International Journal of Soft Computing 2 (5): 617-623, 2007
ISSN: 1816-9503

Measuring Semantic Similarity Between the Concepts Based on an Ontology

[1]P. Selvi and [2]N.P. Gopalan
[1]Department of Computer Science and Engineering, [2]Department of Computer Applications,
National Institute of Technology, Tiruchirappalli-620015, India

Abstract: Ontologies are in the heart of the knowledge management process. Different semantic measures have been proposed in the literature to evaluate the strength of the semantic link between two concepts or two groups of concepts from either two different ontologies (ontology alignment) or the same ontology. This study proposes a method for measuring semantic similarity/distance between terms. This measure combines strengths and complements weaknesses of existing measures that use knowledge base as primary source. The proposed measure uses a new feature of common specificity (ComSpe) besides the path length feature. The ComSpe feature is derived from_information content of concepts and information content of the knowledge base given a corpus. We evaluated the proposed measure with benchmark test set of term pairs scored for similarity by human experts. The experimental results demonstrated that our similarity measure is effective and outperforms the existing measures. The proposed semantic similarity measure gives the best correlation (0.874) with human scores in the benchmark test set compared to the existing measures.

Key words: Semantic similarity, lexical database, wordNet, specificity, ontology, word similarity, information content, corpus statistics

INTRODUCTION

Semantic similarity measures play important roles in information retrieval and Natural Language Processing. The need to determine the degree of semantic similarity between two lexically expressed concepts is a problem that pervades much of computational linguistics. Measures of similarity or relatedness are used widely in such NLP applications as word sense disambiguation; example based machine translation, determining discourse structure, text classification, summarization and annotation, information extraction and retrieval, automatic indexing, lexical selection (Rubenstein and Goodenough, 1965). Despite the usefulness of semantic similarity measures in these applications, robustly measuring semantic similarity between two words (or entities) remains a challenging task.

There are many methods to compute word similarity (or relatedness) nowadays. Generally speaking, they can be classified into two basic methods: one is based on ontology or a semantic taxonomy and the other is based on collocations of words in a corpus. On the other hand, WordNet (Abney and Light, 1999; Lin, 1998) is particularly well suited for English word similarity measures and many researchers proposed different measures of similarity, which were based on it. These measures vary from simple edge-counting to attempts to factor in peculiarities of the network structure by considering link direction (Wu and Palmer, 1994) (hso), random methods which return random numbers, relative depth or path (random,wup,lch) (Leacock and Chodorow, 1998; Wu and Palmer, 1999) and density (Eneku and Gernnan, 1996). These analytic methods now face competition from statistical and machine learning techniques; but a number of hybrid approaches have been proposed that combine a knowledge-rich source, such as a thesaurus, with a knowledge-poor source, such as corpus statistics (res,lin,jcn) (Leacock and Chodorow, 1998; Miller and Charles, 1991; Reshik, 1999; Jiang and Conrath, 1997). In 2003, Pedersen and Banerjee pointed out the Adapted Lesk (lesk) (Richardson *et al*., 1994) and Patwardhan suggested context vector (vector) (Resnik, 1995; Rubenstein and Goodenough, 1965) to measure the similarity or relatedness of English words.

In this study, we explore the existing semantic similarity measures that use ontology as primary information source and then we propose a new ontology based measure. The proposed measure is a combination measure using ontology structure and corpus-based features that have a great potential in measuring semantic

Corresponding Author: P. Selvi, Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, India-620015

similarity. Moreover, the proposed measure uses a new feature of common specificity (ComSpe) in addition the path length feature. We evaluated the proposed measure with benchmark test set of term pairs scored for similarity by human experts. The experimental results show that our technique is effective producing the best correlation with human scores in the benchmark test set compared with the existing measures. In this study, we use the term "concept node" to denote a concept class represented as a node on ontology and that contains a set of synonymous concepts. The similarity of two concepts belonging to the same node (i.e., synonymous concepts) reaches maximum and the similarity of two concepts is the similarity of two concept nodes containing them.

EXISTING MEASURES OF SIMILARITY

Many techniques have been proposed for evaluating the semantic similarity between two concepts in a HO. They can be classified into two categories: Edge based and node based approaches. These approaches are duals, as the similarity can be defined as 1-distance when values are normalized to [0..1].

The edge based approach is the traditional, most intuitive and simplest similarity measure. It computes the distance between two concepts based on the number of edges found on the path between them. Resnik (1995) introduced a variant of the edge-counting method, converting it from a distance to a similarity metric by subtracting the path length from the maximum possible path length:

$$\text{sim}_{\text{EDGE}}(a,b) = (2\times D) - \text{len}(a,b) \quad (1)$$

Where a and b are concepts in the taxonomy is the maximum depth of the taxonomy and len (a, b) is the shortest path between concepts a and b. Another popular variant of the edge based approach is the metric proposed by Leacock and Chodorow (1998) which scales the shortest path by twice the maximum depth of the taxonomy.

$$\text{sim}_{\text{LEACOCK}}(a,b) = -\log\left(\frac{\text{len}(a,b)}{2\times D}\right) \quad (2)$$

The node-based approach was proposed by Resnik (1995) to overcome the drawbacks of the edgecounting approach, which considers the distance uniform on all edges. Resnik defined the similarity between two concepts as the information content of the lowest common ancestors, LCA(a, b). The Information Content (IC) of a concept c is defined as the negative log likelihood of the probability of encountering an instance of the concept, i.e. IC(c) = -log P(c). The intuition behind the use of the negative likelihood is that the more probable a concept is of appearing, then the less information it conveys. Formally, the similarity is defined as follows.

$$\text{sim}_{\text{RESNIK}}(a,b) = \max_{c\in \text{LCA}(a,b)} \text{IC}(c) \quad (3)$$

While Resnik defined the similarity based on the shared information (Lin, 1998) defined the similarity between two concepts as the ratio between the amount of information needed to state the commonality between these two concepts and the information needed to fully describe them.

$$\text{sim}_{\text{LIN}}(a,b) = \frac{2\times \text{IC}(\text{LCA}(a,b))}{\text{IC}(a)+\text{IC}(b)} \quad (4)$$

Hybrid approaches combine both approaches defined above. Resnik (1995) proposed a combined model that is derived from the edge-based notion by adding the information content as a decision factor. They defined the link strength between two concepts as the difference of information content between them. Following this, Jiang's distance metric is defined as follows:

$$\text{sim}_{\text{JIANG}}(a,b) = \text{IC}(a)+\text{IC}(b)-2\times \text{IC}(\text{LCA}(a,b)) \quad (5)$$

PROPOSED SIMILARITY MEASURE

The knowledge bases may be constructed in a hierarchy that is commonplace in the world. The lexical hierarchy is connected by following trails of super ordinate terms in "is a" or "is a kind of" (ISA) relations. The ISA hierarchical structure of the knowledge base is important in determining the semantic distance between words. Figure 1 shows a portion of such a hierarchical semantic knowledge base.

Given two words w1 and w2, we need to find the semantic similarity of s(w1, w2) for these two words. We

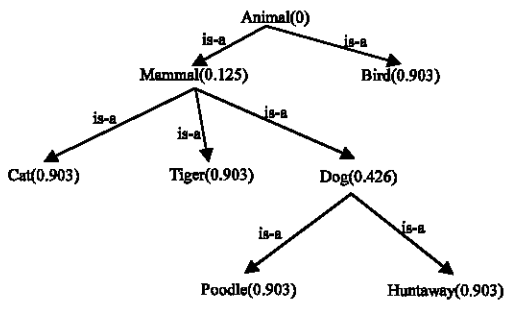

Fig. 1: A simple animal ontology

Table 1: Similarity features of 8 similarity measures

	Chraracteristics		Features	
Similarity measure	Sources	Semantics	Path	Depth
Rada	Ontology	Distance	*	None
Resnik	Ontology+corpus	Similarity	None	Wt*
Leacock-Chodorow	Ontology	Similarity	*	None
Jiang-Cornath	Ontology+corpus	Distance	Wt*	None
Wu-Palmer	Ontology	Similarity	None	*
Lin	Ontology+corpus	Similarity	None	Wt*
Sussna	Ontology	Distance	Wt*	*
Hirst-StOnge	ontology	Relatedness	*	None
Proposed measure	Ontology+Corpus	Distance	Wt*	Wt*

* is denoted for path length or depth length, Wt* is denoted for weighed path or weighted depth, "none" is denoted for the feature is not used by measure

can do this by analysis of the knowledge base, as follows: Words are associated with concepts in the ISA hierarchy. Therefore, we can find the first concept in the hierarchical semantic network that subsumes the concepts containing the compared words. One direct method for similarity calculation is to find the minimum length of path connecting the two concepts containing the two words (Rada *et al*., 1989). For example, Fig. 1 illustrates a fragment of the semantic hierarchy of WordNet (Miller, 1995). The shortest path between cat and dog is cat-mammal-dog, the minimum length of path is 2, the synset of mammal is called the subsumer for the words cat and dog; while the minimum path length between cat and tiger also 2. From the result, a wrong conclusion is that cat, dog and tiger are equally similar but we may know that cat is more similar to tiger than to dog.

Rada *et al.* (1989) demonstrated that this method works well on their much constrained medical semantic nets (with 15,000 medical terms). However, this method may be not so accurate if it is applied to larger and more general semantic nets such as WordNet (Miller, 1995). To address this weakness, the direct path length method must be modified by utilizing more information from the hierarchical semantic nets. It is intuitive that concepts at upper layers of the hierarchy have more general semantics and less similarity between them, while concepts at lower layers have more concrete semantics and stronger similarity. Therefore, the depth of concept in the hierarchy should be taken into account. Moreover, local density of the semantic nets is also a factor that affects the similarity between words. In summary, similarity between words is determined not only by path length but also by depth and density.

The Table 1 summarizes the characteristics and the influential features used by the existing measures along with our proposed measure. All the measures in the Table 1 use either the path or depth feature but not both; therefore, can be grouped into: Path-based measures (Path-length, Leacock and Chodorow and Jiang and Conrath) and depth-based measures (Wu and Palmer, 1994; Resnik, 1999; Lin, 1998). However, no proposed measure takes into account the whole features without use of the corpus.

We propose a new method based on the combined features of weighted path and weighed depth features in one measure as path length and depth length are special cases of weighted path and weighted depth. In weighted path or weighted depth approaches, the links between ontology nodes are not equal in term of strength/weight and link strength can be determined by local density, node depth, information content and link type (Leacock and Chodorow, 1998; Richardson *et al*., 1994).

However, weighted path approach, e.g,. Jiang and Conrath (1997) has limitation as it takes into account individual IC of individual concept nodes; therefore, it is affected by using a small corpus as some words may not occur in small corpora. Thus, such words will always have their similarity with any other word reaches the minimum. Through using path length, we can see the relationships between any presented concepts in the ontology. Therefore, we use node counting for path feature. Beside path length feature, we also use weighted depth as kind of specificity of concept nodes in the measure.

Additional feature: The proposed measure uses a new feature of common specificity (ComSpe) besides the path length feature. The ComSpe feature is derived from information content of concepts and information content of the ontology given a corpus. The LCS node of two given concept nodes determine their common specificity in ontology. We define the common specificity of two concept nodes in ontology based on ontology structure and corpus as follows:

$$ComSpe(w_1, w_2) = IC_{max} - IC(LCS(w_1, w_2)) \quad (6)$$

Where ICmax (ontology information content) is the maximum IC of concept nodes in the ontology. The *ComSpe* feature determines the common specificity of two

concept nodes in the ontology based on given corpus and ontology structure. The less the common specificity value of 2 concept nodes the more they are share information and thus the more they are similar. When the IC of LCS of two concept nodes (w1 and w2) reaches ICmax, that is,

$$IC(LCS(w_1, w_2)) = IC_{max} \quad (7)$$

then the two concept nodes reach the highest common specificity which equals to zero:

$$ComSpe(w_1, w_2) = 0 \quad (8)$$

The combined semantic distance measure: The contribution of this study is twofold: Introduce the new common specificity feature and the way we combine (non-linearly) the semantic features in the measure. In this study, we discuss these two points and present the new semantic similarity/distance measure.

Each of the two features (*viz.* Path length and ComSpe) is a semantic distance feature and can form a semantic similarity measure by itself. We would like our proposed semantic measure to achieve the following conditions. When path length equals to one (e.g., two concept nodes are the same node in the ontology) the semantic distance value must reach minimum (thus similarity reaches maximum) regardless of *ComSpe* feature as the two concepts are synonymous or identical. Therefore, we use product of semantic distance features. The shorter the path length (shortest path length) between two concept nodes in the hierarchy tree, the more similar they are. Lower level pairs of nodes are semantically closer (more similar) than higher-level pairs. Then the proposed semantic similarity measure will be as follow:

$$SemDist(w_1, w_2) = \log\left((path - 1)^{\alpha} \times (ComSpe)^{\beta} + c\right) \quad (9)$$

Where $\alpha > 0$ and $\beta > 0$ are contribution factors of two features (Path length and (ComSpe (w1, w2)); c is a constant. Path is the path length (shortest path length) of two concept nodes using node counting. If c is zero, the combination is linear and to insure the distance is positive and the combination is non-linear, c must be greater or equal to one (c = 1). When two concept nodes have path length of 1 using node counting (Path = 1), then they have a minimum semantic distance (i.e., maximum similarity) that equals to zero regardless of common specificity feature.

IMPLEMENTATION

Two databases are used in the implementation of the proposed method of semantic similarity; they are WordNet (Miller, 1995) and the Brown Corpus (Franics and Kucera, 1979). Both databases are publicly available and widely used in previously published works. This section first provides a brief description of these two databases, then presents the search in the lexical taxonomy and the statistics from the corpus.

WordNet is an on-line semantic dictionary-a lexical database, developed at Princeton by a group led by Miller (1995). The version used in this study is WordNet 2.0. WordNet partitions the lexicon into nouns, verbs, adjectives and adverbs. Nouns, verbs, adjectives and adverbs are organized into synonym sets, called synsets. A synset represents a concept in which all words have similar meaning. Thus, words in a synset are interchangeable in some syntax. Knowledge in a synset includes the definition of these words as well as pointers to other related synsets.

We also used and inherited existing implemented measures in the Perl module WordNet: Similarity developed by Pedersen and used Resnik's (1995) technique to calculate IC of concept particularly for nouns based on their frequencies.

The Brown Corpus (Francis and Kucera, 1979) of Standard American English was the first of the modern, computer readable, general corpora. It was compiled by Francis and Kucera of Brown University. The corpus consists of one million words of American English texts printed in 1961. The texts for the corpus were sampled from 15 different text categories to make the corpus a good standard reference. The number of texts in each category varies. There are a total of 500 texts, each consisting of just over 2,000 words. Much research within the field of corpus linguistics has been made using these data.

Obtaining information sources: The statistics from the Brown Corpus are used to obtain the information content of a concept. There are some slightly different methods of calculating the concept probabilities in a corpus (Miller *et al*., 1993). In this research, we use Resnik's (1999) method, particularly for noun probability. Each noun that occurred in the corpus was counted as an occurrence of each taxonomic class containing it. We compute the frequency freq(c) of a concept node *c* by counting all the occurrences of the concepts in corpus contained in or subsumed by the concept node c. Then concept node probability is computed directly as:

$$\text{freq}(c) = \sum_{w \in words(c)} \text{count}(w) \quad (10)$$

$$p(c) = \frac{\text{freq}(c)}{N} \quad (11)$$

Where N is the total number of nouns in the corpus that are also present in WordNet. The information content of concept *c* is then given by:

$$IC(c) = -\log p(c) \quad (12)$$

The benchmark data sets: There are two well-known benchmark test sets of term pairs that were scored by human experts for semantic similarity for general English. The first set (RG) is collected by Rubenstein and Goodenough (1965) and covers 51 subjects containing 65 pairs of words on a scale from "highly synonymous" to "semantically unrelated" (Table 2 contains only subset of this dataset). The second dataset (MC) was collected by Miller and Charles (1991) in a similar experiment conducted 25 years after Rubenstein and Goodenough collected RG set and contains 30 pairs extracted from the 65 pairs of RG and covers 38 human subjects.

EXPERIMENTAL EVALUATION

In this study, we present and discuss the evaluation procedure and the experimental results of the proposed measure. We carried out the experiments with two steps. First, we tune the strategy parameters on the training data set D1. Given the value of a parameter, semantic similarity values of the word pairs are calculated. Then, the correlation coefficient between the computed semantic similarity values and the human ratings of Rubenstein-Goodenough's is calculated. Thus, a set of correlation coefficients is obtained by changing the value of the strategy parameters. The parameters resulting in the greatest correlation coefficient are considered as the optimal parameters for that particular strategy. Second, the identified optimal parameters are used to calculate semantic similarity for word pairs in test data set D0. Again, the correlation coefficient between computed similarity values and human ratings of Rubenstein-Goodenough's is calculated for words pairs in D0. This correlation coefficient is used to judge the suitability of the particular strategy comparing to other strategies and previously published results.

The Table 3 shows part of this dataset that can be found in WordNet. We used above training dataset to

Table 2: Semantic Similarity of Human Ratings and Basic Measures for Test Set D0 (Top 15 pairs)

Word peir	RG rating	Information content	Length	Depth
Fruit-furnace	0.05	1.8563	6	2
Autograph-shore	0.06	0	30	0
Automobile-wizard	0.11	0.9764	11	0
Mound-stove	0.14	2.9062	6	2
Grin-implement	0.18	0	30	0
Asylum-fruit	0.19	1.8563	6	2
Asylum-monk	0.39	0.9764	10	0
Graveyard-madhouse	0.42	0	12	1
Boy-rooster	0.44	2.3852	11	1
Cushion-jewel	0.45	1.8563	6	2
Asylum-cemetery	0.79	0	9	1
Grin-lad	0.88	0	30	0
Shore-woodland	0.90	1.5095	5	1
Boy-sage	0.96	2.5349	5	2
Automobile-cushion	0.97	2.9062	7	3

Table 3: Semantic Similarity of Human Ratings and Basic Measures for Training Set D1

Word peir	RG rating	MC replica	Resnik replica	Information content	Length	Depth
Cord-smile	0.02	0.13	0.1	1.1762	12	0
Rooster-voyage	0.04	0.08	0	0	300	0
Noon-string	0.04	0.08	0	0	30	0
Glass-magician	0.44	0.11	0.1	1.0105	8	0
Monk-slave	0.57	0.55	0.7	2.9683	4	2
Coast-forest	0.85	0.42	0.6	0	6	1
Monk-oracle	0.91	1.1	0.8	2.9683	7	2
Lad-wizard	0.99	0.42	0.7	2.9683	4	2
Forest-graveyard	1.00	0.84	0.6	0	7	1
Food-rooster	1.09	0.89	1.1	1.0105	12	0
Coast-hill	1.26	0.87	0.7	6.2344	4	3
Car-journey	1.55	1.16	0.7	0	30	0
Crance-implement	2.37	1.68	0.3	2.9683	4	3
Brother-lad	2.41	1.66	1.2	2.9355	4	2
Bird-crane	2.63	2.97	2.1	9.3139	3	5

Table 4: Results of absolute correlations of the proposed measure with human ratings using the training dataset with different parameter values

	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 3$	$\alpha = 3$
	$\beta = 1$	$\beta = 1$	$\beta = 1$	$\beta = 1$	$\beta = 1$
Parameter values	$c = 1$	$c = 1$	$c = 1$	$c = 2$	$c = 3$
Semantic distance	0.703	0.739	0.749	0.735	0.734
Correlation	0.734	0.71	0.872	0.874	0.873

Table 5: Absolute correlations with RG human ratings and WordNet 2.0 for four combination-based measures

Measure	Correlation with RG
Combined approach	0.874
Resnik	0.830
Jiang and conrath	0.854
Lin	0.853

train for optimal parameters of the proposed measure. Table 4 shows some experiment results using two corpora. When $\alpha = 3$ and $\beta = 1$ the performances of SemanticDist are very close and reach highest correlations with human scores (Table 4). We also observe from the results in Table 5 that should be greater than to get higher correlations. This implies that the Path feature contributes more to the semantic similarity than the *ComSpe* feature Eq. 9. The testing was conducted using the RG test set (65 pairs)and Brown Corpus.

The results of the RG experiments in Table 4 shows that our measure produces good and stable performance with this set. Furthermore, the correlation results in Table 4 shows that the proposed measure can perform well in any corpus sizes and reach very good correlations with RG dataset. Furthermore, we also investigate performances of other information-based measures on two corpora using the RG dataset and WordNet 2.0 and the results are in Table 4.

Based on the benchmark data set, the optimal parameters for the proposed measure are: $\alpha = 3$; $\beta = 1$. The experimental results demonstrated that our measure significantly outperforms published measures and is close to individual human judgement (with a correlation of 0.874). The results in Table 5 shows clearly that our SemDist measure, outperforms the other information-based measures. Moreover, Resnik gives a good stability in performance using the corpus compared with Jiang and Conrath and Lin.

CONCLUSION

This study presented word similarity measures from a new perspective. The proposed measure combines all the strengths of some traditional approaches. In a similar manner to other researchers, we carried out experiments on a benchmark set of word pairs with human similarity ratings. The best correlation against Rubenstein-oodenough's human similarity ratings in literature has been 0.8484 (Francis and Kucera, 1979; Miller *et al*., 1991), while ours is 0.874. The experimental results demonstrated that our measure significantly outperforms previous published measures. Our measure uses a new feature (ComSpe) that contributes well to the performance given by scaling the IC of the least common subsumer of two given concepts to the maximum IC of the ontology. Furthermore, the proposed measure can be adaptive to get optimum performance in specific domain by effective training strategy and can perform well in any corpus size.

REFERENCES

Abney, S. and M. Light, 1999. Hiding a Semantic Class Hierarchy in a Markov Model, Proc. ACL Workshop Unsupervised Learning in Natural Language Processing, pp: 1-8.

Eneko Agirre and German Rigau, 1996. Word sense disambiguation using conceptual density. In: Proc. 16th Int. Conf. Computational Linguistics, Copenhagen, pp: 16-22.

Francis, W.N. and H. Kucera, 1979. Brown Corpus Manual-Revised and Amplified, Department of Linguistics, Brown University, Providence, R.I.

Jiang, J. and D. Conrath, 1997. Semantic similarity based on corpus statistics and lexical taxonomy, In Proceedings of International Conference on Research in Computational Linguistics, Taiwan.

Leacock, C. and M. Chodorow, 1998. Combining local context and WordNet similarity for word sense identification, In Fellbaum, pp: 265-283.

Lin, D., 1998. An information-theoretic definition of similarity, In: Proceedings of the 15th International Conference on Machine Learning, Madison, WI.

Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross and K. Miller, 1993. Introduction to WordNet: An online Lexical Database, in Five Papers on WordNet, CSL report, Cognitive Science Laboratory, Princeton University.

Miller, G. and W.G. Charles, 1991. Contextual Correlates of Semantic similarity. Language and Cognitive Processes, 6: 1-28-1991.

Miller, G.A., 1995. WordNet: A Lexical Database for English, Comm.ACM., 38: 39-41.

Rada, R., H. Mili, E. Bichnell and M. Blettner, 1989. Development and Application of a Metric on Semantic Nets, IEEE. Trans. Sys. Man and Cybernetics, 9: 17-30.

Resnik P., 1995. Using information content to evaluate semantic similarity, In: Proc. 14th Int. Joint Conf. Artificial Intelligence, Montreal, pp: 448-453.

Resnik, P., 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Applications to Problems of Ambiguity in Natural Language, J. Artificial Intelligence Res., 11: 95-130.

Richardson, R., A.F. Smeaton and J. Murphy, 1994. Using WordNet as a Knowledge Base for Measuring Semantic Similarity, Working paper CA-1294, School of Computer Applications, Dublin City University, Dublin.

Rubenstein, H. and J.B. Goodenough, 1965. Contextual Correlates of Synonymy. Comm. ACM., 8: 627-633.

Wu, Z. and M. Palmer, 1994. Verb Semantics and Lexical Selection, In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico.