Proceedings of the International Symposium on Matching and Meaning, Michael Chan and Fiona McNeill (Eds.),
at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

The Role of Taxonomy Properties in Information Content Metrics

Raúl Ernesto Menéndez-Mora [12] and **Ryutaro Ichise** [1]

Abstract. Assessing the semantic similarity between words is a generic problem in many research fields such as artificial intelligence, biomedicine, linguistics, cognitive science and psychology. The difficulty of this task lies in how to find an effective way to simulate the process of human judgment of word similarity by combining and processing a number of information sources. This paper presents an ongoing research where some properties of the WordNet taxonomy are used in the semantic similarity computation process. A new metric for calculating the Information Content (IC) of word senses is proposed as an alternative to the corpus based model. In an experimental evaluation the performance of the new metric is tested through their use in some IC WordNet based semantic similarity measures on a standard 28 word pair dataset.

1 Introduction

Essential for human categorization and reasoning, semantic similarity of words has become a topic of many research fields such as artificial intelligence, biomedicine, linguistics, cognitive science, and psychology. In general, semantic similarity is extensively used in a variety of applications, like words sense disambiguation [17], detection and correction of malapropisms [1], information retrieval [5], automatic hypertext linking [3] and natural language processing. Several applications to the field of artificial intelligence are discussed in [19]. However, despite numerous practical applications today, its theoretical foundations lie elsewhere, in cognitive science and psychology where it was the subject of many investigations and theories.

Semantic similarity measures have been recently applied and developed in biomedical ontologies, e.g. the Gene Ontology. They are mainly used to compare genes and proteins based on the similarity of their functions rather than on their sequence similarity [14]. Let take another example of peer-to-peer networks [4] into which semantic similarity also has found its way. Assuming a shared taxonomy among the peers to which they can annotate their content, similarities among peers can be inferred by computing similarities among their representative concepts in the shared taxonomy. This way, the more two peers are similar, the more efficient it is to route messages toward them. Numerous similar applications is the reason for the increasing interest in this subject, whose ultimate goal is to mimic human judgment with respect to the similarity of word pairs.

Semantic similarity of words is often represented by the similarity between the concepts associated with the words. Several methods have been developed to compute word similarity. Mostly operating on the taxonomic dictionary WordNet and exploiting its hierarchical structure. Nodes based methods which generally use as its information source the Information Content (IC) of the concept, have a weakness. The conventional way of measuring the IC of word senses is to combine knowledge of their hierarchical structure from an ontology with statistics on their actual usage in text as derived from a large corpus, doing IC based methods corpus dependent.

The increasing need for better measures has led us to this study in the hope of upgrading existing semantic similarities measures. In particular, we exploits some foundations of the Intrinsic IC metric [20] approach for developing a novel metric for computing the IC. The metric, completely derived from WordNet without the need for external resources from which statistical data is gathered, is applied to recently developed semantic similarity measures [11].

The paper is structured as follows. The next section reviews some background knowledge and related work. Section 3 presents our approach as well as the novel IC metric. Section 4 discusses the results of the experiment, and section 5 summarizes our work, draws some conclusions, and outlines future tasks.

2 WordNet Similarity Measures

2.1 WordNet

A number of semantic similarity computation methods operate on the taxonomic dictionary WordNet exploiting its hierarchical structure. WordNet [2] is a machine-readable lexical database that is organized by meanings, and it was developed at Princeton University. Synonymy[1], hyponymy[2], meronymy[3] and other arbitrarily typed semantic relationships between concepts are represented in this lexical network of English words.

WordNet, as an ontology, is intended to model the human lexicon, and psycholinguistic findings were taken into account during its design. It is classified as a light-weight ontology, because it is heavily grounded on its taxonomic structure that employs the IS-A inheritance relation; and as a lexical ontology, because it contains both linguistic and ontological information [10]. Figure 1 shows a fragment of WordNet's structure where solid lines represent is-a links and dashed lines indicate that some intervening nodes were omitted to save space. Nouns, verbs, adjectives, and adverbs are each organized into networks of synonym sets (*synsets*) that each represent one underlying lexical concept and are interlinked with a variety of relations. A polysemous[4] word will appear in one synset for each of its

[1] National Institute of Informatics 2-1-2 Hitotsubashi Chiyoda-ku, Tokyo, 101-8430 Japan. {menendez, ichise}@nii.ac.jp
[2] Facultad de Informática y Matemática, Universidad de Holguín Ave. XX Aniversario, Piedra Blanca, Holguín, 80100 Cuba. raul@facinf.uho.edu.cu

[1] The semantic relation that holds between two words that can (in a given context) express the same meaning.
[2] The semantic relation of being subordinate or belonging to a lower rank or class.
[3] The semantic relation that holds between a part and the whole.

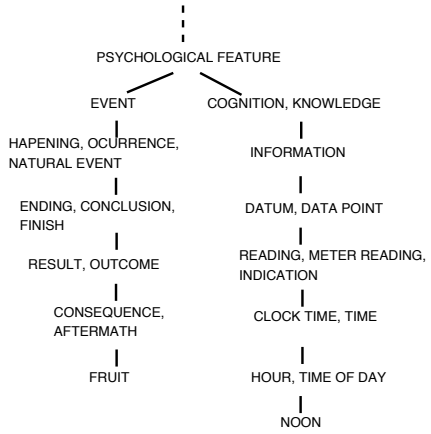

Figure 1. Fragment of WordNet taxonomy.

senses. The backbone of the noun network is the subsumption hierarchy (*hyponymy/hypernymy*), which accounts for close to 80% of the relations in WordNet.

2.2 WordNet based Semantic Similarity Measures

Based on WordNet and depending on the features taken into consideration, semantic similarity measures can be classified into two different types: *edge-based* and *node-based semantic similarity measures*.

Edge based semantic similarity measures tend to focus on structural semantic information like the distance between nodes, i.e., the number of edges separating two nodes. So, distance of the shortest path between the concepts' nodes, depth of the concepts' nodes as well as the depth of their *lowest common superset*[5]*, or subsumer (lcs)* [21] are the main features considered by this type of metrics. Some measures which classify under this category are: Wu & Palmer [21], Leacock & Chodorow [7] and Li [8] semantic similarity measure.

On the other hand, node based semantic similarity measures compute the similarity between two nodes by associating a weight to each of them. From the perspective of information theory this weight stand for the information content (IC) of the concept represented by the node. Some of the semantic similarity measures which classify under this category are: Resnik [16], Jiang & Conrath [6], Lin [9] and Pirró & Seco [15] semantic similarity measure. In the next subsection we will deeper in how this value is computed.

2.3 Information Content

Previous information theoretic approaches for computing semantic similarity [16, 6, 9] obtain the needed IC values by statistically analyzing a corpus. They associate probabilities to each concept in the taxonomy based on word occurrences in a given corpus. The IC value is then obtained by considering the negative log likelihood:

$$IC(c) = -\log p(c) \tag{1}$$

where c is some concept in WordNet and $p(c)$ is the probability of encountering c in a given corpus. Resnik [16] was the first to consider the use of this formula for the purpose of semantic similarity judgments.

Seco et al. [20] proposed a different approach where they decided to use WordNet as a statistical resource with no need for external ones. Their method of obtaining IC values rests on the assumption that concepts with many hyponyms convey less information than concepts that are leaves. So, the more hyponyms a concept has the less information it expresses, otherwise there would be no need to further differentiate it:

$$IC(c) = 1 - \frac{\log\left(hypo(c) + 1\right)}{\log\left(\max_{wn}\right)} \tag{2}$$

where the function $hypo$ returns the number of hyponyms of a given concept and $\max_{wn}$ is a constant that is set to the maximum number of concepts that exist in the taxonomy.

2.4 Commonalities and Differences in Similarity Measures

Let introduce now the measures we will used during the experiment. Relying on the fact that most of the WordNet based semantic similarity measures just take into consideration semantic commonalities among concepts for computing their values, Menéndez and Ichise [11] proposed a featured based model for computing the semantic similarity between words where the strength of semantic differences were also exploited:

$$Sim(c_1, c_2) = \alpha * Comm(c_1, c_2) - \beta * Diff(c_1, c_2) \tag{3}$$

where $Comm(c_1, c_2)$ stands for *semantic commonalities*, $Diff(c_1, c_2)$ the *semantic differences*, and α and β tuning factors ($0 \leq \alpha,\beta \leq 1$) that represent the importance of the commonalities and differences in the model. By applying their model they proposed five semantic similarity measures as modifications of traditional WordNet based metrics from which we will be using two.

The modified Resnik measure considers the *semantic commonalities* to be the information content of the *lcs* and the *semantic differences* to be the information content encompassed by concepts, minus the one already considered in the *lcs*.

$$Sim'_{res}(c_1, c_2) = \begin{array}{r} \alpha * IC(lcs(c_1, c_2) \\ -\beta * (IC(c_1) + IC(c_2) \\ -2 * IC(lcs(c_1, c_2)) \end{array} \tag{4}$$

Menéndez an Ichise observed the modified Jiang & Conrath similarity expression $Sim'_{j\&c}(c_1, c_2)$ they obtained was identical to the modified Resnik expression (Equation 4), and it was a generalization of the Pirró & Seco similarity measure [15].

$$Sim_{P\&S} \subset Sim'_{res}(c_1, c_2) = Sim'_{j\&c}(c_1, c_2) \tag{5}$$

$$Sim_{P\&S}(c_1, c_2) = \begin{cases} 3IC(lcs(c_1, c_2)) & \\ \quad -IC(c_1) - IC(c_2) & if\, c_1 \neq c_2 \\ 1 & if\, c_1 = c_2 \end{cases} \tag{6}$$

Equation 7 represents the modified Lin measure:

$$Sim'_{lin}(c_1, c_2) = \alpha * \frac{2*IC(lcs(c_1,c_2))}{IC(c_1)+IC(c_2)} - \beta * \frac{IC(c_1)+IC(c_2)-2*IC(lcs(c_1,c_2))}{IC(c_1)+IC(c_2)} \tag{7}$$

[4] Polysemy: The ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings.

[5] For example, in Figure 1, the *lcs* between the concepts *fruit* and *noon* is the concept *psychological feature*.

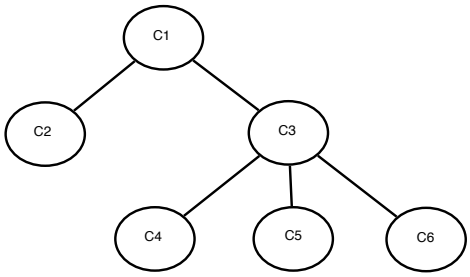

Figure 2. Abstract taxonomy.

Despite proposing a general model, Menendez and Ichise just experimented with the semantic differences [11]. In the following sections we will propose a novel metric for computing the IC which we will use in Pirró & Seco measure as well as in the modified Resnik and Lin measures. We will also study the performance of these semantic similarity measures with the new IC metric.

3 Method

3.1 Approach

As pointed out before, similarity metrics grounded on IC obtain IC's values for concepts by statistically analyzing large corpora and associating a probability to each concept in the taxonomy based on its occurrences within the considered corpora. From a practical point of view, this approach has two main drawbacks: (i) it is time consuming and (ii) it heavily depends on the type of corpora considered. Toward mitigating these drawbacks Seco et al. [20] proposed the Intrinsic Information Content (Equation 2). The IC values of concepts rest on the assumption that the taxonomic structure of WordNet is organized in a meaningful and structured way, where concepts with many hyponyms convey less information than concepts that are leaves, that is, the more hyponyms a concept has the less information it expresses.

From Seco et al. perspective, concepts that are leaf nodes are the most specific in the taxonomy so the information they express is maximal. This means it does not matter how deep is the leaf in the taxonomy. However, our approach, although founded in the same ideas, is different. We stand the depth in the taxonomy it is also an important factor to consider. *The deeper a concept is found in a taxonomy means the amount of previous knowledge is larger and it should bear a higher value of information content.*

For example, in Figure 2 concepts *C2* and *C4* should not have the same IC value, but since they both have the same number of hyponyms under Seco et al. approach they have equal values of IC. We have to clarify that, as well as generally assumed in WordNet, we are considering each of the edges hold the same amount of importance, which indeed it is not true. In the next subsection we introduce the mathematical foundation behind our idea.

3.2 A Novel Information Content Metric

As we made clear in the previous section, our corpora independent IC metric will be a function of the *number of hyponyms* ($hypo(c)$) and the *concept's depth* ($depth(c)$). For example in Figure 2 we say $Sim(c_2, c_1) > Sim(c_4, c_1)$, but if we use Seco et al. intrinsic IC's approach [20] this is not true:

$$\begin{aligned} hypo(c_2) &= hypo(c_4) \\ IC(c_2) &= IC(c_4) \\ 3IC(c_1) - IC(c_2) - IC(c_1) &= 3IC(c_1) - IC(c_4) - IC(c_1) \\ Sim_{P\&S}(c_2, c_1) &= Sim_{P\&S}(c_4, c_1) \end{aligned}$$

To avoid the previous situation we define some properties which should be fulfilled by our IC metric:

$$IC(c) \propto depth(c)$$

$$IC(c) \propto \tfrac{1}{hypo(c)}$$

$$\begin{aligned} IC(c_1) = IC(c_2) \Leftrightarrow \quad & hypo(c_1) = hypo(c_2) \\ & \wedge depth(c_1) = depth(c_2) \\ IC(c_1) < IC(c_2) \Leftrightarrow \quad & hypo(c_1) > hypo(c_2) \\ & \wedge depth(c_1) = depth(c_2) \\ IC(c_1) < IC(c_2) \Leftrightarrow \quad & hypo(c_1) = hypo(c_2) \\ & \wedge depth(c_1) < depth(c_2) \end{aligned}$$

The Information Content metric would be:

$$IC_{hd}(c) = 1 - \frac{\log\left((hypo(c) + 1) * (max_{depth} - depth(c) + 1)\right)}{\log\left(max_{wn} * max_{depth}\right)} \tag{8}$$

where function $hypo$ returns the number of hyponyms of a given concept, max_{wn} is a constant that is set to the maximum number of concepts that exist in the taxonomy, function $depth$ returns the depth of a given concept and max_{depth} represents the maximum depth of the corresponding taxonomy. The denominator, which is equivalent to the value of the most informative concept, serves as a normalizing factor in that it assures that IC values are in the range $[0, 1]$. The above formulation guarantees that the information content decreases monotonically. Moreover, the information content of the imaginary top node of WordNet would yield an information content value of 0.

4 Experiments and results

4.1 Experimental Settings

The purpose of the experiment was to evaluate the new information content metric. We use it into some information content based similarity measures (Equations 4, 6 and 7) and compare the results through the performance of those similarity measures with and without our new information content metric. As a baseline for subsequent comparisons we started with the results presented in [11]. We used the human judgments of Pirró and Seco experiment [15] (P&S in the following) for the word pairs on the Miller and Charles dataset (M&C in the following).

Unfortunately, there is a distinct lack of standards for evaluating semantic similarities, which means that the accuracy of a computational method for evaluating word similarity can only be established by comparing its results against human common sense. That is, a method that comes close to matching human judgments can be deemed accurate. Moreover, some datasets for this evaluation are commonly used. In particular, the Rubenstein and Goodenough dataset (R&G in the following) and Miller and Charles dataset (M&C) are standards dataset for evaluating semantic similarities.

In 1965, Rubenstein and Goodenough [18] obtained "synonymy judgments" of word pairs by hiring 51 subjects to evaluate 65 pairs of nouns. The subjects were asked to assign a similarity from 0 to 4, from "semantically unrelated" to "highly synonymous". Miller and

Charles [12], 25 years later, extracted 30 pairs of nouns from the R&G dataset and repeated their experiment with 38 subjects. The M&S experiment achieved a correlation of 0.97 with the original experiment of R&G. Resnik [16], in 1995, replicated the M&C experiment with 10 computer science students, obtaining a correlation of 0.96. Pirró and Seco [15] (P&S) in 2008 also recreated the R&G experiment this time with 101 subjects, and arrived at a correlation coefficient of 0.972 for the full dataset ($P\&S_{full}$).

As mentioned above, we used the judgments of the P&S experiment for the word pairs of the M&S dataset. However, we considered only 28 word pairs of the 30 used in the M&S experiment: a word missing in WordNet made it impossible to compute ratings for the other two word pairs. All the evaluations were performed using WordNet 3.0 [2] and the Brown Corpus was used for the calculation of the corpora dependent's information content metric. The computation used Pedersen's WordNet::Similarity Perl module [13] as the core. Table 1 shows the parameters values we used for the computation during the experiment.

Table 1. Values of the parameters for the corpus independent information content computation.

Parameter	Taxonomy	Value
$\max_{wn}$	Noun	82115
	Verb	25047
$\max_{depth}$	Noun	20
	Verb	14

For the modified measures, we varied the importance of the difference's factor β in the range of [0,1] and calculated human judgments of the P&S experiment as done in [11]. Nevertheless, we also varied the importance of the semantic commonalities factor α in the range of [0,1] and that's why our results for the modified Resnik measure are better than the one presented in [11].

4.2 Results and Discussion

Table 2 is a summary of the evaluations in the current state of the research. It compiles the correlation values of three similarity measures (Sim'_{lin}, Sim'_{res} and $Sim_{P\&S}$) for the traditional corpus based information content, the intrinsic IC and our proposed metric (IC, IIC and IC_{hd} respectively).

Table 2. Maximum values of correlation obtained using different IC metrics

	IC	IIC	IC_{hd}
Sim'_{lin}	0.858	0.879	0.878
Sim'_{res}	0.831	0.897	0.895
$Sim_{P\&S}$	0.831	0.891	0.892

Analyzing the obtained data we realize of the competitiveness of our new IC metric compared to the intrinsic information content. Furthermore it does outperform the traditional corpora based IC metric for all the similarity measures tested. At this point of the research we still have to find out a better function for describing the relationship between the depth, the number of hyponyms and the IC value of a concept. Due to the "odd" behavior of the function $hypo$ in the domain of English words with a wide range of possible values ($hypo(c) \in [0, 82115]$) it is difficult to properly model the relation between *IC*, *hypo* and *depth*. So we have to model how does behaves an information content metric given two concepts (c_1 and c_2) under the following conditions:

$$hypo(c_1) > hypo(c_2)$$
$$depth(c_1) < depth(c_2)$$

We try to statistically approximate the metric expression through a linear regression but as we expect the relation between $hypo$ and $depth$ functions with IC it is not linear.

5 Concluding Remarks and Future Work

In the present paper we show a new approach for computing IC values in a corpora independent way by using taxonomy properties. We proposed a new IC metric which achieves better results than the traditional corpus based IC metric. Our approach solves the weakness of the intrinsic IC metric while keeping the competitiveness of the metric. We plan to apply some machine learning or evolutive computing methods find out and model different relations between the taxonomy properties an the IC metrics, so we can achieve a bigger improvement of the semantic similarity measures.

REFERENCES

[1] Alexander Budanitsky and Graeme Hirst, 'Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures', in *Workshop on WordNet and Other Lexical Resources, 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, (2001).

[2] *Wordnet: An Electronic Lexical Database*, ed., Christiane Fellbaum, Bradford Books, 1st edn., 1998.

[3] Stephen J. Green, 'Building hypertext links by computing semantic similarity', *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, **11**(5), 713–730, (1999).

[4] Jin Hai and Chena Hanhua, 'Semrex: Efficient search in a semantic overlay for literature retrieval', *Future Generation Computer Systems*, **11**(6), 475–488, (2008).

[5] Angelos Hliaoutakis, Giannis Varelas, Epimenidis Voutsakis, Euripides G. M. Petrakis, and Evangelos E. Milios, 'Information retrieval by semantic similarity', *Int. Journal on Semantic Web and Information Systems (IJSWIS)*, **2**(3), 55–73, (2006).

[6] Jay J. Jiang and David W. Conrath, 'Semantic similarity based on corpus statistics and lexical taxonomy', in *Int. Conf. on Research in Computational Linguistics*, pp. 19–33, (1997).

[7] Claudia Leacock and Martin Chodorow, 'Combining local context and wordnet similarity for word sense identification', in *WordNet: A Lexical Reference System and its Application*, pp. 265–283, (1998).

[8] Yuhua Li, Zuhair A. Bandar, and David McLean, 'An approach for measuring semantic similarity between words using multiple information sources', *IEEE Transactions on Knowledge and Data Engineering*, **15**(4), 871–882, (2003).

[9] D. Lin, 'An information-theoretic definition of similarity', in *15th Int. Conf. on Machine Learning*, pp. 296–304, (1998).

[10] Laurent Mazuel and Nicolas Sabouret, 'Semantic relatedness measure using object properties in an ontology', in *The Semantic Web - ISWC 2008*, volume 5318/2008 of *LNCS*, pp. 681–694, (2008).

[11] Raúl Ernesto Menéndez-Mora and Ryutaro Ichise, 'Effect of semantic differences in wordnet-based similarity measures', in *23rd Int. Conf. on Industrial, Engineering & Other Applications of Applied Intelligent Systems: IEA-AIE 2010*, (2010). Accepted.

[12] G.A. Miller and W.G. Charles, 'Contextual correlates of semantic synonymy', *Languages and Cognitive Processes*, **6**(1), 1–28, (1991).

[13] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi, 'Wordnet::similarity - measuring the relatedness of concepts', in *19th National Conf. on Artificial Intelligence (AAAI-04)*, pp. 1024–1025, (2004).

[14] Catia Pesquita, Daniel Faria1, Andre O. Falcao, Phillip Lord, and Francisco M. Couto, 'Semantic similarity in biomedical ontologies', *PLoS Computational Biology*, **5**(7), (2009).

[15] Giuseppe Pirró and Nuno Seco, 'Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content', in *On the Move to Meaningful Internet Systems: OTM 2008*, pp. 1271–1288, (2008).
[16] Philip Resnik, 'Using information content to evaluate semantic similarity in a taxonomy', *Int. Joint Conf. on Artificial Intelligence*, **14**(1), 448–453, (1995).
[17] Philip Resnik, 'Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language', *Artificial Intelligence Research*, **11**, 95–130, (1999).
[18] Herbert Rubenstein and John B. Goodenough, 'Contextual correlates of synonymy', *Communications of ACM*, **8**(10), 627–633, (1965).
[19] Nuno Seco, *Computational Models of Similarity in Lexical Ontologies*, Master's thesis, University College Dublin, 2005.
[20] Nuno Seco, Tony Veale, and Jer Hayes, 'An intrinsic information content metric for semantic similarity in wordnet', in *European Conference on Artificial Intelligence*, pp. 1089–1090, (2004).
[21] Zhibiao Wu and Martha Palmer, 'Verb semantics and lexical selection', in *32nd Annual Meeting of the Association for Computational Linguistics*, pp. 133–138, (1994).