A Review of Semantic Similarity Measures in WordNet[1]

Lingling Meng[1], Runqing Huang[2] and Junzhong Gu[3]

[1]*Computer Science and Technology Department, Department of Educational Information Technology, East China Normal University, Shanghai, 200062, China*
[2]*Shanghai Municipal People's Government, Shanghai, 200003, China*
[3]*Computer Science and Technology Department, East China Normal University, Shanghai, 200062, China*

llmeng@deit.ecnu.edu.cn, runqinghuang@gmail.com, jzgu@ica.stc.sh.cn

Abstract

Semantic similarity has attracted great concern for a long time in artificial intelligence, psychology and cognitive science. In recent years the measures based on WordNet have shown its talents and attracted great concern. Many measures have been proposed. The paper contains a review of the state of art measures, including path based measures, information based measures, feature based measures and hybrid measures. The features, performance, advantages, disadvantages and related issues of different measures are discussed. Finally the area of future research is described..

Keywords: *semantic similarity, path-based measures, information based measures, feature based measures, hybrid measures, WordNet*

1. Introduction

Semantic similarity measure is a central issue in artificial intelligence, psychology and cognitive science for many years. It has been widely used in natural language processing [1], information retrieval [2, 3], word sense disambiguation [4], text segmentation [5], question answering [6], recommender system [7], information extraction [8, 9] and so on. In recent years the measures based on WordNet have attracted great concern. They show their talents and make these applications more intelligent. Many semantic similarity measures have been proposed. On the whole, all the measures can be grouped into four classes: path length based measures, information content based measures, feature based measures, and hybrid measures. The paper discusses the features, performance, advantages and disadvantages of different measures and makes some suggestions in future research finally.

The remainder of this paper is as follows: WordNet is introduced in Section 2. Different Semantic similarity measures are presented in Section 3. Section 4 provides a comparison and evaluation of the state of art measures, analyzing the features, performance, advantages and disadvantages. A summary and future research is described in Section 5.

2. WordNet

WordNet is the product of a research project at Princeton University [10]. It is a large lexical database of English. In WordNet Nouns, verbs, adverbs and adjectives are organized by a variety of semantic relations into synonym sets (synsets), which represent one concept. Examples of semantic relations used by WordNet are synonymy, autonomy, hyponymy,

[1] The work in the paper is supported by Shanghai Scientific Development Foundation (Grant No. 11530700300) .

member, similar, domain and cause and so on. Some relations are used for word form relation and others for semantic relation. These relations will be associated with words and words to form a hierarchy structure, which makes it a useful tool for computational linguistics and natural language processing. It is commonly argued that language semantics are mostly captured by nouns or noun phrases so that most of the researches focus on noun in semantic similarity calculating. There are four commonly used semantic relations for nouns, which are hyponym/hypernym (is-a), part meronym/part holonym (part-of), member meronym/member holonym (member-of) and substance meronym/substance holonym (substance-of). For example, apple is a fruit (is-a) and keyboard is part of computer (part-of). Hyponym/hypernym (is-a) is the most common relation, which accounts for close to 80% of the relations. A fragment of is-a relation between concepts in WordNet is shown in Figure1. In the taxonomy the deeper concept are more specific and the upper concept are more abstract. Therefore vehicle is more abstract than bicycle and conveyance is more abstract than vehicle. Entity is the most abstract concept.

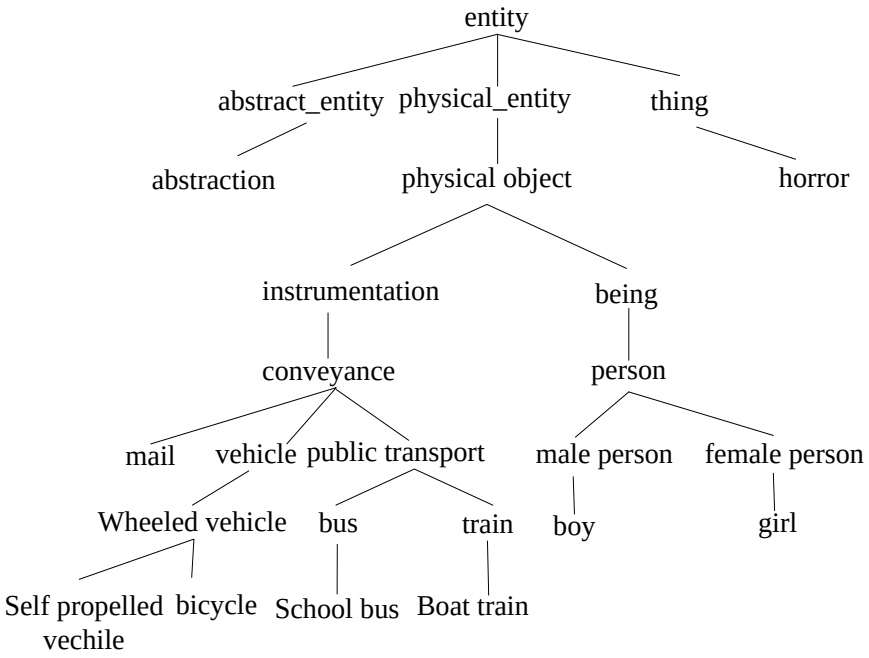

Figure 1. A Fragment of is-a Relation in WordNet

Generally the result obtained from hyponym/hypernym relation is regard as similarity between concepts. And the result obtained from others, such as part-of relation is regard as the relatedness between concepts. In this paper, we are only concerned about the similarity measure based on nouns and hyponym/hypernym relation of WordNet.

Before discussion, Definition of related concept in the following measures is necessary：

(1) len(c_i,c_j): the length of the shortest path from synset c_i to synset c_j in WordNet.

(2) lso(c_i,c_j): the lowest common subsumer of c_1 and c_2

(3)depth(c_i): the length of the path to synset c_i from the global root entity, and depth(root)=1.

(4) deep_max: the max depth(c_i) of the taxonomy

(5) hypo(c): the number of hyponyms for a given concept c.

(6) node_max: the maximum number of concepts that exist in the taxonomy.

(7) sim (c_i,c_j): semantic similarity between concept c_i and concept c_j.

For two compared concepts c_i and c_j in taxonomy as in Figure.1, the length of the shortest path from concept c_i to concept c_j can be determined from one of three cases.

Case1: c_i and c_j are the same concept, thus c_i, c_j and lso(c_i,c_j) are the same node. We assign the semantic length between c_i and c_j to 0, ie.len(c_i,c_j)=0.

Case2: c_i and c_j are not the same node, but c_i is the parent of c_j. thus lso(c_i,c_j) is c_i. We assign the semantic length between c_i and c_j to 1, ie.len(c_i,c_j)=1..

Case3: Neither c_i and c_j are the same concept nor c_i is the parent of c_j, we count the actual path length between c_i and c_j, therefore 1<len(c_i,c_j)<= 2*deep_max.

Based on the above definitions and cases, we discussed the following measures.

3. Semantic Similarity Measures based on WordNet

Semantic similarity measures might be used for performing tasks such as term disambiguation [4], as well as text segmentation [5], and for checking ontologies for consistency or coherency. Many measures have been proposed. On the whole, all the measures can be grouped into four classes: path length based measures, information content based measures, feature based measures, and hybrid measures.

3.1. Path-based Measures

The main idea of path-based measures is that the similarity between two concepts is a function of the length of the path linking the concepts and the position of the concepts in the taxonomy.

3.1.1 The Shortest Path based Measure: The measure only takes len(c_1,c_2) into considerate. It assumes that the sim (c_1, c_2) depend on how close of the two concepts are in the taxonomy. In fact this measure is a variant on the distance method [3, 11]. It is based on two observations. One is that the behavior of conceptual distance resembles that of a metric. The other is that the conceptual distance between two nodes is proportional to the number of edges separating the two nodes in the hierarchy [28].

$$sim_{path}(c_1, c_2) = 2 * deep\_\max - len(c_1, c_2) \quad (1)$$

From formula (1) it is noted that,

(1) For a specific version of WordNet, deep_max is a fixed value. The similarity between two concepts (c_1, c_2) is the function of the shortest path len(c_1,c_2) from c_1 to c_2.

(2) If len(c_1,c_2) is 0, sim_{path}(c_1,c_2) gets the maximum value of 2*deep_max. If len(c_1,c_2) is 2*deep_max, sim_{path} (c_1,c_2) gets the minimum value of 0. Thus, the values of sim_{path} (c_1, c_2) are between 0 and 2*deep_max.

(3) len(mail, vehicle) = len(self-propelled vehicle, bicycle) = 2, therefore, sim_{path} (mail,vehicle) = sim_{path} (self-propelled vehicle, bicycle).

3.1.2 Wu & Palmer's Measure: Wu and Palmer introduced a scaled measure [12]. This similarity measure takes the position of concepts c_1 and c_2 in the taxonomy relatively to the position of the most specific common concept $lso(c_1,c_2)$ into account. It assumes that the similarity between two concepts is the function of path length and depth in path-based measures.

$$sim_{WP}(c_1, c_2) = \frac{2 * depth(lso(c_1, c_2))}{len(c_1, c_2) + 2 * depth(lso((c_1, c_2))} \quad (2)$$

From formula (2) it is noted that,

(1) The similarity between two concepts (c_1, c_2) is the function of their distance and the lowest common subsumer($lso(c_1,c_2)$).

(2) If the $lso(c_1,c_2)$ is root, $depth(lso(c_1,c_2))=1$,$sim_{WP}(c_1,c_2) >0$; if the two concepts have the same sense, the concept c_1, concept c_2 and $lso(c_1,c_2)$ are the same node. $len(c_1,c_2)=0$. $sim_{WP}(c_1,c_2) = 1$; otherwise $0<depth(lso(c_1,c_2))< deep\_max$, $0<len(c_1,c_2)< 2*deep\_max$, $0<sim_{WP}(c_1,c_2) < 1$. Thus, the values of $sim_{WP}(c_1,c_2)$ are in (0, 1].

(3) len(mail, bicycle) = len(wheeled vehicle, bus) = 4, and lso(mail, bicycle) = lso(wheeled vehicle, bus) = conveyance, therefore sim_{WP}(mail, vehicle)= sim_{WP} (self-propelled vehicle, bicycle).

3.1.3 Leakcock& Chodorow's Measure: Leakcock and Chodorow took the maximum depth of taxonomy into account and proposed the following measure [13]:

$$sim_{LC}(c_1, c_2) = -\log \frac{len(c_1, c_2)}{2 * deep\_\max} \quad (3)$$

From formula (3) it is noted that,

(1) As same as formula (1) the similarity between two concepts (c_1, c_2) is the function of the shortest path $len(c_1,c_2)$ from c_1 to c_2.

(2) When c_1 and c_2 have the same sense, $len(c_1,c_2) =0$. In practice, we add 1 to both $len(c_1,c_2)$ and 2*deep_max to avoid log (0). Thus the values of $sim_{LC}(c_1,c_2)$ are in (0, log(2*deep_max+1)]

(3) As mentioned in section 3.1.1, len(mail, vehicle) = len(self-propelled vehicle, bicycle) = 2, therefore, sim_{LC} (mail,vehicle) =sim_{LC} (self-propelled vehicle, bicycle).

3.1.4 Li's Measures: Li's measure is intuitively and empirically derived[1]. It is based on the assumption that information sources are infinite to some extend while humans compare word similarity with a finite interval between completely similar and nothing similar. Intuitively the transformation between an infinite interval to a finite one is non-linear [28]. Therefore the measure combines the shortest path and the depth of concepts in a non-linear function:

$$sim_{Li}(c_1, c_2) = e^{-\alpha * len(c1,c2)} \frac{e^{\beta * depth(lso(c1,c2))} - e^{-\beta * depth(lso(c1,c2))}}{e^{\beta * depth(lso(c1,c2))} + e^{-\beta * depth(lso(c1,c2))}} \quad (4)$$

From formula (4) it is noted that,

(1) Formula (4) will monotonically increasing with respect to $depth(lso(c_1,c_2))$ and decreasing with $len(c_1,c_2)$.

(2) If $len(c_1,c_2)=0$ and $depth(lso(c_1,c_2))\rightarrow deep\_max$, $sim_{Li}(c_1,c_2)\rightarrow 1$.

if $len((c_1,c_2)\rightarrow 2*deep\_max$ and $depth(lso(c_1,c_2))=1$, $sim_{Li}(c_1,c_2)\rightarrow 0$.

Thus the values of $sim_{Li}(c_1,c_2)$ are in $(0, 1)$.

(3) Parameter α and β need to be adapted manually for good performance. In the experiment in [14], $\alpha = 0.2$ and $\beta=0.6$ respectively.

The measures above mentioned are based only on the positions of the concepts in the taxonomy, assuming that links between concepts represent distances. All the paths have the same weight. However, from Figure1 it should be noted that the density of concepts throughout the taxonomy is not constant. As we know in the hierarchy the more general concepts correspond to a smaller set of nodes than the specific concepts. An example is the distance between mail and vehicle, their distance is 2, and the distance between self-propelled vehicle and bicycle is also 2. The two pairs will have the same similarity values, which is not reasonable. In fact self-propelled vehicle and bicycle is more similar than mail and vehicle.

3.2. Information Content-based Measure

It assumed that each concept includes much information in WordNet. Similarity measures are based on the Information content of each concept. The more common information two concepts share, the more similar the concepts are.

3.2.1 Resnik's Measure: In 1995 Resnik proposed information content-based similarity measure [14]. It assumes that for two given concepts, similarity is depended on the information content that subsumes them in the taxonomy.

$$sim_{Resnik}(c_1,c_2) = -\log p(lso(c_1,c_2)) = IC(lso(c_1,c_2)) \quad (5)$$

From formula (5) it is noted that,

(1) The values only rely on concept pairs' lowest subsumer in the taxonomy.

(2) lso(mail, vehicle)=lso(mail, bicycle)=conveyance, therefore sim_{Resnik}(mail, vehicle) = sim_{Resnik} (mail, bicycle)=IC(conveyance)

3.2.2 Lin's Measure: Lin proposed another method for similarity measure [15].

$$sim_{Lin}(c_1,c_2) = \frac{2 * IC(lso(c_1,c_2)))}{IC\ (c_1) + IC\ (c_2)} \quad (6)$$

It uses both the amount of information needed to state the commonality between the two concepts and the information needed to fully describe these terms.

From formula (6) it is noted that,

(1) The measure has taken the information content of compared concepts into account respectively. As $IC(lso(c_1,c_2)) <= IC(c_1)$ and $IC(lso(c_1,c_2)) <= IC(c_2)$, therefore the values of this measure vary between 1 and 0.

(2) lso(mail, bicycle)=lso(bicycle, school bus)=conveyance; if IC(mail)=IC(bicycle)= IC(school bus), then sim_{Lin}(mail, bicycle)=sim_{Lin} (school bus, bicycle).

3.2.3 Jiang's Measure: Jiang calculated semantic distance to obtain semantic similarity [16]. Semantic similarity is the opposite of the distance.

$$dis_{Jiang}(c_1, c_2) = (IC(c_1) + IC(c_2)) - 2IC(lso(c_1, c_2)) \quad (7)$$

From formula (7) it is noted that,

(1) The measure has taken the IC of compared concepts into account respectively.

(2) As mentioned in 3.2.2, lso(mail, bicycle)=lso(bicycle, school bus)=conveyance; if IC(mail)=IC(bicycle)= IC(school bus)=1, then sim_{Jiang} (mail, bicycle)=sim_{Jiang} (school bus, bicycle)

(3) The value is semantic distance between two concepts. Semantic similarity is the opposite of the semantic distance.

In Information content-based similarity measures, each of the measures in Section 3.2 attempts to exploit the information contained at best to evaluate the similarity between the pairs of concepts. Therefore how to obtain IC is crucial, which will affect the performance directly. Generally there are five methods. The First one is to obtain IC through statistical analysis of corpora [14], from where probabilities of concepts occurring are inferred. 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) can be quantified as negative the log likelihood, −log p(c), which means that as probability increases, IC decreases.

$$IC(c) = -\log p(c) \quad (8)$$

Probability of a concept was estimated as:

$$p(c) = \frac{freq(c)}{N} \quad . \quad (9)$$

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 is included, which implies that if c_1 is-a c_2, then $p(c_1) < p(c_2)$. Thus the more abstract the concept is, the higher its associated probability and the lower its information content.

$$Freq(c) = \sum_{w \in W(c)} count(w) \quad (10)$$

The measure is simple, unfortunately, it relies on corpora analysis, and sparse data problem is not avoided. In order to overcome this drawbac,, Nuno proposed a hyponyms-based IC obtained method. WordNet is used as a statistical resource to calculate IC values. It regards IC value of a concept as the function of the hyponyms it has [21].For a concept, the more hyponyms it has, the more abstract it is. That is to say concepts with many hyponyms convey less information than concepts that are leaves. Root node is the least informative and leaf nodes are the most informative in the taxonomy. IC(root) is 0 and IC(leaf) is 1.When we traverse from the leaf nodes to the root node, IC will decreases monotonically and range from 1 to 0. The method is simple and corpora independent. However two concepts with the same number of hyponyms will have the same IC values and all the leaves will have the same IC values too, although they all in different of depth in the taxonomy. eg.IC(mail)=IC(bicycle).

$$IC(c) = \frac{\log(\frac{hypo(c)+1}{node\_\max})}{\log(\frac{1}{node\_\max})} = 1 - \frac{\log(hypo(c)+1)}{\log(node\_\max)} \quad (11)$$

The third one is based on the assumption that taxonomical leaves represent the semantic of the most specific concepts of a domain in WordNet, the more leaves a concept has the less information it expresses [22].

$$IC(c) = -\log(\frac{\frac{|leaves(c)|}{|subsumers(c)|}+1}{\max\_leaves+1}) \quad (12)$$

Where, let C be the set of concepts of the ontology, for a given concept c, leaves(c)={l∈C|l∈hyponyms(c) ∧ l is a leaf}. Subsumers(c)={a∈C | c≤a }∪{c}, c≤a means that c is a hierarchical specialization of a. Max_leaves represents the number of leaves corresponding to the root node of the hierarchy. This method does not take the depth of leaves into account, thus concepts with the same number of leaves will have the same IC values. eg. IC(vehicle)=IC(wheeled vehicle).

The fourth assumes that every concept is defined with sufficient semantic embedding with the organization, property functions, property restrictions and other logical assertions [23]. The IC value is the function of relations and hyponyms. One weight factor is used to adapt the each part's contribution.

$$IC(c) = \rho \cdot IC_{rel}(c) + (1-\rho) \cdot IC_{Nuno}(c)) \quad (13)$$

$$IC_{rel}(c) = \frac{\log(rel(c)+1)}{\log(rel\_\max+1)} \quad (14)$$

$$\rho = \frac{\log(total\_rel+1)}{\log(rel\_\max) + \log(node\_\max)} \quad (15)$$

Where rel(c) denotes the number of relations of concept c. And rel_max represents the total number of relations.

The last one assumes that each concept is unique in the taxonomy and IC value is the function of concept's topology, which can distinguish different concepts effectively and get more accurate IC value. It was defined as [29]:

$$IC(c) = \frac{\log(depth(c))}{\log(deep\_\max)} * (1 - \frac{\log(\sum_{a \in hypo(c)} \frac{1}{depth(a)}+1)}{\log(node\_\max)}) \quad (16)$$

Where for a given concept c, a is a concept of the taxonomy, which satisfies a∈ hypo(c). If c is root, deep (root) is 1 and log (deep(c)) is 0. If c is a leaf, hypo(c) is 0. Then,

$$\sum_{a \in hypo(c)} \frac{1}{depth(a)} = 0 \quad (17)$$

And

$$IC(c) = \frac{\log(depth(c))}{\log(deep\_\max)}) \tag{18}$$

Because sparse data problem is not avoided in Corpora-dependent IC Metric, corpora-independent IC Metric is popular.

3.3. Feature-based Measure

Different from all the above presented measures, feature-based measure is independent on the taxonomy and the subsumers of the concepts, and attempts to exploit the properties of the ontology to obtain the similarity values. It is based on the assumption that each concept is described by a set of words indicating its properties or features, such as their definitions or "glosses" in WordNet. The more common characteristics two concepts have and the less non-common characteristics they have, the more similar the concepts are. One classical measure is Tversky's model, which argues that similarity is not symmetric. Features between a subclass and its superclass have a larger contribution to the similarity evaluation than those in the inverse direction. It is defined as [17]:

$$sim_{Tversky}(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + k|C_1 \setminus C_2| + (k-1)|C_2 / C_1|} \tag{19}$$

Where C_1, C_2 correspond to description sets of concept c_1 and c_2 respectively, k is adjustable and $k \in [0,1]$.

From formula (19) it is noted that,

(1) the values of $sim_{Tversky}(c_1,c_2)$ vary from 0 to 1.

(2) $sim_{Tversky}(c_1,c_2)$ increases with commonality and decreases with the difference between the two concepts.

3.4. Hybrid Measure

The hybrid measures combine the ideas above presented. In practice many measures not only combine the ideas above, but also combine the relations, such as is-a, part-of and so on. A typical method is proposed by Rodriguez. The similarity function includes three parts: synonyms sets, neighborhoods and features [18, 19]. The similarity value of the each part is assigned to a weight, and then summed together. As stated before, in the paper we only concern on the hybrid measures based on is-a relation. Taking information content based measures and path based measures as parameter is commonly used. Generally one or more weight factors which can be adapted manually, are used to adapt the each part's contribution Zhou has proposed a measure, expressed by [20]:

$$sim_{zhou}(c_1, c_2) = 1 - k(\frac{\log(len(c_1, c_2) + 1)}{\log(2 * (deep\_\max - 1))}) - (1-k) * ((IC(c_1) + IC(c_2) - 2 * IC(lso(c_1, c_2))/2) \tag{20}$$

From formula (20) it is noted that,

(1) both IC and path have been taken into considerate.

(2) parameter k needs to be adapted manually for good performance. If k=1, formula (20) is path-based; if k=0, formula (20) is IC-based measure. In the experiment in [21] k=0.5.

4. Comparison and Evaluation

Table1. Comparison of Different Semantic Similarity Measures

category	Principle	measure	features	advantages	disadvantages
Path based	function of path length linking the concepts and the position of the concepts in the taxonomy	Shortest path	count of edges between concepts	simple	two pairs with equal lengths of shortest path will have the same similarity
		W&P	path length to subsumer, scaled by subsumer path to root	simple	two pairs with the same lso and equal lengths of shortest path will have the same similarity
		L&C	count of edges between and log smoothing	simple	two pairs with equal lengths of shortest path will have the same similarity
		Li	non-linear function of the shortest path and depth of lso	simple	two pairs with the same lso and equal lengths of shortest path will have the same similarity
IC based	The more common information two concepts share, the more similar the concepts are.	Resnik	IC of lso	simple	two pairs with the same lso will have the same similarity
		Lin	IC of lso and the compared concepts	take the IC of compared concepts into considerate	two pairs with the same summation of $IC(c_1)$ and $IC(c_2)$ will have the same similarity
		Jiang	IC of lso and the compared concepts	take the IC of compared concepts into considerate	two pairs with the same summation of $IC(c_1)$ and $IC(c_2)$ will have the same similarity
Feature based	Concepts with more common features and less non-common features are more similar	Tversky	compare concepts' feature, such as their definitions or glosses	take concept's feature into considerate	Computational complexity. It can't works well when there is not a complete features set.
Hybrid method	combine multiple information sources	Zhou	combines IC and shortest path	well distinguished different concepts pairs	parameter to be settled, turning is required. If the parameter can't be turned well it may bring deviation.

Different semantic similarity measures have different characteristic. Path based measures take the path length linking the concepts and the position of the concepts into considerate. They use link or edge as parameter to refer to the relationships between concept nodes. Most measures are simple. But local density of pairs can't be reflected. IC based measures based on the assumption that the more common information two concepts share, the more similar the concepts are. The measures are effective. However they can't reflect structure information of concepts, such as the distance. Feature based measure assumes that concepts pair with more common features and less non-common features are more similar. However it can't work well when there is not a complete feature set. Hybrid method combines multiple information sources and can distinguish different concepts pairs. But one or more parameters are needed and turning is required, too.Table1 presents the comparison.

There is not a standard to evaluate computational measures of semantic similarity. Generally there are three kinds of methods.

The first one is a theoretical examination of a computational measure for those mathematical properties thought desirable, such as whether it is a metric whether its parameter-projections are smooth functions, and so on.

The second one is compare the measure by calculating the coefficients of correlation with human judgments [20, 21]. Although it is difficult to obtain a large set of reliable, subject independent judgments, it is a popular method. The similarity values of human judgments are deemed to be correct, which gives the best assessment of the performance of a measure. Two dataset are commonly used. One is provided by Rubenstein and Goodenough (1965) [24]. Rubenstein and Goodenough obtained "synonymy judgment" from 51 human subjects on 65 pairs of words ranged from "highly synonymous" to "semantically unrelated", and the subjects were asked to rate them, on the scale of 0.0 to 4.0. The other is provided by Miller and Charles. In their study 30 pairs were taken from the original 65 pairs [24], 10 from the high level(between 3 and 4), 10 from the intermediate level (between 1 and 3), and 10 from the low level (0 to 1) [25].

The third one is application-oriented evaluation [26, 27]. If some particular application system requires a measure of semantic similarity, we can compare the performance of different measures to find most effective one, while holding all other aspects of the system constant.

5. Summary

This paper reviews various state of art semantic similarity measures in WordNet based on is-a relation. Path based measures, information content based measures, feature based measures and hybrid measures are discussed. We analyses the principles, features, advantages and disadvantages of different measure. Further more, we present the commonly used IC metric in information content based measures. Finally we discuss how to evaluate the performance of a similarity measure. In fact there are no absolute good performance measures. Different measures will show different performance in different applications. In specific application, whether a measure will hold all other aspects of the system well is another factor. In addition WordNet is a common sense ontology. There are many other domain-oriented ontologies. How to effectively solve the heterogeneous problem, and apply the measures in cross-ontology is needed in further research.

References

[1] Y. Li, Z. A. Bandar and D. McLean, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", IEEE Transactions on Knowledge and Data Engineering, vol. 15, Issue 4, **(2003)** July-August, pp. 871 – 882.
[2] R. K. Srihari, Z. F. Zhang and A. B. Rao, "Intelligent indexing and semantic retrieval of multimodal documents", Information Retrieval, vol. 2, **(2000)**, pp. 245-275.
[3] R. Rada, H. Mili, E. Bicknell and M. Blettner, "Development and Application of a Metric on Semantic Nets", IEEE Transactions on Systems, Man and Cybernetics, vol. 19, Issue 1, **(1989)** January-February, pp. 17 - 30.
[4] 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.
[5] H. Kozima, "Computing Lexical Cohesion as a Tool for Text Analysis", doctoral thesis, Computer Science and Information Math., Graduate School of Electro-Comm., Univ. of Electro- Comm., **(1994)**.
[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 and 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] J. Atkinson, A. Ferreira and E. Aravena, "Discovering implicit intention-level knowledge from natural-language texts", Knowl.-Based Syst., vol. 22, no. 7, **(2009)**.
[9] 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.
[10] C. Fellbaum, ed., "WordNet: An electronic lexical database", Language, Speech, and Communication. MIT Press, Cambridge, USA, **(1998)**.
[11] H. Bulskov, R. Knappe and T. Andreasen, "On Measuring Similarity for Conceptual Querying", Proceedings of the 5th International Conference on Flexible Query Answering Systems, **(2002)** October 27-29, Copenhagen, Denmark.
[12] 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.
[13] 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.
[14] 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.
[15] D. Lin, "An information-theoretic definition of similarity", Proceedings of the 15th International Conference on Machine Learning, **(1998)** July 24-27; Madison, Wisconsin, USA.
[16] 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.
[17] A. Tversky, "Features of Similarity", Psycological Review, vol. 84, no. 4, **(1977)**.
[18] M. A. Rodriguez and M. J. Egenhofer, "Determining Semantic Similarity among Entity Classes from Different Ontologies", IEEE Trans. on Knowledge and Data Engineering, vol. 15, no. 2, **(2003)**.
[19] H. Dong, F. K. Hussain and E. Chang, "A Hybrid Concept Similarity Measure Model for Ontology Environment", Lecture Notes in Computer Science on the Move to Meaningful Internet Systems: OTM 2009 Workshops, vol. 5872, **(2009)**.
[20] Z. Zhou, Y. Wang and J. Gu, "New Model of Semantic Similarity Measuring in Wordnet", Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering, **(2008)** November 17-19, Xiamen, China.
[21] N. Seco, T. Veale and J. Hayes, "An intrinsic information content metric for semantic similarity in WordNet", Proceedings of the 16th European Conference on Artificial Intelligence, **(2004)** August 22-27; Valencia, Spain.
[22] D. Sánchez, M. Batet and D. Isern, "Ontology-based information content computation", Knowl.-Based Syst., vol. 24, no. 2, **(2011)**.
[23] Md. 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.
[24] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy", Communications of the ACM, vol. 8, no. 10, **(1965)**.

[25] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity", Language and Cognitive Process, vol. 6, no. 1, **(1991)**.
[26] E. Blanchard, P. Kuntz, M. Harzallah and H. Briand, "A Tree-Based Similarity for Evaluating Concept Proximities in an Ontology", In Proc. 10th conference of the International Fedederation of Classification Societies , Springer, **(2006)**, pp. 3-11.
[27] A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of lexical semantic relatedness", Computional Linguistics, vol. 32, no. 1, **(2006)**.
[28] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. M. Petrakis and E. E. Milios, "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.
[29] L. Meng, J. Gu and Z. Zhou, "A New Model of Information Content Based on Concept's Topology for measuring Semantic Similarity in WordNet", International Journal of Grid and Distributed Computing, vol. 5, no. 3, **(2012)** September, pp. 81-94.

Authors

Lingling Meng

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.

Runqing Huang

Runqing Huang has a PhD from Shanghai Jiao Tong University. He works in Shanghai Municipal People's Government, P. R. China. His present research interests include modeling strategic decisions, economic analysis, e-governance, electronic government and Logistics.

Junzhong Gu

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