ISBN 978-93-84468-20-0
Proceedings of 2015 International Conference on Future Computational Technologies (ICFCT'2015)
Singapore, March 29-30, 2015, pp. 197-203

Sense-based Information Retrieval System by using Jaccard Coefficient Based WSD Algorithm

Su Mu Tyar[1], and Myo Min Than[2]
[1,2]Department of Information Technology, Yangon Technological University, Myanmar
[1]09sumutyar@gmail.com, [2]dr.myominthan79@gmail.com

Abstract**:** *In many natural language processing (NLP) applications such as machine translation, content analysis, and information retrieval (IR), word sense disambiguation (WSD) is an essential technique. In spite of having long history relationship between WSD and retrieval system, there are still challenges in intelligible retrieval and ambiguous query words. To overcome theses challenges, new Jaccard coefficient based WSD algorithm is proposed with the help of WordNet and Corpus lexical resources for automatically identifying the correct meaning of an ambiguous word . In this study, the glosses of Synonyms, Hypernyms and Hyponyms synsets of WordNet and Corpus that encoded senses of each word are considered in queries disambiguation and senses indexing.*

Keywords: *Word Sense Disambiguation, Information Retrieval, Jaccard coefficient, WordNet, Corpus*

1. Introduction

Due to 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.

The ambiguity in the user query 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 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.

So, the proposed system is implemented as the sense-based information retrieval system by using Jaccard coefficient based WSD algorithm. 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. The proposed sense-based information retrieval system has been semantically performed over the words to increase the precision of the IR system. To support the semantic search, this system uses WordNet version 3.0 and Corpus as the lexical resources. For information technology (IT) domain, this system is proposed to retrieve user query relevance technology information.

The rest of the paper is organized as follows: section 2 presents related works. Background theory about the sense-based information retrieval (IR) is presented in section 3. The proposed system design, explanation and experimental results are described in section 4. Finally, conclusion is described in section 5.

2. Related Work

D. Subarani [1] presented the concept-based information retrieval from Tamil text documents. Semantics has been introduced at various linguistic levels, word level, sentence level and document content extraction level and at various stage of information retrieval such as query and document representation, and indexing, to improve the information retrieval from text documents. Domain ontology that has been created with knowledge based, and word sense disambiguation are used to support semantic search in Tamil document repositories.

P. O. Michael, S. Christopher and T. John [2] 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.

Y. Liu, P. Scheuermann and X. Zhu [3] proposed a text classification method based on word sense disambiguation. This algorithm is applied to Brown Corpus. The sense-based text classification algorithm is an automatic technique to disambiguate word senses and then classify text documents. If this automatic technique can be applied in real applications, the classification of e-documents must be accelerated dramatically. It must be a great contribution to the management system of Web pages and digital libraries, etc.

According to literature and concepts pointed out from the previous works, this work is intended to provide a sense-based information retrieval system by using Jaccard coefficient based word sense disambiguation (WSD) algorithm.

3. Sense-based Information Retrieval (IR)

Information retrieval (IR) system is able to accept a user query, understand from the user query what the user requires, search a database for relevant documents, retrieve the documents to the user, and rank the documents according to their relevance [4]. Before the documents in a collection are used for retrieval, some pre-processing tasks are usually performed. For traditional text documents (no HTML tags), the tasks are stopword removal, stemming, and handling of digits, hyphens, punctuation, and cases of letters [5].

The sense-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 Jaccard coefficient based WSD algorithm.

3.1. Word Sense Disambiguation

Word sense is one of the meanings of a word. Words are having different meanings based on the context of the word usage in a sentence. Word sense disambiguation (WSD) is used to find the correct meaning of the sense or the word. WSD is usually performed on one or more texts although in principle bags of words, i.e., collections of naturally occurring words might be employed [9].

WSD can be viewed as a classification task: word senses are the classes, and an automatic classification method is used to assign each occurrence of a word to one or more classes based on the evidence from the context and from external knowledge sources such as Thesauri, Ontology, Machine readable dictionaries (MRD) and WordNet. Among them, this system is used WordNet within WSD for finding semantically related words [7, 8]. Word sense disambiguation 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 [6].

3.1.1. WordNet and Corpus

WordNet and corpus are the external knowledge sources. These are also used as the source of the synsets. The basic relationship between words in the external knowledge source 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 [10]. For instance, one of the synsets of ‘bank’ is {depository financial institution, bank, banking concern, banking company} and its gloss is (a financial institution that accepts deposits and channels the money into lending activities; "he cashed a check at the bank") [3].

3.1.2. Jaccard Coefficient-based WSD Algorithm

Jaccard coefficient-based word sense disambiguation (WSD) algorithm is shown in Figure 1.

Algorithm: Generate correct sense of the ambiguous word.
w : the word in the sentence
aw : the ambiguous word
s : sense of the word
***Preprocessing*:** Segment the input sentence
Remove stopwords from this sentence
***Building context vector*: For** all w_i in the sentence
Look up in the WordNet.
If w_i has only one s **then**
Let w_i be disambiguous word.
else retrieve all s_i of aw_i from WordNet and English corpus.
For all s_i of aw_i
- build context vector using gloss of s_i from WordNet and English corpus;
- remove stopword in each context vector;
EndFor
EndFor
***Disambiguation* : For** all s_i of aw_i
- calculate similarity between the input vector A and each context vector B

$$sim(A,B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sum_{i=1}^{n} (A_i)^2 + \sum_{i=1}^{n} (B_i)^2 - \left(\sum_{i=1}^{n} A_i \times B_i \right)}$$

If s = max score (s_i)
return the correct sense s of the ambiguous word aw;
EndFor

Fig. 1: Jaccard Coefficient-based WSD Algorithm

3.2. Similarity Measure Method

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

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

3.2.1. Sense Weighting Scheme in Document

To calculate the weight of sense within document, SF (sense frequency) and IDF (inverse document frequency) are used. The sense frequency within document is as follows:

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

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:

$$idf_i = \log \frac{N}{df_i} \quad (3)$$

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 (ws_{ij}) within document is as follows:

$$ws_{ij} = sf_{ij} \times idf_i \tag{4}$$

3.2.2. Sense Weighting Scheme in Query

The weight of the sense within query is as follows:

$$ws_{iq} = \left[0.5 + \frac{0.5 sf_{iq}}{\max\{sf_{1q}, sf_{2q}, ..., sf_{|v|q}\}} \right] \times \log \frac{N}{df_i} \tag{5}$$

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.

4. Proposed System Design

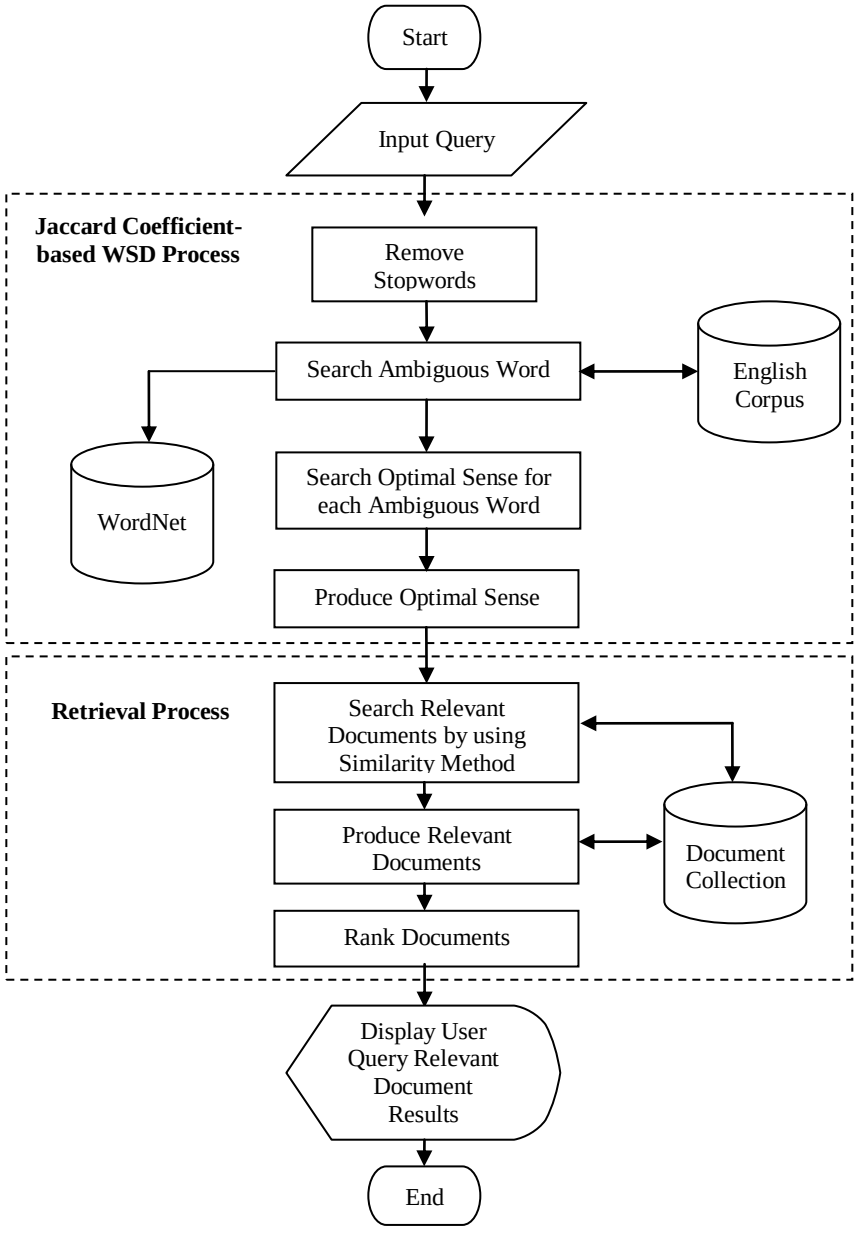

Fig. 2: Proposed Sense-based Information Retrieval System Design

The proposed sense-based information retrieval (IR) system design is shown in Figure 2. In this system, there are two main processes: word disambiguation and information retrieval process. Jaccard coefficient method is applied in disambiguation the ambiguous words. In the information retrieval process, similarity measure method for retrieving the relevant documents in the database. First of all, user input query is accepted to remove the stopwords as a pre-processing step. And then, the ambiguous word, the word that has more than one meaning, is chosen with the help of English corpus and WordNet. After that, the optimal sense for each ambiguous word is determined by applying Jaccard coefficient-based word sense disambiguation (WSD) algorithm. In the optimal sense searching process, WordNet and English corpus are used as the external knowledge resources. After getting the optimal sense for each ambiguous word, documents which have the user require information from the document database collection. For relevant documents searching process, disambiguous user query is used. Finally, the most relevant documents are retrieved to the user according to the ranking process.

4.1. Explanation of the Proposed System

The aim of this study is to improve the performance of Information Retrieval (IR) system. Three documents are considered in the document collection as a sample. These three documents are shown in Figure 3.

Document1

In the learning process, learners are able to acquire knowledge. Learning methods include traditional learning, e-learning, blended learning, mobile learning and personalized learning.

Document2

Knowledge acquisition involves the acquisition of knowledge from human experts, books, documents, sensors, or computer files.

Document3

Knowledge acquisition is the process used to define the rules and ontologies required for a knowledge-based system.

Fig. 3: Sample Documents

Firstly, the user query is accepted and determined whether each query word is ambiguous or disambiguous word. If the user query is "the learning process" for the sample system. The correct meaning of the ambiguous word is determined with the help of the user query input vector. The correct sense of the word is manipulated by calculating the similarity between the user query input vector and each context vector that used the gloss of each sense from the WordNet and English Corpus. Sample ambiguous words and their senses are shown in Table 1.

TABLE I: Sample Ambiguous Words and Their Senses from WordNet and English Corpus

Ambiguous Word	No: of senses	Sense 1	Sense 2	Sense 3	Sense 4	Sense 5
learning	8	acquisition	eruditeness	acquire	study	memorize
information	5	info	knowledge	accusation	data	entropy
process	9	procedure	operation	summons	outgrowth, appendage	treat

The correct sense is regarded depends on the ranks of similarity value between each context vector and input vector. The sense that has the highest similarity result is assumed the correct meaning of the ambiguous word, disambiguous query as a final step of WSD process. Both kind of user query are shown in Figure 4.

User Input Query: the learning process

Disambiguated User Query: learning acquisition process operation

Fig. 4: User Query

Although the disambiguated user query is used in sense-based IR system, the keyword-based IR system applies the simple user query in this study to compare the efficiency. According to the results, the sense-based IR is much more efficient than the keyword-based IR. Retrieval results are shown in Table 2.

TABLE II: Retrieval Results of the Sample documents

ID	Retrieval Results by using Disambiguated User Query	Retrieval Results by using User Input Query
1	Document 1 (Relevant)	Document 1 (Relevant)
2	Document 2 (Relevant)	Document 2 (Irrelevant)
3	Document 3 (Relevant)	Document 3 (Relevant)

4.2. Performance Analysis

To access the "accuracy" or "correctness" of the proposed system, "precision" method is used.

- Precision: the percentage of retrieved documents that is relevant to the query. It can be defined as follows:

$$\text{precision} = \frac{\left|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}\right|}{\left|\{\text{retrieved documents}\}\right|} \quad (6)$$

For the above sample, the experimental results of the proposed system are shown in Figure 5. In spite of the same accuracy result for document 1, higher accuracy results are gained by using the disambiguous query in information retrieval process.

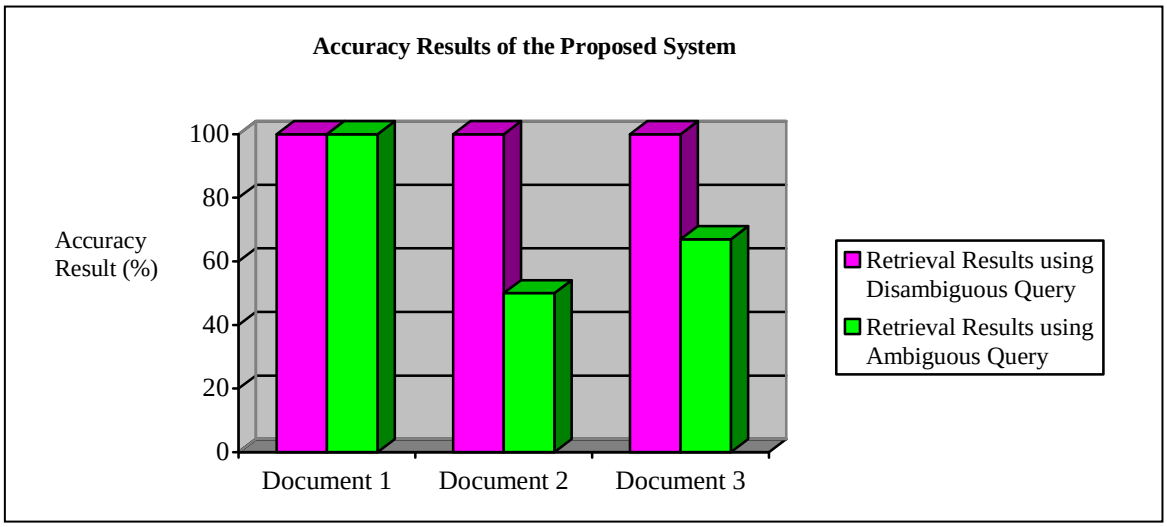

Fig. 5: Experimental Results of the Proposed System

5. Conclusion

The proposed sense-based information retrieval system is developed based on the semantic oriented technology. Sense-based IR system used disambiguous query word is much more efficient than most other IR systems used user direct query. The glosses of Synonyms, Hypernyms and Hyponyms synset of WordNet and Corpus external knowledge resources are used in determining the disambiguous query word. To search the optimal sense of ambiguous query word, this study proposed Jaccard coefficient based WSD algorithm with the help of both external knowledge resources. Working flow of correct sense searching is the calculating similarity value between the user query input vector and each context vector that used the gloss of each sense from the WordNet and English Corpus. This study is useful not only to develop the intelligence information retrieval system but also to understand WSD system for applying any other application area of natural language processing field.

6. References

[1] D. Subarani, "Concept Based Information Retrieval from Text Documents", Dept. of Computer Sciences, SLN College of Sciences, Tirupathi, India, *IOSR Journal of Computer Engineering (IOSRJCE)*, PP 38-38, July-Aug, 2012.

[2] P. O. Michael, S. Christopher and T. John, "Word Sense Disambiguation in Information Retrieval Revisited", The University of Sunderland, Informatics Centre, Canada, 2003.

[3] Y. Liu, P. Scheuermann and X. Zhu, "Using WordNet to Disambiguate Word Senses for Text Classification", *International Conference on Computational Science*, Springer-Verlag Berlin Heidelberg, pp. 780-788, 2007.

http://dx.doi.org/10.1007/978-3-540-72588-6_127

[4] D. Glockner, "Fuzzy Information Retrieval", University Hagen, Department of Computer Science, 2005.

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

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

[7] R. Guzman-Cabrera, P. Rosso and M. Montes-y-Gomez, "Semi-supervised Word Sense Disambiguation Using the Web as Corpus", Universidad de Guanajuato, Mexico, 2009.

http://dx.doi.org/10.1007/978-3-642-00382-0_21

[8] R. Navigli, "Word Sense Disambiguation: A Survey", *ACM Computing Surveys*, Vol. 41, No. 2, Article 10, Italy, February, 2009.

[9] 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.

[10] 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.