International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 11 | Nov 2023 www.irjet.net p-ISSN: 2395-0072

Semantic similarity measurement- A theoretical study of various approaches

Asha Chandran T [1], Sajina K [2]

[1] *Lecturer In Computer Engineering, Department of Computer Engineering, Govt. Womens' Polytechnic College, Kayamkulam, Kerala, India*
[2] *Lecturer In Computer Engineering, Department of Computer Engineering, Govt. Polytechnic College Neyyattinkara, Kerala, India*

Abstract - *Semantic similarity has a wide variety of applications in Natural Language Processing [NLP]. Approaches in semantic similarity measurement falls into categories like dictionary-based, corpus-based, knowledge-based, semantic indexing based etc. This paper aims to study these approaches in detail. Comparison of short sentences or phrases is given more emphasis since they draw more attraction than paragraphs or large documents..*

Key Words: Semantic similarity, WordNet, Concept similarity, Concept nodes, Path length, Information Content, Least Common Subsumer(LCS)

1. INTRODUCTION

Semantic similarity between sentences is a measurement of the semantic relationship, that is, how much they resemble in their meanings. Semantic similarity has a wide variety of applications in areas such as text comparison, text compression, dictionaries and reverse dictionaries, plagiarism detection, ranking search results and many more. Accuracy of results of many of these applications depends heavily on the accuracy score of the semantic similarity score. It is observed that comparison of short sentences or phrases gain more interest than comparing paragraphs or even large documents.

In this article, we aim to present and compare various methods of comparing similarity between sentences. Sentences are not mere grouping of words; the particular ordering of words is also significant. While comparing sentences, methods of comparing the word order such as parse tree depths, bipartite graphs etc. are also considered.

1.1 WordNet

For comparing semantic similarity, researchers make use of many lexical databases. WordNet is one of the most popular databases widely used as a dictionary and a thesaurus. It is a general purpose ontology developed by Princeton University to model the lexical knowledge of English language. WordNet includes four Parts-of-Speech in English such as noun, verb, adjective and adverb. It compiles each of these lexemes into synonym sets or synsets. A synset is a specific meaning of a word. The semantic similarity measurement approaches that use WordNet take the synsets of words or phrases under consideration and compare the glossary in their synsets.

WordNet depicts various relationships predominantly hypernym-hyponym(Is-A relation), holonym-meronym (Part-of relation) etc. These hierarchies will have a more general node. Among these relationship hierarchies, the 'is-a' relationship covers almost 70 percent of the corpus structure.

2. Approaches for Similarity measurement

Various approaches have been developed over years to compute semantic similarity among concepts commonly expressed as short sentences. We look in detail into each of these approaches.

2.1 Structure-based Methods

These methods rely on the structure of the corpus like WordNet. To determine the semantic relatedness between two concepts, these methods take the path length between the concepts under consideration in the WordNet structure.

2.1.1 Shortest path method [1,2]

In this approach, the distance between two concepts in the corpus is measured. Remember that there may be many paths, via different parent nodes, between two concept nodes in the hierarchy. The path length based methods take each of these paths, compute the path length or edge count of each path and the shortest among them is selected as the distance between the two concepts. The smaller the distance, the more semantically similar the concepts will be.

The similarity between two concepts c1 and c2 is calculated as

Sim(c1,c2)=(2*max_depth) –(shortest_path_length(c1,c2))

where max_depth is the depth of the hierarchy.

2.1.2 Weighted edge method [3]

This method uses the sum of weights of the edges in the path between two concepts in the hierarchy. A link or edge is assigned a weight based on two factors

(i)Depth of the hierarchy

(ii)Density of the taxonomy at a level and strength of connotation between parent and child nodes

This method is essentially an enhanced version of the shortest path method.

2.1.3 Leacock- Chodrow method [4]

This is also another method that uses shortest path length between two concepts and the maximum depth of the hierarchy.

$$\text{SimLC}(c1, c2) = \frac{-\ln(shortest\_path\_length(c1, c2))}{2 * \max\_depth}$$

2.1.4 Hirst - St.Onge Method [5]

Most of the semantic relatedness is applicable to '*is-a'* relations for nouns. HSO method is not limited to nouns, but applicable to other relations also. This method classifies concepts into upward, downward and horizontal relations. The upward movement indicates a more general 'is-a' relation whereas the downward movement indicates more specific concepts. Horizontal relations share the similar amount of specificity.

In this approach, number of deviations in the path connecting the two concepts is taken into account. If two concepts are similar, there will not be many deviations in the path connecting them.

The similarity between two concepts c1 and c2 is calculated in HSO method as

Sim_{HSO}(c1,c2)= C- path_length(c1,c2) –

(k* number_of_deviations)

where C and k are constants whose values are set using experiments.

2.1.5 Wu and Palmer [6]

This method calculates the path length of the concepts to be compared to their nearest common ancestor. The notion behind the idea is that the nearest common ancestor is the most specific common concept between the concepts under consideration. The common parent has the minimum number of edges in the *'is-a'* path with the concepts.

The similarity between two concepts c1 and c2, according to this method is,

$$Sim - wp = \frac{2 * \text{N}}{\text{N1} + \text{N2} + 2 * \text{N}}$$

where N is the depth of the nearest common parent from the root node of the hierarchy and N1 and N2 is the number if edges between nearest common parent with c1 and c2 respectively.

2.2 Information Based Methods

Structure-based measures to count semantic similarity have the major disadvantage that the concepts which are at the same semantic distance may not be equally similar. This happens because the information contained in each concept node is not equivalent. Some nodes may contain more generic information whereas some contain particularly specific knowledge. Structure-based methods mostly consider the edge count between concepts, depth of the hierarchy etc. and do not take into account the information contained in a concept node.

Information content indicates the specificity of a concept. A more specific concept contains a considerable amount of information about a topic. A general concept cannot provide much information.

For example, a concept 'motor car' gives much specific information than the concept 'conveyance' which is a more generic concept.

In this section, we compare various information based methods to compute semantic similarity.

2.2.1 Resnik method [7]

In this method, information content of a node is computed approximately by counting the frequency of that concept in a large corpus. The frequency is used to determine the probability of the concept via a maximum likelihood estimate. Negative log of this frequency is considered as a measure of the information content of the concept. Resnik did his experiments with the Brown Corpus.

Information content IC of a concept c is computed as

$$IC(c) = -\ln(P(c))$$

where P(c) is the probability of the concept c.

If the corpus has sense-tagged text, each concept will be associated with a unique sense. In such cases, the frequency of the concept can be easily available. For scenarios with sense-tagging is not present, Resnik suggested counting the number of occurrences of the concept and then dividing it by the number of different senses of that concept.

Semantic similarity of two concepts is proportional to the amount of information content they share. The shared information can be determined by the information content of the lowest common subsumer of these two concepts in the hierarchy.

Thus similarity between two concepts c1 and c2, according to Resnik method, is

$$Sim - Res = IC(LCS(c1, c2))$$

where LCS(c1,c2) is the least common subsume of the concepts c1 and c2 respectively.

2.2.2 Lin's method [8]

This method takes into account both the information content of the concepts under consideration as well as that of the least common subsumer. For this experiment, Lin used the sense-tagged version of SemCor.

Similarity according to Lin's method is

$$Sim - Lin = \frac{2 * IC(LCS)}{IC(c1) + IC(c2)}$$

2.2.3 Jiang-Conrath measure [9]

Lin's method takes the semantic similarity as a ratio of information content of the common parent and the sum of information contents of the concept nodes under consideration. Jiang's method also takes information content of concept nodes and that of the common parent. But, instead of taking the ratio, this method takes the difference of these two values. For checking the similarity, Jiang and Conrath also used the sense-tagged version of SemCor.

Semantic distance between two concepts c1 and c2, according to this method, is

$$Sem - Dis - Jian(c1, c2) = IC(c1) + IC(c2) - (2 * IC(LCS(c1, c2))$$

Accordingly, semantic similarity between c1 and c2 is

$$Sim - Jian(c1,c2) = 2 * (\ln(LCS(c1,c2)) - (\ln(IC(c1)) + \ln(IC(c2)))$$

2.3 Feature-Based Methods

Feature based methods are different from structure based and information based methods in the way that feature based methods do not take into account the taxonomy structure and information content of the nodes and their parents.

Feature based methods assume that each concept is associated with a set of features or properties. An example of feature is the set of definitions or glosses of the concept. Similarity between two concepts is computed as a function of their features. More common features that two concepts share indicate more similarity between them.

In general, we can say that these methods rely more on semantic properties than mere edge counting methods.

2.3.1 Tversky's method [10]

This method assumes that the semantic similarity is not symmetric. That is, the measure of similarity of a concept c1 to another concept c2 is not same as the similarity of c2 to c1. A classic example this model uses is the inheritance relationship. That is, it argues that, the similarity of a subclass to its superclass is more than the similarity of superclass to its subclass.

Semantic similarity of c1 to c2 in this method,

$$Sim - Tve(c1, c2) = \frac{|C1 \cap C2|}{|C1 \cap C2| + k|C1/C2| + (k-1)|C2/C1|}$$

where k is a constant whose value is between 0 and 1and is obtained by observation.

2.4 Hybrid Methods

Hybrid methods combine two or more of the above mentioned methods. The similarity is calculated by taking each method into account, assigning a weight to each method and computing the weighted sum of the methods. The weights can be assigned manually through experimental observations. Hybrid methods may also take into account the relationship between concepts such as 'is-a', 'part-of' etc.

2.4.1 Zhou method [11]

A hybrid method proposed by Zhou takes both the structure-based measure and the information based measure to compute a more accurate similarity measurement. A weight whose value varies between 0 and 1 is chosen manually to assign the contribution of each measure.

$$Sim - Zhou = 1 - k\left(\frac{\ln(shortest - path(c1,c2) + 1)}{\ln(2 * maxdepth - 1)}\right) -$$

$$(1-k) * ((IC(c1) + IC(c2) - 2 * IC(\frac{LCS(c1,c2)}{2})$$

If k=0, this method becomes purely information content based and if k=1, this becomes purely structure-based.

2.5 Comparison of various methods

Category	Method	Metric	Working principle
Structure-based	Shortest path method	Path length and depth of the hierarchy	Shortest path length is subtracted from twice the depth
	Weighted edge method	Weighted path length and depth of the hierarchy	Weight is assigned to each edge based on depth of the hierarchy and density of the taxonomy
	Leacock- Chodrow method	Path length and depth of the hierarchy	Logarithmically conditioned shortest path length is divided by twice the depth of the hierarchy
	Hirst - St.Onge Method	Path length and deviations in the path	Less number of deviations in the path is a measure of more similarity
	Wu and Palmer method	Depth of the hierarchy and distance to the nearest common parent	Distance from the nearest common parent is a measure of the similarity
Information - based	Resnik method	Frequency of occurrence in the corpus	Negative log of the frequency of occurrence is a measure of the information content of the concept node. Information content of the nearest common parent is a measure of the similarity
	Lin's method	Frequency of occurrence of the concepts and that of their nearest common parent	Information content of the nearest common parent divided by that of the concept nodes is a measure of the similarity
	Jiang's measure	Frequency of occurrence of the concepts and that of their nearest common parent	Information content of the nearest common parent subtracted from that of the concept nodes is a measure of the similarity
Feature-based	Tversky's method	compare concepts' feature, such as their definitions or glosses	Common characteristic features(e.g., glossary) of the concepts is a measure of similarity
Hybrid method	Zhou's method	Path length, information content and depth of the hierarchy	Combines structure-based and information-based methods in a certain proportion to find similarity measurement

3. SUMMARY

This paper aims to theoretically evaluate various semantic similarity measurement methods proposed by researchers from time to time. The methods we studied include structure-based methods, information-based methods, feature-based methods and hybrid methods which incorporate the idea of

structure-based and information-based approaches. The methods compared here are based on the popular general-purpose ontology called WordNet. Various other domain-specific ontologies are available. The comparison is aimed to help researchers to select appropriate measure for their requirements. However, an accurate similarity measurement is a problem yet to be researched further.

REFERENCES

[1]Rada, R., Mili, H., Bicknell, E., and Blettner, M. 1989. Development and Application of a Metric on Semantic Nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1):17-30, January/February

[2] 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

[3] .Richardson, R, Smeaton, A. & Murphy, J. 1994. Using WordNet as a Knowledge Base for Measuring Semantic Similarity Between Words. Technical Report Working paper CA-1294, School of Computer Applications, Dublin City University, Dublin, Ireland.

[4] 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

[5]. Hirst, G. and St-Onge, D. 1998. Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In Proceedings of Fellbaum, pages 305-332

[6]. 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.

[7] 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.

[8] D. Lin, "An information-theoretic definition of similarity", Proceedings of the 15th International Conference on Machine Learning, (1998) July 24-27; Madison, Wisconsin, USA.

[9] 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.

[10] A. Tversky, "Features of Similarity", Psychological Review, vol. 84, no. 4, (1977)

[11] Zhou, Z., Wang, Y. and Gu, J., 2008. "New Model of Semantic Similarity Measuring in Wordnet", Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering, November 17-19, Xiamen, China