ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Improving Scalability of R-Dictionary Using Keyword Based Approach

G.Kokila[1], Dr.R.Rajasekaran[2]

P.G Student CSE Dept, V.P.M.M Engineering College for Women, Krishnankovil, Tamilnadu, India[1]

HOD/CSE Dept, V.P.M.M Engineering College for Women, Krishnankovil, Tamilnadu, India[2]

ABSTRACT: Translators and language professionals in general, have long claimed that dictionaries are deficient, especially for regarding access and updating of content. Some authors have also noted that these deficiencies are compounded by the fact that language professionals do not receive (proper) training in dictionary use, and therefore do not fully benefit from them. Electronic dictionaries include new search capabilities, not found in traditional dictionaries that could meet user's needs. Most of the recent work in semantic area suggests the use of Domain specific Ontology to achieve meaningful classification. Hence the design and implementation of a reverse dictionary is described based on the semantic sentence similarity. The objective of this project is to make research in the area of semantic similarity. Semantic similarity is a widely adopted approach to language understanding in which the meaning of a text A is inferred based on how similar it is to another text B. Its scope is typically used in information retrieval with more efficiency and to build a machine that could communicate in natural language.

KEY WORDS: Reverse Dictionary (RD), Text mining, Information Retrieval, Semantic Similarity, Glosses.

I. INTRODUCTION

The use of electronic dictionaries has many advantages over the traditional paper dictionary. However, access to the lexicon and terminology of a dictionary presents certain difficulties, partly due to the lack of user knowledge (even among language experts such as translators) about how a dictionary can be queried to access this kind of information, and partly due to the diversity of ways a dictionary can be consulted (in different areas of the dictionary, with different operators, in widely varied interfaces) which vary from one dictionary to another. Electronic dictionaries are available in both online and offline modes.

Electronic dictionaries and search possibilities: The development of new technologies and the Internet have progressively modified the concept of the dictionary. Many paper dictionaries have been converted to electronic formats, such as CD-ROM, while others are available online. Electronic dictionaries can be classified in various types according to different criteria. In this classification, the author distinguishes between newly developed electronic dictionaries (*new development*) and electronic versions of paper dictionaries (*based on paper*) (Gross 1997, Jacquet-Pfau 2002: 90). Nesi (2000a: 140) states that fully electronic dictionaries are more effective than electronic dictionaries adapted from paper versions: "electronic dictionaries would be most effective if they were designed from scratch with computer capabilities and computer search mechanisms in mind". Electronic dictionaries can be easily updated (Kay 1984: 461, Carr 1997: 214, Harley 2000) and allow a quicker, more precise and exhaustive search, in which a variety of search criteria can be combined.

Types of searches includes

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

In the list of entries: The dictionary shows an alphabetical list of all the dictionary entries starting with the word or sequence of characters introduced by the user. *In the entries:* The dictionary retrieves a list of words that are alphabetically similar to that introduced by the user when the searched word is not found in the dictionary. *In the definitions:* The dictionary generates a list of words whose definitions contain all the words introduced in the dictionary with the operator '&' or some of the words introduced with the operator '|'. *(REVERSE DICTIONARY)Anagrams:* The dictionary retrieves words that result from a combination of all the letters introduced.

Reverse Dictionary is an electronic dictionary where a user submits a sentence or phrase or series of words to a search engine and that search engine then produces that word that the person was defining. Also known as a "rictionary". Follows search type based on in the definition. Hence it is concept based dictionary. The input is set of keywords or terms in the documents. Each keyword has both the property of syntactic similar and semantic similar. *Semantic similarity or semantic relatedness* is a concept whereby a set of documents or terms within term lists are assigned a metric based on the likeness of their meaning / semantic content.

Concretely, this can be achieved for instance by defining a topological similarity, by using ontology's to define a distance between words (a naive metric for terms arranged as nodes in a directed acyclic graph, like a hierarchy, would be the minimal distance in separating edges between the two term nodes), or using statistical means such as a vector space model to correlate words and textual contexts from a suitable text corpus (co-occurrence).

Semantic Sentence Similarity is to find similarity between two sentences. Both semantic and syntactic information make contributions to the meaning of a sentence. Sentence compression may involve recognizing the relationship between parts of a sentence. The relationship may be implicit or explicit. The range value is 0 to 1.

The challenges are databases are rapidly growing due to the increasing amount of information available in electronic form is one the challenge; computing the semantic similarity is another challenge this is referred to *Concept Similarity Problem (CSP).*

Some other challenges are information is unstructured/semi structured; large textual data base and Noisy data for *(e.g.)* spelling mistakes; very high number of possible "dimensions" *(but sparse);* all possible word and phrase types in the language; word sense ambiguity for e.g. Apple *(the company)* or apple *(the fruit).*

II. SYSTEM STUDY

Reverse Dictionary problem description is very simple. There are two main problems.

First, the user input is unlikely to exactly match the definition of a word in the dictionary database. For example, a user may enter the concept or sense phrase (query) *"big building"* when looking for word phrase (result) such as *"castle"* whose dictionary definition be *"large building"* which is conceptually similar but does not contain any of the same words in user input.

Second, the response efficiency needs to be similar to that of forward dictionary lookups.

Hence the problems are as follows upon receiving a search concept; the reverse dictionary consults the forward dictionary at its disposal and selects those words whose definitions are similar to this concept. Those words then form the output of reverse dictionary lookup. This problem is referred as concept similarity problem (CSP).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Estimating the semantic similarity of concepts is an important problem. Results of such studies are reported in a variety of fields, including psychology, natural language processing, information retrieval, language modeling and database systems.

Some related works are helpful in understanding the problem of multiword similarity. One area of such work addresses the problem of finding the similarity of multiword phrases across a set of documents in Wikipedia. The documents contain sufficient contextual information (at least 50-100 words) for similarity matching, which fits within the traditional notion of "short documents" in IR research. In another area of related work, i.e., passage retrieval [6], [7], the concept of interest is modeled as a multiword input query. Work in this area attempts to identify relevant passages in large-text documents.

For a RD, semantic similarities must be computed between multiword phrases. In contrast, in the reverse dictionary scenario, the "documents" considered for similarity are very short dictionary definitions (often consisting of fewer than twenty words), which contain very little contextual information. The lack of contextual information in the RD case adds to the difficulty of addressing this problem.

In the system [1], this system use five pre-create database for dictionary.
One of the databases is RMS database, which contains a table of reverse mapping for each word. In case, new word is add to database the possible RMS is not updated automatically this is drawback of this system.

III. PROPOSED SYSTEM

In RD, semantic similarity is calculated based on sentence phrase which is effective as compared with similarity based on each word. And finding passage similarity is unnecessary since RD contains only fewer words. It reduces the CSP effectively. The drawback of [1] is overcome from this proposed system. This paper is in proceeding.

Architectural design and implementation methodology

The overall architectural design of r-dictionary is somewhat similar to the text mining process because reverse dictionary is process based on the user input *(consider as keyword)* which is a text only.

There are three elements for processing the reverse dictionary: a query, a resource and a result.
The *query* is the word or phrase introduced by the user in the interface of a resource.

The *resource* is the resource or part of the resource in which the word or phrase is searched.

The *result* is the element obtained when a query is searched in a resource.

Solution overview has four modules which are text preprocessing; finding sentence similarity; ranking; analysis of scalability.

Text preprocessing includes stop word removal and stemming. The input is user sentence which is denoted as sense phrase (S).S is query element.

First remove all stop words. Since these stop words do not add specificity to the input, removing them does no harm.
Table 1: List of stop words

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

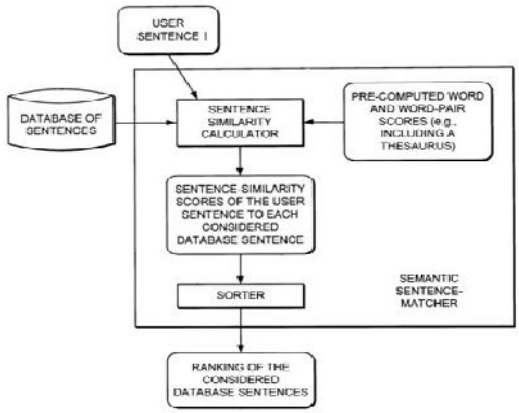
Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

a, be, that, this, the, an, after, before, beside, who, very, then, how, where, else, if, with, of, and, us, during, onto, without, than, some, to, too, in, on, among, into, from, during, etc.

Apply stemming for converting the each word in the sentence to their base form. For that porter stemmer is used. From [2], the porter stemmer is effective when compared with other stemming algorithm.



The output of the text preprocessing is referred as the sense phrase 1(S1) which is input for finding sentence similarity. The S1 is compared with the database sentences *(i.e. meaning of all words)* for finding sentence similarity there are many algorithms some of them are listed in the table below. The Flowchart will represent the how semantic similarity is measuring.

Table 2: Semantic similarity measures

HSO	Two lexicalized concepts are semantically close if their WordNet synsets are connected by a path that is not too long and that "does not change direction too often".
LCH	This measure relies on the length of the shortest path between two synsets for their measure of similarity. They limit their attention to IS-A links and scale the path length by the overall depth D of the taxonomy
WUP	The Wu & Palmer measure calculates relatedness by considering the depths of the two synsets in the WordNet taxonomies, along with the depth of the LCS

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

For finding semantic similarity SEMILAR API is used. SEMILAR API offers a variety of similarity methods based on WordNet (Fellbaum, 1998), Latent Semantic Analysis (LSA; Landauer et al., 1998), Latent Dirichlet Allocation (LDA; Blei et al., 2003), BLEU (Papieni et al., 2002), Meteor (Denkowski and Lavie, 2011), Pointwise Mutual Information (PMI) (Church et al., 1990), methods that use syntactic dependencies, optimized lexico-syntactic matching methods based on Quadratic Assignment, methods that incorporate negation handling, etc. Some methods have their own variations which, coupled with parameter settings and user's selection of preprocessing steps, could result in a huge methods space. SEMILAR provides a framework for the systematic comparison of various semantic similarity methods.

I choose similarity measure based on sentence.

Approaches for finding semantic sentence similarity

Semantic sentence similarity can be finding in two ways.

One is to expand word-to-word similarity (i.e. use similarity of word in one sentence to a word in another sentence and by some means calculate sentence level similarity score) and another approach is to use the semantic representations of sentences to calculate the sentence similarity directly.

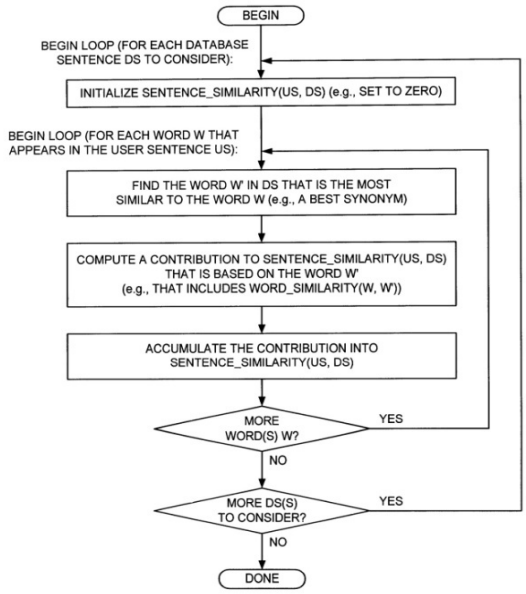


Figure: 1 Flowchart

Sentence-to-Sentence

Measure based on sentence directly

- Corley Mihalcea Comparer
- BLEU Comparer
- METEOR Comparer

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

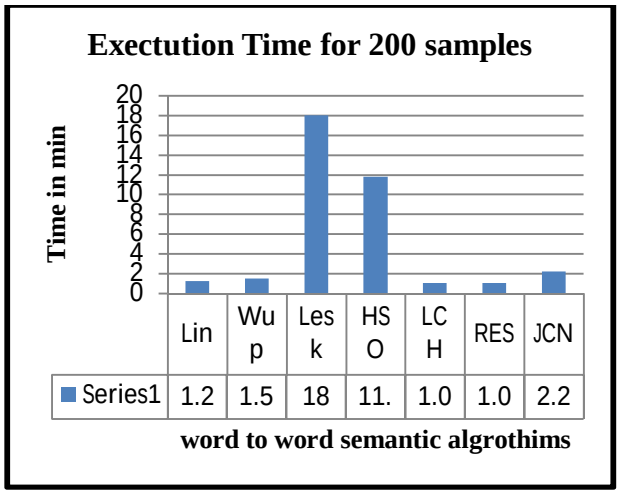
Word-to-Word

The Table 1 measures come under this category.

Then the semantic measures is calculated based on word-to-word is more efficient. Then the measures are sorted and ranked with database sentences, here ranking is done based on relevance and then the corresponding words are considered as candidate words (W). W are listed which is result element of the RD.

IV. ANALYSIS

By using the semantic measures we can improve the relevant measure score. We have to measure the scalability property and compared with existing RD's. For RD, the performance measure execution time of based on word-to-word semantic algorithms are calculated.



In each algorithm, the execution time is large for sample of 200 glosses. The original dataset contains one lakh words the execution time is upto three hours this time.

When user sentence (S) is compared with considered all database sentences purpose we can introduce the clustering concept. We create a cluster based on semantic sentence similarities in the dictionary database. This leads to reduction of database size and computation of similarity measures and time is minimized by introducing the classification.

Measure of RD

Precision: the percentage of retrieved documents that is in fact relevant to the query (i.e., "correct" responses)

Recall: the percentage of documents that are relevant to the query and were, in fact, retrieved.

The following graph shows the precision and recall score.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

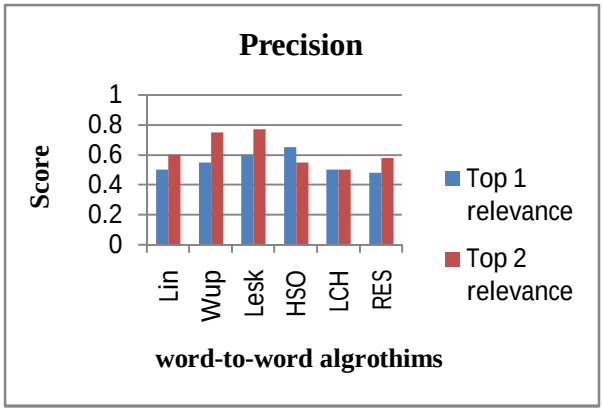
International Journal of Innovative Research in Computer and Communication Engineering

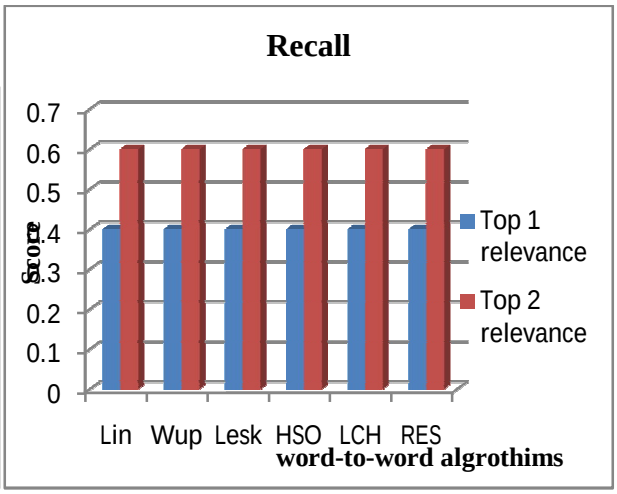
(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014





V. CONCLUSION

Thus RD is effectively created with the idea of semantic similarity. Sentence similarity is achieved by finding word-to-word algorithms.

VI. FUTURE WORK

We have an idea of introducing the wild card characters in user query and eager to develop for bilingual languages.

REFERENCES

[1] Ryan Shaw., Anindya Datta., Debra VanderMeer and Kaushik Dutta "Building a Scalable Database-Driven Reverse Dictionary" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013

[2] Corley and Rada Mihalcea "Measuring the Semantic Similarity of Texts" Courtney *"Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment Association for Computational Linguistics"*.

[3] H. Cui, R. Sun, K. Li, M.-Y. Kan, and T.-S. Chua, "Question Answering Passage Retrieval Using Dependency Relations," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 400-407, 2005.

[4] Jin Feng, Yiming Zhou, Trevor Martin "Sentence Similarity based on Relevance".

[5] Johan Carlberger, Hercules Dalianis, Martin Hassel, Ola Knutsson NADA-KTH Royal Institute of Technology. "Improving Precision in Information Retrieval for Swedish using Stemming" Technical Report IPLab-194, TRITA-NA-P0116, Interaction and Presentation Laboratory, Royal Inst. of Technology and Stockholm Univ., Aug. 2001.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu "BLEU: a Method for Automatic Evaluation of Machine Translation"

[7] Lin Li, Xia Hu, Bi-yun Hu, Jun Wang, Vi-ming Zhou Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009. "Measuring sentence similarity from different aspects"

[8] Satanjeev Banerjee, Ted Pedersen "Extended Gloss Overlaps as a Measure of Semantic Relatedness"

[9] Sneha Jha and H. Andrew Schwartz and Lyle H. Ungar "Penn: UsingWord Similarities to better Estimate Sentence Similarity".

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

[10] SHAN Jian-fang, LIU Zong-tian, ZHOU Wen "Sentence Similarity Measure Based on Events and Content Words".

[11] Snehasis Neogi, Partha Pakray, Sivaji Bandyopadhyay and Alexander Gelbukh "JU_CSE_NLP: Multi-grade Classification of Semantic Similarity Between Text Pairs" "*First Joint Conference on Lexical and Computational Semantics"*.

[12] Thanh Ngoc Dao, Troy Simpson "Measuring Similarity between sentences".

[13] Vasile Rus Mihai Lintean "A Comparison of Greedy and Optimal Assessment of Natural Language Student Input Using Word-to-Word Similarity Metrics".

[14] V.Hatzivassiloglou (et.al) "Detecting Text Similarity over Short Passage" Proc. Joint SIGDAT Conf. Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 203-212, June 1999.

[15] Yuntong Liu, Yanjun Liang *"Journal of Theoretical and Applied Information Technology"*. "A sentence semantic similarity calculating method based on segmented semantic comparison"

[16] C. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval. Cambridge Univ. Press, 2008.

[17] G. Miller, C. Fellbaum, R. Tengi, P. Wakefield, and H. Langone, "Wordnet Lexical Database," http://wordnet.princeton.edu/wordnet/download/, 2009.

[18] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language," J. Artificial Intelligence Research, vol. 11, pp. 95-130, 1999.

[19] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. ACM Press, 2011.

[20] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity," Proc. Nat'l Conf. Artificial Intelligence, 2006.