Jaccard Coefficient-based Word Sense Disambiguation Using Hybrid Knowledge Resources

Su Mu Tyar
Department of Information Technology
Yangon Technological University
Yangon, Myanmar
09sumutyar@gmail.com

Thanda Win
Department of Information Technology
Yangon Technological University
Yangon, Myanmar
Thanda80@gmail.com

***Abstract*—Word Sense Disambiguation (WSD) has become a popular method for solving the ambiguous meaning of the words in Information Retrieval (IR) field area. Under the Natural Language Processing (NLP) community, WSD has been described as the task which able to select the appropriate meaning among the ambiguous meanings to a given word. Among three approaches, supervised based, unsupervised based and knowledge based approaches to WSD, this paper focuses on both supervised based and knowledge based approaches by proposing new Jaccard coefficient-based WSD algorithm to overcome the vocabulary miss match problem. WordNet and corpus external knowledge resources are utilized as the sense repositories by linking up with the new WSD algorithm to consider additional semantic for WSD. According to sample testing, IR system with new WSD algorithm attains more about 20 percent of total accuracy rate than traditional IR system.**

Keywords—word sense disambiguation; information retrieval; Jaccard coefficient; WordNet; corpus

I. Introduction

In the information age, the accuracy of some IR systems had been declining because ambiguities of word and vocabulary miss match problem between query and document collections. All natural languages contain words that can mean different things in different contexts. For example, “bank” has different meanings based on the context of the word usage in the following two sentences: the fisherman jumped off the bank and into the water and the plane took a bank to the left and then headed off towards the mountains. Moreover, a relevant document to a query may be weekly retrieved if it does not contain a word that exactly matches to a query. Thus, WSD [1] under the NLP community had been tested for combined usage of IR system to reduce IR drawbacks such as word ambiguity and vocabulary miss match problems. This research intends to reduce the vocabulary miss match between query and document collections by proposing new Jaccard coefficient based WSD algorithm under the NLP community. WordNet and corpus lexical resources are applied in the new algorithm for the disambiguation of query word effectively. Besides, the three semantic relations, synonym, hypernym and hyponym, provided by WordNet and corpus lexical resources, are also considered completely to be able to all possible considerations of the glosses of senses deal with these three semantic relations in word disambiguation process. The new proposed algorithm is how much preciousness in IR system is proved with a huge number of document collections by comparing the traditional IR method. The remaining sections of the paper are organized as follows: why motivated to implement new effective IR technique is discussed in section II. The research methodology of disambiguation, information retrieval system and WSD’s application are mentioned in section III briefly. Section IV is about new proposed WSD algorithm based on Jaccard coefficient similarity method. The performance analysis between traditional IR system and enhanced IR system applied Jaccard coefficient-based disambiguation task and the detail explanation of accuracy in percentage are clearly expressed in section V. This paper is concluded in section VI briefly.

II. Related Work

D. Subarani [2] 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. Semantically correct sense is able to be picked up when using concept based algorithm than without using this algorithm. P. O. Michael, S. Christopher and T. John [3] demonstrated the relative performance of an IR system using word sense disambiguation 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 [4] 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 is able to be accelerated dramatically. It is a great contribution to the management system of Web pages and digital libraries, etc. D. Duy and T. Lynda [5] proposed a sense-based approach for semantically indexing and retrieving biomedical information. Two word sense disambiguation methods: Left-To-Right word sense disambiguation and Cluster-based word sense disambiguation are used for retrieving correct sense. This

978-1-4673-7863-5/15/$31.00 ©2015 IEEE

approach of indexing and retrieval exploits the poly-hierarchical structure of the Medical Subject Headings (MeSH) thesaurus in disambiguation medical terms in documents and queries. For long query, both disambiguation methods get the improving rate 7.48% over baseline. But 5.61% for WSD-1 and 4.77% for WSD-2 improved rate are achieved respectively over baseline. Satanjeev Banerjee and Ted Pedersen [6] developed Lesk's dictionary based word sense disambiguation algorithm by employing lexical database WordNet instead of using standard dictionary source for the glosses. Five percentage improved accuracy result is attained when varying Lesk algorithm used as benchmarks in testing with sample data from SENSIBAL-2 word sense disambiguation exercise. According to literature and concepts pointed out from the previous works, new IR technique is explored based on Jaccard coefficient similarity method in disambiguation of word sense by proposing efficient algorithm. The accuracy contrasts between IR applied this new Jaccard coefficient based WSD algorithm and traditional IR is significant. Both of these two IR systems are tested with the same documents collection.

III. Information Retrieval System With Disambiguation Method

Traditional IR system performs retrieval based on the presence or absence of keyword in the documents. [13]. There is drawback when these keywords are different but similar meaning. Although query keywords are not found in the documents, some of documents should be retrieved. Therefore, we emphasize for sensible method by applying unambiguous query words in retrieving documents without using keyword matching techniques.

IR system used WSD[7],[8] algorithm is able to sense for retrieving synonym words, even though vocabularies do not match exactly due to query words disambiguation task. There are two main processes in disambiguation task: the former task is the determination of word sense and the later one is the assignment of each occurrence of a word to an appropriate sense respectively. Before this disambiguation step, the stopword removal process is performed previously as an efficient way. Thus, user input query is accepted to remove the stopwords as a pre-processing step firstly. In second task, the remaining query words are constructed input vector respectively.

External knowledge resources' aids are obviously achieved starting from this step to next four steps. The ambiguous word is chosen with the help of WordNet and corpus to extract all sense of each ambiguous word. And then, context vector is built depending on the gloss of each sense to choose the correct sense by calculating the similarity between context vector and input vector. Ahead of last task, the optimal sense for each ambiguous word is determined by applying Jaccard coefficient-based WSD algorithm. In the optimal sense searching process, WordNet [7], [8], [9] and corpus [7], [8], [10] are used as the external knowledge resources. After getting the optimal sense for each ambiguous word, documents which have the user required information from the document database collection. For relevant documents searching process, unambiguous user query is also used by adding sense with the maximum similarity measurement to user query. Jaccard coefficient similarity method [11] is applied for measuring the similarities of query word and the word in the document database collection. Besides, it is also useful in deciding the most relevant sense depends on the context vector. Finally, the most relevant documents are retrieved according to the ranking process and then the result is displayed to the user. Fig. 1 illustrates the flowchart of information retrieval used Jaccard coefficient-based word sense disambiguation method.

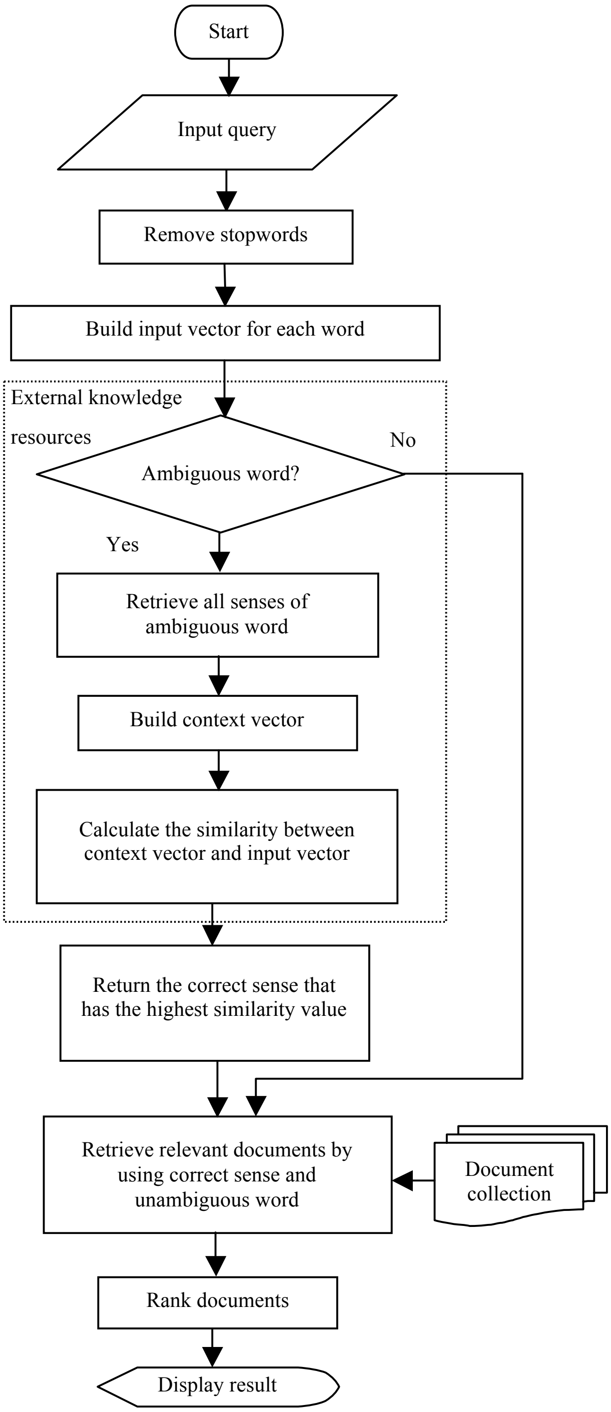
Start
Input query
Remove stopwords
Build input vector for each word
External knowledge resources
Ambiguous word?
No
Yes
Retrieve all senses of ambiguous word
Build context vector
Calculate the similarity between context vector and input vector
Return the correct sense that has the highest similarity value
Retrieve relevant documents by using correct sense and unambiguous word
Document collection
Rank documents
Display result

Fig. 1. Retrieving information by disambigurating word sense

IV. Jaccard Coefficient-Based Word Sense Disambiguation Algorithm

In our WSD algorithm, the three semantic relations, synonym, hypernym and hyponym, are used for external lexical resources to be widely considerable senses in disambiguating. Besides, Jaccard coefficient-based similarity method is also applied in disambiguation step for calculating the optimal sense. There are three main steps in new WSD algorithm such as pre-processing step, building context vector step and disambiguation step as follow:

```
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 WordNet and Corpus
    If $w_i$ has only one $s$ then
        Let $w_i$ be unambiguous word
  else retrieve all $s_i$ of $aw_i$ from WordNet and Corpus
        For all $s_i$ of $aw_i$
            - build context vector using gloss of $s_i$ from WordNet and Corpus
        - remove stopwords in each context vector
        EndFor
EndFor
Disambiguation:
For all $s_i$ of $aw_i$
    - calculate similarities between 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 ($si$)
    return the correct sense $s$ of the ambiguous word $aw$;
EndFor
```

Fig. 2. Sample test results in both IR system (precision in %)

A. Vector Space Model for Proposed IR

Vector representations of documents are created by (SF-IDF) scheme for retrieving documents that are similar to the query. The sense frequency within document (sf_{ij}) is calculated with the raw frequency count.

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

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 (idf_i) is determined with the document counts.

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

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:

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

where, w_{ij} is the weight of the sense s_i in document d_j. The weight of the sense within query is as follows:

$$w_{iq} = \left[0.5 + \frac{0.5 sf_{iq}}{\max\{sf_{1q}, sf_{2q}, ..., sf_{|v|q}\}} \right] \times \log \frac{N}{df_i} \quad (4)$$

where, w_{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.

B. Jaccard Coefficient Similarity in Proposed Algorithm

The similarity between document vector d_j and query vector q is measured with Jaccard coefficient method for the new IR system.

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

V. Performance Analysis

For performance evaluation, precision [12] method is used to access the "accuracy" or "correctness" of the proposed system. Precision is the percentage of retrieved documents that is relevant to the query. Precision-recall method is used to calculate whether the retrieved documents are relevant or not. It is defined as follows:

$$\text{precision} = \frac{|\{\text{relevantdocuments}\} \cap \{\text{retrieveddocuments}\}|}{|\{\text{retrieveddocuments}\}|} \quad (6)$$

All relevant documents are retrieved and tested whether all retrieved documents are relevant to user query by using precision as the main comparing metric to calculate the performance. According to equation (6), the experimental results of the proposed system are shown in Fig. 3. For each experiment, retrieving model is built with Microsoft word file type respectively. Table II and III are the statistics of the proposed IR with Jaccard coefficient WSD and traditional IR. Each row in Table I, II and III is the domain name applied for experimenting proposed in Fig.2. In Table II and III, each three row is number of tested queries for every four domain 25, 50 and 100 respectively. The column AVG NO of RDoc is the list of retrieved documents from each domain respectively. The column AVG NO of RRDoc is the counting of documents that are relevant among the retrieved documents. The last column P in both IR systems, precision, is defined as AVG NO of RRDoc divided by AVG NO of RDoc. Both two column of Precision in Table I is shown the precision value in percentage for both IR systems and every domain.

The precision value increases about 11 percentages in proposed IR system when experimenting with the query of web usage mining. Similarly, 16 percent accuracy rate improves in testing with clustering domain, 28 percentage of accuracy result is higher in testing about information theory respectively. Classification domain testing in both IR systems also achieves higher accuracy rate. According to experimental results, the retrieving technique without using unambiguous query get lower accuracy rate significantly than new proposed IR system.

In conclusion, although unambiguous words querying is not too much strange in both IR systems, ambiguous words querying is not able to sense for vocabulary miss match case in traditional IR system. The vocabulary miss match problem can be significantly overcome by disambiguating query words with the Jaccard coefficient based WSD algorithm shown in Fig. 2.

The ambiguous query is firstly considered to disambiguate based on new WSD algorithm. Then, new unambiguous query is used in retrieving these sample experiments. As a result, all accuracy results get a satisfied level because vocabulary miss match problem between query word and the word in the document is able to overcome by considering new unambiguous query word with the help of proposed WSD algorithm. The better accuracy results, as shown in Table I, are gained for all experiments by using the unambiguous query in information retrieval process.

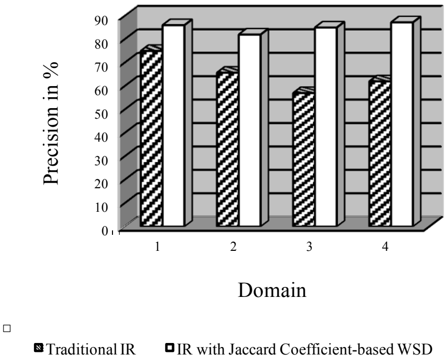

Fig. 3. Sample test results in both IR system (precision in %)

TABLE I. STATISTICS OF TRADITIONAL IR

Domain	NO: of Query	AVG NO: of RDoc	AVG NO: of RRDoc	(P) in Traditional IR
Web Usage Mining	25	373	280	0.75
	50	680	510	
	100	1005	754	
Clustering	25	527	346	0.657
	50	782	514	
	100	811	533	
Information Theory	25	687	392	0.57
	50	1016	579	
	100	1247	711	
Classification	25	417	259	0.62
	50	698	433	
	100	943	585	

TABLE II. STATISTICS OF IR WITH JACCARD COEFFICIENT BASED WSD

Domain	No: of Query	AVG NO: of RDoc	AVG NO: of RRDoc	(P) in IR with WSD
Web Usage Mining	25	373	320	0.86
	50	680	585	
	100	1005	864	
Clustering	25	527	432	0.82
	50	782	641	
	100	811	665	
Information Theory	25	687	584	0.85
	50	1016	864	
	100	1247	1060	
Classification	25	417	364	0.872
	50	698	609	
	100	943	822	

TABLE III. SAMPLE TEST RESULT (IN %)

Domain	Precision	
	IR system with WSD	*Traditional IR*
Web Usage Mining	86%	75%
Clustering	82%	65.7%
Information Theory	85%	57%
Classification	87.2%	62%

VI. Conclusion

This research proposes new Jaccard coefficient based WSD algorithm for disambiguation of word sense by applying hybrid external knowledge resource. WordNet and corpus encode the concepts in terms of set of synonyms and provides many semantic relations. While one or two semantic relations are mostly considered in other studies, this new algorithm is focused on the consideration of the gloss of synonym, hypernym and hyponym semantic relation of WordNet and corpus in unambiguous task for more sensible ranges. The more semantic relations are considered, the more vocabulary miss match problems can be solved as all of possible glosses for every sense related with these three semantic relations are considered in unambiguous process. Moreover, hybrid approach to WSD methodology is proposed by combining knowledge based and supervised based approaches to get high accuracy rate efficiently. Precision recall method, the most widely used for IR performance evaluation, is used in proving the accuracy rate of the new method. Performance of IR used new Jaccard coefficient based WSD algorithm and that of traditional IR are compared in proving how much new algorithm more accurate than tradition. There is higher effectiveness about 20 percent of average precision value with this new algorithm than traditional IR.

Acknowledgment

The preferred spelling of the word he author would like to thank to Dr. Aye Myint, Rector, Yangon Technological University, for his encouragement and invaluable permission for carrying out the research work. The author wishes to express gratitude and appreciation to Dr. Myo Min Than, Pro-rector, Computer University (Pyay), for his supporting, for his mentoring, and for his numerous suggestions concerned with this research.

References

[1] A. Pohl, "Improving the wikipedia miner word sense disambiguation algorithm", the Federated Conference on Computer Science and Information Systems, 2012, pp. 241–248.

[2] 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), July-Aug, 2012, pp. 38-38.

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

[4] 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, 2007, pp. 780-788.

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

[6] Satanjeev Banerjee and Ted Pedersen, "An Adapted Lesk Algorithmfor Word Sense Disambiguation Using WordNet", University of Minnesota, Duluth, MN 55812 USA.

[7] R. Navigli, "ACM Computing Surveys", Vol. 41, No. 2, Article 10, February 2009.

[8] E. Agirre and P. Edmonds, "Word Ssnse Disambiguation: Algorithm and Application", University of Basque Country, 2006.

[9] R. Mihalcea and D. Moldovan, "Semantic indexing using wordnet senses" ACL-2000 workshop on Recent advances in natural language processing and information retrieval, 2000, pp. 35-45.

[10] S. Gauch and J. Wang, "A corpus analysis approach for automatic query expansion,"The Sixth International Conference on Information and Knowledge Management, 1997, pp. 278-284.

[11] A. Islam, E. Milio and V. Keselj, "Comparing word relatedness measures based on google n-grams" COLING 2012 : Posters, 2012, pp 495-506.

[12] Buckley and Voorhees, "Evaluation in information retrieval", Cambridge University Press, 2009.

[13] Rajiv Khosla, Robert J. Howlett, L. C. Jain, "Knowledge based intelligent information and engineering systems", Melbourne, 2005.