantic orientation of an evaluative word. Examples of CVS negations in Japanese are grammatical structures such as: *-nai* (not-), *amari -nai* (not quite-), *-to wa ienai* (cannot say it is-), or *-te wa ikenai* (cannot+[verb]-). Intensifiers are: *totemo-* (very-), *sugoku-* (-a lot), or *kiwamete-* (extremely). In this research we focused mostly on negations, since they have an immediate and significant influence on the meaning of emotive expressions. We applied Contextual Valence Shifters to change the valence polarity of emotive expressions in utterances containing CVS structures. Our hand-crafted database of CVS contains 71 negation structures.

However, using only the CVS analysis we would be able to find out about the appropriate valence of emotions conveyed in the utterance, but we would not know the exact emotion type. Therefore, to specify the emotion types in such utterances we applied the idea of the two-dimensional model of affect. To specify the emotion types after changing polarity with CVS, we applied the idea of the 2-dimensional model of affect [24] which assumes that all emotions can be described in 2-dimensions: the emotion's valence polarity (positive/negative) and activation (activated/deactivated). An example of positive-activated emotion could be "excitement"; a positive-deactivated emotion is, e.g., "relief" (see Figure 4).

Emotion types distinguished by Nakamura [22] were mapped on this model and their affiliation to one of the spaces determined. The emotion types with ambiguous affiliation were mapped on two possible fields. When a CVS structure is discovered, ML-Ask changes the valence polarity of the detected emotion. The appropriate emotion after valence changing is determined as the one with valence polarity and activation parameters different to the contrasted emotion (note arrows in Figure 4).

4.2 CAO - Emoticon Analysis System

Since one of the most popular strategies of expressing emotions in online communication is using emoticons, we supported ML-Ask with emoticon analysis system CAO. It is a system for estimation of emotions conveyed through emoticons developed by Ptaszynski and colleagues [18]. Emoticons are - sets of symbols widely used in text-based online communication to convey emotions. CAO, or *emotiCon Analysis and decOding of affective information*, extracts an emoticon form an input (a sentence) and determines specific emotion types expressed by it using a three-step procedure. Firstly, matching the input with a predetermined raw emoticon database containing over ten thousand emoticons. The emoticons, which

could not be estimated with only the database are automatically divided into semantic areas, such as representations of "mouth" or "eyes", based on the idea of *kinemes*, or minimal meaningful body movements, applied from the theory of kinesics [19, 20]. The areas are automatically annotated according to their co-occurrence in database. The annotation is firstly based on eye-mouth-eye triplet. If no triplet was found, all semantic areas are estimated separately. This provides hints about potential groups of expressed emotions giving the system a coverage of over 3 million possibilities. CAO is used as a supporting procedure in ML-Ask to improve performance of the affect analysis system in utterances, which do not include emotive expressions, like in the example (2) below.

- (1) *Kyo wa nante kimochi ii hi nanda !*
Today:TOP ML:nanteMX:joy day:SUB ML:nanda ML:!
Translation: Today is such a nice day!
- (2) *Iya~, sore wa sugoi desu ne- ! ^o^*
ML:iya~ this :TOP ML:sugoi COP ML:ne- ML:!
Translation: Whoa, that's great! ^o^ MX:joy
- (3) *Akirame cha ikenai yo !*
MX:dislike ML:cha CVS:cha-ikenai{→joy} ML:yo ML:!
Translation: Don't cha give up!
- (4) *Hitoribocchi nante iya da ^^*
MX:sadness ML:nante-da MX:dislike COP ML:^^
Translation: Being alone sucks...

4.3 Affect Analysis Results

At first we calculated the number of all emotive entries among both sets of data. There was 956 emotive samples among 1,495 harmful (63.95%) and 1,029 among 1,504 non-harmful (68.42%) entries. The difference was not high and therefore the number of emotive entries cannot be considered as a highly distinctive feature. However, we made the first assumption, that harmful data are less emotively emphasized than non-harmful. This would be a reasonable assumption, since cyber-bullying is often based on irony or sarcasm, which is not highly emotive, however deliberately uses some amount of emotive information to slander the object of sarcasm. To confirm this thesis we performed other comparisons.

We calculated emotive values of all emotive entries. Although the number of emotive entries and emotive value, both relate to the idea of emotiveness of a corpus in general, they are not directly related. One can easily imagine the difference by imagining a corpus of many slightly emotive entries (high number of emotive entries, but low average emotive value) and another corpus with a small number of highly emotive entries (low number of emotive entries, but high average emotive value). The approximated emotive value for harmful and non-harmful data was 1.47 and 1.5, respectively. Here also the difference is not high, although, since both values are not directly related, this can moderately support the thesis.

Next, we took a closer look on the extracted emotemes. There are four groups of emotemes distinguished in ML-Ask: i) interjections, ii) exclamations, iii) vulgarities and iv) mimetic expressions (*gitaigo* in Japanese) - arranged in order of their emotional weight. Distribution of the extracted emotemes within both entry sets is represented in Table 8. There were relatively more emotemes with high emotive weight in non-harmful data and more low weight emotemes in harmful data, which is another confirmation of the thesis set on the

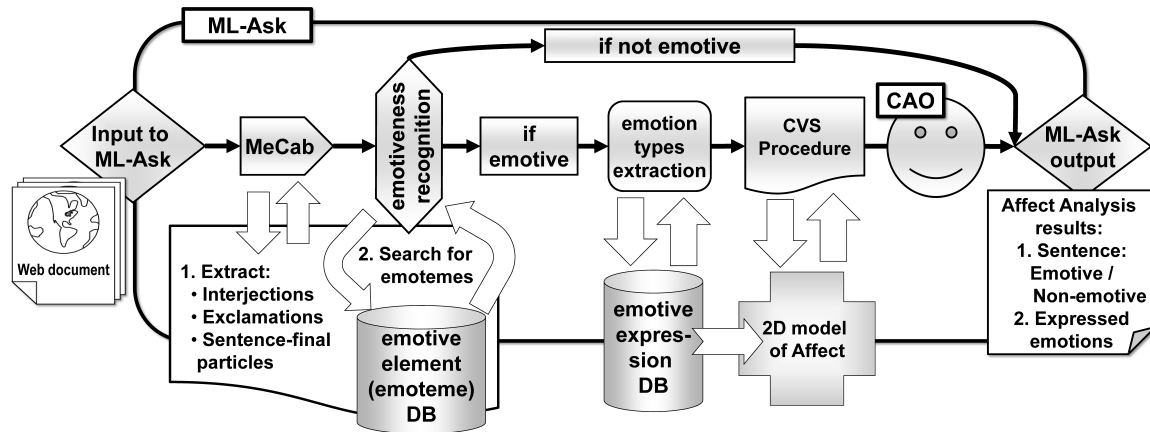


Figure 5. FFlow chart of ML-Ask system.

beginning of this section. The biggest difference between the number of extracted emotemes was for found for vulgarities, which can be regarded as a distinctive feature of cyber-bullying entries. What is also important is the fact that this provides an experimental proof for the assumption made in the design of the machine learning based system for cyber-bullying detection, which is based on a vulgarity lexicon. A similar difference, although lower, also appeared in mimetic expressions, which could be included in further study on enlarging lexicon for the system. As for interjections and exclamations, although they appeared in a large number in both datasets, more of them appeared in non-harmful entries. This could be caused by the fact that these two emoteme types are used to express emotive attitude in a straightforward way. Therefore, although there certainly are cyber-bullying cases where the victims are slandered straightforwardly, not emotional and cold sarcasm is also an often phenomenon.

Table 8. Distribution of the extracted emotemes within both entry sets.

type of emoteme	type of data	
	non-harmful	harmful
interjections	859	784
exclamations	284	174
vulgarities	8	149
mimetic expressions	7	23

Another comparison was made with the closer study of only emotive utterances. ML-Ask is capable of detecting emotive utterances with a very high reliability, although it specifies particular emotion types with low Recall (although with high Precision) [16], because its lexicon (Nakamura's dictionary, cf. [22]) is said to be out of date. Therefore there a certain number of samples is always described as emotive but with no specified emotion types. Such a phenomenon is reasonable from the linguistic point of view, since there are many sentences that are emotive, although the emotion they convey depends on their context.

Therefore we first compared how many were there specified vs unspecified emotive utterances. The result was 13.18% vs 86.82% for cyber-bullying and 11.95% vs 88.05% for the normal entries. The higher ratio of specified emotion types in cyber-bullying data might suggest that people more often use traditional emotive expressions to slander people than they do to express emotions usually. How-

ever, low difference comparing to the non-harmful entries makes this statement weakly reliable.

Next, we compared the number of particular emotive expressions extracted by ML-Ask. The results have been represented in Table 9. The analysis of the number of emotion types represented by the extracted emotive expressions revealed interesting tendencies, although here as well the strength of the proof is low and therefore unequivocal. However, anger scored much higher in the cyber-bullying data than in the normal data, which is a reasonable result, since Web site entries meant to slander others are expected to express anger more often. Fear on the other hand scored very low in cyber-bullying data. This is also reasonable, since it is difficult to bully others by expressing one's fears. Dislike scored slightly higher in the harmful data. As for the positive emotion types, e.g., joy scored higher in normal dataset, which was expectable. On the other hand, fondness scored unexpectedly higher in cyber-bullying dataset. Detailed analysis revealed that people would often express strong sarcasm with the use of positive expressions. Some examples of such entries have been represented in Table 10.

Table 9. The number of particular emotive expressions extracted by ML-Ask from both datasets.

type of data			
non-harmful		harmful	
emotion type	no. extr.	emotion type	no. extr.
dislike	49	dislike	56
joy	32	fondness	49
fondness	21	joy	12
relief	18	relief	9
fear	11	anger	8
gloom/sadness	7	gloom/sadness	7
surprize	6	excitement	3
excitement	5	fear	3
anger	3	shame	3
shame	0	surprize	3

Finally, we compared tendencies in the annotated emotion types with regard to the two-dimensional affect space. The results are represented on Table 11.

In the valence dimension, negative emotions were annotated most often on harmful data, and positive emotions, on non-harmful data, which is a reasonable and predictable result. The differences were not

Table 10. Four examples of cyber-bullying entries gathered during Online Patrol from the unofficial school BBS Web site. The upper three represent expressing strong sarcasm despite of the use of positive expressions in the sentence. Emotemes - underlined; Emotive expressions - bold type font.

>>104 <u>Senzuri</u> koite shinu <u>nante</u>, sonna hageshii <u>senzuri</u> <u>sugee</u> <u>naa</u>. "Senzuri masutaa" toshite isshou agamete yaru yo. Analysis results: Emotemes: <u>nante</u>, <u>sugee</u>, <u>naa</u>, -yo, <u>senzuri</u>; Emotive expression: ageru (worship, respect) → fondness ; Translation: >>104 Dying by 'flicking the bean'? Cannot imagine how one could 'flick the bean' so fiercely. I'll worship you forever, as a 'master-bator'.
2-nen no tsutsuji no onna <u>meccha</u> busu suki na hito barashimashouka? 1-nen no ano ko desu yo ne? kimogatteru <u>n</u> de yamete agete kudasai Analysis results: Emotemes: <u>meccha</u>, -yo, -ne, -nde, <u>kimogaru</u>; Emotive expressions: suki (like) → fondness ; Translation: Ya wanna know who likes 2nd-grade ugly azalea girls? Its that 1st-grader isn't it? He's looks disgusted, so leave him mercifully in peace.
Aitsu wa <u>busakute</u> se ga takai dake no onna, <u>busakute</u> se takai dake ya no ni <u>yatara</u> otoko-zuki <u>meccha</u> <u>tarashi</u> de <u>panko</u> <u>anna</u> <u>onna</u> owatteru Analysis results: Emotemes: <u>yatara</u>, <u>meccha</u>, <u>busai</u>, <u>tarashi</u>, <u>panko</u>; Emotive expressions: -zuki (-lover; an amateur of-) → fondness ; Translation: She's just tall and apart of that she's so freakin' ugly, and despite of that she's such a cock-loving slut, she's finished already.
Shinde kureeee, daibu kiraware-mono de yuumei, subete ga <u>itaitashii</u>... Analysis results: Emotemes: -eee (syllable prolongation), ... (clipsis); Emotive expressions: kiraware-mono (disliked by others) → dislike, <u>itaitashii</u> (pathetic, pitiful) → gloom, sadness ; Translation: Please, dieee, you're so famous for being disliked by everyone, everything in you is so pathetic

that obvious, however, after exclusion of fondness, which, as mentioned above, was often used mostly in sarcasm, the differences became clearer. The smallest difference was observed in emotion types which can be classified as both positive or negative, as their valence usually depends on the particular context.

As for the dimension of activation, non-harmful data was annotated as more vivid in the groups of deeply activated and deeply deactivated emotions. On the other hand, harmful data was annotated more often on emotions with moderately activated emotions, which provides another proof for the thesis set on the beginning of this section.

5 CONCLUSIONS AND FUTURE WORK

In this paper we presented a research on cyber-bullying, a new social problem that emerged recently, together with the development of social networking portals, etc. Cyber-bullying consist in sending messages containing slanderous expressions, harmful for other people, or verbally bullying other people in front of the rest of online community. In Japanese society, on which we focused in particular, this problem is particularly vivid on unofficial school Web sites. To handle the problem, teachers and PTA members perform voluntarily Online Patrol to spot and delete the online entries harmful for other people. Unfortunately, there is already an enormous number of cyber-bullying cases and keeps growing. Therefore the Online Patrol becomes an uphill task when performed manually.

Therefore we started this research to create an artificial Online Patrol agent. As the first step, we created a machine learning-based system for cyber-bullying detection and evaluated it. At first, we manually gathered a lexicon of vulgar words distinctive for cyber-bullying entries. To recognize the vulgar words, but written in an informal or jargonized way, we calculated word similarity with Levenshtein distance. As the result, with the threshold equal 2 the system was able to correctly determine the similar words with a 85% of Precision. The Support Vector Machine model trained for discrimination between harmful and non-harmful data achieved the results of 79.9% of Precision and 98.3% of Recall (with balanced F-score = 88.2%). Finally the rank setting of the online entries according to their harmfulness was set using T-score. We were able to eliminate undesirable bias in the rank setting, unfortunately, the procedure is not yet ideal.

Since new vulgar words appear frequently, we need to find a way to automatically extract new vulgarities from the Internet to keep the lexicon up to date. There is also a need for more experiments with

the system including its different variations and improvements (ex. Levenshtein distance threshold optimization).

Looking for clues for further improvement of the system, we performed comparative affect analysis of the cyber-bullying data and normal entries. As a result, we noticed that, that the harmful data were less emotively emphasized than non-harmful. The thesis is reasonable, since the harmful entries are written with premeditation and aim not in expressing ones own emotions, but in evoking in other online community members negative emotions against victim of the cyber-bullying. The results of comparing different dimensions of emotional emphasis suggested the thesis was true, although the reliability of the proof was not satisfactory and further analysis on more robust data is necessary. Another discovery, although an expected one, was that positive emotions appeared more often in non-harmful data and negative emotions appeared more often in harmful data. However, detailed analysis revealed that, especially for fondness, the expressions of positive emotions are often used in strong sarcastic meaning. Therefore there is a need to analyze the data taking into consideration also other dimensions than the valence and the activation of an emotion. As one of such means we plan to apply Ptaszynski et al's [25, 26] system for contextual affect analysis verifying whether an emotion expressed in an utterance is appropriate for its context. We assume this will help in developing a sufficient model of formalization of cyber-bullying activities.

The problems concerning online security have been escalating ever since the birth of the Internet. Some of them are widely known to the society, such as spam e-mails or hacking, however more and more such problems, including cyber-bullying, appear by the day and social consciousness about them is not yet sufficient. There have been developed solutions for some of these problems (e.g., automatic spam e-mail detection, firewall protection, etc.), while others, like cyber-bullying, only keep escalating. Recognizing such problems and developing the remedy is one of the most urgent matters in the field of Artificial Intelligence today.

ACKNOWLEDGEMENTS

This research is partially supported by: a Research Grant from the Nissan Science Foundation; the GCOE Program founded by Japan's Ministry of Education, Culture, Sports, Science and Technology; the Research Project on Development of a Internet Word Book System for Dynamic Following of Information Transition (Project Number: 20500833).

The authors thank Mr. Motoki Matsumura from Human Rights Research Institute Against All Forms for Discrimination and Racism-MIE for providing data from unofficial school Web sites.

Table 11. Comparison of tendencies in annotation of emotion types with regard to the two-dimensional affect space.

valence (polarity) of emotions								
negative emotion			negative/positive (both possible)			positive emotion		
emotion type	non-harmful	harmful	emotion type	non-harmful	harmful	emotion type	non-harmful	harmful
gloom/sadness	7	7	excitement	5	3	joy	32	12
fear	11	3	shame	0	3	fondness	21	49
anger	3	8	surprise	6	3	relief	18	9
dislike	49	56				SUM	71	70
SUM	70	74	SUM	11	9	SUM (fondness excluded)	50	21

activation of emotions								
deactivated emotion			moderately activated (deactivated/activated)			activated emotion		
emotion type	non-harmful	harmful	emotion type	non-harmful	harmful	emotion type	non-harmful	harmful
gloom/sadness	7	7	joy	32	12	shame	0	3
relief	18	9	fondness	21	49	excitement	5	3
			dislike	49	56	fear	11	3
						anger	3	8
						surprise	6	3
SUM	25	16	SUM	102	117	SUM	25	20

REFERENCES

- [1] L. Robinson. Debating the events of September 11th: Discursive and interactional dynamics in three online fora. *Journal of Computer-Mediated Communication*, 10(4), article 4 (2005).
- [2] L. Leets. Responses to Internet hate sites: Is speech too free in cyberspace? *Comm. Law and Policy*, vol. 6(2), pp. 287-317 (2001).
- [3] H. Chen, J. Qin, E. Reid, W. Chung, Y. Zhou, W. Xi, G. Lai, T. Elhourani, A. Bonillas, F.-Y. Wang, and M. Sageman. The dark web portal: Collecting and analyzing the presence of domestic and international terrorist groups on the web. In: *Proc. 7th IEEE Int. Conf. Intelligent Transportation Systems*, Washington, DC, pp. 106-111 (2004).
- [4] A. Abbasi, and H. Chen. Affect Intensity Analysis of Dark Web Forums. In: *IEEE Intelligence and Security Informatics*, pp. 282-288 (2007).
- [5] G. A. Gerstenfeld, D. R. Grant, C. P. Chiang. Hate online: A content analysis of extremist Internet sites. *Anal. of Soc. Issues and Pub. Policy*, vol. 3(1), pp. 29-44 (2003).
- [6] Ministry of Education, Culture, Sports, Science and Technology. 'Netto jou no ijime' ni kansuru taiou manyuuru jirei shuu (gakkou, kyoin muke) ["Bullying on the net" Manual for handling and the collection of cases (directed to school teachers)] (in Japanese). Ministry of Education, Culture, Sports, Science and Technology (2008).
- [7] B. Belsey. Cyberbullying: An Emerging Threat for the "Always On" Generation. http://www.cyberbullying.ca/pdf/Cyberbullying_Presentation_Description.pdf
- [8] J. W. Patchin & S. Hinduja. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence and Juvenile Justice*, 4(2), 148-169 (2006).
- [9] S. Hinduja, & J. W. Patchin. *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Thousand Oaks, CA: Corwin Press (2009).
- [10] H. Watanabe, W. Sunayama. *Denshi keijiban ni okeru yuuza no seishitsu no hyouka* [User's nature evaluation on Bulletin Board System] (in Japanese). *IEICE Technical Report*, 105(652), 2006-KBSE, pp. 25-30 (2006).
- [11] V. I. Levenshtein. Binary Code Capable of Correcting Deletions, Insertions and Reversals. *Doklady Akademii Nauk SSSR*, Vol. 163, No. 4, pp. 845-848 (1965).
- [12] H. Minoru, E. Hirohide. *Nihongo OCR bun ni okeru eiji, katakana no superu ayamari teisei-hou* [Spelling Correction Method for English and Katakana in Japanese OCR Text] (in Japanese). *Transactions of Information Processing Society of Japan*, 38(7), pp. 1317-1327 (1997).
- [13] K. Ryozo, H. Koudai, S. Tatsuya. *ProductionRule wo mochiita shisutemu hyougen to koshou shindan e no ouyou* [Modeling and Fault Diagnosis of Controlled Plant based on Production Rule] (in Japanese). *The Robotics and Mechatronics Conference 2005*, p. 16 (2005).
- [14] V. Vapnik. *Statistical Learning Theory*, Springer (1998).
- [15] T. Hirotooshi, M. Takafumi, H. Masahiko. *Support Vector Machine ni yoru tekisuto bunrui* [Text Categorization Using Support Vector Machines] (in Japanese). *IPSJ SIG Notes*, 98(99), pp.173-180 (1998).
- [16] M. Ptaszynski, P. Dybala, R. Rzepka and K. Araki. Affecting Corpora: Experiments with Automatic Affect Annotation System - A Case Study of the 2channel Forum -. In *Proceedings of The Conference of the Pacific Association for Computational Linguistics 2009 (PACLING-09)*, pp. 223-228 (2009).
- [17] T. Kudo, 'McCab: Yet Another Part-of-Speech and Morphological Analyzer', 2001. <http://mecab.sourceforge.net/>
- [18] M. Ptaszynski, P. Dybala, R. Rzepka and K. Araki. Towards Fully Automatic Emoticon Analysis System ('o'), In *Proceedings of The Fifteenth Annual Meeting of The Association for Natural Language Processing (NLP-2010)*, Tokyo (2010).
- [19] R. L. Birdwhistell, *Introduction to kinesics: an annotation system for analysis of body motion and gesture*, Univ. of Kentucky Press (1952).
- [20] R. L. Birdwhistell, *Kinesics and Context*, University of Pennsylvania Press, Philadelphia (1970).
- [21] M. Ptaszynski. Boisterous language. Analysis of structures and semiotic functions of emotive expressions in conversation on Japanese Internet bulletin board forum - 2channel -. (in Japanese). M.A. Dissertation, UAM, Poznan (2006).
- [22] A. Nakamura. *Kanjo hyogen jiten* [Dictionary of Emotive Expressions] (in Japanese). Tokyodo Publishing, Tokyo (1993).
- [23] A. Zaenen and L. Polanyi. Contextual Valence Shifters. In *Computing Attitude and Affect in Text*, J. G. Shanahan, Y. Qu, J. Wiebe (eds.), Springer Verlag, Dordrecht, The Netherlands, pp. 1-10 (2006).
- [24] J. A. Russell. A circumplex model of affect. *J. of Personality and Social Psychology*, 39(6):1161-1178 (1980).
- [25] M. Ptaszynski, P. Dybala, W. Shi, R. Rzepka and K. Araki. Towards Context Aware Emotional Intelligence in Machines: Computing Contextual Appropriateness of Affective States. In *Proceedings of Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09)*, Pasadena, California, USA, pp. 1469-1474 (2009).
- [26] M. Ptaszynski, P. Dybala, W. Shi, R. Rzepka and K. Araki. Contextual Affect Analysis: A System for Verification of Emotion Appropriateness Supported with Contextual Valence Shifters. *International Journal of Biometrics*, 2(2), pp. 134-154 (2010).

An idea of a web-crowd based moral reasoning agent

Radoslaw KOMUDA¹, Michal PTASZYNSKI², Rafal RZEPKA² and Kenji ARAKI²

Abstract. In this paper we firstly discuss some of the presumptions about human conscience. By a detailed guidance throughout our ongoing research on morality judgment categories we make an attempt to summarize it and its usefulness for creating an explicit moral reasoning agent with system based on the wisdom of the web-crowd. With special remarks on giving it a mathematically calculative structure and chances created by this approach alteration.

1 INTRODUCTION

Even though we like to consider robots as “mechanical”, “artificial” – we still are keen to believe that theories developed for humans might work just as fine for machines as it does for us. And this has lately lead to the discussion about most useful philosophical system (with Kantianism and utilitarianism in the lead) for the Machine Ethics research. We believe that this approach is a promising path but moreover we think that field of science with such an interdisciplinary character cannot be limited to philosophical deliberations. That is why in our research we try to draw from other disciplines like psychology [1].

While working on a project of a survey for arithmetical morality calculation and with promising results in our preliminary experiments on human respondents we asked a question about making machines able of self-filling the questionnaire. This paper is the answer.

Internet is brimful of social networking (Facebook, MySpace) and video sharing (YouTube) websites. People nowadays like to share their moments with friends by both instant, short messages (Twitter) and long, more balanced ones (blogs) or photos (Flickr, Picasa). They comment on recent news which seem to be flooding the Web. We believe that these unfathomable Web 2.0 features may be hiding even more immense crowd wisdom potential that can be used in Machine Ethics research.

2 CATEGORIES FOR DIFFERING GOOD (+) FROM WRONG (-)

With a common sense knowledge you can easily tell that a rewarded action has to be good and if punished – one has made something wrong. But of course our lives are far more complex

and since assassins get paid (reward) for killing people (wrong) and world is full of accidental (undeserved) victims (punishment) we need more categories or a more sophisticated moral reasoning manner. A ready to use package of essential issues to consider while recognizing good from wrong we found in Lawrence Kohlberg’s theory of human moral development.

2.1 KOHLBERG’S THEORY AND ITS USEFULNESS FOR MACHINE ETHICS RESEARCH

Lawrence Kohlberg was an American psychologist whose research on human morality resulted in a theory [3] of its successive changes in aspects by which we consider an action good or wrong.

He discovered that when we are young – we firstly look on the punishment that the action will cause. The worst punishment (-), the more wrong has to be the action itself. If a child gets grounded for a day for killing his hamster and for a week for eating sweets before dinner – it is going to consider having snack worst than killing. Furthermore, this might lead to a misapprehension in which suffering of accidental victims would be considered adequate to their fault. Surprisingly, thinking only about the reward (+) yet indicates second stage of our moral development on which we do not care for being punished as long as we get what we care about – we will eat the candy again even though we are aware of the threat of the punishment. Although both share same, self-interested (-) concerns about consequences (first category) – they undoubtedly make us ask about Actor’s motivation (second category) in search for more altruistic (+) behaviors such as stealing a car to drive somebody to the hospital. But yet let us not connect this with Actor’s good (+) or evil (-) intentions (third category) which allow us to distinguish unintentional killing from murder.

Later we wish to ask about Actor’s reputation (fourth category) with him being criticized (-) or acclaimed (+). Also some social factors like direct reaction to the act itself (fifth category) – disapproving (-) or appreciating (+) it and improvement (+) or deterioration (-) of relationships (sixth category). Differing these categories might be even harder and judging by them – more complex to recall some law and social discrepancy when it comes to cases of lynching.

Finally we get back to things that can be easily judged like telling if the law or etiquette (seventh and eighth category) has been broken (-) or kept (+).

¹ Faculty of Theology, Nicolaus Copernicus University, 87-100 Torun, Poland. Email: komuda@stud.umk.pl.

² Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan.
E-mail: {ptaszynski, kabura, araki}@media.eng.hokudai.ac.jp.

3 ARITHMETICAL MORALITY CALCULATION

Though conscience is often referred to be the "voice within" and the "inner light" [2] – we are keen to believe that it is a matter of reason and because of that we can assume judging the moral quality of an act a logically successive process. That as any other deduction may be eventually expressed in (by all means – not "reduced to") the language of mathematics with – as for our research – results of '1' and '-1' for acts morally good and wrong respectively and '0' for the neutral ones. The remaining issues are the categories and scale in which the action is ought to be rated.

4 QUESTIONNAIRE FOR MORALITY CALCULATION

The basic idea for our chart is judging an action by the categories distinguished in part three of this paper in a basic scale from '-5' (–) to '5' (+). We believe that by a proper mean calculation machine should be eventually able not to tell good from wrong but also solve more complex cases.

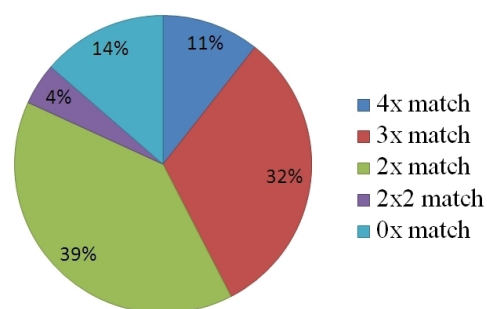
4.1 QUESTIONED SITUATIONS

For our research we have prepared a set of twelve situations that we ask respondents to judge: stealing a car, saving a man, saving men, son marries the woman he loves, being killed, killing a sick dog, kill own dog, man's strong hit, being robbed 10\$, being robbed 100\$, killing a man accidentally, aving a man accidentally. This set includes not only events completely opposite like saving man's life and killing him but also contain situations with different intensity, e.g. reflecting the difference in ratings due to the intensity of the acts: robbing 10\$ or 100\$.

In our preliminary experiments to test the credibility of the questionnaire we asked four respondents familiar with our research (to avoid misunderstanding) to fill it in for all of the situations. Figure 1 illustrates achieved matches with almost 40% of questions about categories presented in section 2.1 getting a semi-accordance (A, A, B, C) and more then every third – a 75% agreement (A, A, A, B) on a "-5" to "+5" scale.

4.2 ACTION'S EMOTIVENESS

Even though intensity of an act is important – we think that the key and most useful factor for our research is act's emotiveness and therefore we included situations of killing a dog with certain modifiers like "own dog" or "sick dog". But – we must not forget that field of Machine Ethics is not about making robots capable of searching information about an action to find direct, modifying circumstances but only judge the action on the basis of provided knowledge [4].



- 4x match, e.g. A, A, A, A - 14 / 132
- 3x match, e.g. A, A, A, B - 42 / 132
- 2x match, e.g. A, A, B, C - 52 / 132
- 2x2 match, e.g. A, A, B, B - 6 / 132
- 0x match, e.g. A, B, C, D - 18 / 132

Figure 1. Diagram presenting preliminary query matches.

Nowadays people are more keen to comment news by which they are moved. Anonymity in the Net allows and encourages to be frank with expressing our stand and by that – changes the Web into a vast source of different opinions on various topics.

4.3 PARTICIPANT CATEGORIZATION

When judging a situation we must not forget that it can always be seen from three points of view. Already recalled Actor – direct doer of the action, Object – on which the action is being performed and an Observer watching the event. For formality let us state that, e.g. 'hitting' is more pleasant than "being hit" and "winning a lottery" is far more fun than "seeing winning a lottery". That is why for our research we have decided to use a two dimension Cartesian coordinate system (see: Figure 2.) with x-axis responsible for positive and negative action consequences and y-axis picturing both pain and pleasure rate what makes seeing the difference far more clearly.

Second concern involves determining the nature of action's participants because while "kicking a ball" is rather acceptable – "kicking a cat" may not. We want a machine to be provided with a system using such categories and already had been able to make our program capable of abstract these nuances.

4.4 WEB MINING THE EMOTIONS

Next step for developing web-crowd based moral reasoning Agent concept was to adapt previous ideas [5, 6] for the new usage – making the Agent capable of not only finding proper data but also organizing and summarizing it to give a reasoned moral judgment. One of the most important changes made by this for hitherto approach was expansion of the categorization idea. Since we prefer practical approach, we wish to present results of our other experiments in mechanized search for emotions associated with actions.

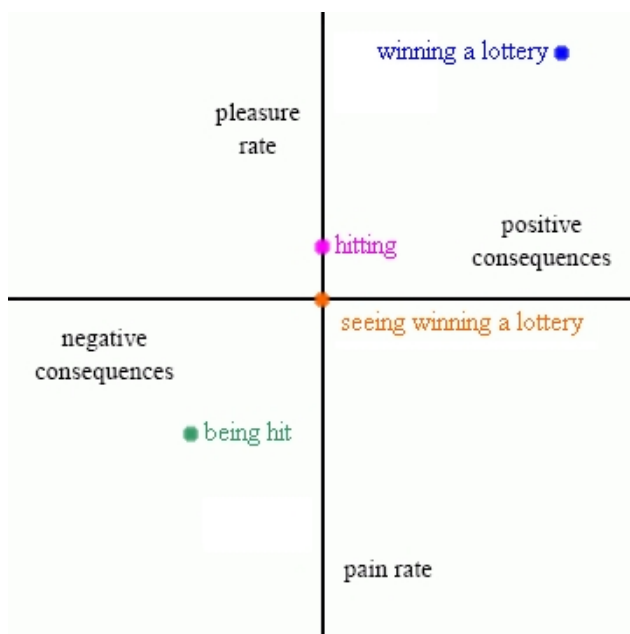


Figure 2. Cartesian coordinate system illustrating differences discussed in section 4.3.

Below tables show top approximate emotions extracted from sentences in Japanese found on the web with searched phrases:

Emotions:	Results:
fear	32 / 97
excite	16 / 97
sadness	15 / 97

Table 1. Results of a query: “has killed man”.

Table 1 shows the results for a query “has killed man” for which our program was able to find 97 opinions in our blogger's utterances data-base and out of which about 33% find this phrase fearsome.

Emotions:	Results:
dislike	36 / 110
anger	16 / 110
shock	13 / 110

Table 2. Results of a query: “has killed mother”.

For Table 2 we changed Object for a more emotional one what resulted in having 59% showing an unhesitating disapproval.

We believe that with a common sense judgment and a daily life experience one may call these results at least promising. With use of this criteria we make an assumption on possible queries' results and results showing 84% of positive emotions

association with “has marry someone” or relating “has hit a man” with emotions like “sadness” and “anger” make us believe in chances and perspectives of such a moral reasoning Agent.

Emotions:	Results:
joy	88 / 99
like	8 / 99

Table 3. Results of a query: “has marry someone”.

Emotions:	Results:
sadness	17 / 49
anger	10 / 49

Table 4. Results of a query: “has hit a man”.

5 FUTURE WORK

In the near future we wish to release an Internet version of the questionnaire to build a data-base of moral judgments made by human respondents. By comparing it with result given by our Agent we should be able to improve its scale, its specific sensitivity and eventually – its efficiency.

Furthermore we wish to extend our research with emotive modifiers. Not only possessive pronouns like “my” or “your” but also with verbs since our subservient queries in that field showed our Agent being able to differ “losing”, “stealing” and “taking away” one's fortune.

REFERENCES

- [1] R. Komuda. Kohlberg's Theory of Moral Development Stages as a Path to Follow In Machine Ethics Research. In: Language Acquisition and Understanding Research Group (LAU) Technical Reports, Winter 2009, pp. 6-10, Sapporo, Japan. (2009).
- [2] Rosemary Moore. The Light in Their Consciences: The Early Quakers in Britain 1646–1666. Pennsylvania State University Press, University Park (2000)
- [3] Kohlberg, Lawrence. Essays on Moral Development, Vol. I: The Philosophy of Moral Development. San Francisco, CA: Harper & Row. (1981).
- [4] Anderson, M.; Anderson, S. L. *Machine Ethics: Creating an Ethical Intelligent Agent*. [in:] AI Magazine, vol. 28, number 4, 15-26 (2007).
- [5] Wenhan Shi. *Discovering Emotive Content in Utterances Using Web-mining* (in Japanese). Hokkaido University. (2008).
- [6] Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka and Kenji Araki: “Disentangling emotions from the Web. Internet in the service of affect analysis”. Proceedings of the Second International Conference on Kansei Engineering & Affective Systems (KEAS'08), pp 51-56, Nagaoka, Japan. (2008).

Learning a Grounded Language Model for Human-Robot Interaction

Antonio Chella and Haris Dindo and Daniele Zambuto¹

Abstract. Human-robot interaction is one of the most active fields in robotics research. Robots are expected to populate our homes and offices in the near future. However, their linguistic capabilities are still far from being satisfactory. Most systems lack the capability to adapt and expand themselves beyond the preprogrammed set of communicative behaviors, which makes them unsuitable in real-world applications.

In this paper we will present a system in which a robotic agent can learn a grounded language model by actively interacting with a human user. The model is grounded in the sense that meaning of the words is linked to a concrete sensorimotor experience of the agent, and linguistic rules are automatically extracted from the interaction data. The system has been tested on the NAO humanoid robot and it has been used to understand and generate appropriate natural language descriptions of real objects. The system is also capable of conducting a verbal interaction with a human partner in potentially ambiguous situations.

1 Introduction

Visually grounded human-robot interaction is recognized to be an essential ingredient of socially intelligent robots, and the integration of vision and language increasingly attracts attention of researchers in diverse fields: from robotics and artificial intelligence to computational linguistics and neuroscience [3]. However, in this paper we will not concentrate on the various facets of human-robot interaction research, but concretely on the problem of how to associate spoken words with their perceptually grounded meanings while acquiring a simple language.

The aim of this work is to investigate the lexical acquisition problem, namely how can a robot be bootstrapped into communication and what are the necessary prerequisites for robots in order to learn a language? In particular, the focus is set on grounded systems that learn to generate and understand contextualized spoken descriptions of objects in visual scenes. There are two basic problems the robots need to solve as they acquire a language:

1. Identify the meaning of words grounded in perceptual data;
2. Infer a rudimentary grammar for further understanding and interaction.

The process of lexical acquisition in infants seems to be innately driven by the principle of reference: words refer to objects, actions, and attributes of the environment. Observational learning may be used to deduce word meanings from cross-situational experiences. Joint attention plays an important role in learning terms of reference.

Infants are more likely to connect words with their referents when engaged in joint attention with their caregivers [2].

A well-known problem in observational learning, is the Quine's paradox: an infinite number of possible meanings can be inferred from a finite set of utterance-context pairs. A likely solution to this problem is that all infants have certain biases which constrain the set of possible meanings of words [4, 1]. For example, the whole object assumption proposes that children will assume a novel label refers to a whole object rather than its parts. The mutual exclusion assumption proposes that they prefer to assign only one label to a concept. These assumptions are considered good strategies for bootstrapping the inference process.

Present approach uses these assumptions to bootstrap the lexical acquisition process. The robot must acquire the possible meanings of words from their non-linguistic (perceptual) input, and determine which co-occurrences are relevant from a multitude of potential co-occurrences between words and entities in the environment while acquiring syntactic rules. Initially, the system will acquire a minimal grounded language model from the paired *description-referent* examples, without any prior semantic and syntactic information, except the main referent of description utterance, from which the system will be able to acquire novel concepts and more complex language models (e.g., spatial clauses).

However, the ultimate goal of the proposed system is to take advantage of acquired concepts and language model to engage in simple dialogue with a human partner. All concepts underlying acquired language model are used to initialize dynamic fluents as predicate calculus terms and update robot's knowledge base representing the state of the world from sensor data. Language acquisition therefore proceed in parallel with concept acquisition. Concepts acquired from lexical acquisition are used to initialize a logic representation of several observable entity of world. For example, the meaning of the word "red" is used to seed the logic representation of the "red" concept. Concepts underlying acquired language model can be considered as independent from language acquisition process and can be reused for other cognitive tasks. Fig. 1 provides an overview of the learning system.

The system has been tested on the NAO humanoid robot and it has been used to understand and generate appropriate natural language descriptions of real objects. The system is also capable of conducting a verbal interaction with a human partner in potentially ambiguous situations, e.g. when a human interlocutor refers to several co-occurring entities (i.e. objects sharing identical perceptual properties) in the world.

¹ University of Palermo, Italy, email: {chella,dindo,zambuto}@dinfo.unipa.it

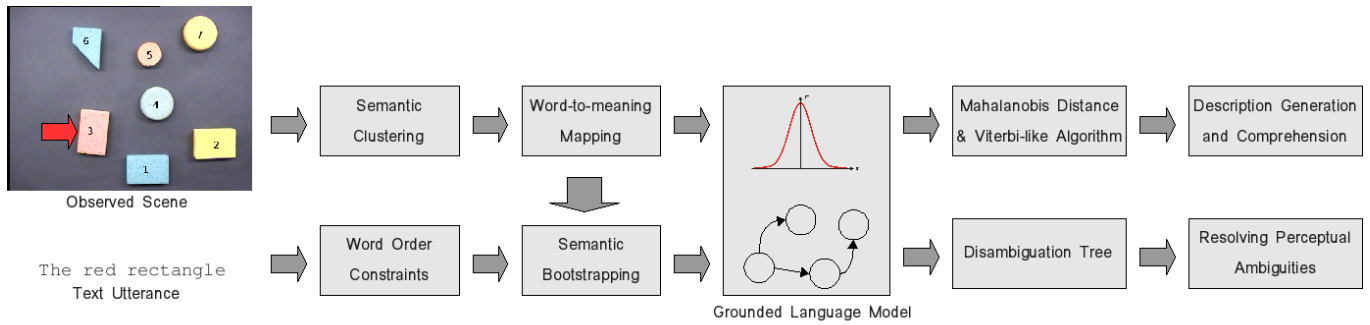


Figure 1. Model of grounded language acquisition: from word learning to simple language and dialogue.

2 Related works

There has been a huge interest in grounded language acquisition in the past years. Probably, the earliest and the most influent work on language analysis is Winograd's SHRDLU [12] in which an artificial agent connected to a simulated robotic arm is capable of understanding questions on the world, execute given actions, and ask for help in case of ambiguous dialogue. However, SHRDLU is based on a purely symbolic representation of the world, and the artificial agent has no cognition about the "real" meaning of a given sentence. It is mostly based on a pure syntactic analyzer, and the semantics is deeply hardwired into the system.

Visual Translator system [7] (VITRA) is a natural language generation system which is grounded directly in perceptual input. From a sequence of digitized video frames low-level sensory processes perform recognition and tracking of visible objects. Authors provide a geometrical reconstruction of the perceived scene. Detailed domain knowledge is used to categorize spatial relations between objects, and dynamic events. Higher level propositions are formed from these representations which are mapped to natural language using a rule-based text planner. In contrast to other works, VITRA is not designed as a learning system.

Roy implemented a system, CELL [9] (Cross-channel Early Lexical Learning), able to learn object names from a corpus of spontaneous infant-directed speech, and to process single and two-word phrases which referred to the colour and shape of objects. CELL seems to be the first model of language acquisition which learns words and their semantics from raw sensory input without any human-assisted preparation of data. In DESCRIBER [10], the same authors address the problem syntactic structure acquisition within a grounded learning framework. Learning algorithms acquire probabilistic structures which encode the visual semantics of phrase structure, word classes, and individual words. Using these structures, a planning algorithm integrates syntactic, semantic, and contextual constraints to generate natural and unambiguous descriptions of objects in novel scenes.

Another related approach is TWIG [5], a word learning system that allows a robot to learn compositional meanings for new words that are grounded in its sensors. TWIG allows a robot (1) to learn the meanings of deictic pronouns, (2) to contrast new word definitions with existing ones, thereby creating more complex definition, and (3) to use word learned in an unsupervised manner for production, comprehension, or referent inference. The techniques that TWIG intro-

duces are extension inference and word definition tree. Its technique are more generally applicable to other word categories, including verbs, prepositions and nouns. In another related approach, authors develop language acquisition system without knowledge of grammar and vocabulary, which learns concrete noun concepts from user utterances and paired images [11]. The system adopts a frequentist approach, in which utterances and images containing the same objects are processed and the most probable noun is chosen as label.

Our work, while not making significant advances compared to the systems at the state-of-the-art, puts more emphasis on a fundamental problem in the language acquisition process, namely the selection of the referent. Our goal is to create robust grounded language learning algorithms, able to build operational knowledge even in absence of important pragmatic information.

3 Model overview

Before going into details of the problems addressed in the present work we provide operational definitions of several terms which are used throughout this article. A *semantic category* (or semantic unit) specifies a range of sensory inputs which can be grouped and associated with a word/symbol. For example, a semantic category might specify a portion of the colour spectrum. Such a semantic category could be used to ground the semantics for a colour term such as "red". A *semantic class* specifies a set of semantic categories grounded in the same sensory channel. For example, a semantic class could be used to associate acquired colour terms (colour class). Finally, a *lexical item* encodes the association between a word and its corresponding semantic category.

In the experimental setting each training sample is comprised of an utterance and a visual context representing the semantics of the word sequence. Utterances consist of phonetic transcripts of spoken sequences recorded by the auditory sensor. Context consists of visual sensory input which co-occurs with the utterance, and usually it contains instances of multiple semantic categories. A key assumption is that any word in the utterance may refer to any semantic category inferred from the co-occurring context.

All semantic categories are derived from visual sensory signals. Feature extractors computes visual features from the sensors (video). Each extracted feature encodes relevant, non-redundant, information from the visual sensory stream about observable proprieties of the world (semantic class). Potential visual features include categories of shape, colour, size, and spatial relation. The context is initially re-

stricted to a given referent (target) object, and any word may potentially be paired with any semantic category which is derived from the same utterance-context pair. A typical teaching scenario is depicted in Fig. 2.

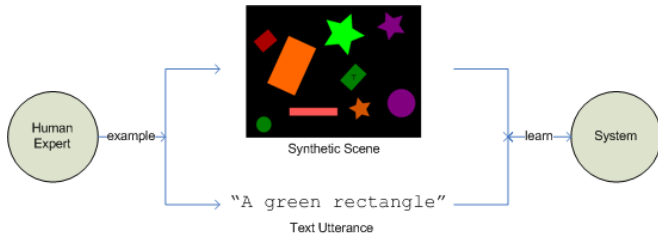


Figure 2. A typical teaching scenario for the acquisition of a visually grounded language.

3.1 Learning how to ground words to perceptual stimuli

Grounding can be considered as the process whereby internal representations are connected to external percepts. This is based on the principle that cognitive agents and robots learn to name entities, individuals and states in the external (and internal) world at the same time as they interact with their environment and build sensorimotor representations of it [6]. Our approach addresses two interrelated questions in lexical acquisition: how semantic categories can be learned and how symbols/words can be associated with appropriate semantic categories. These problems are known as semantic categorization (or *semantic clustering*) and *word-to-meaning mappings* respectively. No *a priori* knowledge about innate semantic categories is assumed by the model; instead, through repeated experience, appropriate categories must be learned from positive examples and the model must associate linguistic units with appropriate semantic categories. This process is depicted in Fig. 3.

Basic problem regarding semantic clustering is how to establish the semantics of individual words, since each word in an utterance may potentially be associated with any subset of co-occurring visual features. For the sake of simplicity, we can assume that each word may be associated with only one of co-occurring semantic categories. However, natural language usually does not provide exhaustive labels of all referents in a scene. In other words, we cannot assume that the absence of a word indicates the absence of the corresponding property. As a consequence, appropriate categories must be learned from the positive examples only.

In the present approach, the meaning of each word (i.e. its semantic category) is treated as a random variable and modeled with a multivariate Gaussian distributions. For example, the meaning of a word associated to the semantic class 'color' is associated with a multivariate Gaussian distribution over red-green-blue colour space. These densities are incrementally estimated from contexts which co-occur with a word. We must quantify the effect of the presence of each word on the distribution of each extracted feature. A semantic distortion metric is used to select appropriate semantic category from several hypothetical ones. It provides the divergence between the distribution of feature values conditioned on the presence of word $p(f^j|w)$ (hypothetical semantic category) and the distribution of feature values conditioned on the absence of word $p(f^j|\bar{w})$ (background dis-

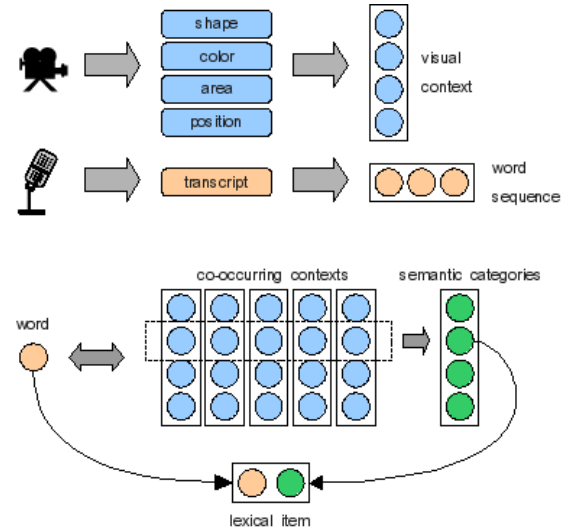


Figure 3. The model must associate words with appropriate semantic categories. These associations must be inferred from collections of experiences. Hypothetical semantic categories are estimated from contexts which co-occur with a word.

tribution). Only the observations f which occurred with the word w are used to compute the unbiased estimates of the word conditioned model. The remaining observations (i.e. visual context of examples in which the word is not present) are used to estimate the background model. These distributions are estimated for each semantic class C_j (shape, colour, size, spatial relation).

We use the distortion between the word-conditioned and background distributions as a measure of association between word and semantic categories. The Bhattacharyya distance measures the similarity of two probability distributions, and the optimal Bayes classification error between any two classes can be bounded by the Bhattacharyya error, ϵ . In our case, the Bhattacharyya error provides a measure of association between words and individual semantic categories. A clustering algorithm estimates hypothetical semantic categories of words from co-occurring contexts and associates each word with semantic categories that maximize the classification error outlined above². Categories below a fixed threshold are discarded, and the word is associated to a special ungrounded class. Ungrounded class gathers all the words that can not be directly grounded in sensory input. In our model, no semantic category (Gaussian density) is associated with these words. The whole process, together with the definition of the involved quantities, is depicted in the Fig. 4.

3.2 Grammar learning

Categories, such as adjective, verbs and nouns, form the basic units for learning the rules of grammar including grammatical relations, cases, and phrase structure configurations. Without syntactic categories a learner will be unable to acquire the rules of the language. Bootstrapping theories provide strategies for deriving syntactic categories from perceptual data. For instance, semantic bootstrapping [8] proposes that the language learner uses semantic categories to seed

² We force the clustering algorithm to associate only one semantic category, without considering multi-class meanings (semantic categories that depend on several classes of perception)

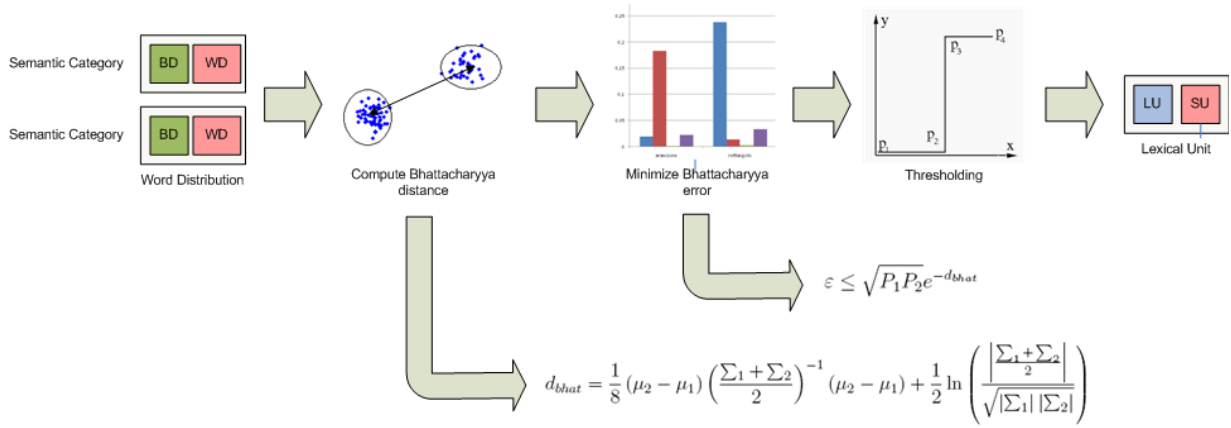


Figure 4. The model must associate words with appropriate semantic categories (BD=background distribution; WD=word distribution). These associations must be inferred from collections of experiences. Hypothetical semantic categories are estimated from contexts which co-occur with a word (LU=lexical unit; SU=semantic unit).

syntactic categories. This theory assumes that the learner has already acquired words and their semantics without the use of any syntax.

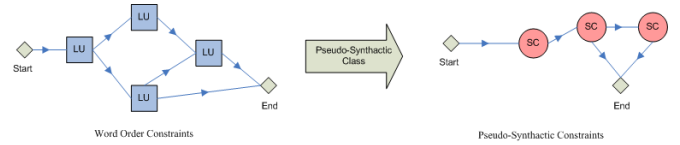
Basic syntactic categories will be acquired directly through words' meaning, and the words associated with similar perceptual categories will be included in same syntactic category. The grammar is modeled as a Markovian chain. We collect word statistics from utterance training corpus encoding: (1) the word transition probability, (2) the probability of beginning an utterance with a word and (3) the probability of ending an utterance with a word. Lexical items generated in the previous phase can be used to cluster words into groups that depend on the associated meaning. We use these groups, and associated probabilities, to construct a more general representation of syntactic constraints in terms of a finite state automaton (FSA) based on semantic classes. These FSA are used as the basis for a deterministic parser which identifies object phrases embedded in complex utterances (Fig. 5a).

3.3 Learning complex spatial concepts

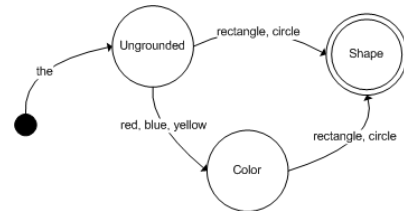
However, complex adjectives, such as “above” and “below”, can not be learned in the same fashion. Instead, we need to take the order of words into account. Treating sentences as unordered set of words tends to lose syntactic contents and its semantic implications. We need to exploit the word ordering constraints and semantic bootstrapping assumption to acquire spatial terms and explore weak generalization mechanisms.

To learn the meaning of spatial terms, we need to find all the objects to which the phrase refers. We consider only those phrases describing spatial relationships between two objects (i.e., “the rectangle above the red circle”). A deterministic parser based on previously acquired FSA, identifies object phrases embedded in complex utterances. We must decide which of two phrases refers to the original target object, and which of the remaining objects in the scene should be linked to the other phrase. We have defined a fitting function which measures the similarity of an utterance to an object based on Mahalanobis distance, which will be explained later. Once the focal objects of the sentence has been determined, we can calculate the features that describe the spatial relationship between these. We must also encode the order of the phrases to learn the difference between spatial terms (i.e., difference between “above” and “below”)

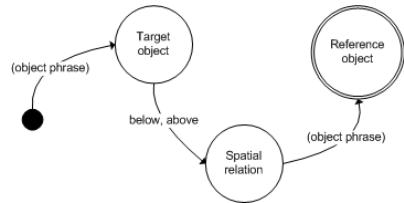
and to calculate correctly the spatial features. The parser replace the object phrases with (ungrounded) labels to facilitate the task. The only information that really matters in our case, is which of the two sentences refers to the target object. Finally, the learning process that associate a meaning to the word describing a spatial feature continues as previously described. New lexical items are simply added to the vocabulary of model, and the learned FSAs are merged into a single syntactic model (Fig. 5).



(a) Main components of the FSA representing a grammar (LU is a lexical unit, SC is a syntactic class)



(b) An example of FSA for simple sentences



(c) An example of FSA for spatial clauses

Figure 5. Finite state automaton (FSA) based on semantic class are used to encode syntactic constraints of object phrase and spatial clauses. These FSAs are used as the basis for a deterministic parser to analyze complex utterances.

3.4 Understanding and generating descriptions

Having learned the probabilistic grounded language model, the task of understanding an utterance can be casted in terms of classical operations on probabilistic graphical models. Thanks to the Mahalanobis distance, we can measure the degree of association ("semantic distance") between a visual feature and a semantic category previously acquired. For example, we consider the case of a sentence like "the red rectangle". The acquired lexical items allow the parser to assign a semantic category (and hence a Gaussian density) to each word of the sentence. Given the visual features of an object, we can calculate the degree of association between sentence and object's visual features as:

$$\sum_i \sqrt{(x - \mu_i) \Sigma_i^{-1} (x - \mu_i)}$$

The object of the scene that minimizes this measure is selected as a possible referent of the sentence. The same procedure can be used for understanding spatial clauses. The deterministic parser guides the extraction of spatial features which depends on the order of object phrases. The referent of the sentence is searched among all possible pairs of objects.

One way to facilitate the understanding of descriptions and to decrease the computational burden is to use logical terms to represent the state of the world (i.e., knowledge about the properties and spatial relationships of objects), and then use query in a logic programming language for inferring the reference object. When the parser begins to analyze an utterance, it queries the system for the state of the world described in the form of a list of all atomic sentences that can be produced from available sensors at the time of query. For each object in the scene, some logical terms are generated through the acquired lexical items. In particular, the system generates a logical term for each visual feature extracted from an object by selecting the lexical item that minimizes the Mahalanobis distance for each semantic class. Deterministic parser has been modified to generate logical query to be submitted to the system. Some examples of possible queries are: "red(X)", "rect(X)", "above(X,Y)", "red(Y)", etc.

The generation of the description of an object is quite simple. We can generate a description as the most likely path of words (in the main FSA) that produces a given observation. Given a set of visual features extracted from an object we generate a description of object through a modified Viterbi-like algorithm. We seek the sequence of words that minimizes the fitting function for that object, and follows the syntactic rules implicitly encoded in the FSA. We need a method to determine the length of a description. The generation algorithm include a stop criterion based on the acquired ending probability.

4 Dialogue and ambiguities

Another problem we have addressed is how to resolve the ambiguities contained in a description through simple verbal interactions. For example, suppose that the robot hears the statement "Grasp the small blue object!" while facing a scene containing several objects having the same perceptual features.

In order to fulfill the task, the robot must disambiguate the context and select one of the referred objects. The process of ambiguity resolution is based on a context-specific human-robot dialogue through specific questions directly related to the properties and spatial relationships of the objects in the observed scene. We have considered only those ambiguities related to perceptual quality shared between objects of the scene. The robot must be able to detect ambiguous

statement and select the least ambiguous question to be asked to. If all the red objects have different shape, the robot can resolve easily the ambiguities asking for the shape of a specific object, i.e. "Is the rectangle?". However, if all items have the same shape, the robot must solve the ambiguity by using relative spatial terms between objects, that usually are more ambiguous than the previous. For example, some objects may share particular spatial relationships with other objects (i.e. more objects below the yellow rectangle), making the question partially ambiguous, or even the phrase could not be fully understood (either by a human or by the robot) because the landmark object is indistinguishable from the target.

Possible ambiguities can be removed through a simple dialogue based on yes/no questions. *Disambiguation tree* reconstruct the robot's decision process in choosing a question to disambiguate a context. They are essentially decision trees: the possible target objects are stored in the leaves, a disambiguation strategy is given by the path from the root to the object's leaf. The interior nodes can be questions about referent's proprieties itself, or relations to other objects. When the robot finds a reference ambiguity, it builds a disambiguation tree that resolves it. Figure 6 shows an example of disambiguation tree. Path to the left after question indicate that the predicate is satisfied (yes answer), while the right branch indicates that it is not (no answer). Each question consists of attempting to satisfy a logical predicate, and appropriate subtree is explored. This process continues until a leaf is reached and the ambiguities is resolved.

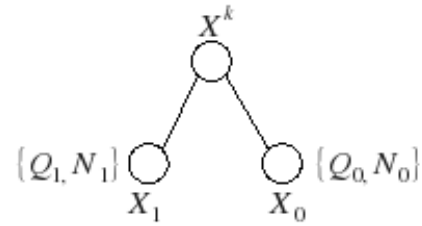


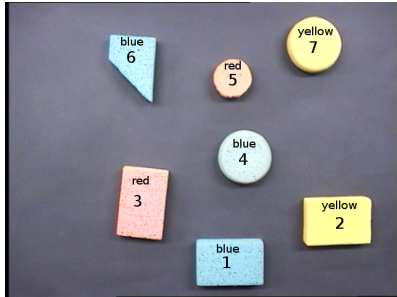
Figure 6. Disambiguation tree reconstruct the robot's decision process in choosing a question to disambiguate a context.

Disambiguation tree are constructed using the output of an ambiguous query (logic queries with more than one solution). The construction method minimizes an entropy-based disambiguation measure. At each decision node, the algorithm chooses a possible question and splits the available objects into two groups. This process then occurs recursively until all objects are selected. Clearly, the algorithm must decide which of these question is most informative. We have used a entropy-based disambiguation measure, characterizing the average amount of uncertainty in a single question. This measure is computed as:

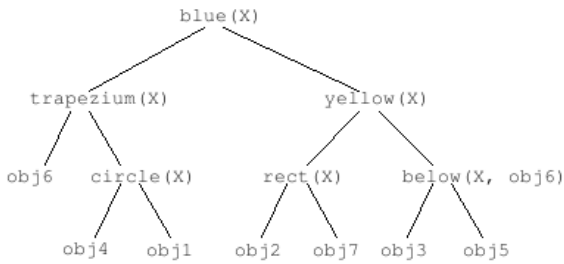
$$\frac{N_0}{N_1} H(X_0) + \frac{N_1}{N_0} H(X_1) + \sum_i C_i(X^k)$$

The first two members of the equation measures the uncertainty related to possible questions in the children nodes. We have calculated the uncertainty linked to the set of possible questions Q in a node, as the entropy calculated on a random variable X generated by this set. The last term measures the uncertainty related to the question selected, depending on the encoding of the question in natural language (i.e. the number of words used) and its contextual ambiguity (i.e. ambiguity in the understanding of the sentence by a listener).

The greedy algorithm selects the question that minimizes this measure based on entropy. The trees generated by the disambiguation algorithm, are well-balanced. Figure 7(b) shows an example of disambiguation tree generated by the algorithm. In this example, the first selected question is the term “blue(X)”. The question “Is blue?” allows the system to divide the objects of the scene into two groups: blue objects and not blue objects. The construction method consider future scenarios, and hence possible question and ambiguities in future contexts, to select appropriate question. The method does not select the questions that have an immediate and obvious answer (i.e. “trapezium(X)”), but more general questions that allow us to isolate and partially resolve the perceptual ambiguity and reduce the number of questions to ask. The subsequent scenarios, as mentioned, are those with a lesser degree of ambiguity.



(a) A sample image from training corpus



(b) A disambiguation tree inferred from the scene

Figure 7. Disambiguation trees are used to resolve the ambiguities contained in a description through simple Yes / No questions. A greedy algorithm generate well-balanced trees by minimizing an entropy-based measure.

5 Experiments

We have performed two set of related experiments in order to test the proposed model. One set involves a quantitative evaluation of the model in the task of describing a scene. Another set is mostly qualitative and involves an experiment on the NAO robotic platform. Both experiments will be described in the following

5.1 Object description: quantitative results

The description task consists of generating phrases which best describe target objects and must be context sensitive since often depend on the other objects in the observed scene. The variation of objects is limited to shape, color, size and position. Each training example is comprised of an utterance and a context representing the semantics of the word sequence. Utterances consist of phonetic transcripts

of spoken sequences recorded by the auditory sensor. Context consists of visual sensory input which co-occurs with the utterance, and usually it contains instances of multiple semantic categories. The experimental setting consists of a set of objects of different shape and color placed on a table. The camera is placed above the table and it ensures a comprehensive view of the scene.

A training corpus from two participants unfamiliar with the project has been collected. Participant were asked to generate two different utterances related to the observed scene such that a listener could later select the same target from the identical scene with the target unmarked. Simple utterances contain reference to exactly one object (target object), whereas complex utterances describe two or more objects and their spatial relations. The reference object of a spatial relation was selected by participants during the data collection task. The training corpus was composed of 356 utterances of which 196 are simple and 160 complex.

Participants were asked to select the object which best fit the description generated by the system. All participants evaluated the same sets of images. Responses were evaluated by comparing the selected object for each image (the target object described by the system) to the actual target object which was selected by participant as the referent of the description. The collected data was used to measure the accuracy of the system-generated description (showed in Table 1). Participants were also asked to select a landmark object and to generate a detailed description of the target object in the scene. Responses were evaluated by comparing the selected objects for each image to the actual target object and landmark object which were selected by the system. Table 2 shows the result of evaluation of system understanding capabilities.

Table 1. Results of an evaluation of human and machine generated descriptions.

Participants	Generated by human	Generated by system
A	87,3 %	80,5 %
B	85,5 %	79,8 %
C	88,3 %	81,3 %
Media	87,1 %	80,5 %

Table 2. Results of an evaluation of machine understanding capabilities.

Participants	Training corpus	Testing corpus
A	90,9 %	87,3 %
B	88,7 %	85,5 %
C	91,1 %	88,3 %
System	86,2 %	78,9 %

The system was able to describe scenes not encountered during the training phase, and to exhibit sequences of words which have never occurred in the training data. The results presented in this section demonstrate the effectiveness of the learning algorithms to acquire and apply grounded structures for the visual description task.

5.2 Human-robot interaction: qualitative results

As previously mentioned, the system has been tested on the NAO robotic platform. NAO is a humanoid robot equipped with Force Sensitive Resistors (FSR) located on the feet, sonars, bumpers, tactile sensors, an IR emitter/receiver, a stereo camera and a pair of microphones. The robot has a number of built-in machine vision modules used in the experimental setup. In addition, we have implemented



Figure 8. An example of human-robot interaction via the learned language model

a set of perceptual and motor schema for basic behaviors such as pointing and grasping.

All concepts underlying acquired language model are used to initialize dynamic fluents as predicate calculus terms and update robot's knowledge base representing the state of the world from sensor data. Only the actions of the robot can modify the values of the fluent associated with the objects. For this reason, the knowledge base is updated after every action of the robot. Obviously there are several logic terms that are constant over time (e.g. color). We have tested the capabilities of the robot to understand the descriptions provided by the users and to conduct a dialogue in case of ambiguities. The robot was given concrete instructions, such as “Point the green object!”, or “Grasp the object to the left of the yellow circle!”³. The previously described disambiguation trees were used to clarify eventual ambiguities induced by the dialogue. An example of dialogue is shown below, while the robots actions are depicted in Fig. 8:

Human: “NAO, grasp the object on the left to the blue one!”
 Robot: Points the yellow rectangle (Fig. 8 left).
 Robot: “Is that the yellow rectangle I am pointing at?”
 Human: “No!”
 Robot: Points the blue circle (Fig. 8 center).
 Robot: “Is that the blue circle I am pointing at?”
 Human: “Yes! That’s right!”
 Robot: Grasps the blue circle (Fig. 8 right).

A set of external observers were judging the goodness of the system with respect to the following factors:

- Naturalness of the robot’s linguistic and motor behavior;
- Differences between the expected and observed behavior.

About ten people were involved in a full-day evaluation session. The overall score was positive in about 80% of collected forms.

6 Conclusions and future works

In this paper we have presented a computational model for the acquisition of a grounded language model to be used in human-robot interaction. The system learns the meaning of the words and their grammatical usage while understanding or generating a sentence. A

³ In the present model, the meaning of verbs “to point” and “to grasp” is hand-coded, and it is not learned by the system. Future releases will address the problem of grounding dynamic terms through the same computational framework.

related computational model constructs a disambiguation tree in order to help the robot conducting a dialogue in real-world settings.

However, a set of important questions still remain to be solved. As presented, the system learns “simple” concepts involving a single perceptual channel. Ongoing work is focused on learning complex concepts from the interaction data. The same computational framework will be employed recursively in order to assign meanings to words by hierarchically describing complex concepts as composed of simpler ones in a Bayesian network. Another issue is related to the process of learning and understanding verbs as words that usually involve an observable action. The work presented here represents the first steps in this direction.

ACKNOWLEDGEMENTS

This work has been partially supported by the EU funded project HUMANOBS: Humanoids That Learn Socio-Communicative Skills Through Observation, contract no. FP7-STREP-231453 (www.humanobs.org).

REFERENCES

- [1] D.A. Baldwin, ‘Understanding the link between joint attention and language’, *Joint attention: Its origins and role in development*, 131–158, (1995).
- [2] P. Bloom, ‘Intentionality and word learning’, *Trends in cognitive sciences*, **1**(1), 9–12, (1997).
- [3] K. Dautenhahn, ‘Socially intelligent robots: dimensions of human-robot interaction’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, **362**(1480), 679–704, (2007).
- [4] C. Fisher, D.G. Hall, S. Rakowitz, and L. Gleitman, ‘When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth’, *The acquisition of the lexicon*, 333–375, (1994).
- [5] K. Gold, M. Doniec, C. Crick, and B. Scassellati, ‘Robotic vocabulary building using extension inference and implicit contrast’, *Artificial Intelligence*, **173**(1), 145–166, (2009).
- [6] S. Harnad, ‘The symbol grounding problem’, *Physica. D*, **42**(1-3), 335–346, (1990).
- [7] G. Herzog and P. Wazinski, ‘Visual TRANslator: Linking perceptions and natural language descriptions’, *Artificial Intelligence Review*, **8**(2), 175–187, (1994).
- [8] S. Pinker, *Language learnability and language development*, Harvard University Press Cambridge, Mass, 1984.
- [9] D. Roy, ‘Grounding Words in Perception and Action: Insights from Computational Models’, *Trends in Cognitive Science*, **9**(8), 389–396, (2005).
- [10] D.K. Roy, ‘Learning visually grounded words and syntax for a scene description task’, *Computer Speech and Language*, **16**(3), 353–385, (2002).

- [11] Y. Uchida and K. Araki, 'Evaluation of a system for noun concepts acquisition from utterances about images (SINCA) using daily conversation data', in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pp. 65–68. Association for Computational Linguistics, (2009).
- [12] T. Winograd, *Understanding natural language*, Academic Press, Inc. Orlando, FL, USA, 1972.

Facial Expressions for Communicating User Feedback

Savandie Abeyratna¹, Galina Paramei¹, Hissam Tawfik² and Rentian Huang²

Abstract. The paper proposes a prototype for an *Affective Interface Feedback System* (AIFS) which aims to convert multi-criteria estimates of customers' satisfaction with a product or service into a non-verbal representation as virtual facial expressions (FEs), or *emoticons*. We focus on developing a system that can provide emotional content that can be unambiguously recognised by any users. A perception experiment was carried out to identify FE features that predominantly characterize the Satisfaction-Dissatisfaction emotion space. Future work is aimed at successfully learning these feature combinations and corresponding prediction factors in order to generate FEs that can convey levels of customer satisfaction to users through the AIFS. The proposed solution is intended to complement current feedback representation methods by adding a mode for more natural, 'at-a-glance' user specific data comprehension.

1 INTRODUCTION

Emotional facial expressions (FEs) are an efficient mode of human non-verbal communication accounting for 55% of interpersonal message conveyed during face to face interactions [1]. FEs can provide feedback in a concise pictorial form whose affective meanings are readily recognizable universally [2]. The psychology behind how people are able to easily decode emotional FEs is still not fully understood. Yet this characteristic of human beings makes FEs a suitable interface candidate for applications aiming to communicate feedback information.

The domain of Customer Relationship Management (CRM) thrives on capturing and understanding customer feedback in order to improve an organisations performance by addressing customer satisfaction. Most commonly data are collected via self-reported surveys, formatted as x -point Likert scales, which aim to assess customers' level of satisfaction or dissatisfaction with a certain product or service [3]. Results obtained using the surveys are traditionally displayed numerically as absolute values or percentages and/or in a graphic form as bar or pie charts. However, in view of a large amount of data collected via surveys there is a need for more concise, means of communicating customers' feedback.

Advances in Affective Computing have given rise to new CRM solutions for measuring and displaying customer satisfaction. It is worth noting that current Affective CRM solutions focus predominantly on developing computer systems that recognize FEs — to reveal by this means the degree of

customer's satisfaction with a product or service in question [4], i.e. an "input related affective research".

There is though a branch of Affective Computing that develops affective interfaces — by taking into account the communication power of FEs and of other non-verbal cues, i.e. an "output related affective research". In current CRM solutions, affective interfaces are mostly in the form of "smileys" with simplistic expressions [5] or cartoon anthropomorphic figures with various expressive postures. However, in all applications the relationship between the certain categories of numerical feedback data, on the one hand, and the smileys ascribed to these categories, on the other, is imposed by system developers. The choices of emotional FEs used in these affective interfaces are not buttressed by psychological research on the perception of the facial emotional content by different user groups, thus challenging the affective interfaces accuracy. In addition, these systems use smiley depictions only of basic emotions [2] while disregarding the continuous (dimensional) nature of FEs [cf. 6] and the relationship between the FE dimensions and the Satisfaction scale used in CRM surveys.

The main aim of the present study is twofold: i) to elaborate the procedure of compressing data using a survey and (ii) to develop an interface for unambiguously conveying feedback on Satisfaction—Dissatisfaction with a product or service in a form of comprehensible emotional facial expressions. Figure 1 illustrates the main research goals which spans from psychological experimentation and analysis to machine learning techniques.

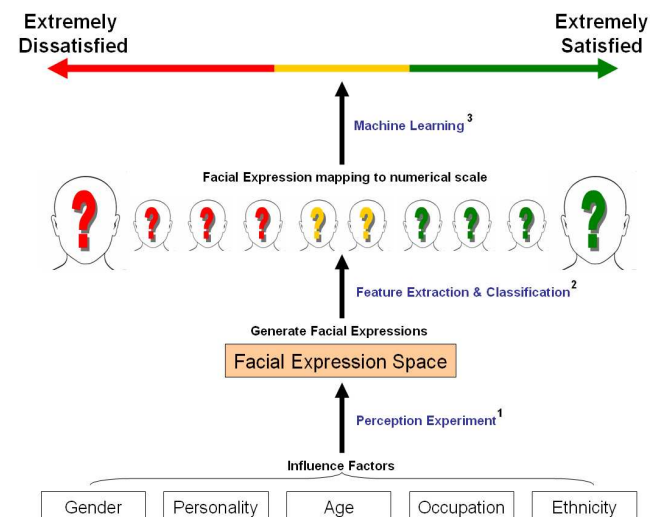


Figure 1. Underlying principles of the *Affective Interface Feedback System*.

¹ Faculty of Science and Social Sciences, Liverpool Hope University, UK

² Faculty of Business and Computer Science, Liverpool Hope University, UK

Email: {abeyras, parameg, tawfikh, huangr}@hope.ac.uk.

The main challenge for the project is ensuring that the information displayed in the FE is unambiguously legible and meaningful to any system user. Thus our vision is to develop an *Affective Interface Feedback System* that would contain individual or group user profiles which will determine the FEs generated to display satisfaction survey results. Psychological perception experiments would provide insight into the influence of gender, personality, age, occupation and ethnicity on the perception of emotional facial expressions (Figure 1²). Machine learning techniques could then be used to generate facial feature patterns depending on user profiles (Figure 1³).

The focus of this paper is directed towards constructing an emotional FE space that represents the satisfaction – dissatisfaction dimensions for different individuals (Figure 1¹), and evaluating simple features of the proposed *Affective Interface Feedback System*.

2 EXPERIMENT

Aim: Finding out which emoticons and facial features can consistently convey certain levels of customers' Satisfaction vs. Dissatisfaction with a product or service to different groups of individuals.

Participants: A total of 48 participants were recruited. Table 1 shows the breakdown of participants according to their ethnicity, gender and age.

Ethnicity	Total	Male	Female	Age			
				16-25	26-35	36-45	46+
European Caucasian	30	15	15	12	7	5	6
Asian+ Other	18	9	9	4	13	1	0
Total	48	24	24	16	20	6	0

Table 1. Distribution of experiment participants

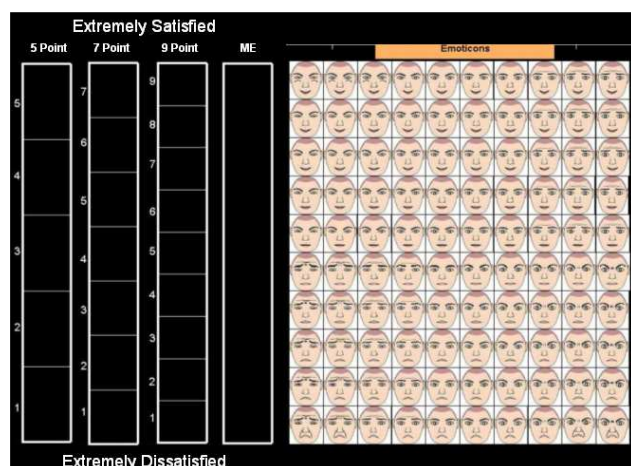


Figure 2. Experimental setup. *Left:* four response formats of the Satisfaction—Dissatisfaction scales, as 5-, 7-, 9-category or magnitude estimation (ME) scale. *Right:* the 10 x 10 emoticon set; Valence (positive, up, vs. negative, bottom) is represented by the vertical dimension; Activity (passive, left, vs. active, right) by the horizontal dimension).

Stimuli: Emoticons, outlined equivalents of Ekman and Friesen's FEs of basic emotions [2], as well as blends of these, with varying degree of the basic-emotion 'parent' pictures. An emoticon set (N=10 x 10) was adopted from [6]. The emoticons vary along two dimensions, **Valence** and **Activity** [7]. In the set, the degree of Valence is conveyed predominantly by the lip curvature and that of Activity by eye opening and eyebrow height.

Procedure: As shown in Figure 2, a Satisfaction—Dissatisfaction scale was presented to participants in four different response formats: 5-, 7- or 9-point category scale or magnitude estimation scale (ME; with positive and negative anchors). The participants were required to choose and drag emoticons from the emoticon set to corresponding divisions of the Likert scales or ten to the ME scale, with not necessarily equal spacing between the emoticon choices. The use of the four different types of response formats was intended for revealing the most informative type of the scale and of number of categories.

3 RESULTS and DISCUSSION

We first identified a set of emoticons representing the two anchors, *Extremely Satisfied* and *Extremely Dissatisfied*, as well as the Neutral value on the Satisfaction—Dissatisfaction scale.

The first finding is that for the *Extremely Satisfied* anchor and the *Neutral* position of the scale, all subjects across all four response formats chose very compact clusters of emoticons, with the **Happy** (high valence).and **Neutral** (medium valence).expressions respectively

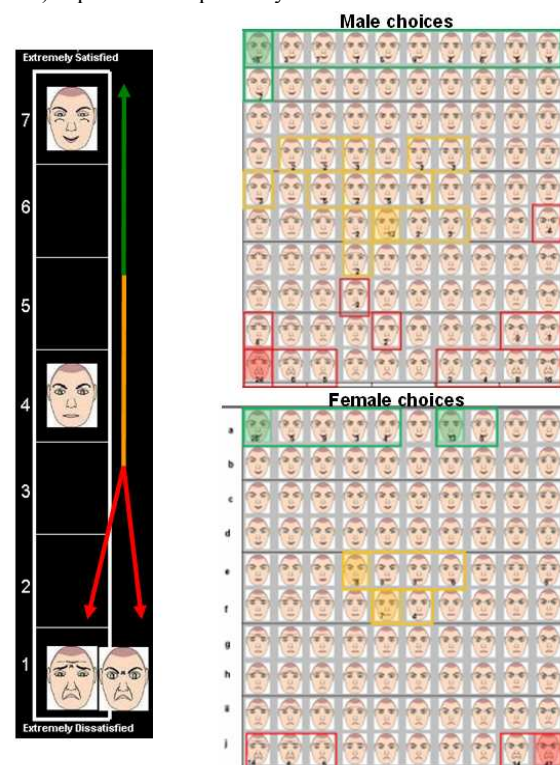


Figure 3. Male (top) and female (bottom) emoticon choices (across all response formats) for the *Extremely Satisfied* (green) and *Extremely Dissatisfied* (red) anchors as well as Neutral value (yellow) on the Satisfaction—Dissatisfaction Scale.

For the *Extremely Dissatisfied* anchor, however, across all response formats the chosen emoticons were split among two clusters of **Angry** (active) and **Sad** (passive) faces, i.e. along the Activity dimension. In addition, this split revealed significant gender difference: Females chose predominantly active **Angry** faces whereas males chose passive **Sad** faces as an expression for their highest level of dissatisfaction (Figure 3).

Such “bifurcation” of the emoticon choices was also observed for the *Extremely Satisfied* anchor but only when the number of emoticons selected (or categories used) was ≥ 7 (Figure 4). These results indicate that (1) users’ Satisfaction—Dissatisfaction scale cannot be represented by emoticons using just one dimension. (2) At least at its negative pole, the Satisfaction—Dissatisfaction scale should reflect by corresponding emoticons two very different behavioural dispositions, active or passive. (3) To disentangle both positive, Satisfaction, and negative, Dissatisfaction, poles, an experimental design should be used that would implement a two-dimensional scheme of evaluation of a product or service. (4) A further experiment would shed light on the revealed gender difference in the emoticon choices. Occupation and/or socio-economic background of participants can also be taken into account for better understanding inter-individual differences in decoding facial expressions.

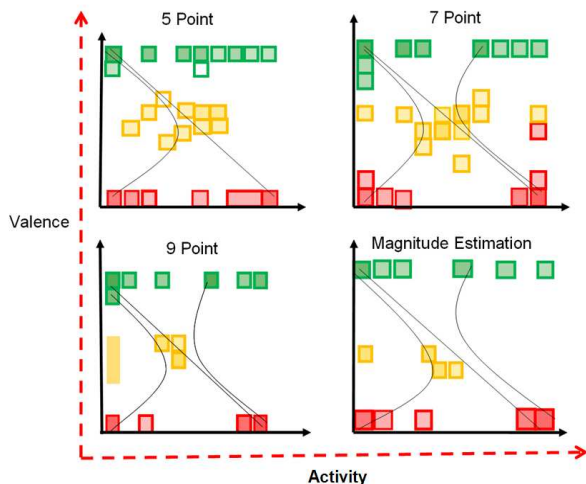


Figure 4. Distribution of selected emoticons for each of the four response formats. Colour coding is as above. The curves illustrate various combinations of the choices that may differ along the Valence - Activity dimensions according to different individuals.

4 PROPOSED SYSTEM

Here we propose a simple prototype for an *Affective Interface Feedback System* (AIFS), instigated by an “output related affective research” pioneered by Chernoff [8] and aiming at representing rich, multi-criteria information via FEs. In the current method virtual FEs, or *emoticons*, identified from the above experiment, are used as a means of communicating feedback between organisations and their customers [2] in order to demonstrate the concept of our research.

As is illustrated in Figure 2, the AIFS uses survey data obtained using x -point Likert scales with two poles as an input. The value of x may vary, thus resulting in a 5-, or 7- point etc.

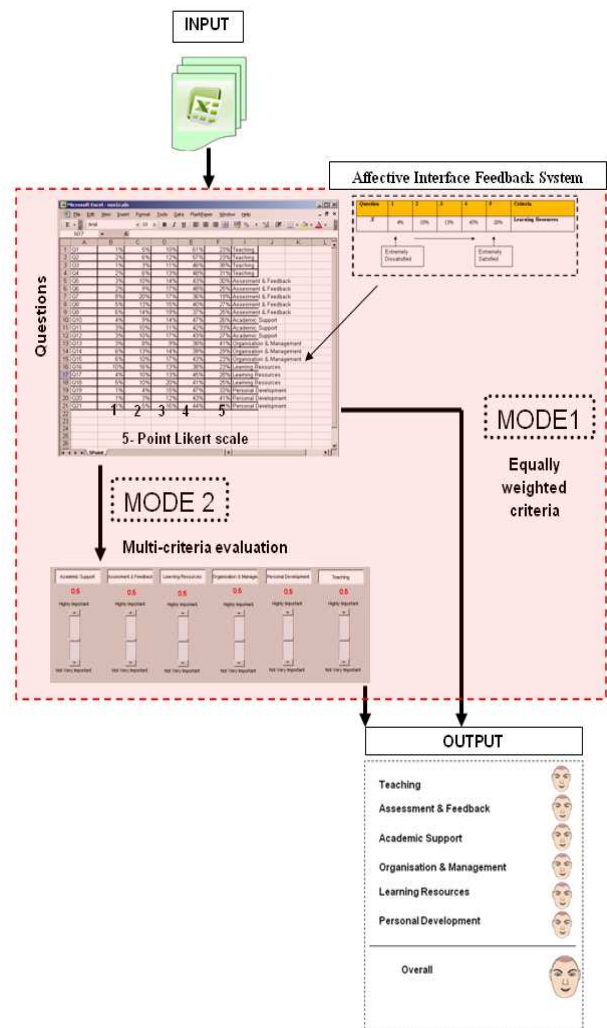


Figure 5. Architecture of the *Affective Interface Feedback System* exemplified by the National Student Survey with six evaluation criteria and showing two modes of multi-criteria data analysis: *Mode 1*: Equal weights; *Mode 2*: User-defined weights.

category scales. The number of evaluation criteria (of questions posed) regarding a product or service can also be varied – depending on the aspects of interest for the organisation. The National Student Survey used by us as an example (Figure 1) implements a 5-point Likert scale and contains 21 questions belonging to six evaluation criteria: Teaching; Assessment & Feedback; Academic Support; Organisation & Management; Learning Resources; Personal Development. Data for all questions is first pooled into these six criteria and represented numerically. Figure 1 indicates the AIFS’s *usability*, *flexibility* and *adaptability*.

The *usability* of the system is addressed via features of the graphical user interface which allows easy navigation through it. The use of an Excel-format data as input adds to the *usability* of the system. The *flexibility* of the AIFS is achieved through variation of the Likert-scale formats: 5-, 7-, 9-point scales or magnitude estimation scale with positive and negative anchors.

Finally, the AIFS's *adaptability* is implemented as two modes of aggregating data in multi-criteria evaluation. For calculating an overall outcome, the AIFS provides two options for data aggregation – either by equally weighting each of the criteria (Mode 1), thus calculating the grand mean, or by allowing the user to define certain weights to individual criteria (Mode 2), calculating the weighted mean.

At its current stage the AIFS converts the numerical data into a pictorial interface in the form of emoticons with a high selection frequency based on the above experiment. Next stages of development are aimed at ensuring the *efficiency* and *accuracy* of the AIFS by incorporating user profile based FE generation based .

5 CONCLUSIONS & FUTURE WORK

In this paper we propose an affective interface which could serve as an add-on to current CRM solutions by communicating customers' feedback data in the form of facial expressions in a user specific manner. The proposed *Affective Interface Feedback System* presented in this paper is at its very early stages and only addresses the adaptability, flexibility and usability of the system. Future work is aimed at improving the efficiency and accuracy of the system. Further psychological experimentation will allow the identification of inter-individual and/or cross-cultural differences that affect individual decoding of emotional FEs. Facial feature extraction (Figure 1²), will allow the classification of user specific feature patterns. Machine learning will be used to train the system to generate FEs depending on user profiles. Providing a user-specific FE as output would contribute to the emotional touch of current affective CRM applications.

6 ACKNOWLEDGEMENTS

The project was supported by HEIF4 grant *E-voting Affective Interface* to HT and GP. We are grateful to Mei Yui Lim for providing us with the emoticon set and her valuable suggestions.

REFERENCES

- [1] P.T. Kuo, I. Chung, J.E. Chen, and C.W. Wu, An emotional robot with facial expressions, *Chung Hua Journal of Science and Engineering*, 5, 9-13 (2007).
- [2] P. Ekman and W.V. Friesen, *Unmasking the Face: A Guide to Recognizing Emotions From Facial Cues*. Prentice-Hall Englewood Cliffs, NJ (1975).
- [3] W. Reinartz, M. Krafft, and W.D. Hoyer, The Customer Relationship Management Process: Its Measurement and Impact on Performance, *Journal of Marketing Research*, 41, 293–305 (2004).
- [4] R. W. Picard, Perceptual user interfaces: affective perception, *Communications of the ACM*, 43: 50-51 (2000).
- [5] D. Tetteroo, Communicating emotions in instant messaging, an overview. In *Proceedings of the 9th Twente Student Conference on IT, Enschede, The Netherlands, 23rd June 2008*
<http://referaat.cs.utwente.nl/new/paper.php?paperID=400>
- [6] M.Y. Lim and R. Aylett, An emergent emotion model for an affective mobile guide with attitude, *Applied Artificial Intelligence Journal*, 23: 835-854 (2009).
- [7] J.A. Russell, A circumplex model of affect. *Journal of Personality and Social Psychology*, 39: 1161–1178 (1980).
- [8] H. Chernoff, The use of faces to represent points in k -dimensional space graphically, *Journal of the American Statistical Association*, 68: 361-368 (1973).

Forgetful and Emotional: Recent Progress in Development of Dynamic Memory Management System for Conversational Agents

Michal Ptaszynski¹ Pawel Dybala¹ Rafal Rzepka¹ Kenji Araki¹

Abstract. Thinking of the ways to improve naturalness and adequacy of utterances in conversational agents, the authors propose a dynamic database management system. The system borrows some features of the forgetting mechanisms in humans. The core of the system, forgetting and recalling algorithms, depend on the frequency of usage of context units and their emotive values derived from evaluative reasoning about both sides of interaction - the agent and the user.

1 INTRODUCTION

Forgetting is a process in which parts of knowledge become rearranged, inaccessible or inactive [1, 2]. In neuromedical terms, forgetting is a result of the fact that as living and growing old beings, we cannot keep the neural connection net of our memory in a perfect shape forever. Every new neural connections made in our brain start decaying right after being settled [3].

The research about chronological fading of memories was started by Ebbinghaus [4] in an experiment with remembering series of syllables, which revealed that fading of memories is inversely proportional to the time expired.

Zitman [5] discovered, during his observations of patients with mood and anxiety disorders, that memory has much in common with emotions, which he confirmed in experiments on hormonal transfers. Nuray Luk [6] supported similar thesis in her research on the role of emotions in language acquisition. Wolfe [7], although using neurobiological nomenclature, stated similarly, that the brain at first seeks to create meaning through the establishment or refinement of existing neural networks (or actualization of the database of our memory) and further, during the process of learning affects it with emotions. Therefore memories with stronger emotive affection are harder to forget.

People usually find the forgetting process a disadvantage or a defect in the human mind. However, as the research on human brain continues, more features of our way of thinking undergo reevaluation. When Markovitch and Scott [2], stated that forgetting *"is a very useful process which facilitates effective knowledge acquisition"*, and that *"mechanisms of forgetting (...) merit study alongside those of acquisition since it is the two together which constitute learning"*, they tried to give forgetting a logical reasoning. However, it's the results of the experiments performed by Kahn et al. [8] that show that forgetting not only is not a defect, but on the contrary, actually helps people organize memory and remember important things.

Therefore, to improve naturalness and adequacy of utterances in conversational agents, the process of forgetting could be applied as well. We propose its application in the form of a system for dynamic management of the agent's expanding database resources. The idea of such a system is based on the following assumptions:

- Memory is an expanding database;
- Forgetting is a crucial element in the process of learning, and
- is beneficial in organizing knowledge;
- Forgetting process is strongly dependent on:
 - a) frequency of connections in use, and
 - b) emotional value of connections

As we assume, final implementation of such system should help process large text databases for conversational agents and effect in two major improvements: cutting down the time costs of wide context processing and allowing the agent to produce more natural utterances.

2 SYSTEM DESCRIPTION

In a conversational agent supported with an expanding database (DB, see Figure 1) all interactions with the user are archived (the archivization can be set as user-specific for better calibration of personalization algorithms). When a user encounters the agent, a new conversation starts. After every user input the agent queries the database searching for adequate context keywords (n-grams, conversational procedures) to produce an utterance. If there are none the agent queries the Internet for adequate associations. Every entry after a query is treated as a separate Context Unit (CU). For better distinction the context units are divided into Context Units obtained from a dialog with a user (dCU) and those in the form of an association lists gathered from the Internet (aCU).

However, since the database expands with every new CU, soon the agent would have to process enormous number of data. As is stated Araki [9], this is one of the difficult problems in processing context-bound sentences. To avoid this authors propose a system with an implemented algorithm of forgetting and recalling of the data archived in CUs according to the actual needs.

2.1 Forgetting algorithm

As stated above, forgetting is a process where parts of knowledge become inaccessible or inactive, which in humans contributes to better management of memory. Since there are context connections in the DB used more often and those used rarely, it is reasonable to base

¹ Graduate School of Information Science and Technology, Hokkaido University, {ptaszynski,paweldybala,kabura,araki}@media.eng.hokudai.ac.jp

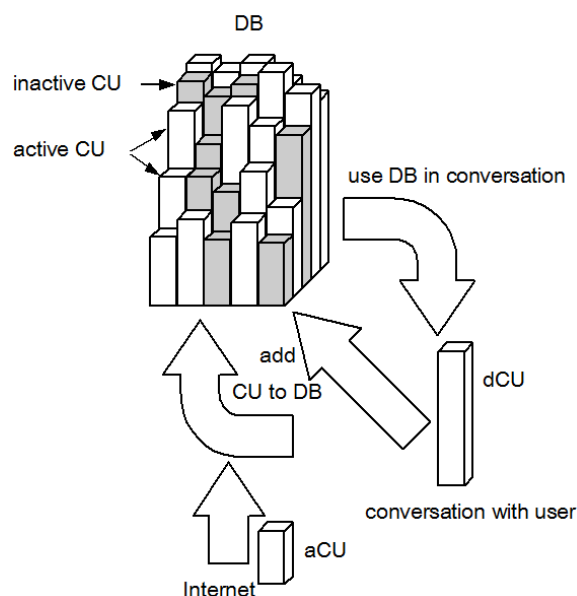


Figure 1. Generalized diagram of the system.

the forgetting process on the frequency of use of the connections. That is, if a connection is not used for a specified amount of time, an event would be marked as inaccessible. Moreover, forgetting in humans is based also on the emotive strength of neuronal connections, which could be represented by using an affect analysis system to add annotations of emotions to the archived CUs during the process of gathering information.

The contents to be forgotten by the agent are selected and the final decision whether to shade a CU is based on the values representing the two features:

1. frequency of CU usage, and
2. emotive strength of CU.

2.2 Recalling algorithm

Although there have been attempts of creating an algorithm of forgetting before with some good ideas, it seems that the attention was always concentrated only on forgetting during the process of learning. Markovitch and Scott [2], as well as later Ishikawa [10], by forgetting meant that the unused links of a large amount of knowledge are forgotten by the means of physically deleting them from the database. Unfortunately, this way the notion of recalling information from the faded database was ignored.

In the proposed algorithm the recalling process works simultaneously with the process of gaining new knowledge (CUs). When the user talks to the agent, a new conversation starts and the agent searches through its database for an appropriate context data to generate a response. If a match is established and the response generated, the connection to the used context is renewed. If not, the agent skips to the process of gaining knowledge from the Internet. When a new associative context unit (aCU) is obtained, it is compared to the inactive forgotten archives. If the new gained aCU associates with an inactive (forgotten) parts of knowledge in the database (dCU), those parts are activated and their connections renewed. If the aCU contains only new data and does not associate with any other CU, it is

added and archived in the database as a new active CU.

3 CONDITIONS FOR IMPLEMENTATION

In the usual indexing methods for textual databases [11, 12], the process of indexing and reevaluation of indexes is usually based on chronology of the established link and the frequency of its reappearance in queries. In the proposed method, apart from this type of indexing, the data is indexed also with information about its emotional value. As there are at least two participants of interaction - the user and the agent, to set such a value there was a need to develop two algorithms for evaluative reasoning about: 1) the user and 2) the agent.

3.1 Evaluative reasoning about user

The algorithm for evaluative reasoning about the user, is based on the assumption that the user can produce either appropriate or inappropriate utterance. To evaluate what is appropriate we based this algorithm on Ptaszynski et al's [13] method for verification of contextual appropriateness of emotions. In their method they used a system for affect analysis ML-Ask to recognize user's emotions and a Web mining technique as an emotion appropriateness verifier (both described in details below).

When ML-Ask discovers an emotive utterance and successfully specifies an emotion type, the Web mining technique extracts emotions frequently associated with the context of the utterance. The verification is based on the comparison of the emotions discovered by ML-Ask in user's utterance and the Web mining procedure results.

If the two match, the emotions expressed by the user are determined as appropriate. In such situations, the conversational agent equipped with this method chooses a dialog strategy to sympathize with the user. However, if the verification procedure estimates that the expressed emotions are inappropriate for the context, the agent undertake different measures e.g., helps the user manage his/her emotions. Four examples of the output of this method are shown below in Table 1.

The conversations with the agent are archived in the database with annotations of both features - emotions expressed by the user and their appropriateness. Primarily, this data is used in future interactions to provide the agent hints about which conversation strategy to choose, according to the appropriateness of the user emotions [13]. In the database management system the annotations of those two features on one whole context unit (CU) form fluctuations of the two features in the particular CU. The number of emotive utterances and emotional state fluctuations is further used as the first part of information used by the forgetting-recalling algorithm - evaluative reasoning about the user.

Table 1. Four examples of the appropriateness verification procedure.

Sentence (Translation)	ML-Ask output	Web-mining [appropriateness]
<i>Shiken ni goukaku shite ureshii!</i> (I'm so happy I passed the exam!)	joyful	joy, happiness [appropriate]
<i>Ano yaro ga kuruma ni hikarete, sukkiri!</i> (I'm so happy that bastard was hit by a car!)	joyful	fear, sad [inappropriate]
<i>Kanojo ni furarete kanashii...</i> (I'm so depressed since my girlfriend left...)	sad, depressed	sadness, gloom [appropriate]
<i>Iisutaa ga kuru kara kanashii ne...</i> (I'm so depressed for the coming Easter...)	sad, depressed	joy, happiness [inappropriate]

3.2 Evaluative reasoning about agent

The information obtained from the affect analysis system ML-Ask is also used to estimate the user's attitude to the agent and therefore to perform evaluative reasoning about the agent.

The results of affect analysis of each utterance provide information on how many user utterances were emotive. Furthermore, the emotions extracted from the user's emotive utterances form a vector on which the emotional states of the user changed during the conversation. This information is then processed as follows (described in detail by Ptaszynski in [14]).

Firstly, if numerous user's utterances were determined as emotive, we assume the user was emotionally involved in the conversation. Emotional involvement in a conversation suggests a tendency toward easier familiarization with the interlocutor [15]. Therefore we can assume that during the user's conversation with an agent, the machine interlocutor is considered to be more human-like the more emotionally emphasized the user's utterances are.

However, this does not yet mean a positive familiarization. The conversation could become emotional also when the interlocutors quarrel. This could happen in a case where the agent makes the user angry. However, if the user agrees to participate in a quarrel with an agent, this could also mean that the user finds the agent's linguistic capabilities to be comparable to himself. Therefore, the information obtained about the general emotional level of the conversation could be interpreted as signifying how much the user finds the agent worth talking to, including familiarity and the user's opinion about the agent's linguistic skills.

Secondly, analysis of specified emotion types conveyed by the user in the whole conversation provides information on the user's particular emotions during the conversation. If the emotions were positive or changing from negative to positive while talking, the general attitude toward the agent is considered to be positive. If the emotions were negative or changing from positive to negative, the attitude is classified as negative. The general attitude toward an agent is calculated as a ratio of conversations with positive tendencies to negative tendencies.

The information acquired this way (the user's general engagement in conversation and orientation of his/her attitude to the agent) provide an overview of the user's sentiment about the agent and is utilized as the second type of evaluative reasoning - about the agent.

4 SUB-SYSTEMS USED

To fulfill the implementation conditions for the database management system, and eventually perform both types of the evaluative reasoning we applied a set of sub-systems developed during this research project.

4.1 Affect analysis system - ML-Ask

ML-Ask is a system for affect analysis of utterances in Japanese. It is language-based and was constructed by Ptaszynski and colleagues [16] as an automatic affect annotation system for large corpora. The ML-Ask system uses a two-step procedure:

1. Analyzing the general emotiveness of utterances by calculating the emotive value;
2. Recognizing the particular emotion types in utterances described as emotive.

ML-Ask is based on the idea of two-part classification of realizations of emotions in language into:

- 1) *Emotive elements* or *emotemes* (ML), which indicate that emotions have been conveyed, but do not detail what specific emotions there are. This group is realized by such subgroups as interjections, mimetics, and vulgar language. Examples are: *sugee* (great!), *wakuwaku* (heart pounding), *-yagaru* (a kind of verb vulgarization);
- 2) *Emotive expressions* (MX); parts of speech that describe emotional states in emotive sentences. This group is realized by such parts of speech as nouns, verbs or adjectives. Examples are: *aijou* (love), *kanashimu* (sadness), and *ureshii* (happiness).

On textual input provided by the user, two features are computed in order: the emotiveness of an utterance and the specific type of emotion. To determine the first feature, the system searches for emotive elements in the utterance to determine whether it is emotive or non-emotive. In order to do this, the system uses MeCab for morphological analysis and separates every part of speech [17]. MeCab recognizes some parts of speech we define as emotemes, such as interjections, exclamations or sentence-final particles, like *-zo*, *-yo*, or *-ne*. If these appear, they are extracted from the utterance as emotemes. Next, the system searches and extracts every emoteme based on the system's emoteme databases (containing 907 items in total). All of the extracted elements mentioned above (exclamations from MeCab, emotemes and emoticons) indicate the emotional level of the utterance.

Secondly, in utterances classified as emotive, the system searches for the expressions describing emotional states using an emotive expression lexicon [18]. This determines the specific emotion type (or types) conveyed in the utterance. An example of analysis performed by ML-Ask (system output) is shown below.

- (1) *Kyo wa nante kimochi ii hi nanda !*
Today:TOP ML:nanteMX:joy day:SUB ML:nandaML:!
Translation: Today is such a nice day!
- (2) *Iya~, sore wa sugoi desu ne~ !*
ML:iya~ this :TOP ML:sugoi COP ML:ne~ ML:!
Translation: Whoa, that's great!
- (3) *Hitoribocchi nante iya da ~~~*
MX:sadness ML:nante-daMX:dislike COP ML:~~~
Translation: Being alone sucks...

4.2 Emoticon analysis system - CAO

CAO is a system for estimation of emotions conveyed through emoticons developed by Ptaszynski and colleagues [19]. Emoticons are - sets of symbols widely used in text-based online communication to convey emotions. The CAO, or *emotiCon Analysis and decoding of affective information* system extracts an emoticon from an input (a sentence) and determines specific emotion types expressed by it using a three-step procedure. Firstly, matching the input with a predetermined raw emoticon database containing over ten thousand emoticons. The emoticons, which could not be estimated with only the database are automatically divided into semantic areas, such as representations of "mouth" or "eyes", basing on the idea of *kinemes*, or minimal meaningful body movements, from the theory of kinesics [20, 21]. The areas are automatically annotated according to their co-occurrence in database. The annotation is firstly based on eye-mouth-eye triplet. If no triplet was found, all semantic areas are estimated separately. This provides hints about potential groups of expressed

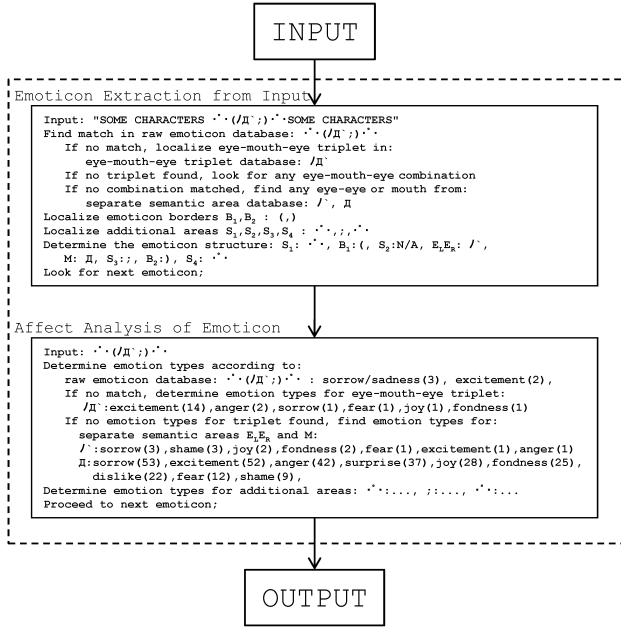


Figure 2. Description of main procedures in CAO.

emotions giving the system a coverage of over 3 million possibilities.

CAO is used as a supporting procedure in ML-Ask to improve performance of the affect analysis system.

4.3 Web mining technique

Shi and colleagues developed a technique for extracting emotive associations from the Web [22]. It takes as an input a sentence and in the Internet searches for emotion types associating with the sentence contents. Ptaszynski interprets this as an online commonsense reasoning about what emotions are the most natural and appropriate to appear within a certain context of an utterance [13]. The technique is composed of three steps: **a)** phrase extraction from an utterance; **b)** morpheme modification; **c)** extraction of emotion associations.

In the first step, an utterance after being analyzed morphologically by part-of-speech analysis tool for Japanese - MeCab [17], and n-gram phrases for further processing are composed using parts of speech separated by MeCab as unigrams. Secondly, the list of n-gram phrases ending with a verb or an adjective modified grammatically by adding causality morphemes. They distinguished five most frequently used morphemes stigmatized emotively in the Japanese language: *-te*, *-to*, *-node*, *-kara*, *-tara*, for which the coverage of the Web exceeded 90%. Finally, the modified phrases are queried in Google search engine with 100 snippets for one morpheme modification per query phrase. This way a maximum of 500 snippets for each queried phrase is extracted from the Web and cross-referenced with the emotive expression lexicon [18]. The higher hit-rate an expression had in the Web, the higher was the naturalness of the emotion type it represents.

However, as the Web mining process was time consuming, we decided to construct a robust emotion object database to make the Web mining technique into a stand-alone system.

Table 2. Example of emotion association extraction from the Web and its improvement by blog mining procedure.

Sentence: <i>Konpyuuta wa omoshiroi desu ne. (Computers are so interesting.)</i>		
Extracted emotion type	Type extracted / all extracted types(Ratio)	
liking	79 / 284	(0.287)
surprise	30 / 284	(0.105)
excitement	30 / 284	(0.105)
fear	29 / 284	(0.102)
...	...	...

```

<emotion type: haji [shame] (202678 EmObj;
4006378 SemCat)>
<emotive expression: sekimen [turn red/feel
ashamed] (2316 EmObj for this EmoExp)>
Sentence<20 Japanese characters>:
"Chakku wo shimeru no wasureta to ki ga
tsuitara sekimen shita..."
["I turned all red when I noticed I
had a fly open..."]
EmObj: <Chakku shimeru no wasureta
to ki ga tsuita> [Notice to
have a fly open]
CausInf: <tara> [because]
EmoExp: <turn red/feel ashamed>
...
<emotion type: yorokobi [joy] (6123947 EmObj;
120068119 SemCat)>
<emotive expression: ureshii [happy]
(1078312 EmObj)>
...
<emotive expression: ureshimi [joyfulness]
(5 EmObj)>
...
...
  
```

Figure 3. A structure of the emotion object database with examples.

4.4 Emotion object database - INFOE

From a blog corpus containing over 350 million of sentences we extracted and implemented a robust ontology-type database of emotion objects - INFOE (*I kNow Formal Objects of Emotions*), containing 19,459,167 of unique emotion objects. Their distribution among emotion types is shown in table 3. A structure of the database including some examples is described in Figure 3. The data in the database is formalized in three ways, using:

1. Statistics of the length of Emotion Object phrases;
2. Syntactical POS tagging and dependency structure;
3. Calculating the number of semantic categories of words appearing in the database;

4.4.1 Sentence Length Based Statistics

The statistics of the Emotion Object phrase length was calculated for each emotion type (see table 3), as well as for each emotive expression used as a seed in emotion object extraction. This part of information in the database represents statistical probability of emotion affiliation of an input generated by a user during interaction with the conversational agent.

4.4.2 POS Tagging and Dependency Structure

The POS tagging of the database was done by MeCab [17]. Dependency structure analysis is done by CaboCha, an SVM supported

Table 3. Distribution of number of emotion objects and semantic categories among emotion types.

Emotion type (No. of expressions)	Number of extracted:	
	emotion objects	semantic categories
joy (224)	6123947	120068119
relief (106)	3321795	66605209
dislike (532)	2957596	59136703
fondness (197)	2441865	45816845
fear (147)	1184952	23490755
excitement (269)	1104998	23102117
sadness (232)	930698	18911960
surprise (129)	898138	18164806
anger (199)	292500	5966179
shame (65)	202678	4006378

dependency structure analyzer for the Japanese language [23]. The information provided by both syntactical analyzes verifies the probability statistics calculated on the basis of the number of characters.

4.4.3 Semantic Formalization

The semantic formalization of the database is performed using *Bunrui Goiho* - word list including semantic categories of words. The categories include such labels as: "Abstract objects", "Human Activities", "Natural Objects and Phenomena", "Subject of Actions", etc. Each main category consists also of number of sub-categories. The distribution of the number of categories according to their appearance for each emotion type is presented in Table 3.

5 IMPLEMENTATION PROCEDURE AND DIRECTIONS

In the process of database management there are two important phases, generation of new CU and management of the CU already existing in the database. The general description of CU management procedure has been explained in section 2. A procedure for generation of new CUs has been represented in Figure 4.

Every utterance of both, the user and the agent is preprocessed to create a model of the conversation. First the utterance is preprocessed by morphological analyzer MeCab and dependency parser CaboCha to generate information about syntax of the utterance, which represents one conversational strategy. Analysis of all utterances during the conversation create a conversational strategy model of this particular conversation.

Together with the syntactical analysis, analysis of emotions takes place. The utterance is analyzed with affect analysis system ML-Ask to specify the emotion type expressed by the user. After that, the emotion is verified whether it is appropriate for its object (utterance context). This is done first by looking up the appropriate emotions in INFOE database of emotion objects, and when no emotion associations are found, the system uses a Web-mining technique to obtain such associations from the Internet. The two above information (expressed emotion and its appropriateness for the context) make up the model of emotional strategies applied in this conversation.

The conversational strategy model and the emotional strategy model constitute the Context Unit (CU). Every CU is saved in the database preserved in XML format and is assigned a time stamp T and an emotional value E being an approximation of emotive values of user utterances. In the later process of forgetting, the decision of whether to keep the CU active, or deactivate it will be based on a value V derived from a function of these two intermediate values, as

in equation 1. At present, however, it is not yet decided what type of function would be the most appropriate function and what should be the threshold of maintaining activation of the CU.

$$V = f(T, E) \quad (1)$$

In subsequent conversation previous active CUs, will be queried to find a CU matching user utterance. The matched features will be the conversation and the emotional strategies models. The agent's response will be generated using the strategies from an utterance subsequent to the user utterance in the database. Since the query will be made only on the active CUs (a subset of the whole database), the query time will be reduced.

The database management system is planned to be implemented in a counselor-companion agent meant to help users manage their emotions through a non-goal-oriented free conversation, or small talk [29]. We have performed some preliminary experiments in this direction with conversational agents. We found out that by adding modality to the agent's responses makes them more natural [24]. Secondly, assuming that the agent-companion should be also able to induce positive emotions in user, e.g., by a humorous response, we showed that implementation of a joke generator in a conversational agent greatly improves its impression [30]. Finally, we showed that a conversational agent using its conversation history (or CUs) to summarize previous conversations is perceived as more intelligent and makes the impression of listening to the user [31].

6 CONCLUSIONS AND FUTURE WORK

In this paper we presented the description of a dynamic database management system for conversational agents. The system borrows the idea from the process of forgetting in humans and is meant to improve the performance and naturalness of speech in conversational agents. The core of the system, forgetting and recalling algorithms, depend on frequency of usage of context units and their emotive values derived from evaluative reasoning about both sides of interaction - the agent and the user.

During the last several years we have been developing the intermediary systems to be implemented in the forgetting-recalling algorithm. We have first proposed the affect analysis system ML-Ask to estimate the emotive value of user's utterances [16]. Secondly, to improve the performance of the affect analysis system we proposed the emoticon analysis system CAO [19]. These two systems are combined in the algorithm of evaluative reasoning of the agent. Furthermore, a Web mining technique has been proposed to extract emotions associating with the context of an utterance [22]. To improve the performance of this method and shorten the time of processing we created an emotion object database. The affect analysis system together with Web mining technique are finally responsible for evaluative reasoning of the user [13].

As all parts of the systems needed in the implementation conditions of the dynamic database management system have been developed, at this moment we are working on combining the subsystems into the database indexing and management system. In the final implementation the database management system will operate on a database of a conversational agent-companion. The baseline for the conversational agent was already developed by Higuchi and colleagues [24] and improved further by Dybala and colleagues [25].

After combining together all parts of dynamic database management system we plan a test phase, where efficiency and performance of the system with the forgetting-recalling algorithm activated and

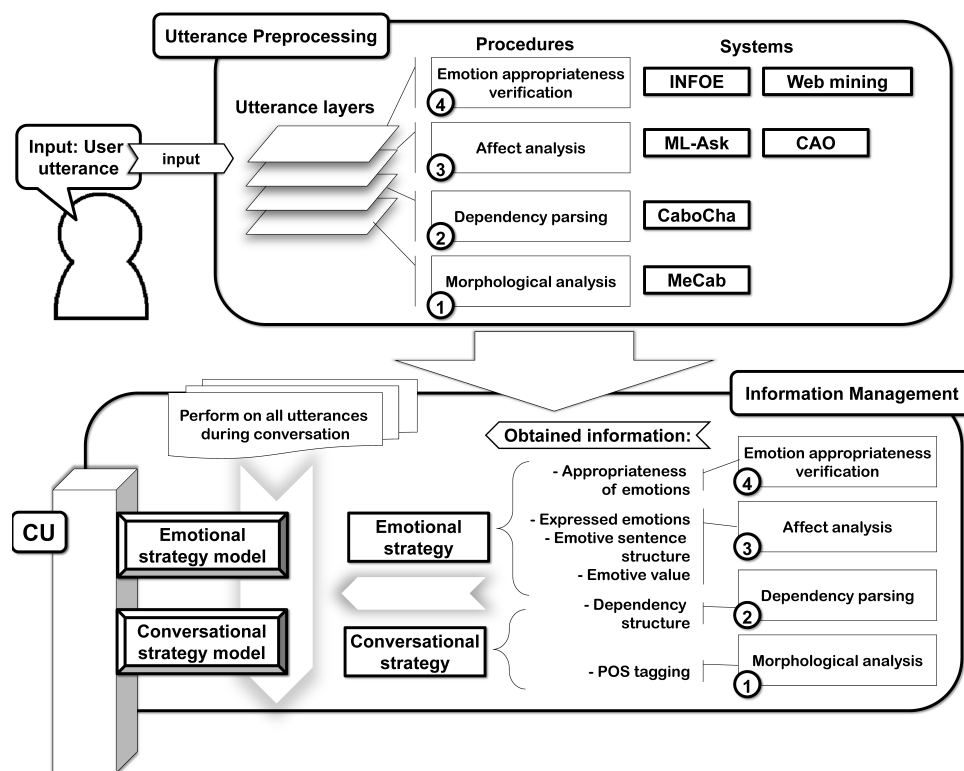


Figure 4. Description of the procedure for CU generation.

deactivated will be compared. In the test phase, naturalness and adequacy of speech will be tested with a standard usability test including several users testing the system for several days.

Fading of the unused context connections would improve agent's performance in processing large database containing context-bound sentences. Since, using the forgetting-recalling algorithms, the database is renewed according to the present needs, the agent after the training phase will operate on an always actual and appropriate database. In the final effect, the agent operating on the database management system presented in this paper should obtain a comparable with humans ability to process language.

Today we are surrounded by various artifacts based on "weak AI" [26, 27]. Card readers, red-eye and handshake effect reduction in our digital cameras, cellular phones, etc. make our lives easier. However, if we wish to create an intelligent interlocutor based on the "weak AI", that is, a program that would be, using Lyons' nomenclature [28], able to send messages informative for the receiver, but without knowing its communicative meaning (as a non-thinking machine), our attention should be paid on creation of an imitation of human way of thinking beginning with its originality, as well as with its all vices and weaknesses. We should not consider the artificial mind as a perfect creation. On the contrary, we should accept it with all its imperfections and defects. It cannot be excluded that some of them, such as forgetting, will eventually turn out to be merits.

ACKNOWLEDGEMENTS

This research is partially supported by a Research Grant from the Nissan Science Foundation and The GCOE Program founded by

Japan's Ministry of Education, Culture, Sports, Science and Technology.

REFERENCES

- [1] J. R. Anderson, *The Architecture of Cognition*, Harvard University Press (1983).
- [2] S. Markovitch and P. D. Scott. The Role of Forgetting in Learning, In *Proceedings of The Fifth International Conference on Machine Learning*. Ann Arbor, MI: Morgan Kaufmann, (1988).
- [3] S. Waddell. Forgetting Those Painful Moments, Worcester, USA: Department of Neurobiology, University of Massachusetts Medical School (2004).
- [4] H. Ebbinghaus. *Memory A Contribution to Experimental Psychology* (1885). (translated by: T. Utsugi. *Kioku nitsuite: jikken shinrigaku he no koken*, Tokyo, Seishinsho Shobo (1978).
- [5] F. G. Zitman. Emotion and memory in mood-anxiety disorders, Schreuder, BJN (2001).
- [6] N. Luk. The Role of Emotions in Language Teaching, In *The Journal of the Imagination in Language and Teaching*, 7 (2002).
- [7] P. Wolfe. The role of meaning and emotion in learning, In *New Directions for Adult and Continuing Education*, 2006(110), pp. 35-41 (2006).
- [8] I. Kahn, N. M. Dudukovic, B. A. Kuhl and A. D. Wagner. Decreased demands on cognitive control reveal the neural processing benefits of forgetting, *Nature Neuroscience*, 10, pp. 908-914 (2007).
- [9] K. Araki, *Shizen gengo shori - kotohajime - kotoba wo oboe kaiwa no dekiru konpyuta* - [Introduction of Natural Language Processing - Can a computer learn language and talk?] (in Japanese), Tokyo, Morikita Publishers (2004).
- [10] M. Ishikawa. A Structural Connectionist Learning Algorithm with Forgetting, *J. of Japanese Society for Artificial Intelligence* 5(5) (1990).
- [11] Google Technology, <http://www.google.com/technology/pigeonrank.html>
- [12] HyperEstrai, <http://hyperestraier.sourceforge.net/index.html>
- [13] M. Ptaszynski, P. Dybala, W. Shi, R. Rzepka and K. Araki. Towards Context Aware Emotional Intelligence in Machines: Computing Con-

- textual Appropriateness of Affective States, In *Proceedings of Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09)*, Pasadena, California, USA, pp. 1469-1474 (2009).
- [14] M. Ptaszynski, P. Dybala, R. Rzepka and K. Araki. An Automatic Evaluation Method for Conversational Agents Based on Affect-as-Information Theory, *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, Special Issue on Emotions, 22(1), pp. 73-89 (2010).
 - [15] C. Yu, P. M. Aoki, A. Woodruff. Detecting User Engagement in Everyday Conversations, In *Proc. 8th Int'l Conf. on Spoken Language Processing (ICSLP)*, Vol. 2, pp. 1329-1332 (2004).
 - [16] M. Ptaszynski, P. Dybala, R. Rzepka and K. Araki. Affecting Corpora: Experiments with Automatic Affect Annotation System - A Case Study of the 2channel Forum -, In *Proceedings of The Conference of the Pacific Association for Computational Linguistics 2009 (PACLING-09)*, pp. 223-228 (2009).
 - [17] T. Kudo. MeCab: Yet Another Part-of-Speech and Morphological Analyzer, 2001. <http://mecab.sourceforge.net/>
 - [18] N. Akira. *Kanjo hyogen jiten* [Dictionary of Emotive Expressions] (in Japanese), Tokyodo (1993).
 - [19] M. Ptaszynski, P. Dybala, R. Rzepka and K. Araki. Towards Fully Automatic Emoticon Analysis System (‘o’), In *Proceedings of The Fifteenth Annual Meeting of The Association for Natural Language Processing (NLP-2010)*, Tokyo (2010).
 - [20] R. L. Birdwhistell. *Introduction to kinesics: an annotation system for analysis of body motion and gesture*, Univ. of Kentucky Press (1952).
 - [21] R. L. Birdwhistell. *Kinesics and Context*, Univ. of Pennsylvania Press, Philadelphia (1970).
 - [22] W. Shi, R. Rzepka and K. Araki. Emotive Information Discovery from User Textual Input Using Causal Associations from the Internet, *FIT-08*, pp.267-268 (2008).
 - [23] T. Kudo, Y. Matsumoto. Japanese Dependency Analysis using Cascaded Chunking, *CONLL 2002*, Taipei (2002).
 - [24] S. Higuchi, R. Rzepka and K. Araki. A Casual Conversation System Using Modality and Word Associations Retrieved from the Web, In *Proceedings of the EMNLP-2008*, pp. 382-390 (2008).
 - [25] P. Dybala, M. Ptaszynski, R. Rzepka and K. Araki. Activating Humans with Humor A Dialogue System that Users Want to Interact With, *IE-ICE Transactions on Information and Systems*, Special Issue on Natural Language Processing and its Applications, E92-D(12), pp. 2394-2401 (2009).
 - [26] J. R. Searle. Minds, brains, and programs, In *Behavioral and Brain Sciences*, 3(3), pp. 417-457 (1980).
 - [27] J. R. Searle. *Mind, language and society: Philosophy in the real world*, New York: Basic Books (1998).
 - [28] J. Lyons. *Semantics*, Cambridge, Cambridge University Press (1979).
 - [29] M. Ptaszynski, R. Rzepka and K. Araki. A Pragmatic Approach to Implementation of Emotional Intelligence in Machines, *The AAI 2009 Fall Symposium on Biologically Inspired Cognitive Architectures (BICA-09)*, Washington, D.C., USA, November 5 - 7, (2009).
 - [30] P. Dybala, M. Ptaszynski, R. Rzepka, K. Araki. Humoroids - Conversational Agents That Induce Positive Emotions with Humor, In *Proceedings of AAMAS*, pp. 1171-1172 (2009).
 - [31] A. Fukuda, K. Araki, R. Rzepka. Effectiveness of Non-task-oriented Dialogue System Using Abstract of Chat Logs, IPSJ SIG Technical Report, 2010-FI-97(11), (2010).

Associative Text Categorisation Rules Pruning Method

Hussein Abu-Mansour¹, Wa'el Hadi², T.L.McCluskey¹ and Fadi Thabtah²

Abstract. In this paper, the problem of rule pruning in associative text categorisation is investigated. We propose a new rule pruning method within an existing associative classification algorithm called MCAR. Experimental results against large text collection (Reuters-21578) using the developed pruning method as well as other known existing methods (Database coverage, lazy pruning) are conducted. The bases of the experiments are the classification accuracy and the number of generated rules. The results derived show that the proposed rule pruning method derives higher quality and more scalable classifiers than those produced by lazy and database coverage pruning approaches. In addition, the number of rules generated by the developed pruning procedure is usually less than those of lazy pruning and database coverage heuristics.

Keywords: Association Rule, Classification, Data Mining, Text categorisation, Rule pruning

1. INTRODUCTION

Associative Classification (AC) also known as classification based on association rule, is an integration of two important data mining tasks, association and classification. Association rule and classification are analogous tasks, with the exception that classification's main aim is the prediction of class labels, while association rule describes correlations among items in a database. Several studies (Liu et al., 1998; Li et al., 2001; Yin and Han, 2003; Thabtah, et al., 2004; Thabtah, et al., 2005; Li et al., 2008; Niu et al., 2009) provide evidence that AC approach is able to derive more accurate classifiers than traditional classification techniques, such as decision trees (Quinlan, 1993; Quinlan, 1998), rule induction (Quinlan and Cameron-Jones, 1993; Cohen, 1995) and probabilistic (Duda and Hart, 1973) approaches. However, this approach suffers from the exponential growth of rules. This is since AC employs association rule during the learning step where all the correlations among the items and the class are discovered in a form of if-then rules. Thus, pruning duplicate and misleading rules becomes essential.

The AC mining approach works as follow. First, all rules satisfying some user-specified parameters (minimum support (minsupp), minimum confidence (minconf)), are generated using an association rule mining algorithm. The number of rules generated might be of the order of several thousands, and many of them are either redundant, i.e. never used, or noisy, i.e. not

discriminative among the classes. Second, the redundant and noisy rules are removed using pruning procedure(s). A significant number of rules can be removed at this stage and the rules left are the interesting ones that form a model (classifier) used later to classify new test cases. Each classifier must have a default rule which is applied when no other classifier rule can be used.

AC algorithms normally derive large sets of rules (Topor and Shen, 2001; Li et al., 2001; Kundu et al., 2008; Li et al., 2008) since (1) classification data sets are typically highly correlated and (2) association rule mining is used in the rule discovery step. As a result, there have been many attempts to reduce the size of classifiers produced by AC approaches, mainly focused on preventing rules that are either redundant or misleading from taking any role in the prediction process of test data cases. The removal of such rules can make the classification process more effective and accurate. It is not always helpful to have a large number of rules when we classify test cases. There is a great chance of having more than one rule contradicting each other in the answer class. We want to have a small number of highly predictive rules.

The aim of this paper is to investigate the rule pruning phase in AC mining in order to reduce the size of the resulting classifiers. Particularly, we develop a new rule pruning method. The developed pruning procedure is implemented within a known AC algorithm called MCAR (Thabtah et al., 2005), and are tested against large text categorisation collection called Reuters-21578 Mod Split (Lewis, 1998). To the best of the authors' knowledge, AC rule pruning on text categorisation has not yet been explored in the literature.

This paper is structured as follows: AC approach and known rule pruning methods are discussed in Section 2. In Section 3, the proposed rule pruning technique is presented. Further, the results that show the impact of pruning on the size of the classifiers and the prediction accuracy are demonstrated in Section 4. Finally, conclusions are given in Section 5.

2. ASSOCIATIVE CLASSIFICATION

2.1 Associative Classification Problem

AC is a special case of association rule in which only the class attribute is considered in the rule's right-hand-side (consequent) (Liu et al., 1998), for example in a rule such as $X \rightarrow Y$, Y must be a class attribute. We follow (Thabtah, et al., 2004) for the definition of the AC problem. A training data set T has m distinct attributes $A_1, A_2, \dots, A_m$ and C is a list of class labels. The number of rows (cases) in T is denoted $|T|$.

Definition 1: A row or a training case in T can be described as a combination of attributes A_i and values a_{ij} , plus a class denoted by c_j .

¹Computing and Engineering Dept, Huddersfield University, UK, Email: {H.Y.Abu Mansour, lee}@hud.ac.uk,

²MIS Dept, Philadelphia University, Jordan. Email: {whadi, ffayez}@philadelphia.edu.jo

Definition 2: An attribute value can be described as a term name A_i and a value a_i , denoted $\langle(A_i, a_i)\rangle$.

Definition 3: An *AttributeValueSet* can be described as a set of disjoint attribute values contained in a training case, denoted $\langle(A_{i1}, a_{i1}), \dots, (A_{ik}, a_{ik})\rangle$.

Definition 4: A *ruleitem* r is of the form $\langle \text{AttributeValueSet}, c \rangle$, where $c \in C$ is the class.

Definition 5: The actual occurrence (*actoccr*) of a *ruleitem* r in T is the number of rows (cases) in T that match the *AttributeValueSet* of r .

Definition 6: The support count (*suppcount*) of *ruleitem* r is the number of rows in T that match r 's *AttributeValueSet*, and belong to the class c of r .

Definition 7: The occurrence of an *AttributeValueSet* i (*occatt*) in T is the number of rows in T that match i .

Definition 8: A *ruleitem* r passes the *minsupp* threshold if $(\text{suppcount}(r)/|T|) \geq \text{minsupp}$.

Definition 9: A *ruleitem* r passes the *minconf* threshold if $(\text{suppcount}(r)/\text{actoccr}(r)) \geq \text{minconf}$.

Definition 10: Any *ruleitem* r that passes the *minsupp* threshold is said to be a *frequent ruleitem*.

Definition 11: A class association rule (CAR) is represented in the form: $(A_{i1}, a_{i1}) \wedge \dots \wedge (A_{ik}, a_{ik}) \rightarrow c$, where the antecedent (rule body/LHS) of the rule is an *AttributeValueSet* and the consequent (RHS) is a class.

Definition 12: We say that a CAR R partial matches a test case t if R contains at least an item in its antecedent that exists in t .

Definition 13: We say that a CAR R fully matches a test case t if all R items are in t .

A classifier is a mapping form $H : A \rightarrow Y$, where A is the set of *AttributeValueSet* and Y is the set of class labels. The main task of AC is to construct a set of rules (model) that is able to predict the classes of previously unseen data, known as the test data set, as accurately as possible. In other words, the goal is to find a classifier $h \in H$ that maximises the probability that $h(a) = y$ for each test case.

2.2 Current Pruning Methods in AC

Several pruning methods have been used effectively to reduce the size of the classifiers in AC, some of which have been adopted from decision trees, like pessimistic estimation (Quinlan, 1987), others from statistics such as Chi-square testing (χ^2) (Snedecor and Cochran, 1989). These pruning methods are utilised during the construction of the classifier, for instance, a very early pruning step, which eliminates ruleitems that do not pass the support threshold, may occur in the process of finding frequent ruleitems. Another pruning approach such as chi-square testing may take place when generating the rules, and post pruning methods, like database coverage (Liu et al., 1998) may be carried out after discovering all potential rules. Throughout this section, we will discuss pruning methods used in AC. It should be noted that we focus in this paper on pruning methods that are solely developed

within AC mining algorithms and not adopted from closely related areas such as statistics and machine learning.

2.2.1 Database Coverage Pruning

The database coverage is a post pruning technique used in AC (Liu et al., 1998), which is usually invoked after rules have been created. If at least one case among all the cases in training data set is fully matched by the rule, the rule is inserted into the classifier and all cases covered are removed from the training data set. The rule insertion stops when either all of the rules are used or no cases are left in the training data set. The majority class among all cases left in the training is selected as default class. The default class is used in case when there are no covering rules. After this process, the first rule which has the least number of errors is identified as the cutoff rule. All the rules after this rule are not included in the final classifier since they often produce errors (Liu et al., 1998). The database coverage method was used first by CBA (Liu et al., 1998) and then latterly by other associative algorithms, including CBA (2) (Liu et al., 2003), CMAR (Li et al., 2001), ARC-BC (Antonie and Zaïane, 2002), CAAR (Xu et al., 2004), ACN (Kundu et al., 2008), Multi-label Classification based on Association Rules (Li et al., 2008), and ACCF (Li et al., 2008).

2.2.2 Redundant Rule Pruning

In AC, all attribute value combinations are considered in turn as a rule's condition, therefore rules in the resulting classifiers may share training items in their conditions and for this reason there could be several specific rules containing many general rules. Rules redundancy in the classifier is unnecessary and in fact could be a serious problem, especially if the number of discovered rules is extremely large.

A pruning method that discards specific rules with less confidence values than general rules called redundant rule pruning, has been proposed in (Li et al., 2001). Redundant rule pruning method works as follows: Once the rule generation process is finished and rules are sorted, an evaluation step is performed to prune all rules such as $I' \rightarrow c$ from the set of generated rules, where there is some general rule $I \rightarrow c$ of a higher rank and $I \subseteq I'$. This pruning method significantly reduces the size of the resulting classifiers and minimises rules redundancy (Li et al., 2001).

Algorithms, including (Li et al., 2001; Antonie and Zaïane, 2004), have used redundant rule pruning. They perform such pruning immediately after a rule is inserted into the compact data structure, the CR-tree. When a rule is added to the CR-tree, a query is issued to check if the inserted rule can be pruned or some other already inserted rules in the tree can be removed.

2.2.3 Database like Pruning (Lazy Pruning)

Some AC techniques (Baralis and Torino, 2000; Baralis, et al., 2004; Baralis et al. 2008) argue that pruning classification rules

should be limited to only “negative” rules (those that lead to incorrect classification). In addition, they claim that database coverage pruning often discards some useful knowledge, as the ideal support threshold is not known in advance. Due to this, these algorithms have used a late database coverage-like approach, called lazy pruning, which discards rules that incorrectly classify training cases and keeps all others.

Lazy pruning happens after rules have been created and stored, where each training case is taken in turn and the first rule in the set of ranked rules applicable to the case is assigned to it. The training case is then removed and the correctness of the class assigned to the case is checked. Once all training data have been considered, only rules that wrongly classified training data are discarded and their covered data are put into a new cycle and the process is repeated until all training data are correctly classified. The results are two levels of rules; the first level contains rules that correctly classified at least one single training case and the second level contains rules that were never used in the training phase. The main difference between lazy pruning and database coverage pruning is that the second level rules that are held in the memory by the lazy pruning are completely removed by the database coverage during rule discovery step. Furthermore, once a rule is applied to the training data, all cases covered by the rule are removed (negative and positive) by the database coverage method.

Experimental tests reported in (Baralis and Torino, 2000; Baralis, et al., 2004; Baralis, et al., 2008) using 26 different data sets from (Merz and Murphy, 1996) showed that methods that employ lazy pruning such as L3 and L3G may improve classification accuracy on average by 1.63% over other techniques that use database coverage pruning

(Liu et al., 1998; Li et al. 2001). However, lazy pruning may lead to very large classifiers, which makes it difficult for a human to understand or be able to interpret. In addition, the experimental tests indicate that lazy pruning algorithms consume more memory than other AC techniques and, more importantly, they may fail if the support threshold is set to a very low value due to the very large number of potential rules.

3. THE PROPOSED RULE PRUNING METHODS

In this section, we discuss the proposed rule pruning method along with an example to illustrate it. Assume that the eleven rules shown in Table 1 are produced by an AC algorithm called MCAR (Thabtah et al., 2005) using minsupp and minconf of 20% and 40%, respectively from the training data set shown in Table 2. Before pruning kicks in the rules must be ranked according to confidence and support values in descending manner. The rule ranking criterion used in MCAR is as follows: (1) The rule with a higher confidence has a higher rank than the others. (2) If confidence values are the same between two or more rules, then the one with a higher support has a higher rank than the others. (3) If the confidence and support values of two or more rules are the same, then the one with fewer numbers of conditions in the antecedent has a higher rank. (4) If all the above criteria are

similar, then the rule with a larger class count in the training data set (majority class) has a higher rank than the others

According to Table 1, Rule-5 is the highest ranked rule since its confidence is the largest one among the rest of the rules. Though Rule-1, Rule-6, Rule-7, Rule-9, Rule-10, and Rule-11 have the same confidence; Rule-1 is ranked higher due to its larger support. Still Rule-6, Rule-7, Rule-9, Rule-10, and Rule-11 have the same confidence and support values; but Rule-6 is ranked higher due to the fact that it has less number of values in its antecedent, and so forth.

3.1 High Precedence Classify Correctly Pruning Method (HPC).

Not all of the rules derived during the learning phase can be used in the prediction step, and therefore some rules will be removed. Figure 1 represents the HPC rule pruning procedure. To describe this method, let's use the example of Tables 1 & 2 in which the first ranked rule (Rule-5) is applied on the training data shown in Table 2. We find that this rule partially matches documents 3 and 4, and the class of this rule is identical to the class labels of these documents. So, this rule is inserted into the classifier, and we remove documents 3 and 4 from the training data set. We proceed to the second ranked rule i.e. Rule-1, and we check the applicability of this rule with the remaining training documents. We find that this rule partially matches

Rule-id	Rule	conf	sup	Class count	Rank
1	real=>sport	0.99	0.286	2	2
2	madrid=>sport	0.67	0.285	2	10
3	stock=>acq	0.75	0.428	3	8
4	share=>acq	0.67	0.285	3	9
5	group=>acq	1	0.285	3	1
6	amman=>general	0.99	0.285	2	3
7	madrid & real=>sport	0.99	0.285	2	6
8	share & stock=>acq	0.67	0.285	3	11
9	group & stock=>acq	0.99	0.285	3	4
10	group & share=>acq	0.99	0.285	3	5
11	group&share&stock=>acq	0.99	0.285	3	7

Table 1: Example of a rule-based model (potential rules)

id	Document	Class	Random Rank
1	real, madrid, stock, loss, share	sport	1
2	real, madrid	sport	5
3	stock, share, group	acq	2
4	stock, share, buy, group	acq	3
5	stock	acq	4
6	iraq, amman, jordan	general	7
7	amman, madrid, real	general	6

Table 2: Example of a training data

documents 1, 2, and 7, but the class label for document 7 is not consistent with that of Rule-1. So, we insert Rule-1 into the classifier since it covers correctly at least one document, and we remove documents 1, and 2. The third ranked rule (Rule-6)

```

Input: Given a set of generated rules  $R$ , and training data set  $T$ 
Output: classifier ( $CI$ )
1  $R' = \text{sort}(R)$ ;
2 For each rule  $r_i$  in  $R'$  Do
3     Find all applicable data cases in  $T$  that match  $r_i$ 's condition
4     If  $r_i$  correctly classifies a training case in  $T$ 
5         Insert the rule at the end of  $CI$ 
6         Remove all cases in  $T$  covered by  $r_i$ 
7     end if
8     If  $r_i$  cannot correctly cover any training case in  $T$ 
9         Remove  $r_i$  from  $R$ 
10    end if
11 end
    
```

Figure 1. HPC pruning method

covers correctly two documents 6 and 7, so it gets inserted into the classifier, and documents 6 and 7 are discarded from the training documents set, and we repeat the same steps for the rest of candidate rules until the training set becomes empty. In this example, the classifier contains just four rules, and the remaining candidate rules are deleted.

The difference between this pruning method and the database coverage (Liu et al., 1998) is that in the HPC method a rule gets inserted into the classifier if it partially covers at least one training case. On the other hand, in the database coverage, a rule must fully match the training case antecedent (rule body) in order to be considered.

Category Name	Training set	Testing Set
Acq	1650	719
Crude	389	189
Earn	2877	1087
Grain	433	149
Interest	347	131
Money-FX	538	179
Trade	369	117
Total	6603	2571

Table 3: Number of documents per category (Reuters-21578)

4. EXPERIMENTAL RESULTS

4.1 Text Data Collection

The benchmark used in the experiments is the Reuters-21578 (Lewis, 1998). The Reuters-21578 is the most widely used text data set in the text categorisation research. We used the ModApte version of Reuters-21578. This split leads to a corpus of 9,174

documents consisting of 6,603 training and 2,571 testing documents, respectively.

We tested our pruning procedures within the MCAR algorithm on the seven most populated categories with the largest number of documents assigned to them in the training data set. The experiments are conducted on 2.8 Pentium IV machine with 1GB RAM, and the proposed methods and MCAR are implemented using VB.Net programming language with a minsupp and minconf of 2%, and 40%, respectively. The minsupp has been set to 2% since more extensive experiments reported in (Liu, et al., 1998; Li, et al., 2001, Thabtah, et al., 2004, Thabtah, et al., 2005) suggested that it is one of the rates that achieve a good balance between accuracy and the size of the classifiers. The confidence threshold, on the other hand, has a smaller impact on the behaviour of any AC method and it has been set to 40%.

Table 3 represents the number of documents for each category in the Reuters-21578 data set. On these documents we performed stop word elimination but not stemming, and we select the top 1000 features using Chi Square (Snedecor and Cochran, 1989).

In the following section, we show the results of the proposed pruning procedures, the database coverage method, lazy method, and the case of no pruning.

4.2 Experimental Results

We performed extensive experiments on the seven most populated categories of the Reuters-21578 test collection to compare our proposed methods, the database coverage (Liu et al., 1998), lazy pruning (Baralis and Torino, 2000), and the case of no pruning with reference to the number of rules generated and the predictive accuracy.

Table 4 shows the number of rules derived from the Reuters text collection when different pruning approaches are implemented within MCAR algorithm (Thabtah, et al., 2004). It is obvious from the numbers shown in Table 4 that in general HPC rule pruning method outperforms lazy pruning, database coverage, and the case of no pruning.

Moreover, the number of rules generated without pruning method on "Acq" class is 80, whereas the number of rules derived using HPC pruning procedure is 28. The additional fifty two rules produced in the case of no pruning may decrease the classification accuracy and increase the prediction time. It is obvious from the numbers shown in Table 4 that algorithms, which use lazy pruning approach, often generate many more rules than those that employ other approaches. In particular, for all classification data sets we considered, MCAR using lazy pruning method produced more rules than other considered pruning heuristics.

One of the principle reasons for generating large number of rules by lazy pruning algorithms is due to storing rules that do even cover a single training data case in the classifier. For example, while constructing the classifier by the MCAR algorithm using lazy pruning method, every rule is evaluated to validate whether or not it covers a training data case. If a rule covers correctly a training case, it then will be added into the classifier, otherwise it will be removed. However, rules, which have been never tested, are also added as spare rules into the classifier. These spare rules

are may raise many problems such as user understandability and maintenance. Furthermore, the classification time may increase since the classification system must iterate through the rule set, and consequently these problems limit the use of lazy associative algorithms.

Class	no Pruning	HPC	Database Coverage	Lazy
Acq	80	28	27	40
Crude	8	5	4	6
Earn	172	15	17	55
Grain	5	5	5	5
Interest	4	1	2	4
Money -FX	23	12	12	15
Trade	9	4	6	8
Total	301	70	73	133

Table 4: MCAR number of rules produced when iffrent pruning approaches are used against Reuter data set

Unlike lazy pruning approach, the database coverage method and our proposed methods eliminate the spare rules and that explains its moderate size classifiers. Specifically, MCAR using our proposed methods and MCAR using database coverage algorithms generate reasonable size classifiers if compared with MCAR using lazy pruning method. This enables domain users to benefit from.

Data	No Pruning	HPC	Database Coverage	Lazy pruning
Acq	80.6	99.9	82.6	82.4
Crude	76.2	96.3	80.4	88.4
Earn	99.6	99.7	99.6	99.6
Grain	90.6	96	91.8	94.8
Interest	3.1	69.5	3.1	3.1
Money-FX	49.7	73.7	49.7	49.7
Trade	93.2	94.9	93.2	93.4
Average	70.4	90.0	71.5	73.1

Table 5: Accuracy per class label of the Reuter data derived by the MCAR algorithms using the different rule pruning methods

Table 5 depicts the classification accuracy (%) derived by MCAR algorithm using the different rule pruning methods on the seven most populated categories of Reuters-21578. We also reported the accuracy derived without pruning in column 2. The accuracy numbers have been generated using a minsupp of 2% and a minconf of 40%. Table 5 indicates that when the HPC pruning is employed, the classification accuracies derived are

better than those of database coverage and the lazy pruning methods. Particularly, the won-tied-loss records of HPC records against no pruning, database coverage and lazy pruning are 7-0-0, 7-0-0 and 7-0-0, respectively. In fact, we achieved on average +19.6%, +18.5%, +16.9% higher prediction rates within MCAR algorithm than no pruning, the database coverage and lazy pruning, respectively on the Reuter text collection.

5. CONCLUSION

In this paper, we proposed a new rule pruning method within associative classification mining. We conducted experiments on seven categories selected from the Reuters-21578 text collection using the developed pruning procedures and existing pruning methods in associative classification. The bases of the comparison are the number of produced rules and the accuracy, and we implemented all pruning methods within a known associative algorithm called MCAR. The experimental results revealed that the HPC method outperformed all other pruning techniques with reference to predictive accuracy and number of rules generated. Particularly, it achieved on average +19.6%, +18.5%, +16.9% higher prediction rates within MCAR algorithm than no pruning, the database coverage and lazy pruning, respectively. In near future, we intend to expand our research to include other rule pruning heuristics in the areas of decision trees, statistics, and rule induction.

REFERENCES

- [1] Antonie M. and Zaiane O. (2004). An associative classifier based on positive and negative rules. Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (pp. 64 - 69), Paris, France.
- [2] Antonie M. and Zaiane O. (2002). Text Document Categorization by Term Association, Proceedings of the IEEE International Conference on Data Mining (ICDM '2002), (pp.19-26), Maebashi City, Japan.
- [3] Baralis E., Chiusano S. and Garza P. (2008). A Lazy Approach to Associative Classification. IEEE Trans. Knowl. Data Eng. 20(2): 156-171.
- [4] Baralis E., Chiusano S. and Garza P. (2004). On support thresholds in associative classification. Proceedings of the 2004 ACM Symposium on Applied Computing, (pp. 553-558). Nicosia, Cyprus.
- [5] Baralis E. and Torino P. (2000). A lazy approach to pruning classification rules. Proceedings of the 2002 IEEE ICDM'02, (pp. 35). Maebashi City, Japan.
- [6] Cohen W. (1995). Fast effective rule induction. Proceedings of the 12th International Conference on Machine Learning, (pp. 115-123). CA, USA.
- [7] Duda R. and Hart P. (1973). Pattern classification and scene analysis. John Wiley & son.
- [8] Kundu G., Islam M., Munir S. and Bari M. (2008). ACN: An Associative Classifier with Negative Rules, Computational Science and Engineering, vol. 0, no. 0, (pp. 369-375), 11th IEEE International Conference on Computational Science and Engineering.
- [9] Lewis D. (1998). Reuters 21578 text categorisation test collection. <http://www.daviddlewis.com/resources/testcollection/s/reuters21578>.

- [10] Li B., Li H., Wu M. and Li P. (2008). Multi-label Classification based on Association Rules with Application to Scene Classification, *icys*, (pp.36-41), 2008 The 9th International Conference for Young Computer Scientists.
- [11] Li W., Han J. and Pei J. (2001). CMAR: Accurate and efficient classification based on multiple-class association rule. *Proceedings of the ICDM'01*, (pp. 369-376). San Jose, CA.
- [12] Liu B., Hsu W. and Ma Y. (1998). Integrating classification and association rule mining. *Proceedings of the KDD*, (pp. 80-86). New York, NY.
- [13] Liu B., Ma Y., Wong, C-K. and Yu. P. (2003). Scoring the data using association rules. *Applied Intelligence*, 18(2003): 119-135.
- [14] Merz C. and Murphy P. (1996). UCI repository of machine learning databases. Irvine, CA, University of California, Department of Information and Computer Science.
- [15] Niu Q., Xia S. and Zhang L. (2009). Association Classification Based on Compactness of Rules, *wkdd*, (pp.245-247), Second International Workshop on Knowledge Discovery and Data Mining.
- [16] Quinlan J. (1998). Data mining tools See5 and C5.0. Technical Report, RuleQuest Research.
- [17] Quinlan J. (1993). C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann.
- [18] Quinlan J. and Cameron-Jones R. (1993). FOIL: A midterm report. *Proceedings of the European Conference on Machine Learning*, (pp. 3-20), Vienna, Austria.
- [19] Quinlan J. (1987). Simplifying decision trees. *International journal of man-machine studies*, 27(1987): 221-248.
- [20] Snedecor W. and Cochran W. (1989). *Statistical Methods*, Eighth Edition, Iowa State University Press.
- [21] Thabtah, F., Cowling, P., and Peng, Y. (2005) MCAR: Multi-class classification based on association rule approach. *Proceeding of the 3rd IEEE International Conference on Computer Systems and Applications* (pp. 1-7).Cairo, Egypt.
- [22] Thabtah, F., Cowling, P., and Peng, Y. (2004) MMAC: A new multi-class, multi-label associative classification approach. *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM '04)*, (pp. 217-224). Brighton, UK. (Nominated for the Best paper award).
- [23] Topor R. and Shen H. (2001). Construct robust rule sets for classification. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 564-569). Edmonton, Alberta, Canada.
- [24] Xu X., Han G. and Min H. (2004). A novel algorithm for associative classification of images blocks. *Proceedings of the fourth IEEE International Conference on Computer and Information Technology*, (pp. 46-51). Lian, Shiguo, China.
- [25] Yin X. and Han J. (2003). CPAR: Classification based on predictive association rule. *Proceedings of the SDM* (pp. 369-376). San Francisco, CA.
- [26] Yoon Y. and Lee G. (2008). Text Categorization Based on Boosting Association Rules, *icsc*, pp.136-143, 2008 IEEE International Conference on Semantic Computing.

Metaphorical and Contextual Affect Detection in an Intelligent Agent

Li Zhang¹

Abstract. Real-time affect detection from open-ended text-based dialogue is challenging but essential for the building of effective intelligent user interfaces. In this paper, we report updated developments of an affect detection model from text, including affect detection from two particular types of metaphorical affective expressions (food and cooking metaphors) and affect detection based on context. The overall affect detection model has been embedded in an intelligent conversational AI agent interacting with human users under loose scenarios. Evaluation for the updated affect detection component is also provided. Our work contributes to the symposium's themes on human-agent interaction based on affect sensing, affective computing and daily life dialog systems.

1 INTRODUCTION

We have reported an affect detection component, EMMA (emotion, metaphor and affect) on detecting simple and complex emotions, meta-emotions, value judgments etc from literal expressions in English in our previous work [1]. Recently, metaphorical language has drawn our attention since it has been widely used to provide effective vivid description. Fainsilber and Ortony [2] commented that “an important function of metaphorical language is to permit the expression of that which is difficult to express using literal language alone”. In our previous application, we have also identified several metaphorical affective expressions from the collected transcripts (such as animal metaphor (“X is a rat”) and affects as external entities metaphor (“joy ran through me”) [1, 3, 4]) and conducted initial development on affect sensing from these metaphorical phenomena.

The work presented here reports further developments on affect detection from another two particular English metaphorical expressions with affect implication, which include food metaphor (“X is walking meat”, “X has a pizza face”) and cooking metaphor (“the lawyer grilled the witness on the stand”, “I knew I was cooked when the teacher showed up at the door”) based on the syntactical analysis of metaphorical phenomena shown in previously collected transcripts in our study and another metaphor online dictionary². The transcripts used in our work were previously created by the testing subjects (secondary school students) in previous user study [1]. Also, since context plays an important role in the interpretation of the affect conveyed by the user during the interaction, we have used Markov chain modelling (trained by the collected previous

transcripts) and dynamic programming to recommend the most possible affect conveyed in the current input when the affect detection component interprets ‘neutral’ solely based on the analysis of the input itself during the proper improvisation (in our study, we use ‘proper improvisation’ to indicate that users stayed in their characters for creative role-playing after they became familiar with the chosen scenarios and each character).

The new developments have been incorporated into the affect detection component, which has been embedded in an intelligent conversational agent, engaged in a drama improvisation with human users under loose scenarios (e.g. school bullying and Crohn’s disease designed for our application). The affective states detected by this component have been used to produce expressive behaviours for human-controlled characters. The conversational AI agent also provides appropriate responses based on the detected affect from users’ input in order to stimulate the improvisation.

In both scenarios, the AI agent plays a minor role in drama improvisation. It plays a close friend of the bullied victim (the leading role) in school bullying scenario, who tries to stop the bullying and a close friend of the sick leading character in Crohn’s disease scenario who tries to give support to his friend with the decision on his life-changing operation.

We have also analyzed affect detection performance based on previously collected other transcripts from user testing by calculating agreements via Cohen’s Kappa between two human judges, human judges and the AI agent with and without the new development respectively in order to verify the efficiency of the metaphorical and contextual affect sensing.

The content is arranged in the following way. We report relevant work in section 2 and the new developments on affect detection from food and cooking metaphor in section 3. Affect interpretation based on context is discussed in section 4. Newly produced evaluation results of the affect detection component are reported in section 5. Finally we summarize our work and point out future directions for further developments in section 6.

2 RELATED WORK

Textual affect sensing is a rising research branch for natural language processing. ConceptNet [5] is a toolkit to provide practical textual reasoning for affect sensing for six basic emotions, text summarization and topic extraction. Shaikh et al. [6] provided sentence-level textual affect sensing to recognize evaluations (positive and negative). They adopted a rule-based domain-independent approach, but they haven’t made attempts to recognize different affective states from open-ended text input.

Although *Façade* [7], a one act interactive drama, included shallow natural language processing for characters’ open-ended

¹ School of Computing, Teesside University, TS1 3BA, UK. Email: {l.zhang}@tees.ac.uk.

² http://knowgramming.com/cooking_metaphors.htm

Proceedings of the Linguistic And Cognitive Approaches To Dialog Agents Symposium, Rafal Rzepka (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

utterances, the detection of major emotions, rudeness and value judgements is not mentioned. Zhe and Boucouvalas [8] demonstrated an emotion extraction module embedded in an Internet chatting environment (see also Boucouvalas [9]). It used a part-of-speech tagger and a syntactic chunker to detect the emotional words and to analyse emotion intensity for the first person (e.g. 'I' or 'we'). Unfortunately the emotion detection focused only on emotional adjectives, and did not address deep issues such as figurative expression of emotion (discussed below). Also, the concentration purely on first-person emotions is narrow. There has been relevant work on general linguistic cues that could be used in practice for affect detection (e.g. Craggs and Wood [10]).

There is also well-known research work on the development of emotional conversational agents. Egges et al. [11] have provided virtual characters with conversational emotional responsiveness. Elliott et al. [12] demonstrated tutoring systems that reason about users' emotions. They believe that motivation and emotion play very important roles in learning. Virtual tutors have been created in a way that not only having their own emotion appraisal and responsiveness, but also understanding users' emotional states according to their learning progress. Aylett et al. [13] also focused on the development of affective behaviour planning for the synthetic characters. Cavazza et al. [14] reported a conversational agent embodied in a wireless robot to provide suggestions for users on a healthy living lifestyle. Hierarchical Task Networks planner and semantic interpretation have been used in this work. The cognitive planner plays an important role in assisting with dialogue management, e.g. giving suggestions to the dialogue manager on what relevant questions should be raised to the user according to the healthy living plan currently generated. The user's response has also been adopted by the cognitive planner to influence the change of the current plan. The limitation of such planning systems is that they normally work reasonably well within the pre-defined domain knowledge, but they will strike when open-ended input going beyond the planner's knowledge has been used intensively during interaction. The system presented here intends to deal with some of such challenges.

Our work is distinctive in the following aspects: (1) affect detection from metaphorical expressions; (2) real-time affect sensing for basic and complex emotions, meta-emotions, value judgments etc in improvisational role-play situations from open-ended user input; (3) affect detection for second and third person cases ('you', 'she', 'they') (since English belongs to low context languages, it would allow us to interpret affect conveyed by the second and third person cases. However, in high context languages (e.g. Japanese), the interpretation of emotions conveyed by others rather than the experiencing subject will only be regarded as an opinion); and (4) affect interpretation based on context profiles.

3 FURTHER DEVELOPMENT ON METAPHORICAL AFFECT DETECTION

Before we introduce the new developments on affect detection from metaphor and context profiles, we introduce our previous work on affect detection and responding strategy for the AI agent briefly. Our original system has been developed for age 14-16 secondary school students to engage in role-play

situations under loose scenarios in virtual social environments. Without pre-defined constrained scripts, the human users could be creative in their role-play within the highly emotionally charged scenarios – school bullying and Crohn's disease. The AI agent could be activated to interact with human actors by playing a minor bit-part character in both scenarios as mentioned earlier. We have used responding regimes for the conversational AI agent in order to stir up the discussion and stimulate the improvisation, e.g. making responses when its confidence on interpreting affect from users' input is high. The language used by the secondary school students during their role-play is highly diverse with various online chatting features, such as abbreviations (e.g. 'den' (then), 'r' (are)), acronyms (e.g. 'lol' (laughing out loud)) and slang. Our previous work had pre-processing procedures to deal with abbreviations, acronyms, misspellings and slang [1]. Diverse metaphorical language phenomena have also been used to express emotions during the interaction, which challenge any NLP system greatly which intends to pursue logical semantic and syntactic analysis or affect/opinion mining and draw our attention for further study.

Affect detection from food metaphor

Food has been used extensively as metaphor for social position, group identity, religion, etc. E.g. food could be used as a metaphor for national identity. British have been called 'roastbeefs' by the French, while French have been referred to as 'frogs' by the British. It has also been used to indicate social hierarchy. E.g. in certain Andean countries, potatoes have been used to represent poor rural farmers of native American descent and white flour and bread have been used mainly to refer to wealthy European descent³. In our school bullying scenario, the big bully has called the bullied victim (Lisa) names, such as "u r a pizza", "Lisa has a pizza face" to exaggerate that fact that the victim has acne. Another most commonly used food metaphor is to use food to refer to a specific shape. E.g. body shape could be described as 'banana', 'pear' and 'apple'⁴. In our application, "Lisa has a pizza face" could also be interpreted as Lisa has a 'round (shape)' face. Therefore, insults could be conveyed in such food metaphorical expression. We especially focus on the statement of 'second-person/a singular proper noun + present-tense copular form + food term' to extract affect. A special semantic dictionary has been created by providing semantic tags to normal English lexicon. The semantic tags have been created by using Wmatrix [15], which facilitates the user to obtain corpus annotation with semantic and part-of-speech tags to compose dictionary. The semantic dictionary created consists mainly of food terms, animal names, measureable adjectives (such as size) etc with their corresponding semantic tags due to the fact they have the potential to convey affect and feelings.

In our application, Rasp, a statistical-based domain-independent robust parsing system for English, informs the system the user input with the desired structure - 'second-person/a singular proper noun + present-tense copular form + noun phrases' (e.g. "Lisa is a pizza", "u r a hard working man", "u r a peach"). The noun phrases are examined in order to recover the main noun term. Then its corresponding semantic

³ <http://www.answers.com/topic/food-as-metaphor>

⁴ <http://jsgfood.blogspot.com/2008/02/food-metaphors.html>

Proceedings of the Linguistic And Cognitive Approaches To Dialog Agents Symposium, Rafal Rzepka (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

tag is derived from the composed semantic dictionary if it is a food term, or an animal-name etc. E.g. “u r a peach” has been regarded as “second-person + present-tense copular form + [food-term]”. WordNet [16] has been employed in order to get the synset of the food term. If among the synset, the food term has been explained as a certain type of human being, such as ‘beauty’, ‘sweetheart’ etc. Then another small slang-semantic dictionary collected in our previous study containing terms for special person types (such as ‘freak’, ‘angle’) and their corresponding evaluation values (negative or positive) has been adopted in order to obtain the evaluation values of such synonyms of the food term. If the synonyms are positive (e.g. ‘beauty’), then we conclude that the input is an affectionate expression with a food metaphor (e.g. “u r a peach”). The processing procedures are listed in the following for the input “u r a peach”:

1. Rasp recognizes the input has a structure of ‘second-person (you) + present-tense copular form (are) + noun phrases (a peach)’;
2. The newly composed semantic dictionary is used to obtain the semantic tag for the main noun term (‘peach’) in the noun phrase, ‘peach’ -> ‘food term’;
3. Thus the input is regarded as ‘second-person (you) + present-tense copular form (are) + [food-term: peach]’. The system interprets that the input is a food metaphor;
4. The food term, ‘peach’, is sent to WordNet to get its synonyms. The synonyms include special person types, such as ‘beauty’ and ‘sweetheart’;
5. The special slang-semantic dictionary containing special person types and their corresponding evaluation values is used to obtain the evaluation values (positive) of special person types ‘beauty’ and ‘sweetheart’;
6. Since the evaluation values are positive, the system concludes the input expresses affectionate with a food metaphor.

However, in most of the cases, WordNet doesn’t provide any description of types of human beings when explaining a food term (e.g. ‘pizza’, ‘meat’ etc). According to the nature of the scenarios (e.g. bullying) we used, we simply conclude that the input implies insulting with a food metaphor when calling someone food terms (“u r walking meat”, “Lisa is a pizza”).

Another interesting phenomenon drawing our attention is food as shape metaphor. As mentioned earlier, food is often used as a metaphor to refer to body shapes (e.g. “you have a pear body shape”, “Lisa has a garlic nose”, “Lisa has a pizza face”). They might indicate literal truth, but most of which are potentially used to indicate very unpleasant truth. Thus they could be regarded as insulting. We extend our semantic dictionary created with the assistance of Wmatrix [15] by adding terms of physiological human body parts, such as face, nose, body etc. For the user’s input with a structure of ‘second-person/a singular proper noun + have/has + noun phrases’ informed by Rasp, the system provides a semantic tag for each word in the object noun phrase. If the semantic tag sequence of the noun phrase indicates that it consists of a food term followed by a physiological term (‘pizza face’), the system interprets that the input implies insulting with a food metaphor. Our processing could also identify input such as “she is a cake lover/eater” as literal expressions since the main objective terms are not food terms but persons/consumers. In our study, we have also made attempts to classify several metaphorical

phenomena (including food metaphor) from literal expressions no matter if such metaphorical expressions conveyed any affect or not.

However, examples, such as “you have a banana body shape” and “you are a meat and potatoes man”, haven’t been used to express insults, but instead the former used to indicate a slim body and the latter to indicate a hearty appetite and robust character. Other examples such as “you are what you eat” could be very challenging theoretically and practically. They also indicate the direction of the future extension of our current system.

Affect detection from cooking metaphor

Another type of metaphor that has great potential to carry affect implication is cooking metaphor. Very often, the agent himself/herself would become the victim of slow or intensive cooking (e.g. grilled, cooked, etc). Or one agent can perform cooking like actions towards another agent to realize (at least most of the time) punishment or torture. Examples are as follows, “the lawyer grilled the witness on the stand”, “he knew he was cooked when he saw his boss standing at the door”, “he basted her with flattery to get the job”, “She knew she was fried when the teacher handed back her paper” etc.

In these examples, the suffering agents have been regarded as food. They bear the results of intensive or slow cooking. Thus these agents who suffer from such cooking actions carried out by other agents tend to feel pain and sadness, while the ‘cooking performing’ agents may take advantage of such actions to achieve their intentions, such as persuasion, punishment or even enjoyment. The syntactic structures of some of the above examples also indicate the submissive of the suffering agents. E.g. in the examples, passive sentences (“he knew he was going to be grilled when he got home”) have been very often used to imply unwillingness and victimised of the subject agents who are in fact the objects of the cooking actions described by the verb phrases (“X + copular form + passive cooking action”). In other examples, the cooking actions have been explicitly performed by the subject agents towards the object agents to imply the former’s potential willingness and enjoyment and the latter’s potential suffering and pain (“A + [cooking action] + B”).

Thus in our application, we focus on the above two particular types of expressions. We use Rasp to recognize user input with such syntactic structures (‘A + copular form + VVN’, ‘A + VV0/VVD/VVZ + B’). Many sentences could possess such syntactic information (e.g. “Lisa was bullied”, “he grills Lisa”, “I was hit by a car”, “Lisa was given the task to play the victim role”, “he knew he was cooked”, “I steamed it” etc), but few of them are cooking metaphors. Therefore we need to resort to semantic profiles to recognize the metaphorical expressions. Rasp has also provided a syntactic label for each word in the user input. Thus the main verbs were identified by their corresponding syntactic labels (e.g. ‘given’ labelled as ‘past participle form of lexical verbs (VVN)’, ‘likes’ and ‘grills’ labelled as ‘-s form of lexical verbs (VVZ)’ and the semantic interpretation for their base forms is discovered from WordNet [16]. Since WordNet has provided hypernyms (Y is a hypernym of X if every X is a (kind of) Y) for the general noun and verb lexicon, ‘COOK’ has been derived as the hypernym of the verbs described cooking actions. E.g. ‘boil’, ‘grill’, ‘steam’, and

Proceedings of the Linguistic And Cognitive Approaches To Dialog Agents Symposium, Rafal Rzepka (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

'simmer' are respectively interpreted as one way to 'COOK'. 'Toast' is interpreted as one way to 'HEAT UP' while 'cook' is interpreted as one way to 'CREATE', or 'CHEAT' etc. One verb may recover several hypernyms and in our application, we collect all of them. Another evaluation profile [17] is resorted to in order to recover the evaluation values of all the hypernyms for a particular verb. If some hypernyms are negative (such as 'CHEAT') and the main object of the overall input refers to first/third person cases or a singular proper noun ('him', 'her', or 'Lisa'), then the user input (e.g. "he basted her with flattery to get the job") conveys potential negative affect (e.g. pain and sadness) for the human objects and potential positive affect (e.g. persuasion or enjoyment) for the subjects. If the evaluation dictionary fails to provide any evaluation value for any hypernym (such as 'COOK' and 'HEAT UP') of the main verbs, then we still assume that 'verbs implying COOK/HEAT UP + human objects' or 'human subjects + copular form + VVN verbs implying COOK/HEAT UP' may indicate negative emotions both for the human objects in the former and the human subjects in the latter. Two example implementations are provided in the following.

E.g. for the user input "I was fried by the head teacher", the step-by-step processing is as follows:

1. Rasp identifies the input has the following structure: 'PPIS1 (I) + copular form (was) + VVN (fried)';
2. 'Fry' (base form of the main verb) is sent to WordNet to obtain its hypernyms, which include 'COOK', 'HEAT' and 'KILL';
3. The input has the following syntactic semantic structure: 'PPIS1 (I) + copular form (was) + VVN (Hypernym: COOK)', thus it is recognised as a cooking metaphor;
4. The three hypernyms are sent to the evaluation profile to obtain their evaluation values, thus 'KILL' is labelled as negative while others can't be returned with any evaluation values from the profile;
5. The input is transformed into: "PPIS1 (I) + copular form (was) + VVN (KILL: negative)";
6. The subject is a first person case, then from common-sense knowledge, the input indicates the user who is speaking suffered from a negative action and may lead to a negative emotional state (e.g. 'sad' or 'angry').

Another example input is "she grilled him in front of the class". Since the hypernyms of 'grill' only contain 'COOK' (recognized as a cooking metaphor), we cannot trace it from the evaluation profile to return an evaluation value. However, since the input has the following syntactic and semantic structure: "PPHS1 (she) + COOK + PPHO1 (him)", it implies that a third-person subject performed a cooking action towards another human object. The human object ('him') may experience 'sad' (or 'angry') emotional state.

Sometimes, our processing may not be able to recognize the metaphorical phenomena (the cooking metaphor). However it is able to interpret the affect conveyed in the user input. E.g. if we have input "he basted her with flattery to get the job", the recognition process is as follows:

1. Rasp recognises the input has the main syntactic structure of "PPHS1 (he) + VVD (basted) + PPHO1 (her)";
2. 'Baste' (verb base form) is sent to WordNet to obtain its hypernyms, which include 'MOISTEN', 'BEAT', and 'SEW';
3. The hypernyms do not include 'COOK', thus it is not recognised as a cooking metaphor;

4. The hypernyms are sent to the evaluation profile to obtain their evaluation values. We obtain negative evaluation value for the hypernym 'BEAT' from the profile.

5. The input has the following syntactic and semantic structure: 'PPHS1 (he) + VVD (BEAT: negative) + PPHO1 (her)'. It implies that a third-person subject performed a negative action towards another human object.

6. The human object ('her') may experience 'sad' or 'angry' emotional state.

Although the processing sometimes is not ideal, our method still manages to retrieve the affect embedded in the user input. Our processing is indeed at a very initial stage to sense affect from cooking metaphor and only can successfully detect approximate 30% of the cooking metaphorical phenomena shown in the online metaphor dictionary mentioned earlier. However it has pointed out promising directions for affect detection from figurative language. After our efforts to improve the performance of affect sensing from individual turn-taking user input, we focus on affect justification using context profiles. In the future work, we also intend to use metaphor ontology to recognise and generate metaphors.

4 AFFECT SENSING FROM CONTEXT PROFILES

Our previous affect detection has been performed solely based on individual turn-taking user input. Thus the context information has been ignored. However, such information plays a very important role in the interpretation of the affect conveyed during interaction. In this section, we are going to discuss the linguistic contextual indicators, cognitive emotion simulation from communication context and our approach developed based on the above features.

Linguistic indicators for contextual communication

We have noticed some linguistic indicators for contextual communication in the recorded transcripts. One useful indicator is (i) imperatives (such as "shut up", "go on then", and "bring it on"), which are normally used to imply negative or positive responses to the previous speaking characters. Other useful contextual indicators are (ii) prepositional phrases (e.g. "by who?", "in the green room?"), semi-coordinating conjunctions (e.g. "so we are good then"), subordinating conjunctions ("because Lisa is a dog") and coordinating conjunctions ('and', 'or' and 'but'). These indicators are normally used by the current 'speaker' to express further opinions or gain further confirmation from the previous speakers. The following is part of an example transcript from school bullying scenario, which is mainly about the bully, Mayid, picking on the new comer to the school, Lisa. Lisa's friends, Elise and Dave, are trying to stop the bullying. The school teacher, Mrs Parton, also tries to find out what is going on. The italic input from the bully, Mayid, is the improvisation based on context. The linguistic contextual indicators (such as 'too' and 'and') are used to indicate the intention of the big bully of strengthening his threatening towards other characters (especially to Lisa, the bullied victim). In this example, Dave was played by EMMA.

Mayid: im am the bully ha ha ha

(AI) Dave: Hey guys! Lisa, r u all right?

Lisa: i am ok

Mayid: Lisa is mine

Elise: lisa, why r u upset??

Elise: u look ok yesterday

(AI) Dave: Hey m8, r u all right? u look terrible.

Mayid: i run Lisa over and kill her ha ha

Lisa: i am ok

Mayid: *my mum is a bully too*

Elise: i love lisa

(AI) Dave: That sounds great

Mayid: *and my dad*

In addition, (iii) short phrases for questions are also used frequently in the recorded transcripts to convey contextual communication, e.g. “where?”, “who?”, “who is Dave” or “what”. However, except for the potential implication of ‘confusion’, there is rarely any other affective state expressed in such short question style responses.

(iv) The character names are also normally used in the user input to indicate that the current input is intended for a particular character, e.g. “Dave go away”, “Mrs Parton, say something”, “Dave what has got into you?” etc. Very often, such expressions have been used to imply potential emotional contextual communication between the current speaking characters and the named characters. Therefore the current speaking characters may imply at least ‘approval’ or ‘disapproval’ towards the opinions/comments provided by the previous named speaking characters. Finally there are also (i) some other well known contextual indicators such as ‘yeah/yes followed by a sentence (“yeah, yeah, we are good”, “yeah, we will see”, “yeah, we cool Lisa”)', “I think so”, ‘no/nah followed by a sentence’, “me too”, “exactly”, “thanks”, “sorry”, “lol”, “grrrr”, “hahahaha”, “ohhhh” etc. Such expressions are normally used to indicate affective responses to the previous user input.

Since natural language is ambiguous and there are cases that contextual information is required in order to appropriately interpret the affect conveyed in the input (e.g. “go on then”, “and my dad”), our approach reported in the following integrates the above contextual linguistic indicators with cognitive contextual emotion prediction to uncover affect conveyed in emotionally ambiguous input.

Emotion modelling in communication context

There are also other aspects which may influence the affect conveyed in the communication context. E.g. in our application, the affect conveyed by the speaking character himself/herself in the recent several turn-taking, the ‘improvisational mood’ that the speaking character is created, and emotions expressed by other characters, especially by the contradictory ones (e.g. the big bully), have great potential to influence the affect conveyed by the current speaking character (e.g. the bullied victim). Sometimes, the story themes or topics also have potential impact to emotions or feelings expressed by characters. For example, people tend to feel ‘happy’ when involved in discussions on positive topics such as harvest or raising salary, while people tend to feel ‘sad’ when engaged in the discussions on negative themes such as economy breakdown, tough examination or funeral. In our application, although the hidden story sub-themes used in the scenarios are not that dramatic, they are still highly emotionally charged and used as the signals

for potential changes of emotional context for each character. E.g. In the school bullying scenario, the director mainly provided interventions based on several main sub-themes of the story to push the improvisation forward, i.e. “Mayid starts bullying Lisa”, “why Lisa is crying”, “why Mayid is so nasty/a bully”, “how Mayid feels when his uncle finds out his behaviour” etc. From the inspection of the recorded transcripts, when discussing the topic of “why Lisa is crying”, we noticed that Mayid (the bully) tends to be really aggressive and rude, while Lisa (the bullied victim) tends to be upset and other characters (Lisa’s friends and the school teacher) are inclined to show anger at Mayid. For the improvisation of the hidden story sub-theme “why Mayid is so nasty/a bully”, the big bully changes from rude and aggressive to sad and embarrassed (e.g. because he is abused by his uncle), while Lisa and other characters become sympathetic (and sometimes caring) about Mayid. Usually all characters are trying to create the ‘improvisational mood’ according to the guidance of the hidden story sub-themes (provided via director’s intervention). Therefore, the story sub-themes could be used as the indicators for potential emotional context change. The emotion patterns expressed by each character within the improvisation of each story sub-theme could be very useful for the prediction of the affect shown in a similar topic context, although the improvisation of the characters is creative within the loose scenario. It will improve the performance of the emotional context prediction if we allow more emotional profiles for each story sub-theme to be added to the training data to reflect the creative improvisation (e.g. some improvisations went deeper for a particular topic).

Therefore, a Markov chain is used to learn from the emotional context shown in the recorded transcripts for each sub-theme and for each character, and generate other possible reasonable unseen emotional context similar to the training data for each character. Then a dynamic algorithm is used to find the most resembling emotional context for any given new situation from Markov chain’s training and generated emotional contexts. The most resembling emotional context is used to adjust/predict the affect conveyed in the current user input. The recommended affect will also be further justified by the most recent affect conveyed by other characters (e.g. a most recent ‘insulting/rude’ input from Mayid may cause Lisa and her friends to be ‘angry’).

Markov chains are usually used for word generation. In our application, they are used to record the frequencies of several emotions showing up after one particular emotion. In our previous work, we detected 25 affective states including basic (e.g. ‘happiness’) and complex emotions (e.g. ‘embarrassment’ and other affective states described by OCC model), meta-emotions, value judgments etc. We have employed 12 most commonly used emotions (caring, arguing, disapproving, approving, grateful, happy, sad, threatening, embarrassed, angry/rude, scared and sympathetic) out of the 25 affective states in our present work on contextual emotion analysis and prediction. A matrix has been constructed dynamically for the 12 most commonly used emotions in our application with each row representing the previous emotion followed by the subsequent emotions in columns. The Markov chains employ roulette wheel selection to ensure to produce a greater probability to select emotions with higher frequencies than emotions with lower occurrences. This will allow the generation

From the inspection of the previous testing transcripts, we noticed that the human director normally intervened to suggest a topic change (e.g. “find out why Mayid is a bully”). Thus for a testing situation for a particular character, we use the emotion context attached with his/her user input starting right after the most recent director’s intervention and ending at his/her last second input, since such a context may belong to one particular topic. By using the dynamic algorithm, a particular series of emotions for a particular story sub-theme has been regarded as the most resembling context to the test emotional situation and an emotional state is recommended as the most possible emotion for the current user input. The detailed explanation of the approach taken and example runs are in the following. Since the most recent affect histories of other characters may also have impact on the affect conveyed by the current speaking character, the recommended affect by the dynamic algorithm will also be further evaluated by these affect histories.

At the training stage, first of all, the school bullying transcripts collected from previous user testing have been divided into several topic sections with each of them belonging to one of the story sub-themes (including “Mayid starts bullying Lisa”, “why Lisa is crying”, “why Mayid is so nasty/a bully”, “how Mayid feels when his uncle finds out his behaviour” etc). The classification of the sub-theme is mainly based on the human director’s intervention which was recorded in the transcripts. Then we have used two human annotators to mark up the affect of every turn-taking input in the transcripts using context inference. Thus for each character, we have summarized a series of emotions expressed throughout the improvisation of a particular story sub-theme. Since the improvisation is creative under the loose scenario, some of the sub-themes (e.g. “why Mayid is so nasty”) have been suggested for improvisation for one than once in some transcripts and some of the topics (e.g. “why Lisa is crying”) are only shown in a few of the collected transcripts. We made attempts to gather emotional contexts as many as possible (at least 4 emotional contexts) for each character for the improvisation of each sub-theme in order to enrich the training data.

The following is a small portion of one recorded transcript used for the training of the Markov chain. The human annotators have marked up the affect expressed in each turn-taking user input.

DIRECTOR: why is Lisa crying?

Elise Brown [caring]: lisa stop cryin

Lisa Murdoch [disagree]: lisa aint crying!!!!

Dave Simons [caring]: i dunno! y u cryin lisa?

Mayid Rahim [rude]: cuz she dnt realise she is lucky to b alive

Elise Brown [angry]: beat him up! itss onlii fat..he’ll go down straight away

Mayid Rahim [rude]: lisa, y u crying? u big baby!

Mrs Parton [caring]: lisa, r u ok?

E.g. the emotional context for Mayid from the above example is: ‘rude’ and ‘rude’ (we use one letter to represent each emotional label, e.g. the emotional context here is represented as ‘R R’), and in the similar way, the emotional contexts for other characters have been derived from the above example, which are used as the training data for the Markov chain for the topic “why Lisa is crying”. We have summarised

the emotional contexts for each story sub-theme for each character from 4 school bullying transcripts and used them for the training of the Markov chain. The topics we have taken into consideration at the training stage are the following: “Mayid starts bullying.”, “why is Lisa crying?”, “why is Mayid nasty/a bully?” and “how does mayid feel if his uncle knew about his behaviour?”

At the test stage, our affect detection component, EMMA, is integrated with an AI agent and detects affect for each user input solely based on the analysis of individual turn-taking input itself as usual. The above algorithms for context-based affect sensing will be activated when the affect detection component recognizes ‘neutral’ from the current input during the emotionally charged proper improvisation after all the characters have known each other and went on the virtual drama stage. First of all, the linguistic indicators are used to identify if the input with ‘neutral’ implication is a communication contextual input. E.g. we mainly focus on the checking of the five contextual implications we mentioned previously, including imperatives, prepositional phrases and conjunctions, simplified question sentences, character names, and other commonly used contextual indicators (such as ‘yea/yes/yeah’ or ‘yea/yes/yeah followed by a sentence’). If any of the above contextual indicators exists, then we further analyse the affect embedded in the input with contextual emotion modelling reported here.

E.g. we have collected the following transcript for testing.

DIRECTOR: U R IN THE PLAYGROUND (indicating bullying starts)

1. Lisa Murdoch: leave me alone! [angry]

2. Mayid Rahim: WAT U GONNA DU ? [neu] -> [angry]

3. Mayid Rahim: SHUT UR FAT MOUTH [angry]

4. Elise Brown: grrrrr [angry]

5. Elise Brown: im telin da dinna lady! [threatening]

6. Mayid Rahim: go on den [neutral] -> [angry]

7. Elise Brown: misssssssssssss [neu]

8. Elise Brown: lol [happy]

9. Lisa Murdoch: mayid u gna gt banned [threatening]

10. Mayid Rahim: LOL [happy]

11. Mayid Rahim: BY HU [neu] -> [angry]

12. Elise Brown: mayid got a tongue yo! [angry]

The affect detection component detected that Lisa was ‘angry’ by saying “leave me alone!”. It also sensed that Mayid was ‘neutral’ by saying “WAT U GONNA DU (what are you going to do)?” without consideration of context. From Rasp, we obtained that the input is a simplified question sentence. Therefore, it implies that it could be a situation raised by the previous context (e.g. previous user input from Lisa) and the further processing for emotion prediction is activated. Since we also don’t have an emotional context yet at this stage for Mayid (the very first input from Mayid after the director’s intervention), we cannot resort to the Markov chain and dynamic programming algorithms at this stage to predict the affect. However, we could use the emotional context of other characters to predict the affect for Mayid’s current input since we believe that an emotional input from a character, especially from an opponent character, has great potential to affect the emotions expressed by the current speaking character.

In the most recent chatting and affect history, there is only one user input from Lisa after the director’s intervention, which implied ‘anger’. Since Lisa and Mayid have a negative

29 March – 1 April 2010, De Montfort University, Leicester, UK relationship, we predict that Mayid currently experiences negative emotion. Since capitalisations have been used in Mayid’s input, then we conclude that the affect implied in the input could be ‘angry’. However, EMMA could be fooled if the affect histories of other characters fail to provide any useful indication for reasoning of the emotion conveyed in the current user input (e.g. if Lisa implied ‘neutral’ in the most recent input, then the interpretation of the affect conveyed by Mayid would still be ‘neutral’).

EMMA also detected affect for the 3rd, 4th, and 5th user input in the above example (based on individual turn-taking) until it detected ‘neutral’ again from the 6th input “go on den (go on then)” from Mayid. Since it is an imperative mood sentence, the input may imply a potential (emotional) response to the previous speaking character. Sometimes affect expressed in the imperatives are very explicit, e.g. “shut up/it (rude/angry)”. However, there are imperatives which convey affect very implicitly, e.g. the current input, “go on then”. Further processing is required to uncover the affect embedded in it. Thus the emotional context profile for the Mayid character is retrieved, i.e. [angry (the 2nd input) and angry (the 3rd input)]. The Markov chain is used to produce the possible emotional context based on the training data for each sub-theme for the Mayid character.

The following are examples of the generated emotional profiles for the sub-theme “Mayid starts bullying” for the Mayid character:

1. T A A N A A [‘threatening, angry, angry, neutral, angry and angry’]
2. A A A [‘angry, angry, and angry’]
3. N A N A N A [‘neutral, angry, neutral, angry, neutral, and angry’]
4. A A N [‘angry, angry and neutral’]
5. D A A A A N A [‘disapproval, angry, angry, angry, angry, angry, neutral, and angry’]

The dynamic algorithm is used to find the smallest edit distance between the test emotional context [angry and angry] and the training and generated emotional context for the Mayid character for each sub-theme. In the above example, the second and fourth emotional sequences have the smallest edit distance (distance = 1) to the test emotional context and the former suggests ‘angry’ as the affect conveyed in the current input (“go on den”) while the latter implies ‘neutral’ expressed in the current input. Thus we need to resort to the emotional context of other characters to justify the recommended affects. From the chatting log, we find that Lisa was ‘angry’ in her most recent input (the 1st input) while Elise was ‘threatening’ in her most recent input (the 5th input). Since the bully, Mayid, has a negative relationships with Lisa (being ‘angry’) and Elise (being ‘threatening’), the imperative input (“go on den”) may indicate ‘angry’ rather than ‘neutral’. Therefore the affect detection component adjusts the affect from ‘neutral’ to ‘angry’ for the 6th input.

In this way, by considering the linguistic contextual indicators, the potential emotional context one character was in, relationships with others and recent emotional profiles of other characters (especially the leading opponent character), our affect detection component has been able to inference emotion based on context to mark up the rest of the above test example transcript (e.g. Mayid being ‘angry’ for the 11th input). However our processing could be fooled easily by various

diverse ways for affective expressions and creative improvisation (emotional patterns not shown in the training and generated sets). We intend to adopt more linguistic contextual hints, psychological (context-based) emotional theories and better simulation tools (e.g. hidden Markov modelling) for further improvements. Our processing currently only focused on school bullying scenario. We are on our way to extend the context-based affect sensing to other scenarios to further evaluate its efficiency.

5 EVALUATION

We carried out user testing with 220 secondary school students from Birmingham schools and Education Village in Darlington for the improvisation of school bullying and Crohn’s disease scenarios. Generally, our previous statistical results based on the collected questionnaires indicate that the involvement of the AI character has not made any statistically significant difference to users’ engagement and enjoyment with the emphasis of users’ notice of the AI character’s contribution throughout. Briefly, the methodology of the testing is that we had each testing subject had an experience of both scenarios, one including the AI minor character only and the other including the human-controlled minor character only. Such arrangement could not only enable us to measure any statistically significant difference to users’ engagement and enjoyment due to the involvement of the AI minor character, but also provide us the opportunity to compare the performance of the AI minor character with that of the human-controlled one. After the testing sessions, we obtained users’ feedback via questionnaires and group debriefings. Improvisational transcripts were automatically recorded during the testing so that it allows further evaluation of the performance of affect detection component.

Therefore, we produce a new set of results for the evaluation of the updated affect detection component with metaphorical and context-based affect detection based on the analysis of some recorded transcripts of the school bullying scenario. Generally two human judges (not engaged in any development stage) marked up the affect of 150 turn-taking user input from the recorded another 4 transcripts from school bullying scenario (different from the transcripts used for the training of Markov chain). In order to verify the efficiency of the new developments, we provide Cohen’s Kappa inter-agreements for EMMA’s performance with and without the new developments for the detection of the most commonly used 12 affective states. In the school scenario, EMMA played a minor bit-part character (Lisa’s friend: Dave). The agreement for human judge A/B is 0.45. The inter-agreements between human judge A/B and EMMA with and without the new developments are presented in Table 1.

	Human Judge A	Human Judge B
EMMA (previous version)	0.38	0.30
EMMA (new version)	0.40	0.32

Table 1. Inter-agreements between human judges and EMMA with and without the new developments

Although further work is needed, the new developments on metaphorical and contextual affect sensing have improved

Proceedings of the Linguistic And Cognitive Approaches To Dialog Agents Symposium, Rafal Rzepka (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK
EMMA's performance of affect detection in the test transcripts comparing with the previous version.

The evaluation results indicated that most of the improvements (approximately 80%) are obtained for negative affect detection based on the inference of context information. But there are still some cases: when the two human judges both believed that user inputs carried negative affective states (such as angry, threatening, disapproval etc), EMMA regarded them as neutral. One most obvious reason is that some of the previous pipeline processing (such as dealing with mis-spelling, acronyms etc, and syntactic processing from Rasp etc) failed to recover the standard user input or recognize the complex structure of the input which led to less interesting and less emotional context for some of the characters and may affect the performance of contextual affect sensing. We also aim to extend the evaluation of the context-based affect detection using transcripts from other scenarios. Moreover, some of the improvements (nearly 20%) in the updated affect sensing component are made by the metaphorical processing. However, since the test transcripts contained a very small number of metaphorical language phenomena comparatively, we intend to use other resources (e.g. Wallstreet Journal and other metaphorical databases (e.g. ATT-Meta)) to further evaluate the new development on metaphorical affect sensing.

6 CONCLUSIONS & FUTURE WORK

We have made new developments on automatic affect sensing in metaphorical figurative language and employed context profiles for affect interpretation. Although EMMA could be challenged by the rich diverse variations of the metaphorical language phenomena we focused on and other improvisational complex context situations, we believe these areas are very crucial for development of effective intelligent user interfaces and our processing has made promising initial steps towards these areas. Also, further testing is needed to further evaluate the efficiency of metaphorical and context-based affect detection. We also intend to compare our system's affect sensing performance with that of a well-known affect sensing tool, ConceptNet [5], as another way for future evaluation.

REFERENCES

- [1] L. Zhang, J. A. Barnden, R. J. Hendley, M. G. Lee, A. M. Wallington and Z. Wen. Affect Detection and Metaphor in E-drama. *Int. J. Continuing Engineering Education and Life-Long Learning*, Vol. 18, No. 2, 234-252. (2008).
- [2] L. Fainsilber and A. Ortony. Metaphorical uses of language in the expression of emotions. *Metaphor and Symbolic Activity*, 2(4), 239-250. (1987).
- [3] ATT-Meta Project Databank: Examples of Usage of Metaphors of Mind. <http://www.cs.bham.ac.uk/~jab/ATT-Meta/Databank/>. (2008).
- [4] Z. Kövecses. Are There Any Emotion-Specific Metaphors? In *Speaking of Emotions: Conceptualization and Expression*. Athanasiadou, A. and Tabakowska, E. (eds.), Berlin and New York: Mouton de Gruyter, 127-151. (1998).
- [5] H. Liu & P. Singh. ConceptNet: A practical commonsense reasoning toolkit. *BT Technology Journal*, Volume 22, Kluwer Academic Publishers. (2004).
- [6] M. A. M. Shaikh, H. Prendinger & I. Mitsuru. Assessing sentiment of text by semantic dependency and contextual valence analysis. In *Proceeding of AII 2007*, 191-202. (2007).
- [7] M. Mateas. Ph.D. Thesis. Interactive Drama, Art and Artificial Intelligence. School of Computer Science, Carnegie Mellon University. (2002).
- [8] X. Zhe & A. C. Boucouvalas. Text-to-Emotion Engine for Real Time Internet Communication. In *Proceedings of International Symposium on Communication Systems, Networks and DSPs*, Staffordshire University, UK, 164-168. (2002).
- [9] A. C. Boucouvalas. Real Time Text-to-Emotion Engine for Expressive Internet Communications. In *Being There: Concepts, Effects and Measurement of User Presence in Synthetic Environments*. G. Riva, F. Davide and W. IJsselsteijn (eds.), 305-318. (2002).
- [10] R. Craggs & M. Wood. A Two Dimensional Annotation Scheme for Emotion in Dialogue. In *Proceedings of AAAI Spring Symposium: Exploring Attitude and Affect in Text*. (2004).
- [11] A. Egges, S. Kshirsagar & N. Magnenat-Thalmann. A Model for Personality and Emotion Simulation, In *Proceedings of Knowledge-Based Intelligent Information & Engineering Systems (KES2003)*, Lecture Notes in AI. Springer-Verlag: Berlin, 453-461. (2003).
- [12] C. Elliott, J. Rickel & J. Lester. Integrating Affective Computing into Animated Tutoring Agents. In *Proceedings of IJCAI'97 Workshop on Intelligent Interface Agents*, 113-121. (1997).
- [13] R. Aylett, S. Louchart, J. Dias, A. Paiva, M. Vala, S. Woods and L. E. Hall. Unscripted Narrative for Affectively Driven Characters. *IEEE Computer Graphics and Applications* 26(3). 42-52. (2006).
- [14] M. Cavazza, C. Smith, D. Charlton, L. Zhang, M. Turunen and J. Hakulinen. A 'Companion' ECA with Planning and Activity Modelling. In *Proceedings of the 7th International Conference on Autonomous Agents and Multi-Agent Systems*. Portugal, 1281-1284. (2008).
- [15] P. Rayson. Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. Ph.D. thesis, Lancaster University. (2003).
- [16] C. Fellbaum. WordNet, an Electronic Lexical Database. The MIT press. (1998).
- [17] A. Esuli and F. Sebastiani. Determining Term Subjectivity and Term Orientation for Opinion Mining. In *Proceedings of EACL-06*, Trento, IT. 193-200. (2006).

Chain of Events: Multi-stage Approach to Humor and Emotions in HCI

Pawel Dybala¹, Michal Ptaszynski¹, Rafal Rzepka¹ and Kenji Araki¹

Abstract. Humor and emotions are strongly related to each other. Both these topics are recently gaining more and more attention from HCI researchers. Some of existing research projects employ humor generating engines in conversational systems, in order to improve their interaction with humans. Evaluation of such joking systems is often related to users' emotions, i.e. their feelings towards the systems during and after the interaction. However, in human-human interaction, humor is usually a reaction to particular emotional states, and can be even used intentionally, to change interlocutor's emotions (e.g. sadness into joy). Thus, also in HCI it is important to broaden our view and take into consideration not only emotions that follow the humorous act, but also those that precede it. In this paper we propose to distinguish between what we call a "two-stage" approach, in which emotions are seen only as a reaction to humor, and a "multi-stage" approach, in which we consider emotions and humor as an interchanging chain of events, occurring one after another. The multi-stage approach is illustrated with an actual example - our joking conversational system, which 1) analyses users' emotions during the interaction, 2) uses humor as a reaction to users' particular emotional states, and 3) analyses their reactions to it. We briefly present the results of its evaluation experiments, and finally propose an idea of user-adapting humor sense model, in which the humor-emotion chain can be extended even further.

1 INTRODUCTION

In recent years we have seen numerous research projects in the fields of AI (Artificial Intelligence) or HCI (Human-Computer Interaction), focusing on such subjective and problematic (in terms of computing) domains as humor or emotions. The world of science started to realize that in order to create human-like artificial companions for humans, we need to equip them with such abilities as expressing and experiencing emotions and producing and understanding humorous acts.

These two issues are obviously strongly related to each other. Humor can elicit emotions, change them, reduce or increase their intensity. By observing humans' emotive reactions to humor, we can evaluate its quality or timing, as well as assess peoples' sense of humor and their character in general. Thus, it seems natural that these correlations should be used also in HCI, be it in humor generation, its evaluation or any other activity involving these two issues.

Some of existing research projects in HCI do utilize the fact that emotions and humor are so closely related. Most of them focus on emotions' role in evaluation process, i.e. on checking humans' emotive reactions to humor. In other words, it is what happens

AFTER the humorous stimuli that is a subject of interest in most of known research. However, as mentioned above (and presented below in more details – see 2), humor can also be a reaction to something (particular feelings, like sadness or depression), and thus it is also important to take into consideration emotions that occur BEFORE the humorous acts, i.e. to which humor is reaction to. To our knowledge, no existing research in HCI so far has taken this issue into consideration, at least not in an explicit manner.

In this paper we propose to define the relation between humor and emotion as a chain of events, related to each other and occurring consecutively during interaction (between humans as well as in HCI, although in our research project we focus on the latter). We make a distinction between what we call a two-stage approach (like the one described above – humor followed by emotion) and multi-stage approach (in which more links of the chain, like emotions that precede humorous acts, are taken into consideration). We analyse existing research in the field of psychology (see 2.1), summarize current achievements in the field of humor and emotion-related projects in HCI (see 2.2), and describe our work on humor-equipped conversational system for Japanese, in which jokes are told according to users' emotional states. In the system, the multi-stage approach to humor and emotions has actually been implemented and tested experimentally (see 3). Finally, we give an outline of humor sense individualization model we are currently working on (see 4), which will allow us to add one more link to our chain of events.

2 TWO- AND MULTI-STAGE APPROACH TO HUMOR AND EMOTIONS RELATION

As mentioned above, in this paper we perceive the relation between humor and emotions as a chain of events, occurring one after another. Thus, we propose to distinguish between two types of approaches to humor-emotions relation:

- 1) **two-stage approach** – in which only one pair of links is taken into consideration, regardless to their order, i.e. whether it is humor-emotion or emotion-humor (however, to our knowledge, most if not all of existing research, in which this two-stage approach is employed, use the former).
- 2) **multi-stage approach** – in which more links of the chain are taken into consideration (e.g. humor-emotion-humor, emotion-humor-emotion-humor etc.). Most of existing research in which this approach is employed, are limited to three links.

¹ Graduate School of Information Science and Technology, Hokkaido University, Kita-ku Kita 14 Nishi 9, 060-0814 Sapporo, Japan. Email: {paweldybala, ptaszynski, kabura, araki}@media.eng.hokudai.ac.jp

In below sections, we summarize existing research on humor and emotions conducted in two fields: psychology and HCI, discussing which of them represent the two-, and which multi-stage approaches.

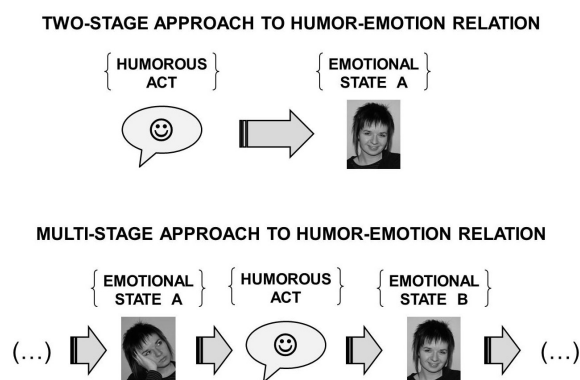


Figure 1. Two- and multi-stage approach to humor-emotions relation

2.1 Psychology

Humor in relation to emotions (and vice versa) has been a subject of psychological research for a long time. Potentially beneficial role of humor as a reaction to stress was discussed by Freud [1], while May [2] stated that humor can help distance oneself to various problems. These (and more) works, however, were of purely theoretical nature, while many others investigated the humor-emotions relation in an experimental manner. Some of them, according to what we proposed above, represent the two-stage approach, whilst other (most) go further and take into consideration more than two (in most cases three) links of the chain, by this representing the multi-stage approach. Most of the research investigate the role of humor as a countermeasure to negative emotions, sometimes using such methods as artificial induction of negative moods in experiment participants, prior to exposing them to humorous stimuli.

Ruch, for instance [3], conducted a study in which he investigated humans' reactions to humorous contents, such as clown performance or funny videos. Not very surprisingly, most participants reacted with smile and laughter, which, states Ruch, are expressions of the emotion of mirth. This research represents the two-stage approach to humor and emotions, as only emotional reaction after humor was investigated here.

In their research, Danzer et al. [4] also used induction in order to elicit dysphoric moods in experiment participants. Those of them who were afterwards exposed to humor (comedy videotape), showed a significant decrease in dysphoria, which dropped back to the initial level. According to our division of approach to humor and emotions, this study represents the multi-stage group, as it investigates three links of the chain: preceding emotions (dysphoria) – humorous stimulus (funny video) – following emotional states (decrease in dysphoria).

Comedy tapes, which seem to be quite a good humorous stimulant, were also used by Moran [5], who discovered that humor can cause significant reduction in anxiety level of experiment participants, and Szabo et al. [6], who showed that it can lead to increase in participants' positive moods and decrease

in emotional distress. These works also represent the multi-level approach.

To induce negative emotions, some researchers use also videos, while other refer to other methods, such as asking experiment participants to solve frustrating or even unsolvable tasks. Trice [7] showed that stress caused by such induction was mitigated by watching funny videos. Yovetich et al. [8], in turn, falsely informed experiment participants that in 12 minutes they will experience painful electric shock. Stress caused by this was smaller for this group of participants that was exposed to humor while waiting for the shock.

Needless to say, psychological research on humor and emotions are not limited to those briefly listed above. However, examples presented in this section show that there is a general tendency in this field rather to use multi-stage than two-stage approach.

2.2 HCI

As in the field of psychology, most of existing research employ the multi-stage approach to humor and emotions, one would expect similar situation also in HCI. Thus, the more surprising is the fact that, to our knowledge, no research hitherto uses this approach, and existing projects tend to take into consideration only two links of the humor-emotions chain.

In the field of computational humor, the most popular is its linguistic genre, including such types of jokes as puns, riddles or word plays. The source of ambiguity in such jokes are such features of language as homophony or polysemy, and this makes them more possible to compute using methods and tools of NLP (Natural Language Processing). Some of existing systems work in a stand-alone manner, i.e. they generate jokes in isolated forms, without integrating them into wider contexts. Other try to place humorous act during interaction between humans and computers. Evaluations of the latter type of applications often involves (or is limited to) investigating users' opinions and feelings and emotions toward the interaction. Thus, in this section, we focus on these systems, which use humor during interacting with humans.

Binsted's JAPE system [9] was a punning riddles generator for English. It would create riddle-type word plays, such as:

-What do you call a murderer with fibre?
-A cereal killer.

The system itself was a stand-alone application, i.e. it generated jokes in isolated forms, such as the one above. However, it was later used by Ritchie et al. [10] to create STANDUP - system described as "a language playground, with which a child can explore sounds and meanings by making up jokes, with computer assistance [10]." Target users of STANDUP were children with CCN (complex communication needs), which can result in lower levels of literacy. In evaluation experiments, they used numerous methods, one of which required users to make comments about the system. Some children stated that they enjoyed the interaction and that they liked the system. In other words, their emotive reactions to humor were investigated, which means that this research represents the two-stage approach to humor and emotions.

Not in all works on humor in HCI jokes are generated by computers. In their research, Morkes et al. [11] investigated the impact of preprogrammed jokes on both CMC (computer mediated communication) and HCI. In order to measure users' reactions to humor, they asked them about their feelings and

impressions after the interactions, like likeability and friendliness of the partner. They also checked participants' mirth responses, i.e. number of times they smiled or laughed during the interaction. Results showed that in conversations which included humor, in comparison to those which did not, the users evaluated their computer partners as more sociable, friendly and likeable. This approach is also of the two-stage type.

In our previous research [12] we also investigated the role of humor in human-computer conversation. We constructed a pun generator for Japanese, which we then implemented into a non-task oriented conversational system (also in Japanese). The system, named Pundalin, would tell jokes in every third turn of the conversation. In order to evaluate its performance, we asked 13 users to perform conversations with it and with a similar system without humor, and then analysed the chat logs using Ptaszynski et al.'s ML-Ask Emotiveness Analysis System [13]. By doing this, we checked how many and what emotions did the users have during the interactions. The results showed, that the system with humor elicited much more emotions in users, most of which were positive [12].

2.3 The Gap

All research in the field of HCI presented in 2.2, including our previous work, represent the two-stage approach to humor and emotions. Thus, we can say that there is a gap between the state of the art in psychology, where most of works in this subject are of the multi-stage type, and current progress in HCI, where all projects so far represent the two-stage approach.

To us, the multi-stage approach seems far more natural, as humor not only causes emotions, but is also a reaction to them. As a matter of fact, it also works in the same manner in the HCI research presented above, as obviously humor in interactions investigated there is a reaction to something. However, this issue has not been investigated explicitly, and thus in HCI emotions preceding humorous acts are still to be taken into consideration.

3 MULTI-STAGE APPROACH TO HUMOR-EMOTIONS RELATION IN HCI – SYSTEM OUTLINE

In this section we briefly describe our research, which represents the multi-stage approach to humor and emotions. As mentioned above, in our previous work [12] we presented a humor-equipped conversational system for Japanese, which tells jokes to make humans feel better. In that research we investigated users' emotions towards the system after the interactions, which means that it represented the two-stage approach.

As a continuation of that research, we constructed a new system, in which we investigate not only users' emotions after the interaction, but also their emotive states during the interaction. The system detects users' emotional states in every turn of conversation, and on this basis decides if a joke should be told to make them feel better.

Below we present elements constituting our system and an outline of its algorithm.

3.1 Components

Our new system, named Multi-PUNDA, constitutes of three components: baseline conversational system, pun generator and emotiveness analysis system.

Baseline conversational system – developed by Takahashi et al. [14], non-task-oriented conversational system for Japanese uses an Internet-based association algorithm to generate response relevant to user's input (sentence). Using query engines such as Google, it generates lists of words commonly associated with input, and then integrates them into sentences. Below we present an example of the agent in action:

User: “*Reizouko no naka ni tsumetai nomimono ga arimasu*” (There's something cold to drink in the fridge)

Agent: “*Tsumetai juusu nara sakki nomimashita yo*” (If you mean cold juice, I have just drunk it!).

For more details, see [14].

Pun generator – engine that generates Japanese linguistic jokes (puns). From users' input (a sentence) it extracts a base word, which is then transformed into list of pun candidates, using phonetic transformation patterns. Next, the candidates are queried in the Internet, and one with the highest hit rate is selected and integrated into a sentence. Below we present an example of the agent in action:

User: *Kaeru daikirai!* (I hate frogs!)

Agent: *Kaeru to ieba tsukaeru yo ne!* (Speaking of frogs, we can use that!)

For more details, see [12].

Emotiveness Analysis System – developed by Ptaszynski et al. [13], ML-Ask system detects emotions from text. It performs three types of analysis: 1) it checks if a sentence is emotive (if it contains emotive elements from database, like interjections, exclamations, or vulgar language), 2) determines the types of emotions, and 3) determines their valence (positive/negative) and activation degree (activated/deactivated). If in the second step, no emotive expressions are found in the input, ML-Ask uses Shi et al.'s [15] web mining technique to extract emotive associations for the sentence. Below we present an example of the agent in action:

Sentence: *Kyou wa atatakai desu ne.* (It's warm today, isn't it?)

Emotive elements: *-ne* (-isn't it)

Emotive expressions: none (⇒use web mining procedure)

Emotions found on the Web: joy; valence: positive

For more details, see [13] and [15].

3.2 The System

The components described above were combined to create a system that tells jokes according to users' emotional states. Basing on existing research (summarized in 2.1), showing that humor can elicit positive emotions and reduce their negativity, we assumed that the system can tell jokes when users feel bad, in order to make them feel better. Also, if their emotional state is neutral, a joke can be told to induce positive feelings.

Thus, the system analyses users' emotions during the conversation – this function is performed by the ML-Ask system – and on this basis it makes a decision if a joke should be told. If

users' emotions are positive or neutral, response is generated by the pun generator, otherwise – by the baseline chatterbot. The system's algorithm is presented in Figure 2.

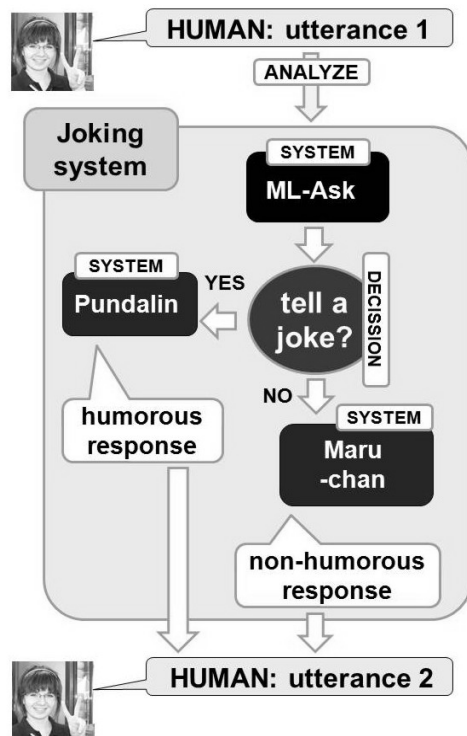


Figure 2. Joking system – algorithm outline

3.3 Emotive Evaluation

Performance of the system described in 3.2 was evaluated in numerous experiments, among them user-oriented and automatic. In the former, we asked 13 users to conduct conversations with two systems (humor-equipped and a similar one without humor), and then asked them to complete a questionnaire, which (among others) included questions about their self-reported emotions towards each system. The results were then compared with our database of emotive expressions in order to check their value – every positive emotion was counted as plus 1, and every negative as minus 1. In the latter, we analysed the chat logs using the ML-Ask system in order to detect emotions in users' utterances – we checked general emotiveness of utterances as well as valence changes during the conversations.

These results are summarized in Table 1.

	baseline			multi-Punda		
emotiveness	91 (average: 7.0 per utterance)			125 (average: 9.6 per utterance)		
valence changes	to +	to -		to +	to -	
	68%	32%		94%	6%	
self-reported emotions	+	-	sum	+	-	sum
	5	14	-9	15	7	+ 8

Table 1. Emotive evaluation-results

3.4 Summary

An overall message coming from these results is that the system with humor elicited generally more emotions in the users, and most of them were positive. However, what is more important for this paper, is that the approach described above is multi-stage, as we take into consideration users emotions both before and after humorous acts. In other words, we treat the humor-emotions relation as a chain of events, which is being monitored during the interaction between humans and our system. To our knowledge, this is the first research in the field of HCI representing such an approach.

4 MULTI-STAGE APPROACH – MORE THAN THREE LINKS

The system presented above represents the multi-stage approach to humor and emotions. That said, the approach is actually three-stage (emotions-humor-emotions), and as such, in our opinion it does not exhaust the subject, as the relation between humor and emotions should not be limited to only three stages. Immediate emotive reactions to humor have also implications in the long run – they can induce further emotion changes and influence the way of perceiving further humorous acts. Thus, it is necessary to broaden our view and take into consideration more links of this chain of events.

In this section we summarize our idea of a user-adapting joking system. To the previous system (presented in 3), we are planning to add a procedure, allowing to build users' humor sense models, which will lead to personalization of the system. As the models are built basing on users' emotions displayed after humorous acts, we can say that in this part of our research we take into consideration more than three stages of humor-emotions relation.

4.1 User-Adapting Joking System

Humor is a very individual treat. We all perceive it differently, and what is funny to one person, may be boring or insulting for someone else. These individual aspects of sense of humor should also be taken into consideration in HCI. Joke telling systems should be able to adapt themselves to each user's personal preferences and his/her type of sense of humor.

To answer that need, we proposed an algorithm of a user-adapting joking system [16]. To the system described in 3, we added an emotiveness analysis based procedure, in which users' reactions to humorous acts are used to build their individual models of sense of humor. Currently we are working on implementation of the algorithm.

As showed on the Figure 3, in the humor sense individualization procedure, users' responses to humor are analysed by ML-Ask in order to check their emotive reactions. Results of the analysis are then used to build a model of users' preferences.

If, for instance, during a conversation with User A, the system detects that he is angry, it tries to make him feel better by telling him a joke about blondes. The user reacts to it with a utterance, in which the system detects irritation. On this basis, the system can state that: 1) User A may not like jokes about blondes, and 2) User A may not like being told jokes when he is angry. These informations are stored in a database, and on this basis User A's humor sense model can be build. The more data the system

acquires about User A's reactions to humor, the more accurate model can it build and the better it can adapt to this particular users' preferences.

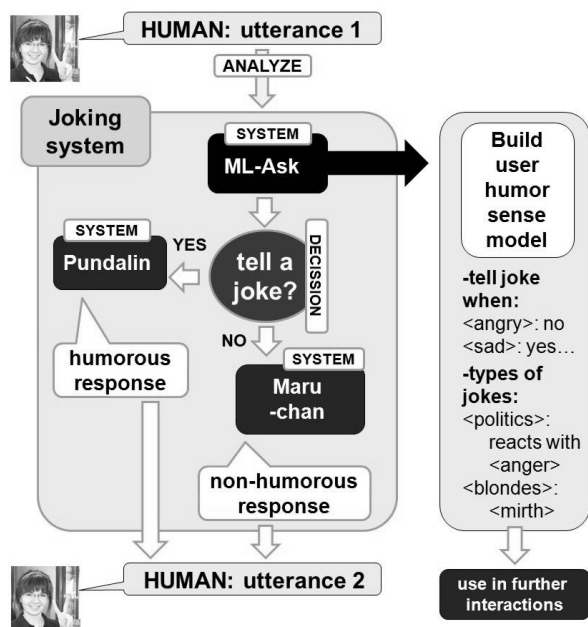


Figure 3. Joking system adapting to users' humor sense – algorithm outline

4.2 More Links of the Chain

After adding the procedure described in 4.1 to the system described in 3, we can see that the approach to humor-emotions relation was broadened. In the new system, equipped with the humor individualization algorithm, we take into consideration not only emotions that occur immediately after the particular humorous act, but also forthcoming humorous acts and users' reactions to them. Thus, the number of links of the humor-emotions chain can be continuously extended, and the more links (or sets of links) we add, the better the system will be, i.e. the better it can adapt to each user's personal needs.

5 CONCLUSION

In this paper we proposed to distinguish between two types of humor-emotions relation: two- and multi-stage. We summarized existing research in the fields of psychology and HCI, stating that in the latter no work so far is representing the multi-stage approach. Next, we presented our works, which represent the multi-stage approach to humor-emotions relation. This brings HCI closer to psychology, in which such an approach is far more common.

Needless to say, there is still much to be done in the subject of humor-emotions relation, and some issues need to be addressed and discussed. For instance, in all works described in this paper, links of the humor-emotions chain occur interchangeably, i.e. humor is always preceded and preceded by emotion(s), and so on. This, however, does not have to be true – emotions, rather

than isolated events, are in many cases by themselves small chains of events. Anger, for example, can be followed by sadness or resignation, or other emotions, and these changes also need to be taken into consideration. Thus, in the multi-stage approach, the relation between humor and emotions does not have to look like this:

(...) emotion – humor – emotion – humor – emotion (...),

but could, for example, look like this:

(...) emotion – humor - emotion – emotion – emotion (...).

This gives us some new, interesting possibilities of analysis, which we are planning to address in the nearest future in our research.

5 ACKNOWLEDGEMENT

The research presented in this paper was partially sponsored by the Global COE Hokkaido University, and a research grant from the Nissan Research Foundation.

REFERENCES

- [1] S. Freud. Humor. *International Journal of Psychoanalysis*, 9, 1-6 (1928)
- [2] R. May. Man's search for himself. New York: Random House (1953)
- [3] W. Ruch. State and trait cheerfulness and the induction of exhilaration: A FACS study. *European Psychologist*, 2(4), 328-341 (1997)
- [4] A. Danzer, J.A. Dale, and H.L. Klions. Effect of exposure to humorous stimuli on induced depression. *Psychological Reports*, 66(3, Pt 1), 1027-1036 (1990)
- [5] C.C. Moran. Short-term mood change, perceived funniness, and the effect of humor stimuli. *Behavioral Medicine*, 22(1), 32-38 (1996)
- [6] A. Szabo, S.E. Ainsworth, and P.K. Danks. Experimental comparison of the psychological benefits of aerobic exercise, humor, and music. *Humor: International Journal of Humor Research*, 18(3), 235-246 (2005)
- [7] A.D. Trice. Alleviation of helpless responding by a humorous experience. *Psychological Reports*, 57(2), 474 (1985)
- [8] N.A. Yovetich, J.A. Dale, M.A. Hudak. Benefits of humor in reduction of threat-induced anxiety. *Psychological Reports*, 66(1), 51-58 (1990)
- [9] K. Binsted. Machine humour: An implemented model of puns, Ph.D. Dissertation, University of Edinburgh, UK (1996)
- [10] G. Ritchie, R. Manurung, H. Pain, A. Waller, R. Black, and D. O'Mara. A practical application of computational humour. *Proceedings of the 4th International Joint Conference on Computational Creativity*, 91-98 (2007)
- [11] J. Morke, H.K. Kernal, and C. Nass. Effects of humor in task-oriented human-computer interaction and computer-mediated communication: A direct test of srcr theory. *Human-Computer Interaction*, 14(4), 395-435 (1999)
- [12] P. Dybala, M. Ptaszynski, R. Rzepka, and K. Araki. Activating Humans with Humor - A Dialogue System that Users Want to Interact With. *IEICE Transactions on Information and Systems Journal, Special Issue on Natural Language Processing and its Applications*, Vol.E92-D, No.12, 2394-2401 (2009)
- [13] M. Ptaszynski, P. Dybala, Shi Wenhan, R. Rzepka, and K. Araki. A System for Affect Analysis of Utterances in Japanese Supported with Web Mining. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Special Issue on Kansei Retrieval*, Vol. 21, No. 2, 30-49 (194-213) (2009)
- [14] M. Takahashi. *Web ni yoru kyoukikinhito to n-gram moderu wo mochiita hatsuwabunnseiseishuhou* (Utterance Generation Method Using Web Search Results and Word n-grams), Bachelor dissertation, Hokkaido University, Sapporo, Japan (2009)

- [15] W. Shi, R. Rzepka, and K. Araki. User Textual Input Using Causal Associations from the Internet, In Proceedings of FIT2008, 267-26 (2008)
- [16] P. Dybala, M. Ptaszynski, R. Rzepka, and K. Araki. Computers with personalized sense of humor. In Proceedings of The 3rd International Symposium on Global COE Program of Center for Next-Generation Information Technology Based on Knowledge Discovery and Knowledge Federation (GCOE-NGIT 2010). Sapporo, Japan, 205-206 (2010)

Evaluation of a Humor Generation System by Real World Application with ¥500,000 to Win

Jonas Sjöbergh and Kenji Araki¹

Abstract. We describe the evaluation of a system for automatic humor generation in Japanese. First we used the traditional “having jokes rated by evaluators” method, with a result of 3.3 on a scale from 1 (boring) to 5 (funny). To complement this evaluation and to see if 3.3 is “good enough” we entered a system generated performance in a competition with a ¥500,000 prize. We also generated responses to arbitrary requests from the audience at a live event. While we did not win the 500,000 yen we did reach the final, and both the performance and the real time generation were well received. Sentiment analysis of blogs covering the event also showed that our system compared well to the other teams. That the system could compete successfully against human made contributions indicates that the score of 3.3 is “good enough” for real world applications of the system.

1 Introduction

Humans often use humor but computers are so far bad at this. Humor makes daily interaction smoother and more enjoyable, helps people deal with troubling events, and generally makes people feel better. We believe it would be good if machines could use humor, though this is a difficult task. Work on computational humor has been done on both humor recognition [6, 3] and humor generation [1, 7], see [2] for a good overview of the field.

Humor is subjective, and many other factors also influence whether something is perceived as funny or not. This makes evaluating and comparing computational humor systems difficult. Most evaluations consist of asking humans to rate how funny they think generated jokes are, which is a reasonable evaluation method but has some problems. It is difficult to compare systems evaluated at different times, under different circumstances, or by different evaluators, since so many factors other than the quality of the joke also influence perceived funniness. It is also difficult to judge if an achieved result is good enough to be useful based on such evaluations, since there are no standard baselines or gold standards to relate to.

We have a humor generation system that has been evaluated in the traditional way of asking evaluators to rate how funny they thought the system output was [4, 5]. The latest version achieved a score of 3.3 on a scale from 1 to 5. As an earlier version scored 2.8 in a different evaluation, it seems likely that the system has improved, but how good is 3.3?

Reported results from other systems are often low, but it is hard to compare with results from evaluations under different conditions. It is also somewhat hard to construct a relevant baseline or upper bound. Also, as far as we know, there are no other systems that generate full performances connecting different jokes etc., though there are several systems generating stand alone jokes. Thus, there is no

similar research to relate to either. It is of course possible to hire professional comedians to create performances similar to what the system generates and then compare the system to human level performance. A quick subjective check of the system output shows that it is quite far from human level performance, though, so spending money on this is currently not that appealing. There are no freely available sources of human made humor that are similar enough to the system output to be useful as comparisons.

Simple baselines for a lower evaluation limit can be constructed by instantiating the patterns used by the system with random words instead of words chosen by the system. Such text is really boring for the evaluators to wade through though, and since we are not allowed to pay our evaluators it is hard enough to get reasonable numbers of participants already.

We did however want to complement this evaluation and preferably relate this score of 3.3 to real world performance. We found an opportunity for this in a competition for silly robots. The competition rules were that the system must have some mechanical parts, have no serious purpose, and must make people laugh. The funniest system, based on a two minute performance and if qualifying for the final also a 15 minute interview like performance chance, wins ¥500,000. We entered the competition with a performance generated using our system as a tool. Would a system scoring 3.3 be good enough to compete against human made contributions? The other contestants would thus be our “gold standard” to compare against, and one evaluation metric would be if our system could do reasonably well in comparison to the other teams or not.

We also tested the system capabilities by generating humorous responses to arbitrary requests from audience members at the live final. Would this score of 3.3 mean the system was robust and general enough to handle unrestricted input in real time and still entertain people?

Since the conditions of the contest were not under our control, a lot of the evaluation measurements would likely be hard to quantify or be rather subjective. What would be a reasonable method of measuring if the hundreds of spectators were entertained or not? We decided that since this was our first attempt at complementary evaluation methods, we would use very simple measure such as “a large part of the audience laughing is good, booing or uncomfortable silence after jokes is not good”. We also complemented this with analysis of blog entries written by spectators discussing what they thought of the event, as a more easily measurable evaluation metric.

2 System

We have built a humor generation system for Japanese made up of modules that generate different types of jokes. It can thus easily be

¹ Hokkaido University, Japan, email: {js, araki}@media.eng.hokudai.ac.jp

extended with new modules or adapted to new tasks by changing how the modules are used. Below we describe the modules and three variations of the system using these differently. More detailed descriptions can be found in [5]. Most modules are fairly simple though modules using sophisticated reasoning etc. to generate jokes could be added when methods that work reasonably well become available.

2.1 Database

This module represents the system reusing jokes it has heard other people tell. We automatically collected a database of word play jokes in Japanese using a few seed jokes, web queries, and pattern matching. A simple example of how this module works is searching the web for occurrences of a joke the system already knows. If two seed jokes (or jokes found earlier by the module) occur in the same context (e.g. an HTML list) the module downloads all other content that occurs in an identical left and right context (e.g. all the other list items). The contexts are determined automatically by finding the longest common left and right contexts of the two known jokes in the page, similar to the method used in [8]. The database has almost 3,000 jokes, mostly word play jokes.

When a joke related to some word is needed, a joke containing this word that has not already been used is returned.

2.2 Proverbs

The proverb module has a database of sayings in Japanese. To generate a joke a proverb is twisted into a new variant by replacing content words with similar sounding dirty words. These are taken from a collection of a few hundred dirty words that have also been grouped into the categories sex related, feces related, and insults. If two or more dirty words from the same category sound similar enough to content words of the same word class, a joke is generated by replacing the original words with these.

The jokes are normally presented in patterns like “Recently my life has felt like *proverb*” “Oh? For me it has been more like *dirty proverb*”. Instead of proverbs, other inputs can also be used, such as in the real time system detailed below.

2.3 Riddles

When a word play riddle is needed for a word, the module checks if it has any similar sounding dirty words. A riddle is formed along the lines of: “A *word* is a *word*, but what kind of *word* is *hint*?”, “What?”, “Answer: *similar sounding dirty word*”. Hints are generated by searching a large corpus for phrases like “a *dirty word* is *hint*”. Hints found are then assumed to be reasonable descriptions if they co-occur more strongly with the answer than with the original word.

The riddle module can also be used with other patterns. When using the system as a tool, it gave jokes like “Give me an example sentence for ‘*content word*’”, replying with an example sentence for a similar sounding dirty word.

2.4 Responses

Japanese stand-up comedy is usually performed in pairs, and the interaction between the straight man and the funny man is important. Generally the straight man must chastise the funny man when he says something stupid or misunderstands something. We have a module that generates responses like these, mainly using a database

of generic responses. This module was only used in the stand-alone system.

2.5 Basic Stand-Alone System

The basic version of the system is completely automatic and generates stand-up comedy performances in Japanese, to be performed by two performers. This system uses the different joke modules to generate jokes and ties them together into a coherent performance by simple transitions like “Speaking of *word from one joke, another joke related to this word*”. It uses text-to-speech to generate sound files and uploads these into two small robots that perform the generated stand-up comedy. This is the version of the system that was evaluated in a previous experiment [5] in the traditional way of having evaluators rate the funniness of the output.

The generated performances normally contain one proverb joke, one riddle joke, and two database jokes. The responses from the straight man in the performance can also lead to more jokes based on these being included in the performance.

2.6 System as a Tool

Since the completely automatic system sometimes makes mistakes, such as generating jokes that are incomprehensible or sentences with broken grammar etc., we decided to use the system as a humor generation tool when competing against human made performances in the silly robot competition. Humor is a very difficult problem, so having a system that can be used as a joke generation tool is still a good first step, we believe.

All the jokes in the performance were generated by the system but we manually selected which jokes to include in the performance, removing jokes that were too hard to understand (using rare or difficult words) and correcting minor mistakes. We also manually reordered the jokes and wrote the transitions between them. Finally, since the text-to-speech output is hard to understand even under good conditions, we expected it to not be enough for a live performance in front of hundreds of audience members. Instead we had two students read the generated output into a microphone and then made it sound “robotic” by using a phaser filter. In the end, we were also given the use of a big monitor, which we used to show subtitles of what the robots were saying.

Database jokes were not used in the competition entry version of the system. We wanted to see if the automatically generated jokes were funny enough to compete against man made contributions and thus did not want to include any man made jokes. This led to pretty much all the jokes in the performance being dirty jokes.

2.7 Real Time System

We also constructed a real time version of the system to be used at the live event if we made it to the final. It takes an input statement and calls the different modules to see if any of them can produce a relevant joke. Since the database is the fastest option, first the system extracts content words from the input and sees if there is any joke in the database containing one of these content words that has not already been used. If so, it outputs one randomly selected such joke. If this fails, the system tries the proverb module, treating the whole input statement as if it was a proverb and then proceeding as described earlier. If this also fails, the riddle module is called with each content word from the input until a riddle joke can be generated. If this too fails, the system outputs statements along the lines of “La la la, I cannot hear you, lalala”.

3 Evaluations

3.1 Traditional Evaluation

As mentioned, we previously evaluated the stand-alone system using the traditional method of having evaluators watch the generated performances and rate them from 1 (boring) to 5 (really funny). 33 evaluators took part, and the total impression of the system was 3.3 on average with very low variation between performances but as can be expected very high variation between different evaluators. We evaluated the different components too, but here we focus only on the impression of the whole system. For more details see the original paper [5].

It is difficult to judge if 3.3 is good enough for any specific use or if it is very funny at all. Another measure is that 45% of the evaluators gave the performances a 4 or a 5 which reasonably indicates that they liked them. Since computational humor systems generally fare quite badly (humor is difficult and the system outputs are thus generally not that funny), we believed that this was a quite good result, but how good? To find out more we tried two other tests for the system.

3.2 Silly Robot Competition

Bacarobo 2008² was the second installment of the Bacarobo silly robot competition. The rules are that each system has to be mechanic in some way, must have no serious purpose, and must make people laugh. The funniest system wins ¥500,000. Bacarobo is held in Japan, and contestants can enter performances in either Japanese or English, though so far only systems for Japanese have made it to the finals. In 2007, 11 systems made it to the live performance final, but in 2008 only 5 teams were to be selected for the final. In the end, they received too many good contributions and put 6 teams in the final.

The competition is organized by *Maywa Denki*³, which is an art performance company. The finals are held in cooperation with *Yoshimoto Kogyo*, a large Japanese entertainment company that employs most of Japan's famous stand-up comedians. The final is held in a *Yoshimoto* theater, and in 2008 it was in the *Yoshimoto ∞Hall* in Shibuya, Tokyo. It seats slightly over 200 guests, who had to pay an entrance fee of 3,000 yen and the event was more or less sold out. The audience was a fairly even mix of (mainly Japanese) men and women. People were of all ages, even a few children which was not so good for our very foul mouthed robot performance.

The first stage of the competition was sending a video including a two minute competition performance and an additional three minutes of time allotted for explaining what your robot is doing. At the live final, each team was given a short presentation and then entered the stage. After performing the previously submitted two minute performance live, a rather long (about fifteen minutes) interview like explanation session followed, the contents of which varied a lot between teams. The jury consisted of four minor celebrities: the company boss of *Maywa Denki*, a manga writer/paper columnist, a movie director, and a professor from Tokyo University. The live event was led by a famous comedian.

3.2.1 Other Teams

We here give a brief description of the five other teams that made it to the final. They are presented in the performance order in the final.

Magaru-ken & Tono-metto consisted of a samurai haircut robot and a sword robot that were used as lie detectors and nervousness

detectors. For instance, the sword is held in the hand and when the holder starts sweating it bends to indicate stress. The performance consisted of a man wearing the haircut robot on his head and holding the sword in his hand being asked questions by a woman in a white lab coat. One question was "Have you ever bought a used school girl swim suit in an Internet auction?", to which he answered "No" but the robots indicated nervousness and possible dishonesty. The other questions were in similar vein. During the interview part of the competition they used the robots on the jury members, asking for instance if they had rented their own movies to raise the rental statistics.

YKRN was a face recognition robot that had the face of a Japanese female celebrity stuck on a movable arm that swung around looking for human faces. When it found a face the eyes lit up and it stopped moving around. For faces of people it recognized it had various programmed reactions, such as trying to kiss any jury member that came close and to completely ignore the comedian presenting the final. This was the winning team.

The third performance was our system, described above.

Kangaery robo was a replica of a thinking man sculpture that had a computer screen instead of a face and one more computer screen showing what the robot itself was seeing. It also did face recognition by recognizing a big green board with a cut out for your face that volunteers from the audience were asked to put to their faces. When it recognized a face it switched the stick figure face it normally had for the recognized face, and changed the face in the image of what the robot was seeing to be the stick figure face instead of the audience member's original face. The name of the system came from the fact that the robot was said to know what people it switched face with were thinking. The robot was asked what the audience member was thinking, and answered with various embarrassing things, typed into the system by a team member back stage.

Push-kun & Yome-push were two trash cans with legs that walked around and spoke to each other and their maker. They also did a dance number that finished with the female robot giving birth to a small baby robot. One of the robots was in the final in the previous Bacarobo contest too.

Tsuneo & his friends consisted of four high school girls and three robots. One robot was a vacuum cleaner stuffed in a shell that looked like a pig that "ate" (by vacuuming it up and throwing it out the other end) glitter fed to it by one of the girls. The second robot was another vacuum cleaner stuck inside a green elephant shaped shell that blew into three party blowout-noisemakers when touched. The last robot was a small sleeping man that snored very loudly and whose stomach grew when snoring. In the end, it exploded.

A lot of performance time was used by the accompanying people performing and interacting with the robots, while the robots themselves were fairly passive.

3.2.2 Results

Our contribution, made by using our system as a tool and containing only computer generated jokes (no jokes from the database) made it to the final but did not win. When the robots performed the generated jokes the jury and audience members laughed at the "correct" places and most people seemed to enjoy the performance though this of course varied from person to person. One jury member laughed incredibly much, for instance.

The main result of this part of the evaluation is the fact that we made it to the final, though we did not win the contest. The individual placements in the final were not given, only which team was the

² <http://bacarobo.com/>

³ <http://www.maywadenki.com/>

overall winner, so a more precise measure than between place 2 and 5 was not available. We believe just reaching the final against man made contributions is a fairly strong result. The other finalists were all very good/funny.

3.3 Live Audience Interaction

After the pre-generated two minute performance at the live event, we also showed the real time capabilities of the system. The system output was text only, and the text was shown on several big monitors around the theater. First the head of the jury requested a joke related to the sponsors of the event. This was set up beforehand. This was so as to set the mood, show how this part of the performance was to transpire, and to guarantee at least one example that the system could handle successfully (the audience did after all have to pay to see the event).

It turned out that we misspelled the sponsor name and input *Oronami shi* (instead of *Oronamin C*). The system part-of-speech and word boundary analyzer (Japanese does not mark word boundaries) found the word *nami* (wave) when misanalysing this. It output a database joke: *Nami to ieba, Konishi Manami no yoko ni, Shimanami Kaidou* (“Speaking of waves, right beside *Konishi Manami* I saw *Shimanami Kaidou*” – the last part of “beside” plus the name of a road sounds like the name of an actress). This got a good reaction.

Next, a small girl in the first row was asked for her favorite word. She said *Manafi*, which is the name of an animated character in a children’s show. No module could generate a response so the failure mode kicked in and *kikoenai, kikoenai, kikoenai* (“Lalala, I cannot hear you.”) was output. This also got a lot of laughs and reactions along the lines of “Very human like reaction!”. This indicates that the system fails somewhat gracefully, and that the recovery options in the system are useful when the system cannot produce sophisticated output.

Prompted for her favorite food the girl said *ringo* (“apples”). The system output: *Ringo ga jimen ni ochita tte iu ton* (“People say the apple fell to the ground.” – the last part sounds like “Newton” in Japanese); another database joke, giving an OK reaction.

Next a jury member asked for a joke regarding “*Touhoushinki*”. We did not know it at the time but it is the name of a Korean boy band. Not knowing the proper Chinese characters to use we input it using phonetics (*hiragana*). The system misanalysed this unknown word as several short words and returned a database joke related to *hou* (in the boy band name it roughly means “direction”, but in the joke found it means roughly “ho” as an interjection in for example “Ho ho ho”). *Hou to ieba: “Houkama ha aitsu kamo!” “Houka, houka”* (Speaking of “ho”: “That guy might be the pyromaniac!” “Hoho, is that so”). This got a good reaction.

Finally we were asked for a Christmas joke and got: *Kurisumasu ha kuri de sumasu* (“I make do with a chestnut for Christmas”); a proper Christmas joke with good reactions.

The results of this part of the evaluation is first that the system was robust enough to output jokes reasonably relevant to the input for all inputs except one. As can be expected, the database joke module was heavily used, mainly because it is called first and is applicable to a wide range of inputs. Large parts of the audience laughed at each joke, indicating that the system not only can output something, but that the output is also funny. Of course, parts of this could be because of things like the system misunderstanding the inputs being funny.

Not only did people laugh a lot at the generated system output, they also seemed very impressed (an “ohhh” of feeling impressed went through the theater) when during the explanations that followed

we mentioned that the system searches the Internet for jokes automatically. That such things can be done seemed to be more impressive than the first two minute performance.

3.4 Sentiment Analysis of Blog Reports

After the contest was over, we downloaded blog postings covering the Bacarobo 2008 event and did sentiment analysis of these. We did the sentiment analysis by hand, since there was not that much text to deal with and since automatic methods are somewhat lacking in accuracy. We ignored postings mentioning only that someone went to the event, since we wanted data on the impression of the different performances. We also ignored press release like postings just mentioning the contest and the name of the winning team. Only postings that actually described the contents of at least one performance from the contest were examined.

For each performance that reached the final we checked four different things: First, how many postings mention the performance in question at all, and second how much text was used for describing the performance (measured in number of characters). The basic idea is that if people care about a performance they will be more likely to mention it and also to write more about it. If a performance is not mentioned, it did not make enough of an impression for the writer to care. Generally, for these two points a higher score is better, though of course having only negative statements written about your system would give high scores without indicating a good performance.

The third and fourth things examined were the number of positive statements (like “the crowd laughed a lot”, “I was impressed”) and the number of negative statements written about the performance. Most postings spent most of the text on fairly factual descriptions of the contents of the performances, with some subjective impressions added at the end or sometimes interspersed throughout, so there were not that many positive and very few negative statements.

What should be considered a positive statement is also somewhat up for debate, since quite a few comments called the performances “stupid”. Since the goal of the competition is to be stupid, it may not be negative in this context, though in many cases it was not very positive either. We did not count statements referring to performances as “silly/stupid” (*baka*) as positive or negative, since most of them probably just played on the name of the competition (*Bakarobo*, “silly robots”). Negative sounding comments like “the system looked scary” are not necessarily negative either. We counted statements as negative only when the writer was clearly criticising the performance, such as by saying “I did not understand what the point was” or “the movements were very awkward”. The results of the sentiment analysis are shown in Table 1. The winning team of Bacarobo 2008 was YKRN.

Table 1. Results of sentiment analysis of blogs. For each team, the number of blogs mentioning the performance, the amount of text (number of Japanese characters) discussing it, and the number of positive and negative statements about the team were checked.

Team	Mentioned	Text Size	Pos.	Neg.
Magaru-ken & Tono-metto	9	8,681	15	1
YKRN	8	9,365	15	2
Our System	9	9,723	25	5
Kangaeru robo	6	8,189	11	0
Push-kun & Yome-push	6	7,209	14	5
Tsuneo & his friends	7	6,698	6	2

Our system is one of the three performances mentioned by the most number of bloggers. It is also the system which has the largest amount of text dedicated to it, and it has the most positive statements. The reason for the high number of positive statements is just one blogger, who seemed to be very impressed with our system, though. Not counting this blogger our system still has the most positive statements, though the difference to the next best performance, the winning team YKRN, is then only 1.

Our system also has the most negative comments, tied with one other performance. The negative comments about our system and performance were almost exclusively that the jokes were too dirty, especially since there were children watching. There was also one negative comment about the t-shirt worn when starting the performance and answering questions from the jury, which while part of the performance is not really a property of the system itself. There were also quite a few comments about some sound problems we had during the beginning of our performance, though we did not count these as negative statements about our system since the fault was in the sound system of the theater and not in our system.

The positive statements were mainly that people were impressed with how advanced the system was (or they thought it was), in relation to the real time interaction with the audience and jury. Other positive statements were that the jokes were very dirty (two people liked that), that the robots were very cute (one comment), that it was nice to see a foreign entry (said by a blog in English), and that one person appreciated that we were prudent or commonsensical enough to also prepare the big monitors for subtitles.

3.5 Discussion

Most results that can be taken from this evaluation are of course quite informal. We do however think they give an interesting complementary view and tell us things not possible to learn from more traditional and strict evaluation methods.

All in all, both the audience and the jury members seemed to enjoy both the pre-generated naughty joke performance, and perhaps especially the real time generation of jokes. Quite likely the latter was considered impressive because it was completely clear that the computer was doing all the work, while the pre-generated performance could just as well have been human made (like the other competition entries), though we did explain that all jokes were computer generated.

Other feedback apart from the fact that we reached the final, and the laughter and general happy impression of the jury and audience members there, include: the other participants (most of them university researchers) being impressed by our system and photographing it back stage (despite it looking like an ordinary laptop); audience members commenting to us after the show that they thought it was funny and that the robots were cute; and the blogs covering the event, analyzed in the previous section.

Reaching the final of a competition and getting good reactions from an audience are of course quite informal ways of evaluating the system. We believe it shows that the system is “funny enough” to be useful though, since the goal of the system is to entertain people. This is much harder to tell from the results of a more traditional though strict and scientific evaluation. The sentiment analysis of blogs covering the event showed that our system was actually one of the systems that received the most mentions (and the most text dedicated to it) and the most positive comments in blog postings about the event. The sentiment analysis results are fairly even among most teams, so this is not a very strong result, but at least it shows that our system is

not markedly worse than any of the human generated performances. Some systems received markedly less coverage than our system.

Our system did however get the most negative comments too (tied with one other system). All of the relevant comments were about the system generating jokes that are too dirty, or at least inappropriate when performing in front of children. Making a more child-friendly version is thus one of our current projects.

Apart from the results already discussed, it was also very interesting for us just to try to apply the system in these settings that were very different from the normal research settings. There were many problems to overcome that would not have been found if we had not entered this competition. Examples include hardware problems when adapting to the requirements of having several hundreds of audience members that had to hear the robot performance at the same time, real time requirements in the audience interaction, trying to adapt to the very varied composition of the audience (children, foreigners with less knowledge of Japanese etc.), and much more. A result that we were quite pleased with ourselves was the simple fact that we managed to pull off the live performance at all.

In conclusion, while there is a lack of baselines or other comparisons to relate the previous system evaluation result of 3.3 to, these informal evaluations showed that people are entertained enough to still be satisfied after having paid a quite high entrance fee (seeing a professional comedian would have been cheaper). It also shows that the system is robust enough to handle unrestricted input. It is not handled perfectly by any means, but well enough to keep people entertained, the main goal of the system. The system thus seems like a quite useful application of computational humor.

4 Conclusions

We evaluated a humor generation system, and a traditional evaluation gave a score of 3.3 on a scale from 1 to 5. Since this score is somewhat hard to relate to real world performance we set out to test if this 3.3 meant that the system was good enough to entertain people by having the system compete against human made contributions in a “silly/funny robot” competition with a ¥500,000 prize. Combining the system with some manual filtering we managed to reach the final. The audience response was good, and a session of fully automatic real time generation of replies to unrestricted requests from the audience also went well. Sentiment analysis of blogs covering the event showed that our system compared very well to the other teams, though we did get the most negative statements too. These were about the system generating jokes that are too dirty.

Though we did not win the 500,000 yen, the experience still showed that the score of 3.3 seems to indicate a performance good enough for real world application of the system.

We believe using complementary evaluations methods like how far you can go in a competition, sentiment analysis of things written about your system, or informal measurements like “does the audience seem to be laughing a lot at the correct times of a live performance” can give interesting results not easily found using traditional evaluation methods. Just applying the system to real world conditions also showed us many things, e.g. the system is fast enough for real time applications and robust enough to handle input from a very varied audience. We also noticed and fixed real world application related problems like scaling the audio output to be heard by hundreds of people at once. Going out into the real world taught us many new things about our system.

Acknowledgements

We would like to thank Shinsuke Higuchi and Hajime Wakahara for doing the voices for the robots. Most of the system was developed in a project funded by the Japanese Society for the Promotion of Science (JSPS).

REFERENCES

- [1] Kim Binsted, *Machine Humour: An Implemented Model of Puns*, Ph.D. dissertation, University of Edinburgh, Edinburgh, United Kingdom, 1996.
- [2] Kim Binsted, Benjamin Bergen, Seana Coulson, Anton Nijholt, Oliviero Stock, Carlo Strapparava, Graeme Ritchie, Ruli Manurung, Helen Pain, Annalu Waller, and Dave O'Mara, 'Computational humor', *IEEE Intelligent Systems*, **21**(2), 59–69, (2006).
- [3] Rada Mihalcea and Carlo Strapparava, 'Making computers laugh: Investigations in automatic humor recognition', in *Proceedings of HLT/EMNLP*, Vancouver, Canada, (2005).
- [4] Jonas Sjöbergh and Kenji Araki, 'Automatically creating word-play jokes in japanese', in *Proceedings of NL-178*, pp. 91–95, Nagoya, Japan, (2007).
- [5] Jonas Sjöbergh and Kenji Araki, 'A complete and modestly funny system for generating and performing Japanese stand-up comedy', in *Coling 2008: Companion volume: Posters and Demonstrations*, pp. 109–112, Manchester, UK, (2008).
- [6] Julia Taylor and Lawrence Mazlack, 'Toward computational recognition of humorous intent', in *Proceedings of Cognitive Science Conference 2005 (CogSci 2005)*, pp. 2166–2171, Stresa, Italy, (2005).
- [7] Hans Wim Tinholt and Anton Nijholt, 'Computational humour: Utilizing cross-reference ambiguity for conversational jokes', in *CLIP 2007*, Camogli, Italy, (2007).
- [8] Richard Wang and William Cohen, 'Language-independent set expansion of named entities using the web', in *Proceedings of ICDM'07*, pp. 342–350, Omaha, U.S., (2007).

Political Question Generation Method Using Minutes of Municipal Councils

Yasutomo Kimura¹, Hideyuki Shibuki², Keiichi Takamaru³
Tetsuro Kobayashi⁴ and Tatsunori Mori²

Abstract. This paper proposes a generation method of political questions which include Japanese noun phrases of the form N_1 no N_2 . We focus on generating a yes-no question whether a user is interested in an utterance of a councilor. An example of political questions shows as follows: *nisan katanso*(N_1) no *haisyutu*(N_2) *nituite kuwasiku siritai desu ka?* (Do you want to know emissions of carbon dioxide?). In this paper, " N_1 " and " N_2 " mean a noun or a compound noun. The form N_1 no N_2 which include the Japanese postpositional "no" has a much broader usage. These expressions of the form N_1 no N_2 cannot often use a question expression for asking the interesting such as "*simin no minasan*(All of citizen)". Therefore, our method checks on an expression whether a user can understand the meaning of that as a question expression. Moreover, our method expands the expression of N_1 when a question expression of N_1 no N_2 is ambiguous. We obtained good results in the experimental result.

1 Introduction

In recent years, it is possible to find a target document by accessing the enormous resources available on the www through information retrieval techniques. However, A user is difficult to find technical documents targeting, for example, political information and judicial precedents[5]. One of the reasons is that a user tends to make an ambiguous query for information extraction. In the case of political information, the utterances of many councilors include political terms such as "*yosan*(budget)", "*zeikin*(tax)" and so on. Using these political terms, a user cannot get retrieve target documents because there are too much target documents which include these political terms such as "*yosan no sakugen*(budgetary cutback)", "*yosan no gendogaku*(budgeted allowance)", "*yosan no saisyutu*(expenditure of the budget)" and so on. Moreover, it is difficult for a user to think about an appropriate phrase query that would fit these existing in documents. To resolve these problems, we consider that an interactive question answering system is a better approach. In the case of interactive question answering, a system asks a user yes-no questions using a phrase that includes language from target documents. When the system creates a yes-no question using phrases obtained from target documents, it is possible to create a concrete representation by expanding the candidate phrase. However, these phrases tend to become too many words because the system fails to recognize an

optimal phrase. In consideration of these problems, we propose a method generatizing a yes-no question, which is both concrete and clear. In our previous work, we proposed an interactive question answering method using a decision tree. This system was created for the political information supporting service between local councilor and inhabitants through the WWW. In this paper, we present a yes-no question generation method and its experimental results.

2 Creating a question to clarify user's interest

In political information processing, some research focuses on the relation between a user's opinion and a political party plan[8]. Uekami et al. made the Vote Match system where each question is manually created[7]. However, in case of local politics in Japan, the number of municipalities is about 1,760, and their cities, towns and villages have their specific problems. Therefore the manual approach is difficult to realize. In addition to that, each municipality has many councilors.

Therefore, we have made political categories for local politics in order to extract characteristics of each local councilor[6]. We investigated suitable expressions for a user to extract the description of minutes corresponding to a political category. As a result, we confirmed that the question expression of " N_1 no N_2 " was more efficient than that using an usual compound noun.

2.1 Question expression including " N_1 no N_2 "

We collected " N_1 no N_2 " expressions from the council minutes of Otaru City, Japan in 2007. In result 14,336 " N_1 no N_2 " expressions were extracted. Table 1 shows an example of " N_1 no N_2 " extraction. However, there are expressions not suitable for the question. Therefore, we consider a method that filters out incorrect expressions.

Table 1. Examples of " N_1 no N_2 "

" N_1 <u>no</u> N_2 "	Freq of Minutes
<i>rijisya no touben</i> (The statement of the director)	97
<i>simin no minasan</i> (All of citizen)	63
<i>teian no riyuu</i> (The explanation of the suggestion reason)	25
<i>situmon no gaiyou</i> (The summary of the question)	24
<i>isi no kakuho</i> (The securement of the doctor)	13

¹ Department of Information and Management Science, Otaru University of Commerce, Japan, email: kimura@res.otaru-uc.ac.jp

² Graduate School of Environment and Information Science, Yokohama National University

³ Department of City Life Studies, Utsunomiya Kyowa University

⁴ National Institute of Informatics

Table 2. Examples of filtering using Japanese Google N-gram

" N_1 no N_2 " in Minutes	Frequency	" N_1 wo N_2 suru" in Japanese Google N-gram	Frequency
<i>teian</i> <u>no</u> <i>riyuu</i> (explanation of the proposed reason)	25	<i>teian</i> <u>wo</u> <i>setsumei</i> <u>suru</u> (to explain the proposed reason)	43
<i>isi</i> <u>no</u> <i>kakuho</i> (securement of doctor)	13	<i>isi</i> <u>wo</u> <i>kakuho</i> <u>suru</u> (to secure doctor)	31
<i>jinkenhi</i> <u>no</u> <i>yokusei</i> (constrain of personal expense)	13	<i>jinkenhi</i> <u>wo</u> <i>yokusei</i> <u>suru</u> (to constrain personal expense)	70
<i>keikaku</i> <u>no</u> <i>sakutei</i> (establishment of plan)	12	<i>keikaku</i> <u>wo</u> <i>sakutei</i> <u>suru</u> (to establish plan)	51

2.2 Related work on " N_1 no N_2 "

There are many studies of the Japanese noun phrase " N_1 no N_2 ". Kurohashi et al. have conducted semantic analysis of Japanese noun phrases[4]. Japanese noun phrases in the form of N_1 no N_2 have a broad usage. For example :

- *watasi* no *kuruma* (my car) possession
- *yakyuu* no *senshu* (player of baseball) category
- *toranpu* no *tejina* (trick of card) instrument

In these usages of "no", we focus on relationship between N_1 and N_2 being AN "object" and A "verbal noun".

Kataoka et al. proposed that if a verb is an abbreviated predicate, a verbal noun modification N_1 ga/wo/. . .⁵ V-suru⁶ N_2 can be paraphrased into N_1 no N_2 [1]. In this paper, we filter N_1 no N_2 in reference to Kataoka's method.

2.3 Filtering using Japanese Google N-gram

As we mentioned in Section 2, using " N_1 no N_2 " it is often difficult to extract a question expression from the city council minutes. In the section 2.1, we mentioned that by using " N_1 no N_2 " it is often difficult to extract a question expression from the city council minutes. Therefore, we filter expressions that are able to be paraphrased from " N_1 no N_2 " to " N_1 wo N_2 suru". For checking of question expressions, we use Japanese Google N-gram[2]. Japanese Google N-gram includes 7-gram phrases which exist more than 20 times in the WWW. We extracted 805 expressions from 14,336 expressions included in the 2007 minutes of Otaru City council. Table 2 shows the result of filtering by Google N-gram.

3 Investigation of optimal expression

In this section, we find the optimal expressions of expanded candidates in the 2007 minutes of Otaru City council. Some of the 805 candidates that are checked by Google N-gram have ambiguous meaning such as "keikaku no sakutei" (development of plan). Therefore, we increased the number of words without being restricted by N_1 to find the optimal expression, which is both concrete and clear.

⁵ Japanese case particle

⁶ Verb

3.1 Making experiment data

In our research, we set 96 political categories in order to classify a city council minutes, and we annotated the 2007 minutes of Otaru City council. In this section, we use 1,667 utterances of the minutes, and their utterances are annotated in only one (category). In order to expand the " N_1 no N_2 ", we use the CaboCha as a syntactic analysis tool[3]. We make expanded candidate expressions, which lengthen the expression step by step toward the front of the utterance. Table 3 shows an example of expanded candidate expressions. In table 3, the expression of "*hiyo no sisan*" (A test calculation of expenses) is possible to paraphrase as follows:

"*hiyo wo sisan suru*" (I calculate expense as a test). ! !

We expand the candidate expressions from right to left until a punctuation mark or the beginning of the sentence is reached. We made 20 questionnaires such as in table 3. 20 questionnaires include 90 expression candidates. In other words, a subject evaluates the optimal expression among about 4 or 5 expression candidates.

3.2 Evaluation point of view

We evaluate candidates by using three points of view which consists of "ambiguous $\iff$ clearness", "redundant $\iff$ concise" and "unreadable $\iff$ readable". The subject performs five phases of evaluation from three points of view. Here, we explain how to decide the optimal expressions of expanded candidates in the 2007 minutes of Otaru City council. First, we calculate an arithmetic mean of the evaluations by 6 subjects. As shown as the equations (1), an arithmetic mean is calculated using three points of view which consists of "ambiguous $\iff$ clearness", "redundant $\iff$ concise" and "unreadable $\iff$ readable". Here, N denotes the number of subjects, and c is a candidate expression.

$$AM_1(c) = \frac{\sum_{i=1}^N \text{ClearnessScore}(i)}{N}$$

$$AM_2(c) = \frac{\sum_{i=1}^N \text{ConciseScore}(i)}{N}$$

$$AM_3(c) = \frac{\sum_{i=1}^N \text{ReadableScore}(i)}{N} \quad (1)$$

$$HM(c) = \frac{3}{\sum_{i=1}^3 \frac{1}{AM_i(c)}} \quad (2)$$

Second, we calculate a harmonic mean from the results of three points of view, as shown the equation (2). Finally, we determine the optimal expression by selecting the highest harmonic mean.

Table 3. An example of the questionnaire

The example of utterance of councilor	
... -omitted- ...	
, taisyou kuiki wo hirogeta baai no puru ni kakaru hiyou no sisan (A test calculation of the expense to affect a pool when I enlarged the object area.)	
... -omitted- ...	
The candidates of the question expression	
<----- Expanding from right to left.	
Q1	hiyou no sisan
Q2	kakaru hiyou no sisan
Q3	puru ni kakaru hiyou no sisan
Q4	baai no puru ni kakaru hiyou no sisan
Q5	hirogeta baai no puru ni kakaru hiyou no sisan
Q6	taisyou kuiki wo hirogeta baai no puru ni kakaru hiyou no sisan
(A test calculation of the expense to affect a pool when I enlarged the object area.)	
Evaluation item	
In this example, a subject evaluates the following questionnaire from Q1 to Q6.	
No.1	Is this question comprehensible or incomprehensible? ☐ It is a comprehensible question. ☐ It is an incomprehensible question. #If you select comprehensible question, answer the following question.
<----->	
No.2	ambiguous 1 2 3 4 5 clearness
No.3	redundant 1 2 3 4 5 concise
No.4	unreadable 1 2 3 4 5 readable

3.3 Results of expanded expression of "N₁ no N₂"

We performed an expression evaluation with six subjects, two male and four female university students. Table 4, shows the expressions of the highest harmonic mean. In this example, we calculated an ungrammatical expression as "0".

Table 5. Experimental result of selection

Method	Accuracy		Ratio of HM	
Baseline	7/20	0.35	61.84 / 74.63	0.8286
Method1	15/20	0.75	66.62 / 74.63	0.8927
Method2	16/20	<u>0.80</u>	70.24 / 74.63	<u>0.9412</u>

3.4 Discussion

In table 4, the expressions marked with "#" were not expanded, and these expressions confirmed 7 patterns. In this evaluation, there are 90 candidates, which include the 20 unexpanded expressions. Of the 90 candidates, more than 2 people evaluated 71 candidates as an ungrammatical expression. The characteristics of the 71 candidates are as follows: 1. The expanded block including the specific particle is not suitable for expression. For example, the specific particles include "ga" and "ha" in Japanese. 2. The expanded block including adverbs is not suitable for expression.

4 Proposed methods of generating question expressions

We propose two methods based on the section 3.4. Method 1 is limited by the specific particles. Method 1 selects the longest expression that does not include following Japanese particles: "de", "ga", "mo", "ha", "kara", "niokeru" and "niyori". Method 2 is limited by adverbs. Method 2 selects the longest expression that does not include any adverb.

5 Experiment

We perform an experiment that evaluates to select the optimal expression of candidates by our proposed method. We perform the evaluation from two points of view. The point uses the harmonic mean shown in table 4. We calculate the ratio between the sum of the harmonic mean of the selected expressions and the sum of the 20 highest results of harmonic mean. The second point calculates the accuracy of the correct expressions that define the highest harmonic mean. We compare our methods with a baseline method that does not expand "N₁ no N₂".

Table 5 shows the experiment which confirm the effectiveness of Method 2.

6 Conclusion

We proposed two methods for selecting the most appropriate expression from expanded candidates for question generation, and confirmed their effectiveness. As future work, we are going to implement the system using our proposed method.

REFERENCES

- [1] Akira Kataoka, Shigeru Masuyama, and Kazuhide Yamamoto, 'Summarization by Shortening a Japanese Noun Modifier into Expression A no B', 409-414, (1999).
- [2] Taku Kudo and Hideto Kazawa, 'Web Japanese N-gram Version 1', (2007).
- [3] Taku Kudo and Yuji Matsumoto, 'Japanese Dependency Analysis using Cascaded Chunking', 63-69, (2002).
- [4] Sadao Kurohashi and Yasuyuki Sakai, 'Semantic Analysis of Japanese Noun Phrases : A New Approach to Dictionary-Based Understanding', 481-488, (1999).
- [5] Rafal Rzepka, Hideyuki Shibuki, Yasutomo Kimura, Keiichi Takamaru, Masafumi Matsuhara, and Koji Murakami, 'Toward Automatic Support For Japanese Lay Judge System - Processing Precedent Factors For Sentencing Trends Discovery', 33-41, (2008).
- [6] Keiichi Takamaru, Hideyuki Shibuki, Yasutomo Kimura, Dai Hasegawa, Hokuto Ootake, and Kenji Araki, 'Extraction of Political Activity of Assemblyman from Minutes of Municipal Assemblies Using the Political Category', B11, (2009).
- [7] Takayoshi Uekami and Tetsuya Sato, 'Estimating the policy positions of political actors: an application of computerized coding to the Japanese policy documents(in Japanese)', 61-73, (2009).
- [8] VOTEMATCH, 'http://www.votematch.co.uk/europe/'.

Table 4. The result of question expression of the highest harmonic mean.

	Question expression of the highest harmonic mean	Harmonic mean
1	<i>puuru ni kakaru hiyou no sisan</i> (A test calculation of the expense to be spent for a pool.)	3.60
2	<i>sijyo wo meguru kankyou no henka</i> (An environmental change concerning the market.)	3.94
3	<i>sa-bisu no koujyo#</i> (Improvement of a service)#	3.41
4	<i>sidou no tekiseina kinou no hoji</i> (Maintenance of a reasonable function of a municipal road.)	3.80
5	<i>bunka geijyutsu ni fureru kikai no kakujyuu to jinzai no ikusei</i> (Expansion of the opportunity to mention culture art and upbringing of the talented person.)	3.57
6	<i>kosodate oyako no koryuu nado wo sokusin suru jigyou no kakujyuu</i> (Expansion of a business to promote an interchange of the child cares.)	4.00
7	<i>75 sai ijyou no koureisya heno kakokuna hokenryou no futan</i> (A burden of a severe burden to a senior citizen older than 75 years old.)	4.17
8	<i>simin karano youbou noaru basu no sinsetsu</i> (A new establishment of a bus route with the demand from a citizen.)	4.30
9	<i>kigyuu no ikusei#</i> (Upbringing of a company)#	3.62
10	<i>gakkou no annzenn ni kakawaru kankyou no henka</i> An environmental change about security of a school.	4.17
11	<i>miyagehin no kounyuu#</i> (Purchase of a souvenir)#	3.38
12	<i>jinzai no touyou#</i> (Promotion of a talented person)#	4.41
13	<i>byougentai ga tainai ni sinnyuu site kansenn site zousyokusi hassei suru kannjya no sousyuu</i> The general term of the disease a pathogen invades the body and is contagious and multiplies, and to develop.	2.24
14	<i>tettai no sidou#</i> (Guidance of a removal)#	3.08
15	<i>kaigoyobou puran no sakusei#</i> (Making of a care prevention plan)#	4.01
16	<i>bu no saihen#</i> (Reorganization of a department)#	3.57
17	<i>hokenshou ga tsukaenai youna sikakusyomeisyo no hakkou</i> (Issue of qualification in case when a health insurance card is not usable.)	3.60
18	<i>akaji no kaisyuu to zaisei no kennzennka</i> (Deficit cancellation and fiscal fitness.)	4.30
19	<i>kigyousai ganri syoukankin no gensyuu ya iji kanri no kouritsuka</i> (Efficiency of redemption cost on revenue bond.)	3.74
20	<i>tiiki no kyouikuryoku wo ikasita yutakana kyouikukatsudou no suisin</i> (The promotion of the rich instructional activity that makes use of local education power in.)	3.72
Sum of harmonic mean. (This result is used to evaluate the experiment.) The expression attached of ”#” is not expanded.		74.63