International Journal of Scientific Research in Science, Engineering and Technology (www.ijsrset.com)
© 2020 IJSRSET | Volume 7 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
DOI : https://doi.org/10.32628/IJSRSET207259

Semantic based Information Retrieval System by using WSD and DICE Coefficient

Thwe[*1], Thi Thi Tun[2], Ohnmar Aung[3]
[*1]Faculty of Computer Science, University of Computer Studies (Taunggyi), Myanmar
[2]Faculty of Computer Science, University of Computer Studies (Myitkyina), Myanmar
[3]Faculty of Computing, University of Computer Studies (Taunggyi), Myanmar

ABSTRACT

In many NLP applications such as machine translation, content analysis and information retrieval, word sense disambiguation (WSD) is an important technique. In the information retrieval (IR) system, ambiguous words are damaging effect on the precision of this system. In this situation, WSD process is useful for automatically identifying the correct meaning of an ambiguous word. Therefore, this system proposes the word sense disambiguation algorithm to increase the precision of the IR system. This system provides additional semantics as conceptually related words with the help of glosses to each keyword in the query by disambiguating their meanings. This system uses the WordNet as the lexical resource that encodes concepts of each term. In this system, various senses that are provided by WSD algorithm have been used as semantics for indexing the documents to improve performance of IR system. By using keyword and sense, this system retrieves the relevant information according to the Dice similarity method.

Keywords : Semantic, WSD, Information Retrieval, Dice.

I. INTRODUCTION

With the advent of web search engines and the proliferation of information on the web, the problem of finding relevant documents has become more visible. The method how to easily retrieve the information and knowledge is needed. In this situation, information retrieval (IR) methods are concerned with the process about the representation, storage, searching and finding of information which is relevant to the user query. Ambiguity in natural language has long been recognized as having a detrimental effect on the performance of text based information retrieval (IR) system. A word can have many different meanings, or senses. For example, “bank” in English can either mean a financial institution, or a sloping raised land. The task of word sense disambiguation (WSD) is to assign the correct sense to such ambiguous words based on the surrounding context.

The disambiguated words are essential for many applications such as information retrieval, information extraction, text summarization, and all tasks in a text mining framework. The word sense disambiguation algorithm is needed for semantic indexing to get the correct sense of the indexed words. Semantic indexing of the document changes from the keyword-based approach to the sense-based approach for effective retrieval. The sense-based information retrieval system eliminates either the possibility of retrieving information that is obtained due to the presence of polysemes of the keywords or the irrelevant information that is retrieved because of non

provision of the correct sense of the word in the searching process.

So, this system is implemented to develop an information retrieval system about technology domain by using concepts (semantics) of the text rather than the keywords. In this system, word sense disambiguation has been semantically performed over the words to increase the accuracy of the IR system. This system also used the WordNet as the lexical resource to support semantic search.

II. RELATED WORK

R. Ackerman [1] proposed Latent Semantic Indexing (LSI), which automatically discovers latent relationships among corpora through singular vector decomposition. However, the method is time-consuming when applied to a large corpus. Vector space search technology can be used on any type information that can be represented in a structured fashion, so it will work equally well on text, images, cryptographic keys, or even DNA.

D. Duy and T. Lynda [2] proposed a sense-based approach for semantically indexing and retrieving biomedical information. Word sense disambiguation (WSD) methods: Left-To-Right WSD and Cluster-based WSD are used for retrieving correct sense. This approach of indexing and retrieval exploits the poly-hierarchical structure of the Medical Subject Headings (MeSH) thesaurus for disambiguating medical terms in documents and queries.

P. O. Michael, S. Christopher and T. John [3] demonstrated the relative performance of an IR system using WSD compared to a baseline retrieval technique such as the vector space model. This disambiguation system was trained and evaluated using Semcor 1.6 which is distributed with WordNet.

III. WORD SENSE DISAMBIGUATION

Word sense is one of the meanings of a word. Words" are having different meanings based on the context of word usage in a sentence. WSD is used to find the correct meaning of the sense or the word [4]. WSD process is essential and useful for many applications. These applications are machine translation, speech processing, text processing, content and thematic analysis, grammatical analysis, and information retrieval and hypertext navigation [5].

A. WordNet

WordNet is used as the source of the synsets. The basic relationship between words in WordNet is the Synonym relation called Synset. Words in the same synset are synonymous in a particular sense. Word sense is the meaning a word can take depending how it is used. For example the word "bank" could mean a financial institution in one sense and a river bank in another sense. Each synset of a word contains one or more words including word itself and has a gloss associated with it.

A gloss for a word sense is the definition of the word in that particular sense and typically includes example sentence(s). For instance, one of the synsets of 'bank' is {depository financial institution, bank, banking concern, banking company} and its gloss is ("he cashed a check at the bank"; "that bank holds the mortgage on my home") [6].

B. Word Sense Disambiguation (WSD) Algorithm

Word sense disambiguation (WSD) algorithm consists of five steps. These are as follows:

- Step 1: Preprocessing
  - Segment input sentence and remove stopwords from input sentence
- Step 2: Identification of monosemous word

For all words in the input sentence,

- If the word has only one sense in the WordNet then this word is defined as the disambiguated word.
- Else this word is defined as the ambiguous.

- End For
- Step 3: Retrieving multi-sense
  - Retrieve all senses of ambiguous word from the WordNet.
  - Collect training data about gloss of synonyms and hypernyms concerning with all senses from the WordNet.
- Step 4: Build context vectors for each sense
  - based on collected training data
  - For all context vectors do
- remove stopwords
  - End For
- Step 5: Disambiguation
  - Calculate the similarity between input vector and each of the context vectors by using dice similarity method.
  - Choose the optimal sense of the target word.

IV. SEMANTIC BASED INFORMATION RETRIEVAL

Semantic-based IR system also retrieves the user query relevant information. But, this IR system must consider the synonyms of the query words as a part of the IR query. Relevant synonyms of the query words in the given context contribute the useful information to the query. These relevant synonyms can be identified with the help of proposed dice coefficient based WSD algorithm.

A. SF-IDF Weighting Scheme

In the semantic-based IR system, the vector space model with sense based implementation (SF * IDF) is used to retrieve documents that are similar to the user query. The sense frequency within document is as follows:

$$sf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \ldots, f_{|v|j}\}} \quad (1)$$

where, f_{ij} is the raw frequency count of sense s_i in document d_j. sf_{ij} is the normalize sense frequency of sense s_i in document d_j. The inverse document frequency is as follows:

$$idf_i = \log \frac{N}{df_i} \quad (2)$$

where, df_i is number of document in which sense s_i appears at least once. N is the total number of document in the system. idf_i is the inverse document frequency of sense s_i . The weight of the sense within document is as follows:

$$ws_{ij} = sf_{ij} \times idf_i \quad (3)$$

where, ws_{ij} is the weight of the sense s_i in document d_j.

B. Sense Weighting Scheme in Query

The weight of the sense within query is as follows:

$$ws_{iq} = 0.5 + \frac{0.5 sf_{iq}}{\max\{sf_{1q}, sf_{2q}, \ldots, sf_{|v|q}\}} \times \log \frac{N}{df_i} \quad (4)$$

where, ws_{iq} is the weight of the sense s_i in query q. sf_{iq} is the raw frequency count of sense s_i in query q.

C. Dice Similarity Measure

To measure the similarity between the document vector d_j and the sense-based query vector q, the similarity measure method is as follows:

$$sim(d_j, q) = \frac{\sum_{i=1}^{|v|} 2 | ws_{ij} \times ws_{iq} |}{\sum_{i=1}^{|v|} |(ws_{ij})|^2 + \sum_{i=1}^{|v|} |(ws_{iq})|^2} \quad (5)$$

V. PROPOSED SYSTEM DESIGN

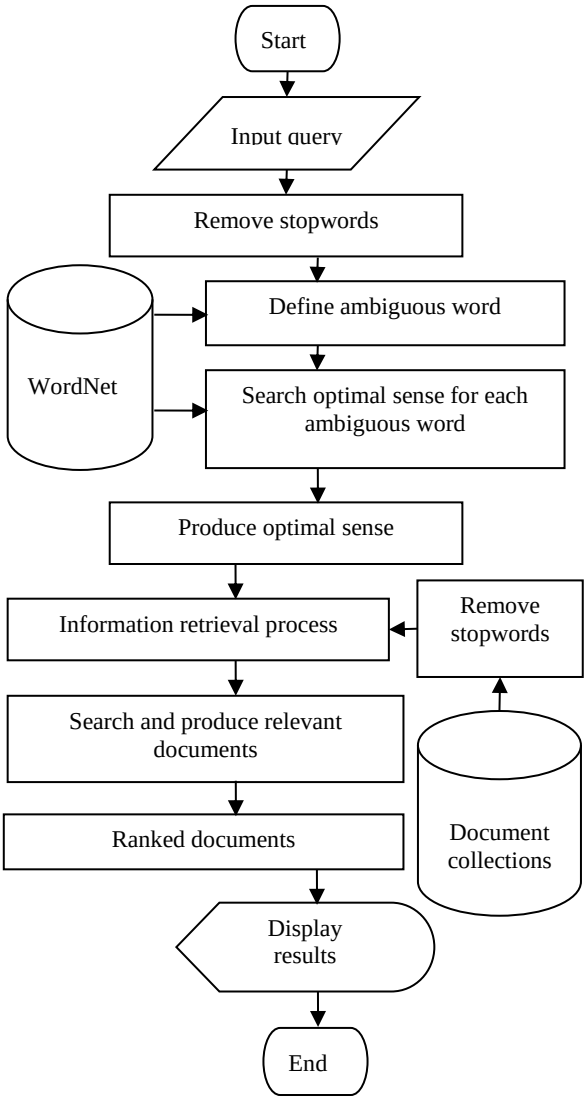

Figure 1: System Flow Diagram

System flow diagram is shown in Figure 1. This system consists of two parts. In the first part, disambiguated words (senses) are produced by using word sense disambiguation algorithm. And then, information retrieval process is performed by using disambiguated words instead of keywords to retrieve user needed information in the second part. At first of the system, the user must input query about required information. After accepting the user query, this system must perform the preprocessing step such as stopwords removal.

And then, this system searches the optimal sense for each ambiguous word within user query according to the word sense disambiguation algorithm. In this algorithm, cosine similarity method and WordNet knowledge resource are used to obtain optimal sense. After producing the optimal sense for each ambiguous word, this system used information retrieval method to retrieve user required information. Finally, this system produced the most relevant documents to the user.

VI. IMPLEMENTATION OF THE SYSTEM

In the word sense disambiguation process, this system disambiguates the input user query (ambiguous query). For disambiguation process, this system uses the WSD algorithm. By using disambiguous query, the performance of information retrieval (IR) system can effectively improve. Word sense disambiguation process is shown in Figure 2.

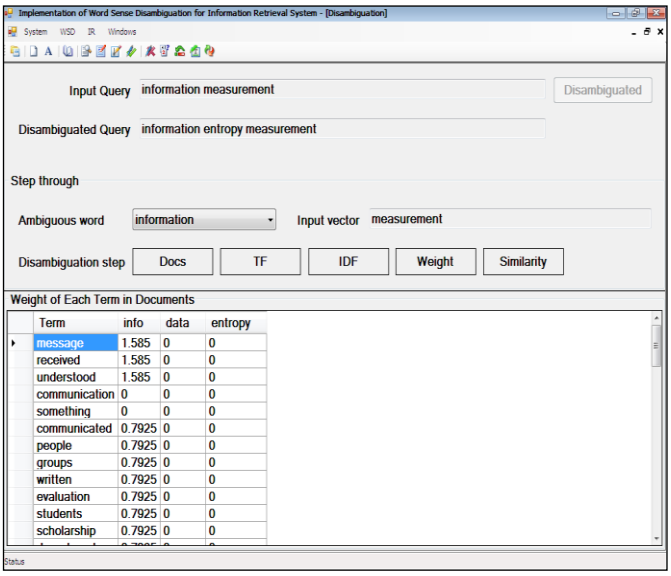

Figure 2: Word Sense Disambiguation Process

In the semantic-based IR process, this system uses the query that is obtained from the WSD process. And then, this system searches the relevant documents that contain user required information by using disambiguous query. After searching, this system displays the relevant documents according to the similarity values. The semantic-based IR process is shown in Figure 3.

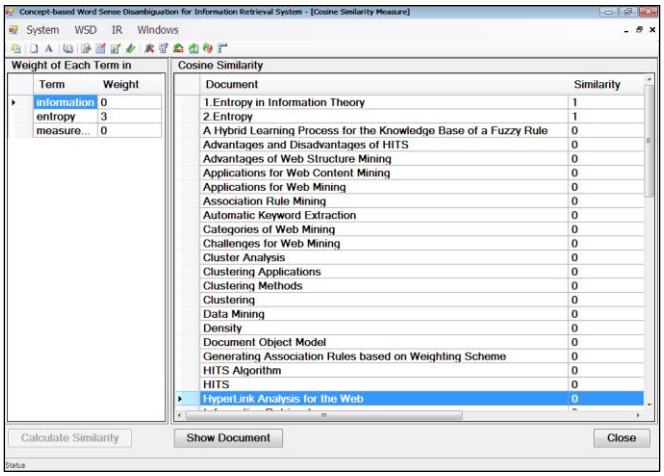

Figure 3: Semantic based IR Process

VII. EXPERIMENTATL RESULTS OF THE SYSTEM

The proposed system is tested by using different queries (ambiguous queries). To show the effectiveness of semantic-based IR, this system retrieves documents that contain user query relevant information by using ambiguous query. As a sample, this system is tested by using ambiguous query "Network about An Interconnected System of People". In this situation, sense-based IR system must first solve different ambiguous word. After disambiguating ambiguous query, the sense-based IR system produces the disambiguous query "Network Web Interconnected Coordinated System People".

To show the performance of semantic-based IR system, this system tested by using different ambiguous query. These queries are shown in Table 1.

TABLE I

USER QUERIES

Query ID	Ambiguous Query	Disambiguous Query
1	network about an interconnected system of people	network web interconnected coordinated system people
2	information measurement	information entropy measurement
3	learning process	learning acquisition process cognitive mental operation cognitive operation
4	collection internet sites web	collection internet sites web World Wide Web WWW
5	mining from the earth	mining excavation earth
6	technology with the art or science of applying scientific knowledge to practical problems	technology engineering engineering science scientific discipline applied science scientific discipline art science scientific discipline applying scientific knowledge practical problems
7	mining to destroy enemy personnel and equipment	mining minelaying destroy enemy personnel equipment

The results of the system is shown in Table 2 and Figure 4.

TABLE III

IR System	Number of Documents				
	Query 1	Query 2	Query 3	Query 4	Query 5
Semantic-based IR System	70	89	78	80	93
Keyword-based IR System	65	60	59	66	80

EXPERIMENTAL RESULTS

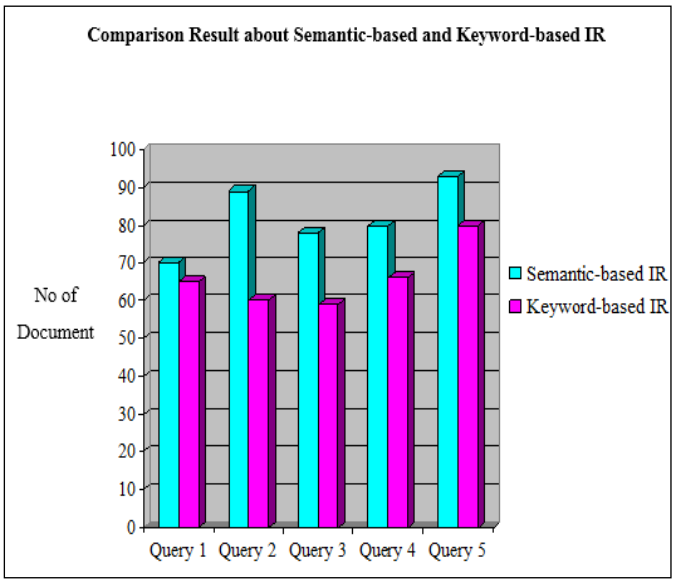

Figure 4: Experimental Results of the System

VIII. CONCLUSION

This system is developed based on the semantic oriented methodology. Thus, this system is useful not only to improve the performance of information retrieval (IR) system but also to find the correct sense of the word by using optimal concept based WSD algorithm. This system also considered content words of the glosses of Synonyms and Hypernyms synset that are associated with the word for finding its correct sense. So, the performance of this system is more precise than other information retrieval system.

IX. REFERENCES

[1]. R. Ackerman, "Theory of Information Retrieval", Florida State University, September, 2003.

[2]. D. Duy and T. Lynda, "Sense-Based Biomedical Indexing and Retrieval", University of Toulouse, Franse, pp. 24-35, 2010.

[3]. P. O. Michael, S. Christopher and T. John, "Word Sense Disambiguation in Information Retrieval Revisited", Proceedings of the 26th Annual International ACM SIGIR conference, pp. 159-166, 2003.

[4]. S. Viswanadha Raju, J. Sreedhar and P. Pavan Kumar, "Word Sense Disambiguation: An Empirical Survey", International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-2, May, 2012.

[5]. I. Nancy and V. Jean, "Word Sense Disambiguation: The State of the Art", Department of Computer Science, Vassar College, 1998.

[6]. A. Bui Muhammad and A. Tambuwal Yusuf, "Query Expansion: Is It Necessary In Textual Case-Based Reasoning?", Nigerian Journal of Basic and Applied Science (NJBAS), 2011.

[7]. B. Liu, "Web Data Mining", Department of Computer Science, University of Illinois at Chicago, USA, Springer-Verlag Berlin Heidelberg, 2007.

Cite this article as :
Thwe, Thi Thi Tun, Ohnmar Aung, "Semantic based Information Retrieval System by using WSD and DICE Coefficient", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 7 Issue 2, pp. 274-279, March-April 2020. Available at doi : https://doi.org/10.32628/IJSRSET207259
Journal URL : http://ijsrset.com/IJSRSET207259