Imperial Journal of Interdisciplinary Research (IJIR)
Vol-2, Issue-5, 2016
ISSN: 2454-1362, http://www.onlinejournal.in

Text Analytic Tools for Semantic Similarity

AbhishekKulkarni[1] ,Chinmay More[2], Mayur Kulkarni[3]& Vishal Bhandekar[4]

[1,2,3&4]Bachelor of Engineering, Department of Computer Engineering,Sinhgad College of Engineering,SavitribaiPhule Pune University,Pune, India.

Abstract:*Semantic similarity is a concept in which the relativeness between texts is based on the likeliness of their meaning. Systems which are based on machine translation use word to word translation at the basic level but they do not always work as intended. Translation can also be achieved using the concept of semantic similarity. A corpus along with its translated text is used. The aim is to achieve translation based upon relatedness between the sentences in the corpus rather than machine translation. This will be beneficial as there is no need to translate each and every text using machine translation and only match the semantics of the words within the sentences with sentences in the corpus to achieve the translation. By using modern means like transliteration, named entities can be translated and the remaining words of the sentences can be matched semantically with the words in corpus to achieve translation of the text.*

Keywords-*Normalizer, Semantic Similarity, Tokenizer,WordNet, Named Entity Recognizer (NER).*

1. INTRODUCTION

Semantic similarity is of great importance in the field of Natural Language Processing, Artificial Intelligence, cognitive science and psychology, both in academic community as well as industry. Accuracy of the system depends upon many factors like normalization, segmentation, etc. It can be used for many purposes like Information Retrieval, document clustering, synonym extraction, etc. The most popular way to compare two objects is the similarity between those two objects. For example, a banana is more similar to fruit rather than a monkey. Obtaining semantic similarity between words or concepts is a central concept in many applications.

In this paper we present an effective way of calculating semantic similarity between sentences along with the development of individual modules which can be used for semantic analysis.

Semantic similarity is the concept of therelatedness between words in terms of meaning. Various methods are available for calculating semantic similarity.

- Corpus based similarity.
- Ontological based similarity.
- WordNet Based similarity.

Corpus contains a predefined set of sentences and their translation to other language. The aim is to match input text with the text in the corpus and achieve translation. There are many cases when the input text and the text in the corpus will not be same but will be same in their meaning.

Some systems cannot translate certain texts correctly.For example, in Microsoft's Bing translator

Input (English): "sign in as sample@sample.com"

Output (Hindi):"संकेतकेरूपमें sample@sample.com है"

Expected output: "sample@sample.com केरूपमेंसाइनइनकरें"

This is incorrect translation as it recognizes and replaces the word directly with its meaning. However, if we have a corpus and we try to match with a particular text or in terms of meaning, then we can easily translate this text.Thus the project aims to develop such module which will work with the corpus and will provide individual tools to the users or developers for semantic analysis.The idea is to find similarity between various sentences.

2. LITERATURE SURVEY

[1] AinuraMadylova, S. G. Oguducu, "A Taxonomy based Semantic Similarity of Documents using the Cosine Measure", Istanbul Technical University, IEEE 2009.

In this paper, semantic similarity is calculated using the cosine similarity measure with the help of vectors which are generated from IS-A taxonomy. The results of proposed method outperform cosine similarity measure.

[2] Tian-Tian Zhu, Man Lan, "Measuring Short Text Semantic Similarity Using Multiple Measurements" International Conference on Machine Learning and Cybernetics, Tianjin, 14-17 July, 2013

In this paper, an improved hybrid semantic similarity algorithm is presented which combines semantic distance-based measure,information content-based measure and attribute-based measure. Here experiment compares similarity ratings of given algorithm with other algorithms on the same datasets and the result shows that this algorithm is better than others.

3. ARCHITECTURE DIAGRAM

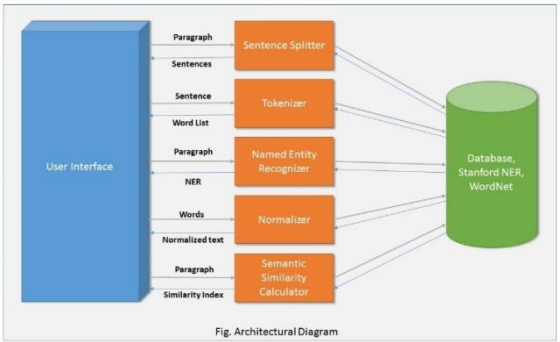

Figure 1. Architecture Diagram

The project is divided into five separate modules. Some modules are acting as a preprocessing modules and Semantic Similarity Calculator module is a main module which calculates the semantic similarity. The given modules are explained as follows:

3.1 Sentence segmentation

Often it may happen that the input text is not a single line or a sentence but can be huge text consisting of multiple sentences. In such case, Segmentation of the sentences needs to be done so that processing can be done on individual sentences. This module breaks the paragraph into sentences. Various cases are handled in this module as it takes care of honorifics, abbreviation, initials, period, exclamation mark and other special characters.

3.2 NER

NER are named entities which are or can be of many types. For example,Ram is a name of a person. Thus it comes under the named entity PERSON. India is a place and hence it comes under LOCATION entity.

The project makes use of Stanford NER for detecting NER in sentences. Stanford NER is of 3-class and 7-class. 3-class NER recognizes PERSON, LOCATION, ORGAIZATIONS and 7-class recognizes TIME, LOCATION, ORGANIZATION, PERSON, MONEY, PERCENT, and DATE. Combined output of class 3 and class 7 NER is used for better result.

3.3 Tokenizer

The main concept is to tokenize sentences into words so that they can be processed. Words like abbreviations should be considered as single token rather than different token.

For example: United States of America should be considered as single token and not separate tokens as United, States, of, America. Stop Words are also removed here.

3.4 Normalizer

The function of the normalizer is to obtain root forms of the words.

For example - root form of ‘Running’is‘Run’.

3.5 Semantic Similarity

The aim is to calculate semantic similarity between sentences. This can be further used for translation of text as the input text which matches with the corpus text can easily be translated.

4. METHOD

Following method is used to calculate the semantic similarity between the two sentences. Take two sentences as sentence1 and sentence2. Consider the length of sentence1as m and the length of sentence2as n. Following are the steps used to achieving Semantic Similarity between sentence1 and sentence2.

4.1Sentence Segmentation

If the input text may contain multiple sentences, then there is a need to segment sentences. Various methods can be applied to segment text like regular expression, NER, neural nets and decision tree. The segmentation tool makes use of various cases for handling sentence segmentation.

4.2 Tokenization

In tokenization, sentences are tokenizing into list of words. These words are considered as individual tokens. In this method, the Stop words are also removed. Stop words are the words which are having less importance and repeating frequently.E.g. for, is, to, an, that, etc. Proper care is taken so than accuracy remains unaffected without increasing recall.

4.3 Perform word stemming

In word stemming, words are brought to their root forms. This method improves the efficiency of the method by increasing recall. Consider two words 'played' and 'playing'. Both this words are having same root form as 'play'. If we do not perform the word stemming, then the relational value between played and playing will reduce and it also affects the overall similarity value.

4.4 Perform part of speech tagging.

In this part of speech tagging, words are tagged according to their types. The different types of words are Noun, Verb, Adjective, Adverb, etc. Tagging is used to understand the context of the word. E.g. - Consider a name as "Ram", Ram can be used as a subject as well object in the sentence. So to understand the context of the sentence, tagging is required.

Up to this, it is the preprocessing required to achieve the semantic similarity. Now the actual method is as follows:

4.5 Semantic Similarity Calculation

Now, relative matrix is created as R[m,n] from two sentences sentence1 and sentence2. The values in the matrix are computed using Wu and Palmer's method of similarity. This value describes how much the two words are similar with each other. R[i, j] is the semantic similarity between the most appropriate sense of word at position i of sentence1 and the most appropriate sense of word at position j of sentence2. So R[i,j] gives the weight of edge connect from i to j.

Now, each i^{th}rowword of the matrix is compared with the j^{th} column's word of the matrix. After comparing, sum of all row to column words values are calculated and vice versacolumn to row word's values are calculated and sum is performed.

The pseudo code for computing similarity of two sentences sentence1 and sentence2 is:

```
Total_X = 0;
Total_Y = 0;
for(inti=0;i<|X|;i++)
{
Max_i = 0;
for(int j=0;j<|Y|;j++)
if(R[i, j] >max_i)
max_i = R[i,j] >max_i;
Total_X += max_i;
}
for(int j=0;j<|Y|;j++)
{
max_j = 0;
for(inti=0;i<|X|;i++)
      if(R[i,j]>max_j)
            max_j=R[I,j]>max_i;
Total_Y += max_j;
}

finalSim = (Total_X + Total_Y) / 2
* (|X| + |Y|);
```

In this way, the overall Similarity between the two sentences is computed using above formula.

4.6 Corpus based suggestion for translation

Corpus is used for translation of text using text from the corpus. Thus the sentences which are similar in terms of their meaning can be used to translate query text.

For E.g.

If there is a sentence stored in the corpus as "Ram goes to school" and its translation in Hindi language. If the user input the query as "Raj goes to school" and wants to translate the sentence, then this sentence as it matches with the corpus sentence semantically can be used for achieving translation.

Highly matched sentences can thus be used to translate the text.

5. FUTURE WORK

- Improve the usability of this project so as to use in translation of multiple languages.
- Improve the accuracy further in translation of text.

6. Conclusion

In this way, Semantic similarity is achieved by using various modules as mentioned above. Due to separate modules, not only semantic similarity is achieved but other tasks like Sentence Segmentation, Named Entity Recognizer, Tokenization, Normalizer, etc. are achieved.

7. Acknowledgement

We would like to thank our Guide Prof. D.P. Salapurkar for the support and guidance she gives us on every step of the project execution. We would also like to thank the project review committee members Prof. G. T. Chavanand Prof. V. R. Manga sir who give us their valuable comments. We would also like to express our gratitude to HOD Dr. P. R. Futane who helped us to accomplish this work.

8. REFRENCES

[1] AinuraMadylova, SuleG¨und¨uz ¨O ˘g¨ud¨uc¨u, "A Taxonomy based Semantic Similarity of Documents using the Cosine Measure", Istanbul Technical University, IEEE 2009

[2] Lu Zhiqiang, Shao Werimin, Yu Zhenhua, "Measuring Semantic Similarity between Words Using Wikipedia" International Conference on Web Information Systems and Mining 2009

[3] Tian-Tian Zhu, Man Lan, "MEASURING SHORT TEXT SEMANTIC SIMILARITY USING MULTIPLE MEASUREMENTS" International Conference on Machine Learning and Cybernetics, Tianjin, 14-17 July, 2013

[4] Kaifeng SUN, Yong JI, Lanlan RUI, Xuesong QIU, "AN IMPROVED METHOD FOR MEASURING CONCEPT SEMANTIC SIMILARITY COMBINING MULTIPLE MATRICES" IEEE 2009

[5] David Croft, Simon Coupland, Jethro Shell, Stephen Brown, "A Fast and Efficient Semantic Short Text Similarity Metric" IEEE 2013

[6]Sazianti Mohd Saad, SitiSakiraKamarudin, "Comparative Analysis of Similarity Measures for Sentence Level Semantic Measurement of Text", 2013 IEEE International Conference on Control System, Computing and Engineering, 29 Nov. - 1 Dec. 2013, Penang, Malaysia

[7] ZENG ZhiHao, HU JiPing, DONG Ting, WANG Yu, "Semantic Web Service Similarity Ranking Proposal Based on Semantic Space Vector Model", 2012 International Conference on Intelligent System Design and Engineering Application

[8] Peter D. Turney, Patrick Pantel, “From Frequency to Meaning:Vector Space Models of Semantics”, Journal of Artificial IntelligenceResearch 20