A New Model of Information Content Based on Concept's Topology for Measuring Semantic Similarity in WordNet[1]

Lingling Meng[1], Junzhong Gu[2] and Zili Zhou[3]

[1]*Computer Science and Technology Department, Department of Educational Information Technology, East China Normal University, Shanghai, 200062, China*

[2]*Computer Science and Technology Department, East China Normal University, Shanghai, 200062, China*

[3]*College of Physics and Engineering, Qufu Normal University, Qufu, 273165, China*

llmeng@deit.ecnu.edu.cn, jzgu@ica.stc.sh.cn, zlzhou999@163.com

Abstract

Information content plays an important role in measuring semantic similarity of concepts. The conventional way of IC obtained is through statistical analysis of corpora. Recently corpora–independent model has attracted great concern in this area. This paper analyzes the state-of-art IC models, highlights important related issues, and presents a novel IC model based on concepts' topology in WordNet. Different from previous work, for a given concept, the depth itself, the number of its hyponyms, and the depth of every hyponym have been taken into considered. Experiment demonstrates that our approach is able to provide more accurate similarity evaluation and achieves significant performance than related works.

Keywords: *IC model, semantic similarity, corpora-independent, concept's topology*

1. Introduction

Semantic similarity measure between concepts is a hot topic for many years in artificial intelligence and cognitive science. It can be dated back to Quillian [1] and the spreading activation algorithm. Nowadays, semantic similarity has been successfully applied in word sense disambiguation [2], information extraction [3, 4], semantic annotation [5], question answering [6], recommender system [7] and so on. Besides this, it also shows its talents in software domain [8] and Bio-Informatics domain [9]. Information content(IC) is an important dimension in measuring semantic similarity between two concepts, and provides an estimation of concept's abstract/specialty, which contributes to better understand concepts' semantic.

This paper presents a novel IC model based on the topology of concepts in WordNet without the help of corpora. Different from previous work [10, 11, 12, 13], for a given concept, the depth of the concept, the number of its hyponyms, and, the depth of every hyponym has been taken into considered in our new model. Experiment shows that our method is able to provide more accurate similarity evaluation and achieves significant performance improvements than related work.

[1] The work in the paper is supported by Shanghai Scientific Development Foundation (Grant No. 11530700300) and Shandong Excellent Young Scientist Award Fund (Grant No. BS2010DX012)

2. Related Work

In the past years many measures have been proposed for measuring semantic similarity in the literature. The measures based on WordNet have attracted great concern recently. WordNet is the product of a research project at Princeton University [23]. It is a large lexical database of English. In WordNet nouns, verbs, adjectives and adverbs are grouped into sets of synsets, which are interconnected via a variety of relations such as meronym/holonym (part-of), hyponym/hypernym (is-a) and so on. For example, an apple is a fruit, and Shanghai is part of China. Hyponym/hypernym (is-a) is the most common relations. An example of is-a relation in WordNet is shown as Figure 1. In the taxonomy the deeper concept is more specific and the upper concept is more abstract. Therefore C_2 is more abstract than C_5. C_5 is more abstract than C_{12}. C_{12} is more abstract than C_{19}. C_1 is the most abstract concept.

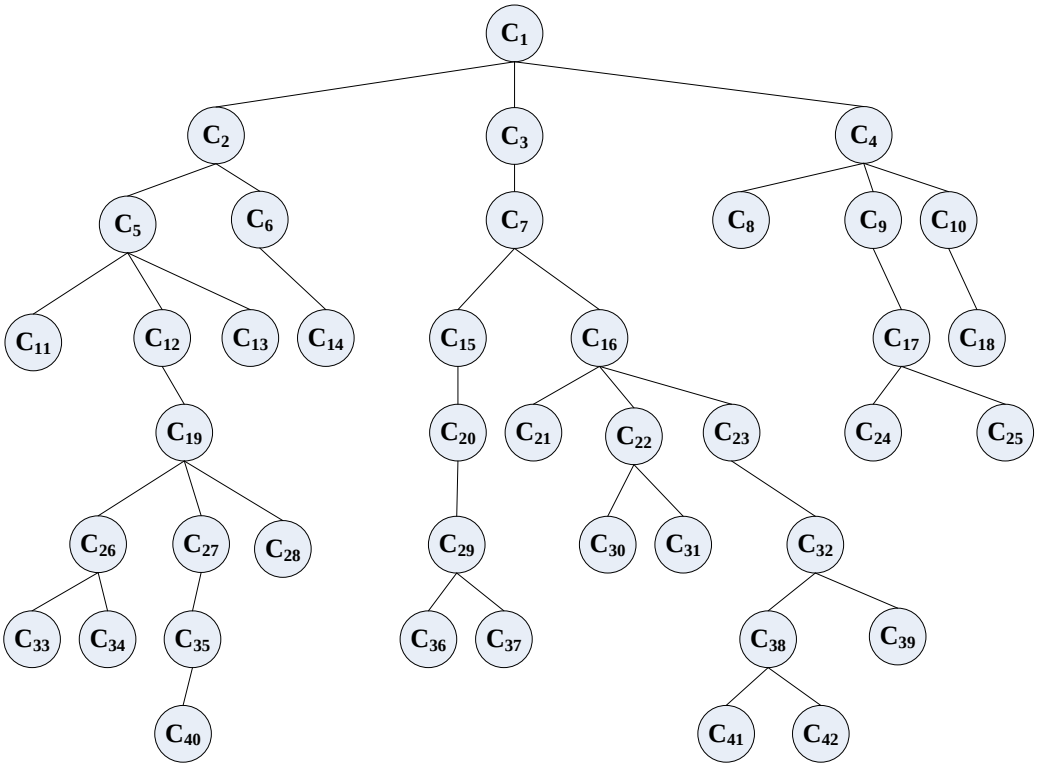


Figure 1. An Example of is-a Relation

These relations are associated with words and words to form a hierarchy structure, which makes it a useful tool for computational linguistics. Because language semantics are mostly captured by nouns and noun phrases, so that we only focus on the similarity measures based on nouns and is-a relations in WordNet. Generally speaking, all the measures based on WordNet can be partitioned into four families [14], that is edge-based approaches (based on how close the two concepts in the taxonomy are) [15, 16], feature-based approaches(base on the properties of the concepts) [17], information-based approaches(based on how much information the two concepts share) [10, 18, 19] and hybrid-based approaches (based on combinations of the previous options) [20]. Information content is an important dimension in information-based similarity measures. In this article, we only concerned about information-based approaches. In what follows, we make a short discussion about it.

2.1. Information Content-based Similarity Measures

Information content-based similarity measure was first proposed by Resnik in 1995 following information theoretic approach [10]. It assumes that, for a concept c in the taxonomy, let p(c) be the probability of encountering and instance of concept c.

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

For two given concepts, similarity is depended on the extent to which they share information in common. The more information two concepts share in common, the more similar they are. In practice, it is indicated by the information content that subsumes them in the taxonomy.

$$sim_{\mathrm{Re}snik}(c_1, c_2) = \underset{c \in common(c_1, c_2)}{MAX} IC_{\mathrm{Re}snik}(c) \tag{2}$$

Where common (c_1, c_2) are the set of concepts that subsume c_1 and c_2.

Lin proposed another information content-based metric in 1998 [19]. He assumed that the similarity between concept c_1 and c_2 depended on not only their shared information, but also their self information, expressed by:

$$sim_{Lin}(c_1, c_2) = \frac{2 * sim_{\mathrm{Re}snik}(c_1, c_2)}{IC_{\mathrm{Re}snik}(c_1) + IC_{\mathrm{Re}snik}(c_2)} \tag{3}$$

Contrary to the above measures, Jiang proposed a measure from different perspective by calculating semantic distance to obtain semantic similarity in 1997 [20]. We get similarity by considering the opposite of the distance.

$$dis_{Jiang}(c_1, c_2) = (IC_{\mathrm{Re}snik}(c_1) + IC_{\mathrm{Re}snik}(c_2)) - 2sim_{\mathrm{Re}snik}(c_1, c_2) \tag{4}$$

2.2. IC Model

It can been seen from formula (2)(3)(4) that information content(IC) is crucial in semantic similarity computation. Some IC models have been proposed. Here, we review some related model that have been used for IC computing and highlights important related issues.

2.2.1. Resnik Model: As stated above in formula (1), in Resnik model, IC(c) can be quantified as negative the log likelihood, −log p(c). Probability of a concept was estimated as follows:

$$p(c) = \frac{freq(c)}{N} \tag{5}$$

Where N is the total number of nouns, and freq(c) is the frequency of instance of concept c occurring in the taxonomy. When computing freq(c), each noun or any of its taxonomical hyponyms that occurred in the given corpora was included.

$$Freq(c) = \sum_{w \in W(c)} count(w) \tag{6}$$

Where W(c) is the set of words subsumed by concept c.

It is noted that,

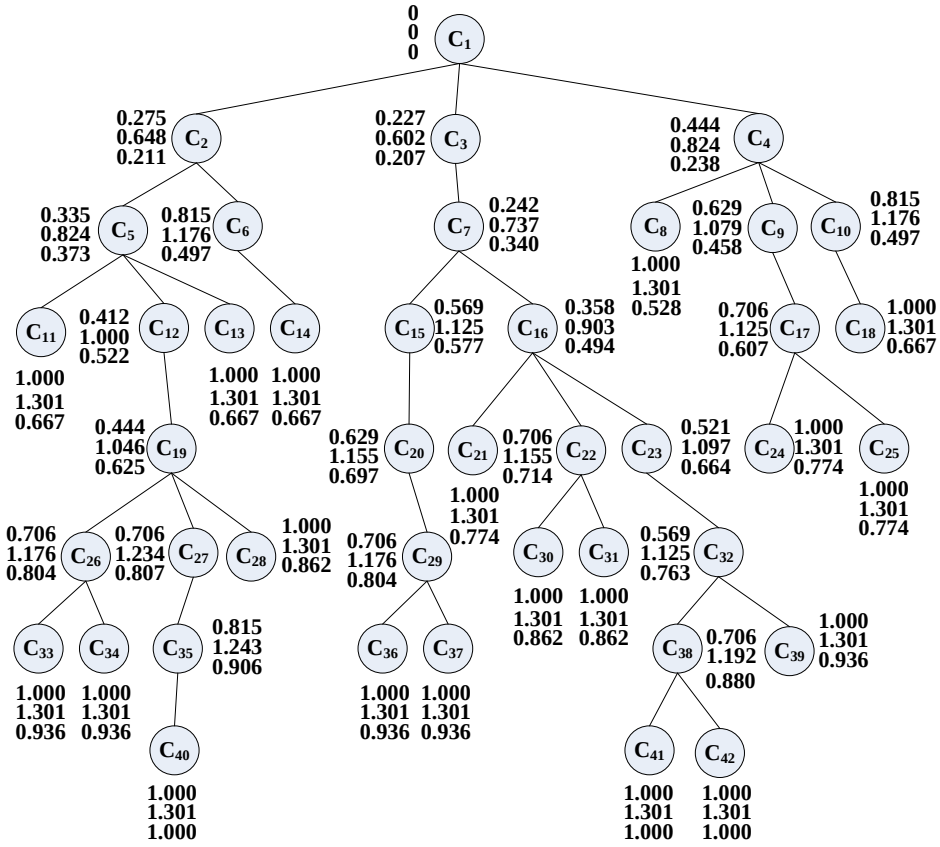
(1) IC is inversely proportional to p(c). When p(c) increases, IC decreases.

(2) It relies on corpora analysis, and sparse data problem is not avoided.

2.2.2. Nuno Model: Nuno model [11] resumed that in WordNet, IC value of a concept is regarded as the function of the hyponyms it has. A concept with more hyponyms expresses less information than the concepts with less ones. It is defined as:

$$IC(c) = \frac{\log(\frac{hypo(c)+1}{\max_{wn}})}{\log(\frac{1}{\max_{wn}})} = 1 - \frac{\log(hypo(c)+1)}{\log(\max_{wn})} \quad (7)$$

Where the function hypo(c) returns the number of hyponyms of a given concept c. And, max_{wn} represents a constant value which is set to the maximum number of concepts that exist in the taxonomy.



Notes:
Line 1: resulted with Nuno model
Line 2: resulted with david model
Line 3: resulted with our new model

Figure 2. IC Values Based on Different Models

From formula (7) and Figure 2, we find that:

(1) IC is inversely proportional to the number of hyponyms that a concept has, and range from 0 to1 (IC (root) =0, IC (leaf) =1).

(2) hypo(C_4)=hypo(C_{19})=7, therefore C_4 and C_{19} have the same IC value 0.444. However, C_{19} is deeper than C_4 in the taxonomy, and it should convey more information than C_4.

2.2.3. David Model: David model [12] based on the assumption that taxonomical leaves represent the semantic of the most specific concepts of a domain, and they are enough to describe and differentiate the concept from any other one, regardless of the amount of inner concepts incorporated in the taxonomy. Formally:

$$IC(c) = -\log(\frac{\frac{|leaves(c)|}{|subsumers(c)|}+1}{\max\_leaves+1}) \quad (8)$$

Let C be the set of concepts of the ontology, for a given concept c,

(1) leaves(c)={l∈ C|l∈ hyponyms(c) ∧ l is a leaf}

(2) max_leaves represents the number of leaves corresponding to the root node of the hierarchy, and subsumers(c) returns the set of subsumers.

(3) subsumers(c)={a∈ C | c≤a }∪ {c}, c≤a means that c is a hierarchical specialization of a.

From formula (8) and Figure 2, it is obviously that,

|subsumers (C_{15})| =|subsumers (C_{17})| =4, and |leaves (C_{15}) |=|leaves (C_{17})|=2. Consequently IC (C_{15}) =IC (C_{17}) =1.125, However, it is noted that hypo (C_{15}) is 4, and hypo (C_{17}) is 2. IC (C_{15}) and I C (C_{17}) should convey different information.

3. A New Model of Information Content Based on Concepts' Topology

Based on previous analysis in section 2, we can see that the IC model could not distinguish different concepts effectively. It is necessary to find a method that will exploit the accurate information of different concepts.

From Figure 2 it is noticed that each concept is unique in the taxonomy. Although two concepts have the same number of hyponyms or leaves, the hyponyms' arrangements are different. Take C_{26}, C_{27} as an example, the hyponyms of C_{26} and C_{27} are 2. The two hyponyms of C_{26} are arranged as siblings side by side. However the two hyponyms of C_{27} are arranged in a line. Therefore we assumed that information content(IC) of a concept is depended by the concepts' topology in the taxonomy. For a given concept c, IC is the function of itself and its hyponyms' arrangements, including concept's depth, the number of its hyponyms, and the depth of each hyponym. In accordance with previous research, the new method is needed to satisfy:

(1) A concept with more hyponyms expresses less information than the concepts with less ones.

(2) The deeper of concept, the more information it conveys.

Hence, the new model is presented, expressed by:

$$IC(c) = \frac{\log(deep(c))}{\log(deep\_\max)} * (1 - \frac{\log(\sum_{a \in hypo(c)} \frac{1}{deep(a)} + 1)}{\log(node\_\max)}) \quad (9)$$

For a given concept c, where deep(c) is the depth of concept c in the taxonomy, deep_max denotes the max depth of the taxonomy, a is a concept of the taxonomy, which satisfies a∈ hypo(c), node_max represents the maximum number of concepts that exist in the taxonomy.

If c is root, deep (root) is 1.Then log (deep(c)) =log (1) =0.

If c is a leaf, hypo(c) is 0. Then,

$$\sum_{a \in hypo(c)} \frac{1}{deep(a)} = 0$$

And,

$$IC(c) = \frac{\log(deep(c))}{\log(deep\_\max)})$$

That is to say, leaves with different depth will have different information content values. For a specific version of WordNet, deep_max is a fixed value. The deeper of a leaf, the more information it expresses.

What's more it can be seen that from formula (9) and Figure 2,

(1) IC (C_1) is 0 and IC (C_{40}) is1.IC values are range from 0 to 1.

(2) If two concepts with the same depth have the same number hyponyms, but their hyponyms are in different depth, they should have different IC values. For example, IC (C_{27}) is bigger than IC (C_{29}), because deep (C_{35}) is equal to deep (C_{36}), and deep (C_{40}) is bigger than deep (C_{37}).

(3) If two concepts with the same number of leaves, but they have different hyponyms, they should have different IC values. For example, IC (C_{17})> IC (C_{15}), even though both C_{17} and C_{15} have two leaves, but hypo (C_{15}) is 2 and hypo (C_{15}) is 4.

(4)Different from IC model in section 2, we think IC is the function of concepts' topology and take every concept's depth, hyponyms into considered. Table 1 presents the comparison among different IC models.

Table 1. Comparison Among Different IC Models

Factors	Whether to consider the factors and the result			
	Resnik model	Nuno model	David model	New model
corpora	Yes	No	No	No
depth(c)	No	No	Yes, depth(c) increase IC increase	Yes, depth(c) increase IC increase
hypo(c)	No	Yes, the number of hypo(c) increase, IC decrease	No	Yes, the number of hypo(c) increase, IC decrease
leaves(c)	No	No	Yes, leaves increase, IC decrease	No
concepts' topology	No	No	No	Yes, depth of hypo(c) increase, IC increase

Next, let's look at the distribution nature of each concept with different IC models, which are illustrated in Figure 3. Obviously, in Figure 3(3), the variance of each concept can be distinguished well. The difference among the concepts is well represented.

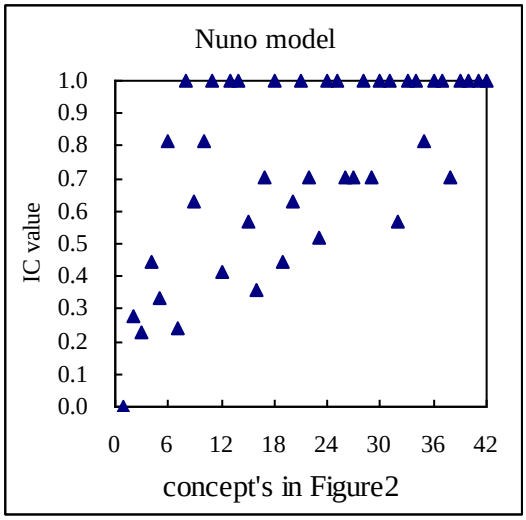


Figure 3(1). Distribution of Concepts with Nuno Model

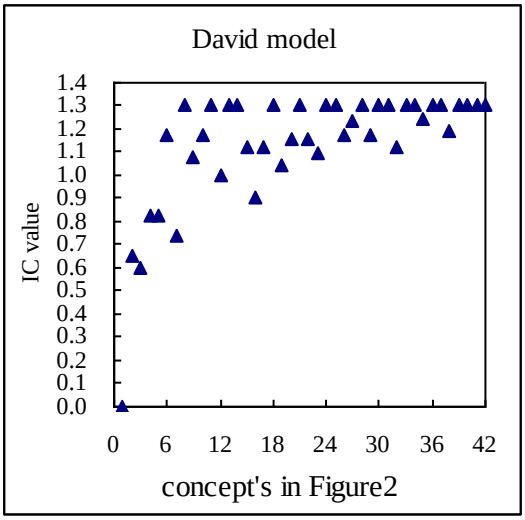


Figure 3(2). Distribution of Concepts with Nuno Model

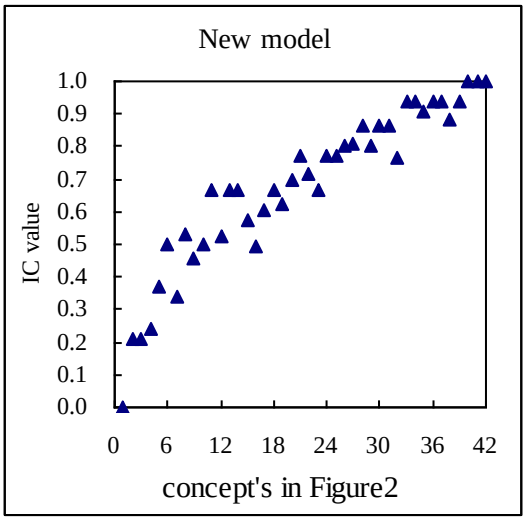


Figure 3(3). Distribution of Concepts with New Model

4. Evaluation

In order to evaluate the behavior of the proposed IC model and compare it with related work, we compare the measures by calculating the coefficients of correlation with human judgments. In this paper, we decide to use classical IC-based measures introduced in section 2.1 and substituted Resnik IC model with Nuno IC model, David IC model and the new IC model respectively.

4.1. Dataset

To evaluate the performance of our new model, a dataset is necessary. One commonly used dataset is provided by Rubenstein and Goodenough (1965) [21]. In Rubenstein and Goodenough's study, 51 undergraduate subjects were given 65 pairs of words, which ranged from "highly synonymous" to "semantically unrelated". The subjects were asked to rate them, on the scale of 0.0 to 4.0.

In order to make fair comparisons, an independent software package of Siddharth Patwardhan and Ted Pederson [22] was used. It is a freely available package allowing the use of WorldNet 3.0 which implements the semantic relatedness measures described by Leacock and Chodorow [15], Jiang and Conrath [19], Resnik [10], Lin [18] etc.

4.2. Words Similarity Calculating Method

Because either or both of the words have more than one sense in WordNet, we took the most similarity pair of sense:

$$sim(w_1, w_2) = \max_{(i,j)}[sim(c_{1i}, c_{2j})] \tag{12}$$

Where c_{1i} is the sense of word1, and c_{2j} is the sense of word2. For each of seven implemented measures, we compute similarity values for the human-rated pairs.

4.3. Results Analysis

Before our analysis, we first compute semantic similarity between pairs of words with the measures introduced in section 2.1 with different IC models, and draw the obtained similarity values in diagrams. For the convenience of expression and comparison, we normalized the values in [0,1]. Theses results are shown in Figure 4.

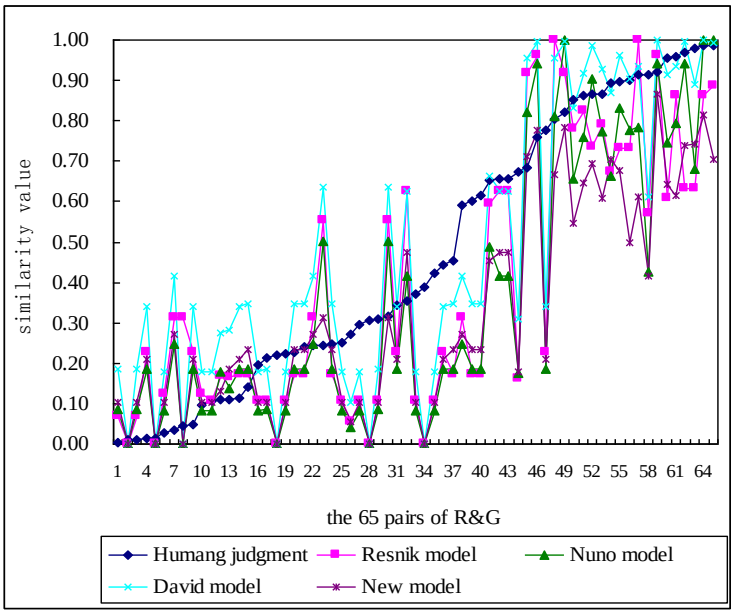


Figure 4(1) Similarity Different IC Models in Resnik's Measure

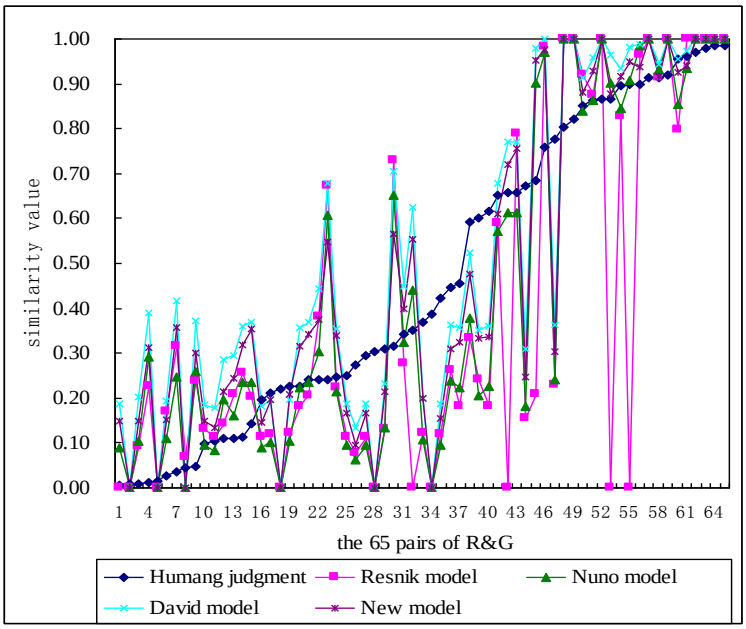


Figure 4(2) Similarity with Different IC Models in Lin's Measure

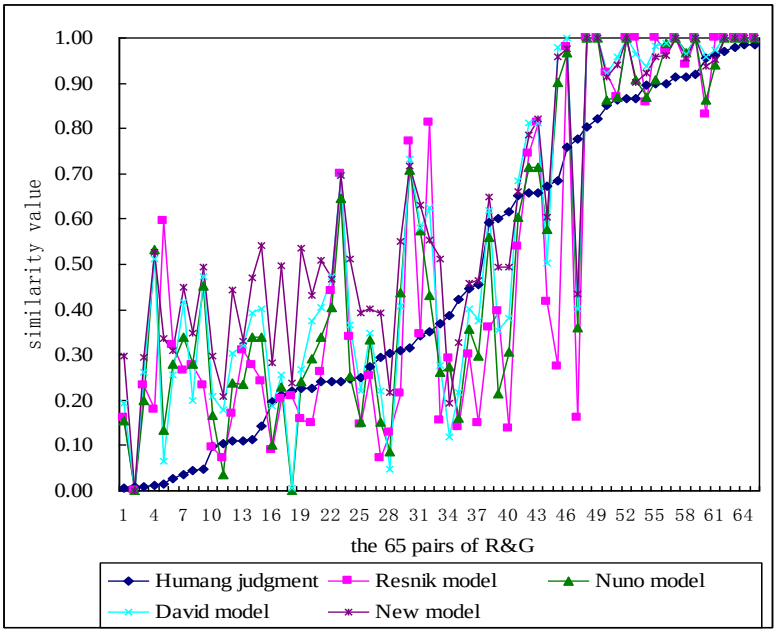


Figure 4(3) Similarity with Different IC Models in Jiang's Measure

In accordance with previous research we compare the chosen measures listed in Section 2.1 with different IC models by calculating the coefficients of correlation with human judgments. The results are presented in Table 2. The first column states the similarity algorithm used in obtaining similarity and the second column to the fifth column is the correlation between algorithm and human judgments.

Table 2. Coefficients of Correlation between Different Measures and Human Judgment

Similarity algorithm	Correlation Coefficient (γ)			
	Resnik model	Nuno model	David model	New model
Resnik	0.8073	0.8400	0.8451	0.8487
Lin	0.7356	0.8643	0.8555	0.8728
Jiang and Conrath	-0.8071	-0.8569	-0.8568	-0.8670

Looking at Table 2, whether in Resnik method, Lin method or Jiang method, coefficients of correlation between human ratings of similarity computed with our proposed model is higher than others. Figure 5 indicates the good performance of our method.

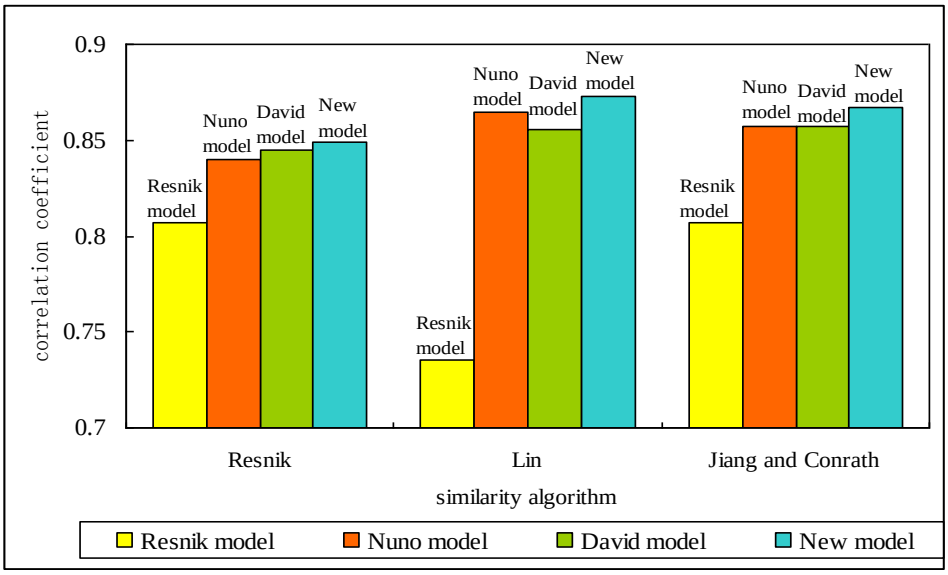


Figure 5. Compare our Proposed Model with Others

5. Discussion and Future Work

This paper presents a new model of information content on concept's topology for measuring semantic similarity in WordNet. Different form previous work, for a given concept, the space structure, including the depth itself, the number of its hyponyms, and the depth of every hyponym have been taken into considered. The results presented in Figure 3 shows that the variance of each concept has been effectively discriminated and get the most accurate value of IC in WordNet.

From Figure 4, we can see that the similarity values of most pairs with our new model are more close to human judgment than other models in Resnik's measure and Jiang's measure. In Lin's measure, the similarity values with our new model are more accurate than other measures all the time.

Further more Table 2 and Figure 5 show that the correlations coefficient values with our model is 0.8487, 0.8728 and 0.8670 in Resnik's measure, Lin's measure and Jiang's measure respectively, which is higher than the values with Resnik's model(0.8073, 0.7356, 0.8071), Nuno model(0.8400, 0.8643, 0.8569)and David model(0.8451, 0.8555, 0.8568). Therefore our model has achieved the best performance.

Besides this the major advantage of this approach is that it does not rely on corpora analysis, thus we can avoid the sparse data problem. In future work, we will put the model into application and plan to use the method in domain-oriented ontology construction and intelligent information retrieval.

Reference

[1] M. R. Quilian, "Semantic memory", In M. Minsky, editor, Semantic Information Processing, MIT Press, Cambridge, MA, **(1968)**.

[2] S. Patwardhan, S. Banerjee and T. Pedersen, "Using measures of semantic relatedness for word sense disambiguation", Proceedings of 4th International Conference on Computational Linguistics and Intelligent Text Processing and Computational Linguistics, **(2003)** February 16-22; Mexico City, Mexico.

[3] J. Atkinson, A. Ferreira and E. Aravena, "Discovering implicit intention-level knowledge from natural-language texts", Knowl.-Based Syst., vol. 22, no. 7, **(2009)**.

[4] M. Stevenson and M. A. Greenwood, "A semantic approach to IE pattern induction", Proceedings of 43rd Annual Meeting on Association for Computational Linguistics, **(2005)** June 25-30; Ann Arbor, Michigan, USA.
[5] D. Sánchez, D. Isern and M. Millán, "Content annotation for the Semantic Web: an automatic web-based approach", Knowl. Inf. Syst., vol. 27, no. 3, **(2011)**.
[6] A. G. Tapeh and M. Rahgozar, "A knowledge-based question answering system for B2C eCommerce", Knowl.-Based Syst., vol. 21, no. 8, **(2008)**.
[7] Y. Blanco-Fernández, J. J. Pazos-Arias, A. Gil-Solla, M. Ramos-Cabrer, M. López- Nores, J. García-Duque, A. Fernández-Vilas, R. P. Díaz-Redondo, J. Bermejo-Muñoz, "A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems", Knowl.-Based Syst., vol. 21, no. 4, **(2008)**.
[8] P. Gomes, N. Seco, F. C. Pereira, P. Paiva, P. Carreiro, J. L. Ferreira and C. Bento, "The importance of retrieval in creative design analogies", Procceedings of the International Joint Conference on Artificial Intelligence, **(2003)** August 9-15;Acapulco,Mexico.
[9] P. W. Lord, R. D. Stevens, A. Brass and C. A. Goble, "Semantic similarity measures as tools for exploring the gene ontology", Procceedings of the 8th Pacific Symposium on Biocomputing, **(2003)** January 3-7; Kauai, Hawaii.
[10] P. Resnik, "Using information content to evaluate semantic similarity", Proceedings of the 14th International Joint Conference on Artificial Intelligence, **(1995)** August 20-25; Montréal Québec, Canada.
[11] N. Seco, T. Veale and J. Hayes, "An intrinsic information content metric for semantic similarity in WordNet", Proceedings of the16th European Conference on Artificial Intelligence, **(2004)** August 22-27; Valencia, Spain.
[12] D. Sánchez, M. Batet and D. Isern, "Ontology-based information content computation", Knowl.-Based Syst., vol. 24, no. 2, **(2011)**.
[13] H. Seddiqui and M. Aono, "Metric of intrinsic information content for measuring semantic similarity in an ontology", Proceedings of 7th Asia-Pacific Conference on Conceptual Modeling, **(2010)** January 18-21; Brisbane, Australia.
[14] G. Varelas, E. Voutsakis and P. Raftopoulou, "Semantic similarity methods in wordNet and their application to information retrieval on the web", Proceedings of the 7th annual ACM international workshop on Web information and data management, **(2005)** October 31 - November 05; Bremen, Germany.
[15] C. Leacock and M. Chodorow, "Combining Local Context and WordNet Similarity for Word Sense Identification", WordNet: An Electronic Lexical Database, MIT Press, **(1998)**, pp. 265-283.
[16] Z. Wu and M. Palmer, "Verb semantics and lexical selection", Proceedings of 32nd annual Meeting of the Association for Computational Linguistics, **(1994)** June 27-30; Las Cruces, New Mexico.
[17] A. Tversky, "Features of similarity", Psycol. Rev., vol. 84, no. 4, **(1977)**.
[18] D. Lin, "An information-theoretic definition of similarity", Proceedings of the 15th International Conference on Machine Learning, **(1998)** July 24-27; Madison, Wisconsin, USA.
[19] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy", Proceedings of International Conference on Research in Computational Linguistics, **(1997)** August 22-24; Taipei, Taiwan.
[20] M. A. Rodríguez and M. J. Egenhofer, "Determining semantic similarity among entity classes from different ontologies", IEEE Trans. Knowl. Data. Eng., vol. 15, no. 2, **(2003)**.
[21] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy", Communications of the ACM, vol. 8, no. 10, **(1965)**.
[22] http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi?version=yes.
[23] C. Fellbaum, editor, "WordNet: An electronic lexical database", Language, Speech, and Communication, MIT Press, Cambridge, USA, **(1998)**.

Authors

Lingling Meng is a PhD Candidate of Computer Science and Technology Department and a teacher of Department of Educational Information Technology in East China Normal University. Her research interests include intelligent information retrieval, ontology construction and knowledge engineering.

Prof. Junzhong GU is Supervisor of PhD Candidates, full professor of East China Normal University, head of Institute of Computer Applications and director of Lab, Director of Multimedia Information Technology (MMIT). His research interests include information retrieval, knowledge engineering, context aware computing, and data mining.

Zili Zhou received PhD degree in 2009 from Computer Science and Technology Department in East China Normal University, majored in ontology and Knowledge engineering. Now he is an associate professor of College of Physics and Engineering in Qufu Normal University and committed to the construction and application of ontology.