International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 01 | Jan -2017 www.irjet.net p-ISSN: 2395-0072

SEMANTIC SIMILARITY BETWEEN SENTENCES

PANTULKAR SRAVANTHI[1], DR. B. SRINIVASU[2]

[1]M.tech Scholar Dept. of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Telangana- Hyderabad, India

[2]Associate Professor - Dept. of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Telangana- Hyderabad, India

---

Abstract - *The task of measuring sentence similarity is defined as determining how similar the meanings of two sentences are. Computing sentence similarity is not a trivial task, due to the variability of natural language - expressions. Measuring semantic similarity of sentences is closely related to semantic similarity between words. It makes a relationship between a word and the sentence through their meanings. The intention is to enhance the concepts of semantics over the syntactic measures that are able to categorize the pair of sentences effectively. Semantic similarity plays a vital role in Natural language processing, Informational Retrieval, Text Mining, Q & A systems, text-related research and application area.*

Traditional similarity measures are based on the syntactic features and other path based measures. In this project, we evaluated and tested three different semantic similarity approaches like cosine similarity, path based approach (wu – palmer and shortest path based), and feature based approach. Our proposed approaches exploits preprocessing of pair of sentences which identifies the bag of words and then applying the similarity measures like cosine similarity, path based similarity measures. In our approach the main contributions are comparison of existing similarity measures and feature based measure based on Wordnet. In feature based approach we perform the tagging and lemmatization and generates the similarity score based on the nouns and verbs. We evaluate our project output by comparing the existing measures based on different thresholds and comparison between three approaches. Finally we conclude that feature based measure generates better semantic score.

Key Words: WordNet, Path based similarity, Features based Similarity, Word Overlap, Cosine similarity, Word order similarity, Semantic similarity.

1. INTRODUCTION

Sentence similarity measures are becoming increasingly more important in text-related research and other application areas. Some dictionary-based measures to capture the semantic similarity between two sentences, which is heavily based on the WordNet semantic dictionary [1].Sentence similarity is one of the core elements of Natural Language Processing (NLP) tasks such as Recognizing Textual Entailment (RTE)[2] and Paraphrase Recognition[3]. Given two sentences, the task of measuring sentence similarity is defined as determining how similar the meaning of two sentences is. The higher the score, the more similar the meaning of the two sentences. WordNet and similarity measures play an important role in sentence level similarity than document level[4].

1.1 Problem Description

Determining the similarity between sentences is one of the crucial tasks in natural language processing (NLP). To estimate the accurate score generated from syntactic similarity to semantic similarity. Computing sentence similarity is not a trivial task, due to the variability of natural language expressions. Measuring semantic similarity of sentences is closely related to semantic similarity between words. In information retrieval, similarity measure is used to assign a ranking score between a query and texts in a corpus [5].

1.2 Basics and background knowledge

In the background we have defined the basic definitions and different strategies that can be used.

1.2.1 WordNet

WordNet is the product of a research project at Princeton University. 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 relations are synonymy, autonomy, hyponymy, member, similar, domain and cause and so on. In this paper, we are only concerned about the similarity measure based on nouns and synonym relation of WordNet.

1.2.2 Semantic Similarity

The semantic similarity sometimes called as topological similarity. Semantic similarity is calculated at document level, term level and sentence level. The document and sentence level is calculated based on the terms which

describe the internal concepts. The measures that have been used to measure sentence similarity fall into two categories: syntactical and lexical.

Syntactic approaches This approach is to detect semantic similarity mostly using syntactic dependency relations to construct a more comprehensive picture of the meaning of the compared texts, identifying whether a noun is considered the subject or the object of a verb.

Lexical similarity

Two main levels for lexical features have been established: explicit level (EL), and implicit level (IL).

1.2.3 Semantic Similarity Measures based on WordNet

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.

1.2.4 Other Related Measures

This section briefly describes some other techniques that ate related to our work. The three major categories of related methods: surface-matching methods, corpus-based methods and query-log methods.

1.2.5 Vector Space Model

We have a Vector Space Model of sentences modeled as vectors with respect to the terms and also have a formula to calculate the similarity between different pair of sentences in this space.

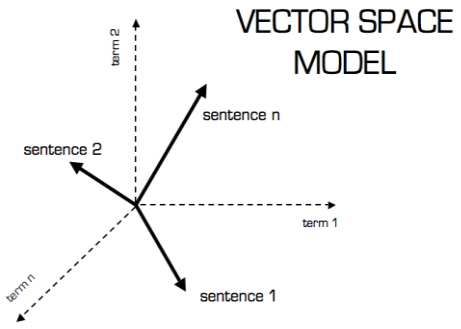

Fig 1.1 Vector space model

2. LITERATURE SURVEY

In the chapter, two different terms are used by different authors or sometimes interchangeably by the same authors to address the same concept: semantic relatedness and semantic similarity.

2.1 Classification of existing similarity measures

The classification is based on how the semantic similarity measure is quantified. The quantification is either based on the ontological structure (eg.,WordNet) or based on the information content[7].

Wu and Palmer Similarity Measure (Wu et al., 1998) Wu and Palmer suggested a new method on semantic representation of verbs and investigated the influence on lexical selection problems in machine translation. Wu and Palmer describe semantic similarity measure amongst concepts C1 and C2. Resnik Measure (1995) Similarity depends on the amount of information of two concepts have in common. Lin extended the Resnik(1995) method of the material content (Lin et al., 1998). He has defined three intuitions of similarity and the basic qualitative properties of similarity. Hybrid approach combines the knowledge derived from different sources of information. The major advantage of these approaches is if the knowledge of an information source is insufficient then it may be derived from the alternate information sources.

The feature based measure 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.

2.2 Proposed Approach

The proposed approach which we are going to develop is to measure the similarity between a pair of sentences. To compute the similarity we follow feature based approach which generates the similarity score in depth of word meaning level and definition level and then comparing the generated results with the previous existing measures for better results. Semantic distance/similarity values of pairs of sentences were calculated using the proposed measure. Therefore, in overall, the proposed measure performs very well and has great potential.

3. ARCHITECTURE

3.1 Architecture Description

This chapter describes the architecture about method we used for measuring sentence similarity based on semantic knowledge database such as WordNet [6].

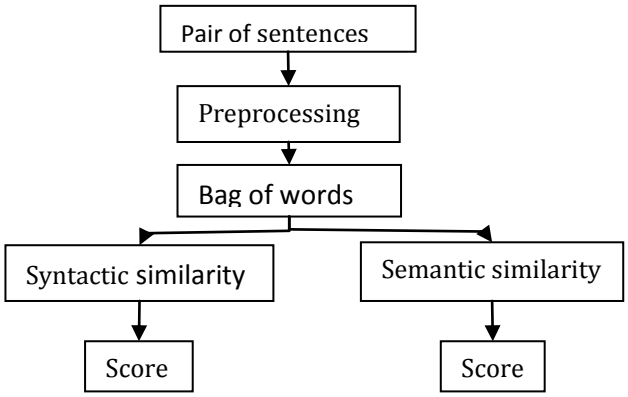

Figure 3.1: Architecture

3.2 Processing steps

3.2.1 Preprocessing

The goal of this phase is to reduce inflectional forms of words to a common base form. In this section the basic preprocessing techniques are discussed.

3.2.2 Tokenization

Tokenization is the task of chopping up sentences into tokens and throwing away punctuation and other unwanted characters. We use WordNet to find relationships between two tokens. The results of the search are the length of the shortest path between the two tokens and depth of the most specific common subsumer of the tokens. Both these values are wrapped in WordNetRelationship. Because the search in WordNet takes a significant time we developed a cache for WordNetRelationship between two tokens (TokenPair), which has speed up the process.

3.2.3 Tagging

Tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context. In our case we tagged the word to noun and verb.

3.2.4 Lemmatization

Lemmatization is a technique from Natural Language Processing which does full morphological analysis and identifies the base or dictionary form of a word, which is known as the lemma.

3.2.5 Syntax Similarity

Syntax similarity is a measure of the degree to which the word sets of two given sentences are similar. A similarity of 1 (or 100%) would mean a total overlap between vocabularies, whereas 0 means there are no common words.

3.2.6 Synsets extraction from wordnet

Synset is a set of synonyms that share a common meaning. Each synset contains one or more lemmas, which represent a specific sense of a specific word. Some relations are defined by wordnet only over lemma. The relations that are currently defined in this way are synonyms, antonyms, derivationally related forms.

3.2.7 Semantic Similarity

Similarity returns a score denoting how similar two word or sentence senses are, based on some measure that connects the senses in is-a taxonomy. The range for each measure is different.

4. IMPLEMENTATION

4.1 Preprocessing

The given pair of sentences or a document is taken as an input and performs the sentence tokenization to extract the sentences and stored in a comma separated value list with number of sentences.
The NLTK Sentence Tokenizer: This tokenizer divides the text into a list of sentences and then word tokenizer is used to divide the sentence into list of tokens. It processes the document and return the stopwords by removing the unuseful data from the document and the retrieved stopwords are used for further analyses.

4.2 Computing sentence similarity approaches

4.2.1 Syntactic similarity approach

Syntactical means structure of the words and phrases. The similarity of two sentences corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity. Cosine similarity is a measure to compute the given pair of sentences are related to each other and specify the score based on the words overlapped in the sentences.
Procedure to compute cosine similarity

To compute cosine similarity between two sentences s1 and s2, sentences are turned into terms/words, words are transformed in vectors as shown in the Table 1. Each word in texts defines a dimension in Euclidean space and the frequency of each word corresponds to the value in the dimension. Then, the cosine similarity is measured by using the word vectors as in below equation.

$$\text{Cos}(s1,s2) = s1 \,.\, s2 \,/\, ||s1||\,||s2|| \qquad \text{(Eq 4.1)}$$

Where $s1 \,.\, s2 = \sum^{n}_{i=1} s1_i \, s2_j$

4.2.2 Semantic similarity approach

Two sentences with different symbolic and structure information could convey the same or similar meaning. Semantic similarity of sentences is based on the meanings of the words and the syntax of sentence. Semantic similarity of sentences is based on the meanings of the words and the syntax of sentence. If two sentences are similar, structural relations between words may or may not be similar. Structural relations include relations between words and the distances between words. If the structures of two sentences are similar, they are more possible to convey similar meanings.

Firstly the given pair of sentences is process for classifying words into their parts of speech(A part-of-speech tagger, or POS-tagger, process a sequence of words, and attaches a part of speech tag to each word. . Parts of speech are also known as word classes or lexical categories.) and labeling them accordingly then these obtained words passed to lemmatizer for identifying the base form of a word known as lemma.
Lemma is used to generated synset from WordNet corpus.

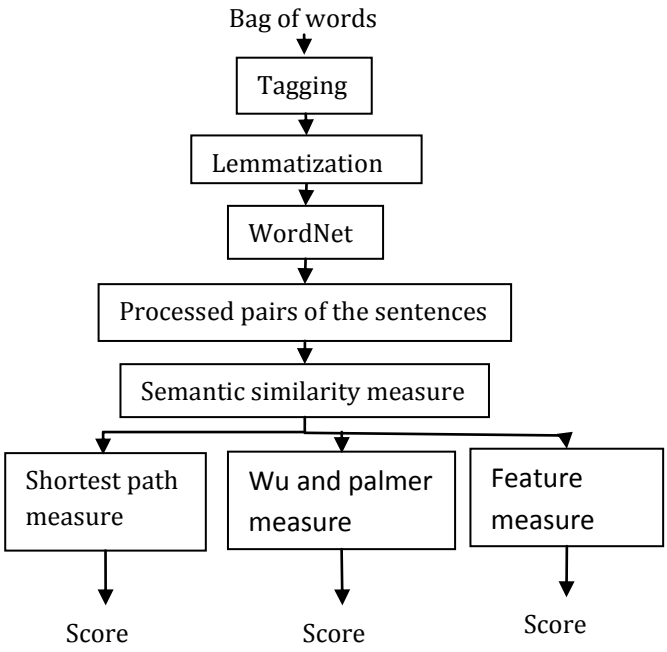

Figure 4.1: Flow of semantic similarity computation

4.2.3 Semantic similarity between words

In order to compare two texts, we must first assign a similarity to pairs of words. Most semantic similarity measures use WordNet [6] to determine similarity between words. We need to find semantic similarity (also relatedness) for two words denoted as rel(w1;w2)[8].

Wu and Palmer [9] defined the similarity measure as words position in the lexical hierarchical structure relative to the position of the most specific common subsumer.
$rel_{W\&P}$ (w1;w2) =2*depth(lcs) / [depth(w1) + depth(w2)].

4.2.4 Similarity score of words and sentences

Let us consider the given two sentences as an input to this process; first the words of two sentences are compared. If the two words of the sentences are matched, it's similarity score is calculated which are based on syntactic level. If the words of the two sentences are not matched, then synsets of the word is extracted from sentnce1and compared with the other word of the sentence2. If the words are matched at synset level then return the score as 1, otherwise return 0. Even the words are not matched, then consider the definition of the word sense of the sentences and compare the similarity score of the sentences which are totally based on semantics. This way we compute how two sentences are similar semantically.

4.2.5 Linguistic measures for similarity

Feature-based approaches assess similarity between concepts as a function of their properties. This is based on the Tversky's model [10] of similarity, which derived from the set theory, takes into account common and non common features of compared terms, subtracting the latter from the former ones. Path based measures are also considered to measure the similarity between sentences based on the concepts. 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.
Finally based on the words generated from the above measures, the syntactic score is calculated and generation of synset results the score of semantic. In this way, the sentence similarity is calculated. The procedural steps are defined in the below algorithm 1.

Algorithm 1 generation of semantic similarity score.
Sim(s1,s2) is computed
Input: pair of sentences
Output: similarity score
For every word of a sentence
Tagging(Nouns & Verbs)
Lemmatization
Synset
For each synset in wordSet
Compute similarity for three measure
//shortest path measure
Spath (s1,s2) = $\sum_{i=1}^{n}$ 2* deep_max –len(w1, w2).
//wu and palmer measure
score = $\sum_{i=1}^{n}$ **[**(2*depth(lcs))/(depth(s1) + depth(s2))**]**.
//feature based measure
Score = $\delta S_s + (1 - \delta)S_r$
Return score
End for
End for

5. EXPERIMENT RESULTS

We use dataset consisting of 1000 pair of sentences derived from the Microsoft research corpus. We ran our algorithm on this dataset. The various similarity score is computed and compared. As pointed out in the introduction, feature based measure give best semantic similarity score between pair of sentences.

Table -1: Sample pair of sentences

S.no	Pair of senetences
1	The problem likely will mean corrective changes before the shuttle fleet starts flying again. He said the problem needs to be corrected before the space shuttle fleet is cleared to fly again.
2	The technology-laced Nasdaq Composite Index .IXIC inched down 1

	point, or 0.11 percent, to 1,650. The broad Standard & Poor's 500 Index .SPX inched up 3 points, or 0.32 percent, to 970.
3	"It's a huge black eye," said publisher Arthur Ochs Sulzberger Jr., whose family has controlled the paper since 1896. "It's a huge black eye," Arthur Sulzberger, the newspaper's publisher, said of the scandal.
4	SEC Chairman William Donaldson said there is a "building confidence out there that the cop is on the beat." "I think there's a building confidence that the cop is on the beat."
5	Vivendi shares closed 1.9 percent at 15.80 euros in Paris after falling 3.6 percent on Monday. In New York, Vivendi shares were 1.4 percent down at $18.29.
6	Bremer said one initiative is to launch a US$70 million nationwide program in the next two weeks to clean up neighborhoods and build community projects. Bremer said he would launch a $70-million program in the next two weeks to clean up neighborhoods across Iraq and build community projects, but gave no details.

Table -2: comparative study of results by above analyzed and implemented measures

S.no	Cosine	Shortest path	Wu and palmer	Feature based
1	0.389	0.333	0.522	0.587
2	0.368	0.066	0.127	0.535
3	0.630	0.001	0.011	0.745
4	0.750	0.114	0.287	0.474
5	0.363	0.110	0.308	0.416
6	0.750	0.111	0.125	0.802

6. CONCLUSIONS

This project 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. Furthermore, we present the commonly used IC metric in Feature based measures. Each of these features covers an aspect of the text on implicit or explicit level. Finally we discuss how to evaluate the performance of a similarity measure. 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 common sense ontology. There are much other domain-oriented ontology. To compute the similarity we follow feature based approach which generates the similarity score in depth of word meaning level and definition level and then comparing the generated results with the previous existing measures for better results. Our proposed semantic similarity approach is better than using the syntactic similarity approaches. In fact, while two sentences are almost identical in terms of their lexical units a slight difference in *numbers*, *temporal constraints*, *quotations' content*, etc, can considerably shift the meaning of a text.

We propose an unsupervised approach to automatically calculate sentence levels similarities based on word level similarities, without using any external knowledge from other ontologies. Our proposed approach based on wordnet ontology which is restricted to domains. In fact, although having a system that annotates these phenomena has many merits in some specific tasks (engineered for a particular purpose), in general domains these phenomena do not occur as often as simple lexical units. We would exchange WordNet

for another knowledge base that has better coverage of words and part of speech classes.

ACKNOWLEDGEMENT

I express my deepest gratitude to my project guide Dr.B.Srinivasu, coordinator of M.Tech in the Computer Science and Engineering department.
A special feeling of gratitude to my wonderful parents, sisters and friends.

REFERENCES

1. Thanh Ngoc Dao, Troy Simpson, 2005: Measuring similarity between sentences.
2. AbdelRahman and Blake, 2012, S. AbdelRahman and C. Blake. Sbdlrhmn: A rule-based human interpretation system for semantic textual similarity task.
3. Agirre *et al.*, 2012: E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre. Semeval-2012 task.
4. Lingling Meng1, Runqing Huang2 and Junzhong Gu3, 2013: A Review of Semantic Similarity Measures in WordNet *International Journal of Hybrid Information Technology Vol. 6, No. 1, January, 2013.*
5. Pilsen, 15. Kvetna, 2012 : Advanced methods for sentence semantic similarity.
6. C. Fellbaum. WordNet: An Electronic Lexical Database. MIT Press, 1998.
7. Lingling Meng1, Runqing Huang2 and Junzhong Gu3, Vol. 6, No. 1, January, 2013:A Review of Semantic Similarity Measures in WordNet1.
8. W. K. Gad and M. S. Kamel. New Semantic Similarity Based Model for Text Clustering Using Extended Gloss Overlaps. In Proceedings of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition, MLDM '09. Springer-Verlag, 2009.
9. 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.
10. A. Tversky, “Features of Similarity”, Psycological Review, vol. 84, no. 4, (1977)

BIOGRAPHIES

Pantulkar Sravanthi received her B.E and M.Tech degree in Computer Science and Engineering from Osmania University.
Her research interests are Natural Language Processing, Data Mining.