Research on the Similarity between Nodes with Hypernymy/Hyponymy Relations based on IC and Taxonomical Structure

Xiaogang Zhang
College of Information Engineering, Tarim University, China
1273966091@qq.com

Lili Sun
Science and Technology Office, Tarim University, China
1316729148@qq.com

Abstract: *The similarity method has an important effect on some tasks of natural language processing, such as information retrieval, automatic translation and named entity recognition. Hypernymy/hyponymy relations are widespread in semantic webs and knowledge graphs, so computing the similarity of hypernymy/hyponymy is a key issue in the text processing field. All measures of both feature-based and IC-based methods have obvious deficiencies. The feature-based method estimated the similarity by the depth of the node, and the IC-based method computed the similarity by the position of the deepest common parent. The deficiency of the feature-based method and IC-based method is that they include one parameter, so the performance is slightly inaccurate and unstable. To address this deficiency, our paper proposed a hybrid method that computes the similarity of hypernymy/hyponymy by a hybrid parameter (dhype(lch)) that implies two parameters: depth of the node and position of the deepest common parent. Compared with several similarity methods, the proposed method achieved better performance in terms of accuracy rate, Pearson correlation coefficient and artificial fitting effect.*

Keywords: *Computing similarity, information content, ontology and knowledge graph, WordNet, hypernymy/hyponymy.*

Received November 16, 2020; accepted August 31, 2021
https://doi.org/10.34028/iajit/19/3/13

1. Introduction

Computing concept similarity is a basic issue in Natural Language Processing (NLP). An excellent similarity method can improve the retrieval rate in information retrieval, promote the translation effect in automatic translation and improve the entity disambiguation effect in Named Entity Recognition (NER). Scholars research similarity with different methods, such as based ontology and information content of WordNet [11], knowledge graph [29], no taxonomic, etc., [2, 14]. WordNet is a widespread semantic web that has been used to compute similarity. In WordNet, all vocabularies are represented by synonym sets. Each set indicates a vocabulary concept and expresses hypernymy/hyponymy relations, part/whole relations and synonym/antonym relations [23]. These semantic relations constitute a semantic network and provide a very good conceptual hierarchy structure [8]. Now, WordNet has been integrated into knowledge graphs DBpedia, YAGO, BabelNet, etc., and research work has been performed, such as machine translation, word discrimination, keyword retrieval, text mapping, information extraction, and entity recognition [3, 9].

Hypernymy/hyponymy relations are widespread in WordNet. An example of hypernymy/hyponymy is shown in Figure 1.

This paper focuses on the similarity estimation with hypernymy/hyponymy relations. In recent years, some scholars have proposed similarity methods, including IC-based methods and feature-based methods. The method of IC-based computed similarity examines the information content in word pairs. In this method, the similarity was determined by the position of the deepest common parent. The feature-based method estimates the similarity according to the structural features of the taxonomy. In this method, the similarity was determined by the depth of the node [28].

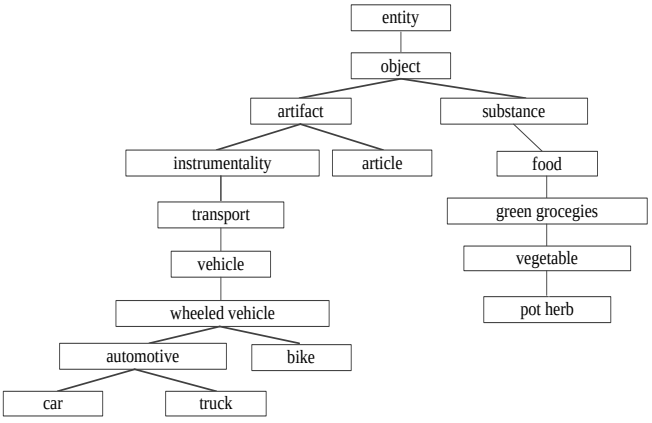

Figure 1. An example of hypernymy/hyponymy.

The advantage of the two methods is that the similarity value was only affected by the surrounding network structure. The deficiency only includes one parameter, and the performance is not stable. In this paper, we proposed a new hybrid method that uses the

parameter dhype(lch(c_1,c_2)) to calculate the similarity. The parameter dhype(lch(c_1,c_2)) includes two parameters dhype and lch, which imply the depth of the node and position of the deepest common parent. So our method has better performance of stability. The list of symbols in this paper is shown in Table 1.

Table 1. List of symbols in similarity computing.

c	The ontology concept node or knowledge graph named entity.
p(c)	The probability which node c appears in a given corpus.
IC(c)	The information content of node c.
hypo(c)	The count of child nodes belonging to c.
max_nodes	The maximum number of the nodes in the classification tree.
depth(c)	The depth of node c.
max_depth(c)	The maximum depth of the classification tree of including c.
len(c_1, c_2)	The shortest path distance between c_1 and c_2 (including itself).
lso(c_1, c_2)	The deepest common parent of c_1 and c_2.
subsumers(c)	The node number from the root to node c along the path of taxonomy.
hypo(lso(c_1, c_2))	The hyponym of the deepest common parent of node-pair c_1 and c_2.
depth(lso(c_1, c_2))	The depth of the deepest common parent of node-pair c_1 and c_2.

Organization of this paper shows as follows. In section 2, we introduce some research, which include some IC models and similarity measures. In section 3, we propose a new method for measuring the concept similarity, and design an experiment which compares the performance between the proposed method, representative methods and artificial data in the M&C dataset. In section 4, we evaluate the proposed method by the Pearson correlation coefficient. In section 5, we summarize this paper and make a plan for future works.

2. Related Works

2.1. Existed Similarity Methods

We review the main methods to compute the similarity of hypernymy/hyponymy relations. These methods have been divided into IC-based methods, distance-based methods, feature-based methods and hybrid methods [28]. The IC-based method computed the concept similarity by examining the information content contained in the word pairs [6]. The distance-based method calculates similarity by computing the distance between nodes (the number of edges linked to two nodes) and transforms the distance into a similarity value [16]. The feature-based method estimated the similarity according to the structural features of the taxonomy, which includes nodes and edges [21]. The hybrid method computed the similarity by merging the advantages of other methods [10].

2.1.1. The IC-based Method

The IC-based methods include Resnik's [18], Jiang and Conrath's [10] and Lin's [13] methods.

Resnik [18] thought that the similarity of a pair of concepts could be determined by the amount of sharing information. He computed the similarity through the information content. He adopted the parameter MSCA (the most specific common abstraction) to compute the similarity. The calculating equation is as follows:

$$sim_R(n_1,n_2) = -\log p\left(lso(n_1,n_2)\right) = IC\left(lso(n_1,n_2)\right) \quad (1)$$

Here, the function $lso(n_1,n_2)$ represents the MSCA of n_1 and n_2 in taxonomy.

Jiang and Conrath [10] computed the semantic distance through the IC sum of two concept nodes subtracting the IC of their MSCA. The distance can be calculated as follows:

$$dist_{JC}(n_1,n_2) = IC(n_1) + IC(n_2) - 2 \times IC\left(lso(n_1,n_2)\right) \quad (2)$$

After a linear transformation, Equation (2) became the following equation [22]:

$$sim_{J\&C}(n_1,n_2) = 1 - \left(\frac{IC(n_1) + IC(n_2) - 2 \times IC\left(lso(n_1,n_2)\right)}{2}\right) \quad (3)$$

Lin [13] believed that the similarity of two concepts could be computed by the ratio of shared information and total information. Lin proposed the equation as follows:

$$sim_{Lin}(n_1,n_2) = \frac{2 \times IC\left(lso(n_1,n_2)\right)}{IC(n_1) + IC(n_2)} \quad (4)$$

For the IC-based method, the most critical issues are how to exactly obtain the IC value of the concept and how to introduce IC into the similarity methods. Thus, computing the IC is the foundation of the IC-based method.

2.1.2. The Distance-Based Method

In addition to the IC-based method, the distance-based method is an important method. This type of method includes Rada's *et al*. [17], Wu and Palmer's [25] and Leacock and Chodorow's [12] methods.

Rada *et al*. [17] thought that the similarity between concepts could be calculated by the minimum path that linked them in the semantic web. Later, they proposed the following equation:

$$dis_{rad}(n_1,n_2) = \min_{\forall i} |path_i(n_1,n_2)| \quad (5)$$

Wu and Palmer's [25] method is another typical method based on the shortest path. They adopted the parameter depth and length to compute the similarity. The corresponding equation is as follows:

$$sim_{W\&P}(n_1,n_2) = \frac{2 \times depth(lso(n_1,n_2))}{len(n_1,n_2) + 2 \times depth(lso(n_1,n_2))} \quad (6)$$

In Equation (6), the function $len(n_1,n_2)$ represents the shortest path distance between n_1 and n_2, the function $lso(n_1,n_2)$ represents the deepest shared parent of n_1 and n_2, and the function $depth(n)$ represents the depth of node n.

The scholar Leacock and Chodorow [12] proposed a nonlinear method to calculate the similarity, which included two parameters: the number of nodes between two concept nodes (including themselves) and the

maximum depth of the classification tree. They proposed the following equation:

$$sim_{L\&C}(n_1,n_2)=-\log\frac{len(n_1,n_2)}{2\times max\_depth_{n\in WordNet}(n)} \quad (7)$$

For the method based on the path distance, in a fixed taxonomy, the path distance between two concepts was farther, and the semantic similarity was smaller [7].

2.1.3. Feature-Based Method

The feature-based method thought that the similarity of two concepts had been determined by the attribute number that two concept nodes shared. For the method based on attribute features, Tversky's [24] method is representative. The equation is as follows:

$$Sim(c_1,c_2)=\theta f(c_1\cap c_2)-\alpha f(c_1-c_2)-\beta f(c_2-c_1) \quad (8)$$

Where parameter $f(c_1\cap c_2)$ is the number of attributes shared by c_1 and c_2; parameter $f(c_1-c_2)$ is the number of attributes that c_1 includes but c_2 does not; parameter $f(c_2-c_1)$ is the number of attributes that c_2 includes but c_1 does not. Parameters "θ", "α" and "β" are all adjustment factors, and their values are determined by the specific task.

In addition to Tversky [24], the scholars Banerjee [5] and Patwardhan [15] proposed their similarity method by calculating attributes.

2.1.4. Hybrid Method

The fourth method is the hybrid method. Some researchers have proposed some methods by mixing distance, IC and attributes. For example, paper [30] uses the shortest path, depth of the nearest common parent node, and node density to compute the similarity.

2.2. Existed IC Models

IC models are usually divided into two categories according to different calculating objects: based on statistical information and based on the ontology taxonomical structure [1].

2.2.1. IC Model based on Statistical Information

This type of model computed the IC value by counting the occurrence frequency of a concept in a given corpus. Resnik [18] is the most representative researcher. He thought that the frequency of concept nodes could be estimated by the term frequency that appeared in Brown Corpus [4]. Resnik's model is as follows [18]:

$$IC(n)=-\log(p(n)) \quad (9)$$

Here, parameter n is a concept node, and function $p(n)$ denotes the probability that n appears in a given corpus. Each term that appeared in the corpus was counted as an occurrence rate of the concept node, which included the term. Function $p(n)$ could be computed as follows [18]:

$$p(n)=\frac{Freq(n)}{N} \quad (10)$$

Here, parameter N is the total number of terms that appeared. Function $Freq(n)$ is computed as follows [18]:

$$Freq(n)=\sum_{\omega\in Word(n)} Count(\omega) \quad (11)$$

Where function $Count(\omega)$ denotes the frequency of word ω appearing in the corpus, and function $Word(n)$ represents a word set subsumed by n.

Theoretically, the advantages of the IC model based on statistical information are high efficiency and suitability for large-scale data processing. Practically, finding a suitable corpus is difficult. Moreover, concepts are included in ontologies, and words are contained in the corpus. To calculate the occurrence ratio, researchers must disambiguate each term in the corpus. Therefore, this type of method is vulnerable to external interference.

2.2.2. IC Models based on Ontology Taxonomy

Unlike the IC model based on statistical information, the IC model based on ontology taxonomy is based on the ontology intrinsic structure. Therefore, this type of model was not affected by external interference, but this model required an organized ontology.

Seco *et al*. [22] were the first researchers to compute IC through an ontology hierarchical structure. They proposed the following equation [19]:

$$IC(n)=1-\frac{\log(|hypo(n)|+1)}{\log(max\_nodes)} \quad (12)$$

Here, function $hypo(n)$ represents the count of child nodes of node n, and parameter max_nodes is the maximum number of concepts in the classification tree. Equation (12) shows that the IC is related to the hierarchical structure, and the IC value of node n can be computed by the hyponym number of node n.

Sanchez *et al*. [20] thought that parameter *subsumer* was an important factor and proposed a new model that adopted the *subsumer* of the leaf node to calculate the IC. The equation is as follows:

$$IC_{David}(n)=-\log\left(\frac{commonness(n)}{commonness(root)}\right) \quad (13)$$

In Equation (13), the function $commonness\ (n)$ is equal to $\sum commonness\ (m)$, which is the commonness of node n. In actual calculation, function $commonness\ (m)$ is equal to $1/subsumers\ (m)$. Parameter m is a leaf node and one of the hyponyms of node n, and function $subsumers\ (m)$ returns the number of nodes from the *root* to node m along the path of taxonomy.

In summary, the method based on the ontology intrinsic structure is more stable than the statistical method because this method does not consider any external information.

3. Propose IC Model and Method

Through reviewing previous works, we note two points. First, the hybrid method, which merges the advantages of other methods, can obviously improve the similarity

performance. Second, the IC model based on the taxonomical structure is more stable than the statistical method.

3.1. Proposed IC Model

Rada *et al*. [17] thought that the length of the minimum path of two concepts could quantify their semantic distance. The equation is as follows:

$$dis_{rad}(n_1,n_2)=|min\_path(n_1,n_2)| \quad (14)$$

Here, parameters n_1 and n_2 represent the noun concept pairs or named entity pairs.

According to information theory, if the depth of the node is deeper, the information content is larger. Pirró and Euzenat [16] calculated the semantic distance of two concepts through the information content of each concept node subtracting their public parts of two information contents. They proposed the following equation:

$$\begin{aligned}&|min\_path(n_1,n_2)|=length(n_1,n_2)\\&\cong\left(IC(n_1)-IC(lso(n_1,n_2))\right)+\left(IC(n_2)-IC(lso(n_1,n_2))\right)\\&=IC(n_1)+IC(n_2)-2\times IC(lso(c_1,c_2))\end{aligned} \quad (15)$$

Where function *IC(lso(n1,n2))* is the IC of the deepest common parent of *n1* and *n2*. As stated above, the relative depth of the node is the minimum distance between the node and *root*, and *root* is the deepest common parent node between *root* and other nodes in the classification tree, i.e., *IC(lso(n, root))=IC(root)*. As *IC(root)≈0*, the depth of concept *n* can be approximated as follows [20]:

$$\begin{aligned}&depth(n)=min\_path(root,n)=length(root,n)\\&\cong IC(root)+IC(n)-2\times IC(lso(root,n))\\&=IC(n)-IC(root)=IC(n)\end{aligned} \quad (16)$$

Similarly, Equation (13) can be improved. Because function *commonness (root)* is equal to *1/subsumers(root)* and *subsumers(root)* is equal to 1 (root itself), the function *commonness (root)* is equal to 1. The improved IC equation is as follows:

$$IC_{new}(n)=-log\left(\frac{commonness(n)}{commonness(root)}\right)=log(subsumers(n)) \quad (17)$$

The IC value was affected by two factors in Equation (17): the depth of node *n* and the number from the *root* to node *n*. Both factors are intrinsic parameters, so this IC model does not interfere with external factors and theoretically achieves better stability.

3.2. Proposed a Similarity Method

Considering information theory and the taxonomy structure, the authors propose a new similarity method as follows:

$$Sim_{new}(n_1,n_2)=2\times log\frac{2\times log(dhype(n_{max\_depth}))}{log(dhype(n_1))+log(dhype(n_2))-2\times log(dhype(lch(n_1,n_2)))} \quad (18)$$

In Equation (18), function *dhype(n)* represents the direct hypernym of node *n*. Function *lch(n)* represents the most specific common abstraction of n_1 and n_2.

Theoretically, the proposed method has three advantages. First, this method does not interfere with external factors because it is based on the ontology taxonomy. Second, compared with other methods, the proposed method was affected by two factors (*dhype* and *lch*) and reduced the count of effect factors. Third, the proposed method reduced the computing difficulty by converting the minimum distance of n_1 and n_2 to a direct hypernym of the most specific common abstraction of n_1 and n_2.

3.3. Experiment and Results

To evaluate the improved IC model and proposed similarity method, we compared the performance of the proposed method and four typical similarity methods.

3.3.1. Data Source and Concept Selection

As a widespread benchmark, the M&C dataset included 30 word pairs determined by professionals. The range of similarity was from irrelevant to identity, according to the scoring [0.0-4.0] [27].

Considering that the M&C dataset only includes words and each word corresponds to a number of concepts in WordNet, this paper transforms the seeking concept into a seeking word. We assume that word w_1 has *m* concepts and w_2 has *n* concepts. When calculating the similarity of w_1 and w_2, we obtain $m\times n$ similarity values. We adopt the largest value of a concept as the word similarity. The equation is as follows:

$$sim(w_1,w_2)=\max_{(i,j)}[sim(c_{1i},c_{2j})] \quad (19)$$

Here, c_{1i} is one of the concepts of word w_1, and c_{2j} is one of the concepts of word w_2.

3.3.2. Experiment Results

In the experiment, we use an open test website that has been widely used to compute the similarity. This website includes Wu and Palmer's [25], Lin's, Leacock and Chodorow's [12], Jiang and Conrath's [10], Resnik's [18] and other algorithms, which can test the word similarity [27]. The similarity of these methods is shown in Table 2.

We adopt a scatter diagram to show the value distribution of Table 2. Using the word pairs and their similarity values as coordinates, in the M&C dataset, the scatter diagram of the similarity values on the artificial test and 5 methods are shown as follows.

Table 2. Similarity scores for different methods in the M&C dataset.

Word-pair	Artificial	W&P's	J&C's	Lin's	L&C's	proposal
autograph-shore	0.0600	0.3077	0.0000	0.0000	1.3863	0.4376
noon-string	0.0800	0.3529	0.0653	0.0923	1.2040	0.763
glass-magician	0.1100	0.5333	0.0604	0.1421	1.6094	0.7704
automobile-wizard	0.1100	0.4545	0.0738	0.1682	1.1239	0.986
mound-stove	0.1400	0.6667	0.0681	0.3143	1.7430	1.1952
coast-forest	0.4200	0.6154	0.0628	0.1181	1.8971	1.2112
boy-rooster	0.4400	0.5600	0.0727	0.2094	1.2040	1.3698
cushion-jewel	0.4500	0.6667	0.0694	0.2572	1.7430	1.522
coast-hill	0.8700	0.7143	0.2187	0.7286	2.0794	1.5244
boy-sage	0.9600	0.6667	0.0680	0.2057	1.8971	1.8944
mound-shore	0.9700	0.7143	0.1672	0.6724	2.0794	1.609
automobile-cushion	0.9700	0.6364	0.0894	0.3812	1.5404	1.4416
crane-rooster	1.4100	0.7586	0.0000	0.0000	1.6094	1.9846
hill-woodland	1.4800	0.6154	0.0592	0.1218	1.8971	1.2112
brother-lad	1.6600	0.7143	0.0830	0.2400	2.0794	2.0272
crane-implement	1.6800	0.7778	0.0784	0.3327	2.0794	1.9458
magician-oracle	1.8200	0.6250	0.0588	0.1828	1.7430	1.796
sage-wizard	2.4600	0.1667	0.0580	0.1809	1.8971	1.8944
oracle-sage	2.6100	0.7059	0.1083	0.5885	1.8971	2.0152
brother-monk	2.8200	0.9565	0.0689	0.2079	2.9957	1.6734
implement-tool	2.9500	0.9412	0.8484	0.9146	2.9957	3.1436
bird-crane	2.9700	0.8800	0.0000	0.0000	2.3026	2.6804
bird-cock	3.0500	0.9565	0.2681	0.7881	2.9957	3.5136
hill-mound	3.2900	1.0000	0.4931	1.0000	3.6889	2.3848
cord-string	3.4100	0.9412	0.6553	0.9188	2.9957	2.1152
midday-noon	3.4200	1.0000	3.5685	1.0000	3.6889	2.8014
glass-tumbler	3.4500	0.5882	0.0626	0.1858	1.6094	2.2612
serf-slave	3.4600	0.8000	0.0000	0.0000	2.3026	1.9552
cemetery-graveyard	3.8800	1.0000	1.0000	1.0000	3.6889	2.679
magician-wizard	3.5000	1.0000	0.0640	1.0000	3.6889	2.222
range	3.8200	0.6923	1.0000	1.0000	2.5650	1.5380

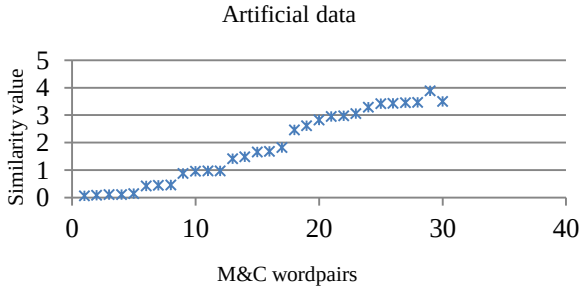

Figure 2. Distribution of the artificial test in M&C.

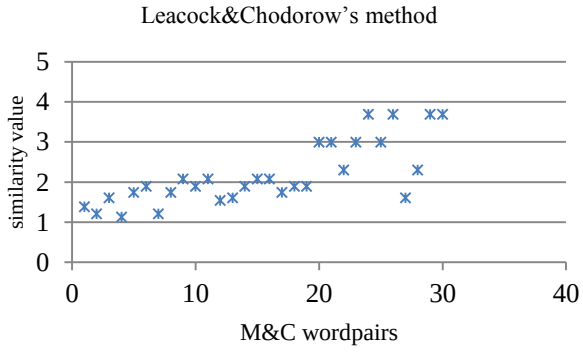

Figure 3. Distribution of Leacock and Chodonow's method in M&C.

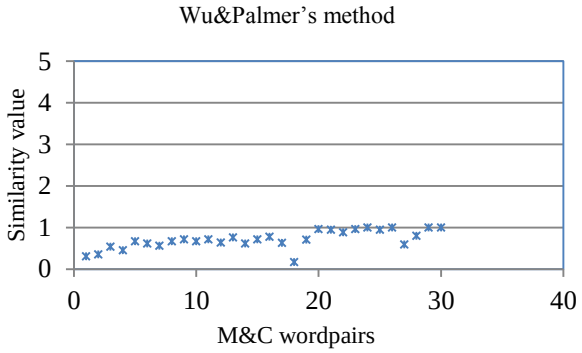

Figure 4. Distribution of Wu and Palmer's method in M&C.

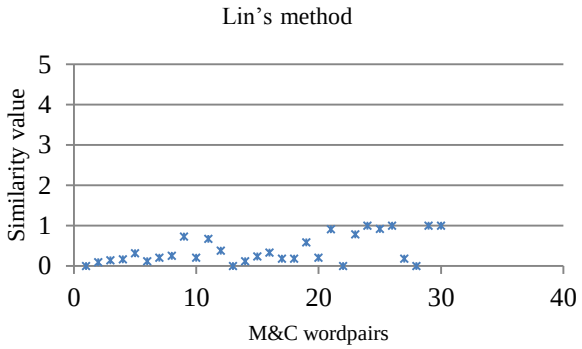

Figure 5. Distribution of Lin's method in M&C.

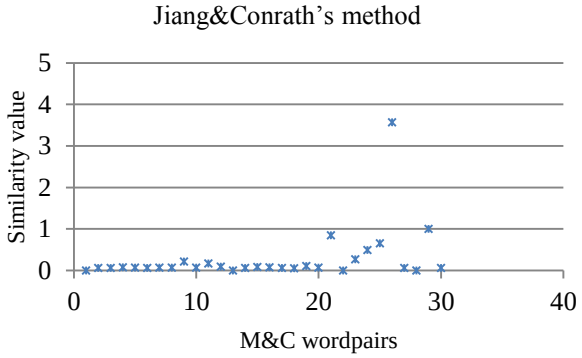

Figure 6. Distribution of Jiang and Conrath's method in M&C.

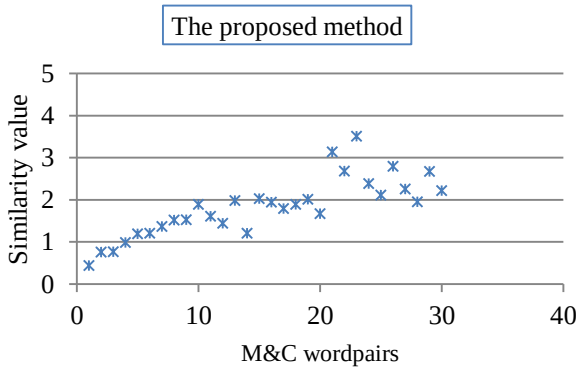

Figure 7. Distribution of the proposed method in M&C.

By comparing Figures 2-7, the scatter distribution of the similarity value of the proposed method is very continuous.

4. Result Analysis and Discussion

4.1. Result Analysis

4.1.1. Fitting Degree

The fitting degree of the proposed method and artificial data is shown in Figure 8:

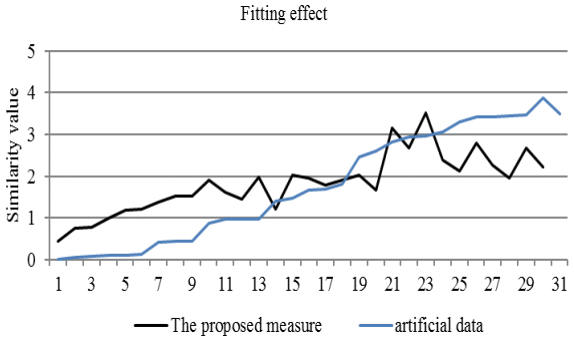

Figure 8. Fitting effect of the proposed method and artificial test in M&C.

Figure 8 shows that the fitting degrees of the proposed method and artificial test are very close, which proves that the proposed method is effective.

4.1.2. Method Evaluation

The Pearson correlation coefficient is an important evaluation metric to evaluate the computing effect of similarity [26]. The evaluating equation is as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2(y_i-\bar{y})^2}} \quad (20)$$

Here, X is the set $(x_1, x_2, \ldots x_n)$, Y is the set $(y_1, y_2, \ldots y_n)$. X is the similarity value computed by a method in the M&C dataset, Y is the similarity value derived from artificial data in the M&C dataset, and Y is the benchmark. x_i represents each term of set X, and $(x_i - \bar{x})$ represents the subtraction between x_i and the mean of x_i. Similarly, y_i represents each term of set Y, and $(y_i - \bar{y})$ represents the subtraction between y_i and the mean of y_i. r_{xy} is the correlation coefficient, and the range is [1, -1].

Using Equation (20), the five representative methods and the proposed method were evaluated in the M&C dataset, and the results are shown in Table 3 (two-sided 0.05 level Pearson correlation):

Table 3. Comparison of the correlation coefficient between representative methods and the proposed method in M&C.

Method	Pearson correlation coefficient score
Artificial Data	1
Wu and Palmer's [25] Method	0.678
Lin's [13] Method	0.543
Jiang and Conrath's [10] Method	0.389

In Table 3, using the artificial data as the benchmark, the correlation coefficient represents the correlation score of each method against artificial data. The correlation score represents the correlation performance of the similarity method. Apparently, the proposed method achieved better performance.

4.2. Discussion

Compared to previous work, five aspects should be addressed in this paper.

First, this paper proposed a new IC model based on the taxonomy structure and information theory. Based on the new IC model, this paper proposed a new similarity method, which computed the concept similarity by parameters *dhype* and *lch*.

Second, Table 2 shows that the proposed method has better performance than other methods in the M&C dataset. This performance is desirable because WordNet is a common ontology, and ontology-based IC models have better independence than the domain corpora [16].

Third, the parameter range is an important benchmark to evaluate the dispersion degree. In Table 2, the last row shows that parameter range of the proposed method reaches 1.5380, which implies that the proposed method has a better dispersion degree than Wu & Palmer's, Jiang and Conrath's and Lin's methods (in W&P's method, range is equal to 0.6923; in L&C's method, range is equal to 2.5652; in Lin's method, range is equal to 1.0000; in J&C's method, range is equal to 1.0000). The parameter range of Leacock and Chodorow's [12] method is lower because the smallest similarity is larger than that of the proposed method (in Leacock and Chodorow's [12] method, "noon-string" is equal to 1.2040; in the proposed method, "autograph-shore" is equal to 0.4376).

The fourth aspect is comparing the method complexity. All six methods compute the logarithms; in general, the complexity is equal. Thus, the main complexity is the difficulty of acquiring parameters. The parameters of the 5 methods are listed in Table 4:

Table 4. Comparison of the parameters of the 5 methods.

Method	Parameters
Wu and Palmer's [25] Method	len(n_1,n_2), lso(n_1,n_2), depth(n)
Lin's [13] Method	lso(n_1,n_2), hypo(n), max_nodes
Jiang and Conrath's [10] Method	lso(n_1,n_2), hypo(n), max_nodes
Leacock and Chodorow's [12] Method	len(n_1,n_2), max_depth(n)
The proposed Method	lch(n_1,n_2), dhyp(n_1,n_2), max_depth

As shown in Table 4, compared with other methods, the proposed method includes three parameters, so the workload does not obviously increase.

Finally, the proposed method has a better fitting degree than the others. As shown in Table 3, based on the evaluation metrics Equation (20), the correlation coefficient of Wu and Palmer's [25] method is equal to 0.678; Leacock and Chodorow's [12] method is equal to 0.792; Lin's method is equal to 0.543; Jiang and Conrath's [10] method is equal to 0.389); the proposed method reached 0.823.

5. Conclusions and Future Work

This paper focuses on the similarity computation of nodes with hypernymous/hyponymous relations. The authors proposed a hybrid method based on information content and hierarchy taxonomy. To evaluate the proposed method, the authors compared the performance of the proposed method, artificial test and four typical methods in WordNet. The experimental results show that the proposed method achieved better performance in the standard dataset.

We summarize the main three contributions of this paper as follows. First, we propose an improved IC model based on information content. Second, we propose a new similarity method based on information content and hierarchy taxonomy. Third, we design an experiment that compares the performance between the proposed method, representative methods and artificial data in the M&C dataset.

Future research should further prove this method in widely used datasets and explore methods of correlation based on ontology and knowledge graphs.

Acknowledgements

The authors gratefully acknowledge the financial support of the National Natural Science Foundation of China (Grant: 61562072).

References

[1] Adhikari A., Singh S., Dutta A., and Dutta B., "A Novel Information Theoretic Approach for Finding Semantic Similarity i+n Word Net," *in Proceeding of TENCON IEEE Region 10 Conference*, Macau, pp.1-6, 2016.

[2] AlMousa M., Benlamri R., and Khoury R., "Exploiting Non-Taxonomic Relations for Measuring Semantic Similarity and Relatedness in WordNet," *Knowledge-Based Systems*, vol. 212, pp. 106565, 2021.

[3] Aouicha M. and Taieb M., "Computing semantic Similarity Between Biomedical Concepts Using New Information Content Approach," *Journal of Biomedical Informatics*, vol. 59, no. 1, pp. 258-275, 2016.

[4] Baker W., "Understanding English as a Lingua Franca-By B. Seidlhofer," *International Journal of Applied Linguistics*, vol. 22, no. 1, pp. 124-128, 2012.

[5] Banerjee S., "Extended Gloss Overlaps As A Method of Semantic Relatedness," *in Proceeding of International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers Inc, pp. 805-810, 2003.

[6] Cai Y., Zhang Q., Lu W., and Che X., "A Hybrid Approach for Measuring Semantic Similarity Based on IC-Weighted Path distance in Word Net," *Journal of intelligent information systems*, vol. 51, no. 1, pp. 23-47, 2018.

[7] Cai Y., Pan S., Wang X., Chen H., Cai X., and Zuo M., "Measuring Distance-Based Semantic Similarity Using Meronymy and Hyponymy Relations," *Neural Computing and Applications*, vol. 32, no. 8, pp. 3521-3534, 2020.

[8] Hengqi H, Juan Y., Xiao L., and YunJiang X., "Review on Knowledge Graphs," *Computer Systems and Applications*, vol. 28, no. 6, pp. 1-12, 2019.

[9] Hussain M., Wasti S., Huang G., Wei L., Jiang Y., and Tang Y., "An Approach for Measuring Semantic Similarity Between Wikipedia Concepts Using Multiple Inheritances," *Information Processing and Management*, vol. 57, no. 3, pp. 102188, 2020.

[10] Jiang J. and Conrath D., "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," *in Proceedings of International Conference Research on Computational Linguistics*, Taipei, pp. 19-33, 1997.

[11] Lastra-Díaz J., Goikoetxea J., Taieb M., Serrano A., Aouicha M., and Agirre E., "Reproducibility Dataset for A Large Experimental Survey on Word Embedding and Ontology-Based Methods for Word Similarity" *Data in Brief*, vol. 26, pp. 104432, 2019.

[12] Leacock C. and Chodorow M., "Combining Local Context and Word Net Similarity for Word Sense Identification," *Word Net: An electronic Lexical Database*, vol. 49, no. 2, pp. 265-283, 1998.

[13] Lin D., "An Information-Theoretic Definition of Similarity," *in Preceding of Fifteenth International Conference on Machine Learning*, San Francisco, pp. 296-304, 1998.

[14] Majumder G., Pakray P., and Avendano D., "Measuring Semantic Textual Similarity Using Modified Information Content of WordNet and Trigram Language Model," *International Journal of Computational Linguistics Research*, vol. 8, no. 4, pp. 171-177, 2017.

[15] Patwardhan S., "Incorporating Dictionary and Corpus Information into a Vector Method of Semantic Relatedness" M.S Thesis, University of Minnesota, 2003.

[16] Pirró G. and Euzenat J., "A Feature And Information Theoretic Framework For Semantic Similarity and Relatedness," *in Proceeding of The Semantic Web–ISWC*, Berlin, pp. 615-630, 2010.

[17] Rada R., Mili H., Bicknell E., and Blettner M., "Development and Application of a Metric in Semantic Nets," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 1, pp. 17-30, 1989.

[18] Resnik P., "Using Information Content to Evaluate Semantic Similarity in A Taxonomy," *in Proceedings of the 14th International Joint*

Conference on, Artificial Intelligence, Morgan Kaufmann Publishers Inc. pp. 448-453, 1995.
[19] Sánchez D., Ribalta A., Batet M., and Serratosa F., "Enabling Semantic Similarity Estimation across Multiple Ontologies: An Evaluation in The Biomedical Domain," *Journal of Biomedical Information*, vol. 25, no.1, pp. 141-155, 2012.
[20] Sánchez D., Batet M., and Isern D., "Ontology-based Information Content Computation," *Knowledge-Based Systems*, vol. 24, no. 2, pp. 297-303, 2011.
[21] Sathiya B. and Geetha T., "A Review on Semantic Similarity Measures for Ontology," *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 4, pp. 3045-3059, 2019.
[22] Seco N., Veale T., and Hayes J., "An Intrinsic Information Content Metric for Semantic Similarity in Word Net," *in Proceeding of European Conference on Artificial Intelligence, Ecai'2004, Including Prestigious Applicants of Intelligent Systems*, Paris, pp. 1089-1090, 2004.
[23] Taieb M., Aouicha M., and Hamadou A., "Computing Semantic Relatedness Using Wikipedia Features," *Knowledge Based Systems*, vol. 50, no. 9, pp. 260-278, 2013.
[24] Tversky A., "Features of Similarity," *Readings in Cognitive Science*, vol. 84, no. 4, 290-302, 1988.
[25] Wu Z. and Palmer M., "Verb Semantics and Lexical Selection" *in Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, pp. 133-138, 2012.
[26] Xiaoli M., Robert R., and Donald R., "Comparing Correlated Correlation Coefficients," *Psychological Bulletin*, vol. 111, no. 1, pp.172-175, 1992.
[27] Yanna W., Zili Z., and Yan H., "The Concept Semantic Similarity Estimation based IC in WordNet," *Computer Engineering*, vol. 37, no. 22, pp. 42-44, 2011.
[28] Zhang X., Sun S., and Zhang K., "A New Hybrid Improved Method for Measuring Concept Semantic Similarity in Word Net," *The International Arab Journal of Information Technology*, vol. 17, no. 4, pp. 433-439, 2020.
[29] Zhu G. and Iglesias C., "Exploiting Semantic Similarity for Named Entity Disambiguation in Knowledge Graphs," *Expert System with Application*, vol. 101, pp. 8-24, 2018.
[30] Zhu X., Li F., Chen H., and Peng Q., "An Efficient Path Computing Model for Measuring Semantic Similarity Using Edge and Density," *Knowledge and Information Systems*, vol. 55, no. 1, pp. 79-111, 2018.

Xiaogang Zhang Associate Professor, College of Information Engineering, Tarim University. Major research: Data Mining, Semantic Computation.

Lili Sun Associate Professor, Office of Academic Research, Tarim University. Major research: Computational linguistics, Data Mining.