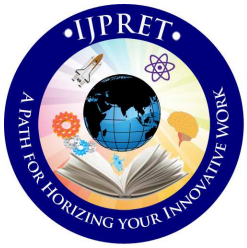
Research Article *ISSN: 2319-507X*
Ankush Maind, IJPRET, 2013; Volume 1(8): 891-901 *IJPRET*



INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

SYNONYM EXTRACTION USING SEMANTIC SIMILARITY MEASUREMENT BETWEEN WORDS

ANKUSH MAIND[1], PROF. ANIL DEORANKAR[2], DR. PRASHANT CHATUR[3]

1. M. Tech. Scholar, Department of Computer Science and Engineering, Government College of Engineering, Amravati, Maharashtra, India.
2. Associate Professor, Department of Computer Science and Engineering, Government College of Engineering, Amravati, Maharashtra, India.
3. Head of Department, Department of Computer Science and Engineering, Government College of Engineering, Amravati Maharashtra, India.

Accepted Date:

27/02/2013

Publish Date:

01/04/2013

Keywords

Semantic similarity,

Synonym Extraction,

ontology

Corresponding Author

Mr. Ankush Maind

Abstract

Semantic similarity is a central concept that extends across numerous fields such as artificial intelligence, natural language processing, cognitive science and psychology. Accurate measurement of semantic similarity between words is essential for various tasks such as, document clustering, information retrieval, community mining and Word Sense Disambiguation. We propose a new approach of semantic similarity measurement using web search engine results and also presented its application in synonym Extraction. Synonym Extraction is a method to find the most perfect meaningful word for a given words from the set of words. This is very useful in online multiple choice type question paper for finding the correct synonym of given words from the option. Here our approach is fully web based approach so it gives the correct answer than the ontology or dictionary based approach. Hence our

INTRODUCTION

The study of semantic similarity between words has been a part of natural language processing and information retrieval for many years. Semantic similarity is a generic issue in a variety of applications in the areas of computational linguistics and artificial intelligence, both in the academic community and industry. Examples include word sense disambiguation [2], detection and correction of word spelling errors (malapropisms) [12], text segmentation [4], image retrieval, multimodal documents retrieval, and automatic hypertext linking [8]. Similarity between two words is often represented by similarity between concepts associated with the two words. A number of semantic similarity methods have been developed in the previous decade; different similarity methods have proven to be useful in some specific applications of computational intelligence. Generally, these methods can be categorized into two groups: Ontology based (dictionary/thesaurus-based) methods and Web based methods.

Our Approach is completely web based because in ontology based approaches faces the limitation of words in the dictionary or ontology, if we use the web search engine each and every updated new words are available. For example in WordNet 2.1 there is synonym for apple is fruit but not as a company but if users are interested in the company then he has to face the failure. Therefore we use here web search engine based approach to measure semantic similarity between words.

So using this approach of measuring the semantic similarity between words one can find the correct synonym for the query word from the set of words. For this firstly we have to measure the semantic similarity between query word and each word from the set of given words then after measuring the semantic similarity between query word and each word from set of words, the pair having highest semantic similarity this will be the correct synonym for the query words.

The remainder of the paper is organized as follows. Section 2 describes detail about the Information Resources such as database “WordNet” and “Web search engine” on which many researchers have done researches. Section 3 gives detail about the proposed method for measurement of semantic similarity

between words. Section 4 describes about the synonym extraction. The paper concludes in Section 5.

INFORMATION RESOURCES

Information Resources are very important factor for measuring the semantic similarity between words. From the starting work of semantic similarity measurement between words many researcher have used WordNet as Information Resource and recently some have used Web search engine.

WordNet

WordNet [10] is a lexical database for the English language. It was created and is being maintained at the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George A. Miller. Differing from other traditional lexicons, it groups words into sets of synonyms called *synsets*, provides short, general definitions, and records the various semantic relations between these synonym sets. WordNet is particularly well suited for similarity measures, since it organizes nouns and verbs into hierarchies of IS-A relations. Figure 1 illustrates a fragment of the WordNet2.1 IS-A hierarchy.

Ontology based methods are Distance based method such as approach by Rada [1], Information content based method was invented by Resnik [2] also by Lin [11] and Distance and Information Content based method was invented by Jiang and Conrath [3].

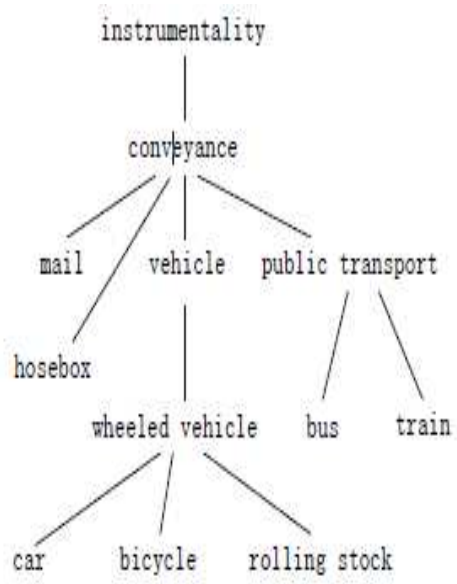


Figure 1. A fragment of WordNet2.1

Web Search Engine

Web search engines provide an efficient interface to the vast information. For the measurement of semantic similarity between words many researcher have been used Web Search Engines results as a resources. Page counts and snippets [9] are two useful information sources provided by most web search engines.

Page count of a query is an estimate of the number of pages that contain the query words. Snippets, a brief window of text extracted by a search engine around the query term in a document, provide useful information regarding the local context of the query term. Semantic similarity measures defined over snippets have been used in query expansion, personal name disambiguation [7], and community mining [8]. Processing snippets is also efficient because it obviates the trouble of downloading web pages, which might be time consuming depending on the size of the pages.

Many researchers have been used snippets as an information source for the semantic similarity measurement between words, few have used only page count as information source and some have used combination of both.

Web search engine based methods include the working of Sahami and Heilman [5] on web Snippets, Cilibrasi and Vitanyi [6] on page counts and also Bollegala, Matsuo and Ishizuka [9] on web snippets and page counts both.

PROPOSED METHOD

The proposed method is fully based on the web search engine results. Here some parameter likes snippets and page counts have been used.

Outline

Following fig. 2 shows the outline of the proposed method. Here firstly one has to use the search engine for searching a given query and the query includes the two words of which we have to calculate the semantic similarity. Because of the vastly numerous documents and the high growth rate of the web, it is time consuming to analyze each document separately. Web search engines provide an efficient interface to this vast information. Page counts and snippets are two useful information sources provided by most web search engines. Then on the basis of this search result of the query following steps has to take which are shown in figure 2.

- Page count-Based Co-Occurrence measures
- Lexical pattern extraction
- Lexical pattern clustering
- Measuring semantic similarity

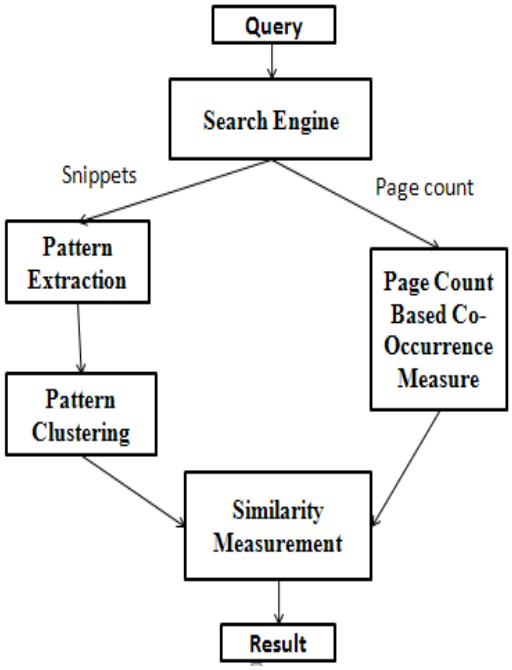


Figure 2. Outline of proposed method

Explanations of all above steps are as follows.

[1] Page count-Based Co-Occurrence measures

In this step we have calculated the page count for given query and for each word in the given query separately also. Page count of a query is an estimate of the number of pages that contain the query words. In general, page count may not necessarily be equal to the word frequency because the queried word might appear many times on one page. Page count for the query "P Q" can be considered as a global measure of co-occurrence of words P and Q. For example, the page count of the query "apple computer" in Google is 300,000,000, whereas the same for "banana computer" is only 4,500,000. The more than 70 times more numerous page counts for "apple computer" indicate that apple is more semantically similar to computer than is banana. Here we have used page count of query words because on the basis of this one can easily conclude about the similarity between words. But in some cases it is failed. Consider the query "P Q" for the search engine. Page counts for the query "P Q" can be considered as an approximation of co-occurrence of two words P and Q on the web. However, page counts for the query P Q alone do not accurately express semantic similarity. For example, Google returns 299,000,000 as the page count for "car automobile," whereas the same is 700,000,000 for "car apple." Although, automobile is more semantically similar to car than apple, page counts for the query "car apple" are more than four times greater than those for the query "car automobile." One must consider the page counts not just for the query "P Q", but also for the individual words P and Q to assess semantic similarity between P and Q. It is simple but using page counts alone

as a measure of co-occurrence of two words presents several drawbacks. First drawback is page count analysis ignores the position of a word in a page. Therefore, even though two words appear in a page, they might not be actually related. Second, page count of a polysemous word (a word with multiple senses) might contain a combination of all its senses. For example, page counts for apple contain page counts for apple as a fruit and apple as a company. Moreover, given the scale and noise on the web, some words might co-occur on some pages without being actually related. For those reasons, page counts alone are unreliable when measuring semantic similarity. Therefore we have compute the four popular co-occurrence measures [8] such as Web Jaccard, Web Overlap (Simpson), Web Dice, and Web Point wise Mutual Information i.e. Web PMI to integrate it with snippets based result for measuring semantic similarity in our approach.

[2] Lexical pattern extraction

In this step we have retrieved the snippets from web search results. Snippets is a brief window of text extracted by a search engine around the query term in a document, provide useful information regarding the local context of the query term. Snippets returned by a search engine for the query of two words provide useful clues related to the semantic relations that exist between two words. Snippets are useful for search because, most of the time, a user can read the snippet and decide whether a particular search result is relevant, without even opening the url. Using snippets as contexts is also computationally efficient because it also avoids the need to download the source documents from the web, which can be time consuming if a document is large. For example, consider the snippet in Figure 3. Here, the phrase 'is a' indicates a semantic relationship between lion and cat. Many such phrases indicate semantic relationships. For example, also known as, is a, part of, is an example of all indicate semantic relations of different types. In the example given below, words indicating the semantic relation between cricket and sport appear between the query words. Replacing the query words by special symbol "!" and "#", one can form the pattern from following example "!, is a large heavy-built social #" .

...lion, is a large heavy-built social cat of open rocky areas in Africa ...

Figure 3. snippet retrieved for the query "lion" AND "cat."

Here we have used this step for extracting the snippets from the web result. Then on these collected snippet, Prefixspan algorithm is used to sub-sequencing the snippet into number of patterns. Here we have collected the patterns from snippets because snippets may contain unnecessary materials which are not needed for describing relation between queried words. These patterns are used for the clustering purpose in next step.

[3] Lexical pattern clustering

After extracting the pattern from the snippet, next step is to cluster the extracted pattern. So to clustering the extracted pattern in proposed method Sequential pattern clustering algorithm [9] is used because it is performed well for more results. Moreover, sorting the patterns by their total word-pair frequency prior to clustering ensures that the final set of clusters contains the most common relations in the data set. This pattern clustering is done on the basis of frequency of patterns because on the basis of this cluster we have computes the feature vector to integrate it with the page count based co-occurrence measure.

[4] Measuring semantic similarity

After completing all above steps, on the basis of co-occurrence measures using page counts and result of feature vector the semantic similarity between queried words are calculated. We have taken feature vector of clusters of lexical patterns from snippets to represent numerous semantic relations that exist between two words. In this section we have integrate both page counts-based co-occurrence measures, and snippets-based lexical pattern clusters to construct a robust semantic similarity measure.

EXPERIMENTAL RESULT

The experimental results include semantic similarities between given two words by using page counts and text snippets retrieved from search engines for given words. For example we have taken here "Apple Computer" as a query to search engine means we have to calculate the semantic similarity between words Apple and Computer here. By using our method following are the steps for measuring semantic similarity between words Apple

and Computer. Before applying the each step firstly we have to give the query "Apple Computer" to Google search engine.

1. In this step Page count for the query is taken from the web search engine which is as follows in Table I.

Table I. Page Count for Query words

Query Words	Page count
Apple	1,750,000,000
Computer	2,820,000,000
Apple Computer	425,000,000

On the basis of this page count we have calculate the Co-Occurrence Measure which are as shown in following Table II.

2. In this step patterns are extracted using the PrefixSpan algorithm from the snippets. Here we have used the 500 snippets and from those snippets we have got the 260 patterns. This all patterns are in between the words "Apple" and "Computer".

Table II. Co-occurrence Measure

Co-occurrence Measure Name	Co-occurrence Measure
WebJaccard	0.102533172
WebOverlap	0.242857142
WebDice	0.185995623
WebPMI	23.03790942

3. Then after extracting the patterns we have clustered it in this step into nine clusters using the Sequential Pattern Clustering Algorithm which are based on relation like is, has, will, being etc. After making the cluster we have calculated the occurrence of each relation in that cluster and then on the basis of this we have calculated the feature vector from that clusters such as follows follows in Table III.

Table III. Feature Vector

Feature Vector No.	Feature Vector Value
1	0.36538461538461536
2	0.015384615384615385
3	0.09230769230769231
4	0.011538461538461539
5	0.03461538461538462
6	0.007692307692307693
7	0.007692307692307693
8	0.046153846153846156
9	0.41923076923076924

4. Then from the step 1 and step 3 we integrate the result of both the steps. From step 1 we have normalised the value greater than 1 into 0 to 1 and then taking the average of all co-occurrence measure which is 0.19044 and from step 3 we have ignored the feature vector no. 1, 2 and 9 because this clusters dose not having any

relation therefore remaining are added and got the result as 0.20000001 and the we have added both result as from step 1 as 0.19044 and from step3 as 0.20000001 and got the result as 0.39044 as a final result i.e. semantic similarity between "Apple" and "Computer". So Semantic Similarity between "Apple" and "Computer" by our method is 0.39044.

SYNONYM EXTRACTION

Synonym extraction is finding the correct meaning to the query word. Here query include only single word. This is also one of the applications of semantic similarity measurement between two words. Here we just have to use the above methods of semantic similarity measurement for each time. Here, we have to measure semantic similarity between query word and each word from the given set of words. Then and Then one can say which one is the correct synonym for the query word. For example we have a question in multiple choice question paper like to find the synonym for the following words "APPLE" and options are FRUIT, COMPANY, TREE and ORGANIZATION. So from this one get confused which one is the correct answer? Because many options are nearly related to APPLE but one is correct but which one? So if we used our approach in this system for finding the correct answer, it will measure the semantic similarity between APPLE-FRUIT, APPLE-COMPANY, APPLE-TREE, and APPLE-ORGANIZATION. Then from this those pair having the more semantic similarity it will be the correct synonym. Here we got the answer like

Table IV. Pair and its Semantic Similarity

Pair	Semantic Similarity
APPLE-FRUIT	0.653254
APPLE-COMPANY	0.362645
APPLE-TREE	0.359000
APPLE-ORGANIZATION	0.308917

This result is calculated using the set of 500 snippets from the web search engine results.

So from this Table IV, pair APPLE-FRUIT having the semantic similarity highest so the correct synonym for the word APPLE is FRUIT. If we use here ontology (WordNet) based approach then there is not all meaning of apple available so we cannot find the semantic similarity between above each pair and hence we will fail to find the correct synonym.

CONCLUSIONS

This paper presented the new approach for semantic similarity measurement between words and its application in synonym extraction. We argue that information source for defining a similarity measure like WordNet are used in many approaches but there is limitation of new words. Hence our approach is web based so one can measure the semantic similarity between any given two words. Another thing is that we have presented here the application of the semantic similarity measurement in the field of synonym extaction along with example. So our approach of measuring semantic similarity between words using the web search engine is very useful for finding the correct synonym of query word.

In future we are trying to extend our work to find the correct synonym for the query having the multiple words.

REFERENCES

1. R. Rada, H. Mili, E. Bichnell, and M. Blettner, "Development and Application of a Metric on Semantic Nets," IEEE Trans. Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 17-30, Jan. 1989.

2. P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy", Proc. 14th Int'l Joint Conf. Artificial Intelligence, 1995.

3. J. Jiang and D. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", Proc. Int'l Conf. Research in Computational Linguistics (ROCLING X), 1997.

4. D. Mclean, Y. Li, and Z.A. Bandar, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", IEEE Trans. Knowledge and Data Eng., vol. 15, no. 4, pp. 871-882, July/Aug. 2003.

5. M. Sahami and T. Heilman, "A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets", Proc. 15thInt'l World Wide Web Conf., 2006.

6. R. Cilibrasi and P. Vitanyi, "The Google Similarity Distance", IEEE Trans. Knowledge and Data Eng., vol. 19, no. 3, pp. 370-383, Mar.2007.

7. D. Bollegala, Y. Matsuo, and M. Ishizuka, "Disambiguating Personal Names on the Web Using Automatically Extracted Key Phrases", Proc. 17th European Conf. Artificial Intelligence, pp. 553-557, 2006.

8. D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring Semantic Similarity between Words Using Web Search Engines", Proc. Int'l Conf. World Wide Web (WWW '07), pp. 757-766, 2007.

9. D. Bollegala, Y. Matsuo, and M. Ishizuka, "A Web Search Engine-Based Approach to Measure Semantic Similarity between Words", IEEE Trans. Knowledge and Data Eng., vol. 23, no. 7, pp. 977-990, July 2011.

10. G. A. Miller, "WordNet: A Lexical Database for English", Comm. ACM, vol. 38, no. 11, pp. 39-41, 1995.

11. D. Lin, "An Information-Theoretic Definition of Similarity", Proc.Int'l Conf. Machine Learning, July 1998.Budanitsky and G. Hirst, "Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures," Proc. Workshop WordNet and Other Lexical Resources, Second Meeting North Am. Chapter Assoc. for computational Linguistics, June 2001.