An Intrinsic Information Content Metric for Semantic Similarity in WordNet

Nuno Seco[1] and **Tony Veale**[1] and **Jer Hayes**[1]

Abstract. Information Content (IC) is an important dimension of word knowledge when assessing the similarity of two terms or word senses. The conventional way of measuring the IC of word senses is to combine knowledge of their hierarchical structure from an ontology like WordNet with statistics on their actual usage in text as derived from a large corpus. In this paper we present a wholly intrinsic measure of IC that relies on hierarchical structure alone. We report that this measure is consequently easier to calculate, yet when used as the basis of a similarity mechanism it yields judgments that correlate more closely with human assessments than other, extrinsic measures of IC that additionally employ corpus analysis.

1 Introduction

Semantic similarity (SS) has for a long time been a subject of intense scholarship in the fields of Artificial Intelligence, Psychology and Cognitive Science. Computational models trying to imitate this human ability date back to Quillian [9] and the spreading activation algorithm.

Nowadays, these computational models of similarity are being included in many software applications with the intent of making these seem more intelligent or even creative (see [2]). The use of SS has also found its way into the Bio-Informatics domain. Recently, Lord [7] studied the effect of using SS strategies when querying DNA and protein sequence databases.

Hence, we present a novel metric of IC that is completely derived from WordNet without the need for external resources from which statistical data is gathered. Experimentation will show that this new metric delivers better results when we substitute our IC values with the corpus derived ones in previously established formulations of SS. These formulations, that make use of IC values, are generally known as Information Theoretic formulas, thus our main focus throughout the paper shall be on these. Nevertheless, when analyzing our results we consider alternative approaches in order to exhaustively evaluate our metric.

2 Information Theoretic Approaches

Previous information theoretic approaches ([4], [10] and [6]) obtain the needed IC values by statistically analyzing corpora. 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:

$$ic_{res}(c) = -log\, p(c) \quad (1)$$

where c is some concept in WordNet and $p(c)$ is the probability of encountering c in a given corpus. Philip Resnik [10] was the first to consider the use of this formula for the purpose of SS judgments. The basic intuition behind the use of the negative likelihood is that the more probable a concept is of appearing then the less information it conveys, in other words, infrequent words are more informative then frequent ones. According to Resnik, SS depends on the amount of information two concepts have in common, this shared information is given by the Most Specific Common Abstraction (MSCA) that subsumes both concepts. In order to find a quantitive value of shared information we must first discover the MSCA, if one does not exist then the two concepts are maximally dissimilar, otherwise the shared information is equal to the IC value of the MSCA. Formally, semantic similarity is defined as:

$$sim_{res}(c_1, c_2) = \max_{c \in S(c_1,c_2)} ic_{res}(c) \quad (2)$$

where $S(c_1, c_2)$ are the set of concepts that subsume c_1 and c_2.

Another information theoretic similarity metric that used the same notion of IC was that of Lin [6], expressed by:

$$sim_{lin}(c_1, c_2) = \frac{2 \times sim_{res}(c_1, c_2)}{(ic_{res}(c_1) + ic_{res}(c_2))} \quad (3)$$

Jiang and Conrath [4] also continued on in the information theoretic vein and suggested a new measure of semantic distance (if we consider the opposite of the distance we obtain a measure of similarity). The most frequently observed version of their distance metric is:

$$dist_{jcn}(c_1, c_2) = (ic_{res}(c_1) + ic_{res}(c_2)) - 2 \times sim_{res}(c_1, c_2) \quad (4)$$

3 Information Content in WordNet

As was made clear in the previous section, IC is obtained through statistical analysis of copora, from where probabilities of concepts occurring are inferred. We feel that WordNet can also be used as a statistical resource with no need for external ones. Moreover, we argue that the WordNet taxonomy may be innovatively exploited to produce the IC values needed for SS calculations.

Our method of obtaining IC values rests on the assumption that the taxonomic structure of WordNet is organized in a meaningful and principled way, where concepts with many hyponyms convey less information than concepts that are leaves. We argue that the more hyponyms a concept has the less information it expresses, otherwise there would be no need to further differentiate it. Likewise, concepts that are leaf nodes are the most specified in the taxonomy so the information they express is maximal. Hence, we express the IC value

[1] Department of Computer Science, University College Dublin, Ireland email: {nuno.seco, tony.veale, jer.hayes}@ucd.ie

of a WordNet concept as a function of the hyponyms it has. Formally we have:

$$ic_{wn}(c) = \frac{log(\frac{hypo(c)+1}{max_{wn}})}{log(\frac{1}{max_{wn}})} = 1 - \frac{log(hypo(c)+1)}{log(max_{wn})} \quad (5)$$

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. 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 $[0, .., 1]$. 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 Empirical Studies

In order to evaluate our IC metric we decided to use the three formulations of SS presented in section 2 and substituted Resnik's IC metric with the one presented in equation 5. In accordance with previous research, we evaluated the results by correlating our similarity scores with that of human judgments provided by Miller and Charles [8]. In their study, 38 undergraduate subjects were given 30 pairs of nouns and were asked to rate similarity of meaning for each pair on a scale from 0 (no similarity) to 4 (perfect synonymy). The average rating for each pair represents a good estimate of how similar the two words are.

In order to make fair comparisons we decided to use a independent software package that would calculate similarity values using previously established strategies while allowing the use of WordNet 2.0. One freely available package is that of Siddharth Patwardhan and Ted Pederson[2]; which implement semantic relatedness measures described by Leacock and Chodorow [5], Jiang and Conrath [4], Resnik [10], Lin [6], Hirst and St. Onge [3], Wu and Palmer [12], the adapted gloss overlap measure by Banerjee and Pedersen [1]. In addition to these we also used Latent Semantic Analysis (LSA) to perform similarity judgments by means of a web interface available at the LSA website[3].

Table 1 presents the similarity obtained using the chosen algorithms and their correlation coefficient (γ) with the human judgments. The first column states the algorithm used in obtaining similarity scores and the second the correlation between the algorithm and human ratings. The last three rows correspond to algorithms using our IC values.

It should be noted that for the sake of coherence of our implementations we normalized and applied a linear transformation to the Jiang and Conrath formula transforming it into a similarity function. The resulting formulation is:

$$sim_{jcn}(c_1, c_2) = 1 - (\frac{ic_{wn}(c_1) + ic_{wn}(c_2) - 2 \times sim_{res'}(c_1, c_2)}{2}) \quad (6)$$

Note that $sim_{res'}$ corresponds to Resnik's similarity function but now accommodating our IC values.

5 Discussion and Future Work

The results obtained using our IC values in the information theoretic formulas seem to have outperformed their homologues which suggests that the initial assumption concerning the taxonomic structure of WordNet is correct. It should be noted that the maximum value obtained, using Jiang and Conrath's formulation, is very close to what Resnik [11] proposed as a computational upper bound. One major advantage of this approach is that it does not rely on corpora analysis, thus we avoid the sparse data problem which is evident in many corpus based approaches.

Future research regarding the Information Content metric will make use of taxonomies other than WordNet, such as the Gene Ontology. This will allow us to conclude if our metric generalizes and can be used with other hierarchal knowledge bases.

[2] This software can be downloaded at http://www.d.umn.edu/ tpederse/.
[3] The web interface can be accessed at http://lsa.colorado.edu/.

Algorithm	γ
Leacock Chodorow	0,82
Hirst St. Onge	0,68
Banerjee and Pedersen	0,37
Wu and Palmer	0,74
LSA	0,72
Resnik	0,77
Lin	0,80
Jiang and Conrath	-0,81
Resnik*	0,77
Lin*	0,81
Jiang and Conrath*	0,84

Table 1. Correlation between human and machine similarity judgments.

REFERENCES

[1] Satanjeev Banerjee and Ted Pedersen, 'Extended gloss overlaps as a measure of semantic relatedness', in *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pp. 805–810, Acapulco, Mexico, (August 2003).

[2] Paulo Gomes, Nuno Seco, Francisco C. Pereira, Paulo Paiva, Paulo Carreiro, José L. Ferreira, and Carlos Bento, 'The importance of retrieval in creative design analogies', in *Procceedings of the International Joint Conference on Artificial Intelligence IJCAI'03 Workshop: "3rd Workshop on Creative Systems"*, (2003).

[3] Graeme Hirst and David St-Onge, 'Lexical chains as representations of context for the detection and correction of malapropisms', in *WordNet: An Electronic Lexical Database*, ed., Christiane Fellbaum, chapter 13, 305–332, MIT Press, (1998).

[4] J. Jiang and D. Conrath, 'Semantic similarity based on corpus statistics and lexical taxonomy', in *Proceedings of the International Conference on Research in Computational Linguistics*, (1998).

[5] C. Leacock and M. Chodorow, 'Combining local context and wordnet similarity for word sense identification', in *WordNet: An Electronic Lexical Database*, ed., Christiane Fellbaum, 265–283, MIT Press, (1998).

[6] Dekang Lin, 'An information-theoretic definition of similarity', in *Proceedings of the 15th International Conf. on Machine Learning*, pp. 296–304. Morgan Kaufmann, San Francisco, CA, (1998).

[7] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble, 'Semantic similarity measures as tools for exploring the gene ontology', in *Procceedings of the 8th Pacific Symposium on Biocomputing*, pp. 601–612, (2003).

[8] George Miller and W.G. Charles, 'Contextual correlates of semantic similarity', *Language and Cognitive Processes*, **6**, 1–28, (1991).

[9] M. Quillian, *Semantic Memory*, 227–270, MIT Press, 1968.

[10] Philip Resnik, 'Using information content to evaluate semantic similarity in a taxonomy', in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–453, (1995).

[11] Philip Resnik, 'Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language', *Journal of Artificial Intelligence Research*, **11**, 95–130, (1999).

[12] Z. Wu and M. Palmer, 'Verb semantics and lexical selection', in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94), Las Cruces, NM, 133-138*, (1994).