International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 1, Januaryr- 2013

Concept-based Information Retrieval by using Optimal Word Sense Disambiguation

K Kaye
Faculty of Information and Communication Technology
University of Technology (Yatanarpon Cyber City)
Pyin Oo Lwin, Myanamar

Win Thandar Aung
Faculty of Information and Communication Technology
University of Technology (Yatanarpon Cyber City)
Pyin Oo Lwin, Myanamar

Abstract

Nowadays, Word sense disambiguation (WSD) is an important technique for many NLP applications such as machine translation, grammatical analysis, content analysis and information retrieval. Information retrieval (IR) is to provide users with documents that will satisfy their information need based on the query. In the 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 optimal concept-based word sense disambiguation algorithm to increase the precision of the IR system about technology domain. This system provides additional semantics as conceptually related words with the help of glosses to the query words and keywords in the documents by disambiguating their meanings. In this system, various senses that are provided by concept-based WSD algorithm have been used as semantics for indexing the documents to improve performance of IR system.

Keywords: Word Sense Disambiguation (WSD), Information Retrieval (IR), WordNet.

1. Introduction

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 has 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. If only documents that containing the relevant sense of a word in relation to a particular query were retrieved, this would improve the precision of the IR system.

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.

Therefore, 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, optimal concept-based 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.

2. Related work

D. Subarani [4] 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 [6] 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.

D. Duy and T. Lynda [3] proposed a sense-based approach for semantically indexing and retrieving biomedical information. Two 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.

3. Word sense disambiguation

Word sense disambiguation (WSD) is an open problem in Natural Language Processing (NLP). Selecting the most appropriate sense of an ambiguous word in a sentence is a central problem in NLP. For representing or understanding the NLP sentence, the correct senses of the content words of the sentence are necessary. WSD identifies the sense of the word used in the sentence or the query when it has multiple meanings. Word sense disambiguation 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 [10].

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 [8, 9].

3.1. WordNet

WordNet is the lexical resource over any other online thesaurus. It encodes concepts in terms of sets of synonyms (called synsets). WordNet provides the user with meaning of a word. Moreover, it also provides the semantic relations such as synonyms, hypernym, hyponyms and antonyms of that word. WordNet divides words into synonym sets or synsets, groups of words that are synonyms of one another. These synsets are then connected by a number of different relations.

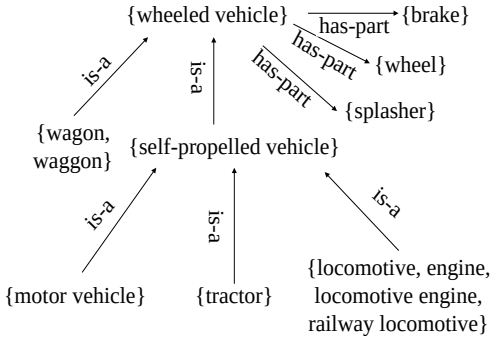

Figure 1. WordNet Semantic Network

The WordNet semantic network is shown in Figure 1. A particular word may appear in several synsets, depending on how many senses it has. These synsets are then inter-connected as a net of synsets by links on a number of different relations such as the following:

- IS-A relation (Hyponym).
  - E.g. Apple is a fruit.
- INCLUDES relation (Hypernym).
  - E.g. Fruits include apple.
- ANTONYM relation.
  - E.g. Boy is an antonym of girl.

When a word is given to the WordNet, a corresponding set of synsets containing all senses of all word is obtained. The disambiguation process aims at choosing the correct sense of the word [4].

3.2. Applications of WSD

Word sense disambiguation process is essential and useful for many applications. These applications are as follows:

- Machine Translation: Sense disambiguation is essential for the proper translation of words depending on the context.
- Information retrieval and hypertext navigation: When searching for specific keywords, it is desirable to eliminate occurrences in documents where the words are used in an inappropriate sense.
- Content and thematic analysis: A common approach to content and thematic analysis is to analyze the distribution of predefined categories of words.
- Grammatical analysis: Sense disambiguation is also necessary for certain syntactic analyses, such as prepositional phrase attachment and in general restricts the space of competing parses.
- Speech processing: Sense disambiguation is required for correct phonetization of words in speech synthesis.
- Text processing: Sense disambiguation is necessary for spelling correction to determine when diacritics should be inserted [5].

4. Information retrieval system

Information retrieval (IR) is the study of helping users to find information that matches their information needs. Technically, IR studies the acquisition, organization, storage, retrieval, and distribution of information. Historically, IR is about document retrieval, emphasizing document as the basic unit. 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 [2].

Documents related to an IR query sometimes contain only the synonyms of the query words instead of the query words themselves. A simple IR system with no knowledge of synonyms fails to recognize the relevance of these documents to the query. So, IR systems must consider the synonyms of the query words as a part of the IR query. However, only relevant synonyms of the query words in the given context contribute useful information to the query. These relevant synonyms cab be identified with the help of a disambiguation algorithm [4].

5. Overview of proposed system

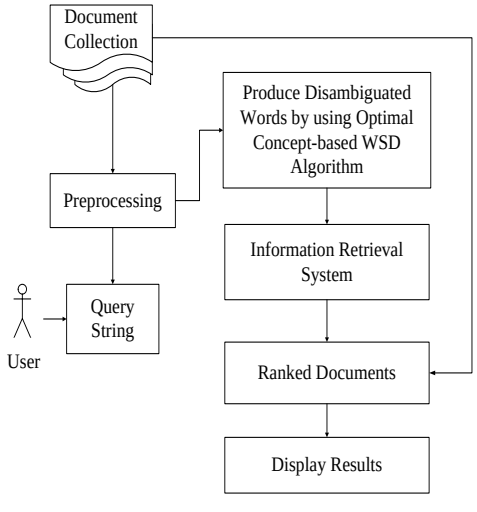

Figure 2. Proposed System Architecture

Proposed system architecture is shown in Figure 2. This system has been developed to retrieve information about technology domain based on their conceptual information. The optimal concept-based word sense disambiguation (WSD) algorithm is proposed in this system. This WSD algorithm is used together with the weighted K-NN (K-Nearest Neighbor) classifier and CBR (Case-based Reasoning) approach to produce the optimal sense of each word.

At first, this system produces various senses of the words within the query and documents by using optimal concept-based WSD algorithm. And then these senses are used for indexing within information retrieval (IR) system.

The effective of keyword-based IR system is decreased by synonyms within query and documents. Synonyms impair the system's ability to find all matching documents because of different meaning words. So, this system is developed to support and improve IR system's ability by using senses of each word.

5.1. Proposed methodology

This system proposed the optimal concept-based word sense disambiguation algorithm and sense-based information retrieval for improving the accuracy of the IR system.

5.1.1. Optimal concept-based word sense disambiguation algorithm. The optimal concept-based word sense disambiguation algorithm includes the conceptually related words and also considers Hypernym synsets. The conceptually related words are taken from the content words of the glosses. The glosses, which are the description of words, are taken from the WordNet.

The steps of this optimal concept-based WSD algorithm are as follows:

1. First the document is preprocessed.
2. And then, the set of disambiguated words (SDW) is assigned as the empty set SDW= {}.
3. At this time, the set of ambiguous words (SAW) is also formed by all the nouns and verbs in the document.
4. If the words within SAW are monosemous words, these monosemous words are removed from SAW and added to SDW.
5. After removing monosemous words, the optimal sense of the remaining ambiguous words are searched by using weighted K-NN classifier and CBR approach.
6. Determination of optimal sense based noun context will identify a set of nouns, which can be disambiguated based on their noun-contexts.
7. Determination of sense of words belonging to same synset as already disambiguated words tries to identify a synonymy relation between the words from SAW and SDW.
8. Determination of sense of words belonging to same synset but not already disambiguated is different

$$prob(k|f_i) = \frac{N(k, f_i)}{N(f_i)}$$

from the previous one, as the synonymy relation is sought among words in SAW.

9. Determination of senses of words belonging to same hypernymy/ hyponomy synset already disambiguated tries to identify words from SAW, which are linked at a distance of maximum 1 with the words from SDW.
10. Default sense attachment determines the remaining words, which could not be disambiguated but are present in the WordNet. Those words are marked with sense #1.

5.1.2. Weighed K-NN classifier. Weighted K-NN is a supervised learning algorithm in which the classification is accomplished by comparing a given test vector with training vector that are similar to it. When an unknown vector is introduced, the weighted K-NN classifier finds *k* most similar training vectors that are closest to the unknown vector. These *k* training records are the *k*-nearest neighbors of the unknown vector. This classifier determines the label of the unknown vector by using its *k* nearest neighbors.

In the weighted K-NN classifier, the distance between two typical vectors $X_1 = (x_{11}, x_{12},\ldots, x_{1n})$ and $X_2 = (x_{21}, x_{22},\ldots, x_{2n})$ is defined as follows:

$$dist(x_1, x_2) = \sqrt{\sum_{i=1}^{n} w_{fi}(x_{1i} - x_{2i})^2} \quad (1)$$

where, w_{fi} is the weight assigned to the feature f_i and x_{ij} is the value of i-th feature in the j-th vector. The weight of the extracted features is as follows:

$$w_{fi} = (\log_{N(k)} N(k, f_i)) * prob(k|f_i) \quad (2)$$

where, N(k, f_i) is the number of paragraphs or sentences in which the feature f_i co-occurs with the *k*-th sense of the ambiguous word. N (k) is the number of paragraphs or sentences in which ambiguous word is in its *k*-th sense. The $prob(k/f_i)$ is as follow:

$$prob(k|f_i) = \frac{N(k, f_i)}{N(f_i)} \quad (3)$$

where, N(f_i) is the number of paragraphs in which f_i occurs [1].

5.1.3. Case-based reasoning (CBR) approach. Case-Based Reasoning (CBR) approach is applied for sense selection with different learning classifier. This approach utilizes knowledge from a past situation to help deal with the complexities of a current problem. CBR life cycle consists of the following four stages:

- Retrieving: Similar previously experienced cases whose problems are considered to be similar are selected.
- Reusing: The cases are reused by either copying or integrating the solution from the retrieved cases.
- Revising: The proposed solution are revised or adapted to try to solve the new problem.
- Retaining: The new solution is stored after being confirmed and validated [7].

5.1.4. Sense-based information retrieval. Sense-based Information Retrieval (IR) is one of the retrieval systems which is browsing through documents and searching for specific information. Sense-based IR is about document retrieval relevant to user queries.

In this system, vector space model with sense based implementation (SF * IDF) is used to retrieve documents that are similar to the user query. In the vector space model, cosine similarity is used to compute the degree of relevance between the user query and document. The cosine similarity method is as follows:

$$\cos ine(d_j, q) = \frac{\sum_{i=1}^{|v|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|v|} w_{ij}^2} \times \sqrt{\sum_{i=1}^{|v|} w_{iq}^2}} \quad (4)$$

where, cosine (d_j, q) is cosine similarity between document d_j and query q. w_{ij} is weight of the sense s_i within document d_j. w_{iq} is weight of the sense s_i within document q.

The sense frequency within document is as follows:

$$sf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j},\ldots,f_{|v|j}\}} \quad (5)$$

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:

$$idf_i = \log \frac{N}{df_i} \quad (6)$$

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:

$$w_{ij} = sf_{ij} \times idf_i \quad (7)$$

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:

$$w_{iq} = \left[0.5 + \frac{0.5 f_{iq}}{\max\{f_{1q}, f_{2q}, \ldots, f_{|v|q})\}}\right] \times \log \frac{N}{df_i} \qquad (8)$$

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

5.2. System flow diagram

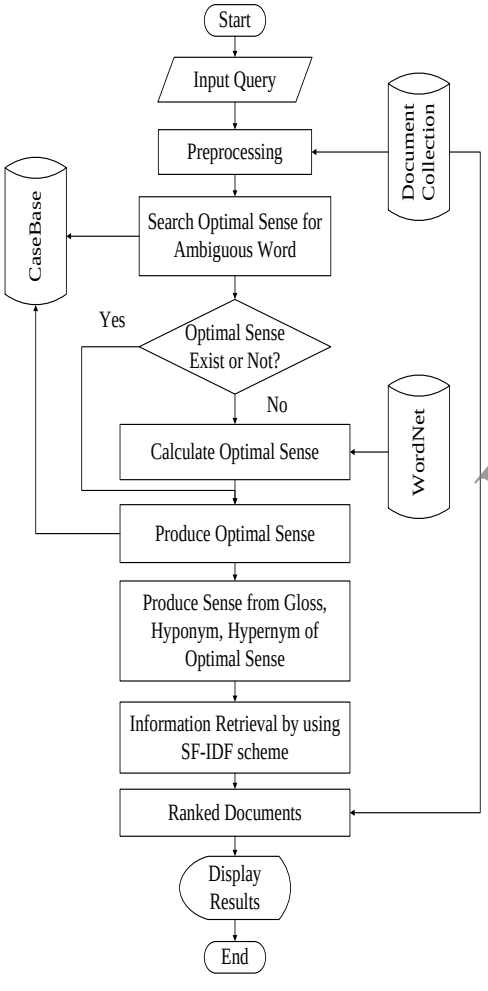

Figure 3. System Flow Diagram

System flow diagram is shown in Figure 3. This system consists of three parts. In the first part, preprocessing step is performed. In next part, disambiguated words (senses) are produced by using optimal concept-based 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 final 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 searchs the optimal sense for each ambiguous word according to the optimal concept-based word sense disambiguation algorithm. In this algorithm, CBR approach, weighted K-NN classifier and WordNet knowledge resource are used to obtain optimal sense. After producing the optimal sense for each ambiguous word, this system used sense-based information retrieval method to retrieve user required information. Finally, this system produced the most relevant documents about technology domain to the user.

5.3. Implementation of the system

This system is implemented to increase the performance of Information Retrieval (IR) system and Word Sense Disambiguation (WSD) algorithms by using concept information. WordNet is also used as knowledge resource.

There are many object oriented programming languages. Among them, this system is implemented by using Microsoft Visual Studio 2010, C# programming. In this system, information (documents) about technology domain is used as application area.

According to the result of this system, the performance of sense-based IR system is more accurate than the performance of keyword-based IR system.

5.4. Performance analysis

To access the "accuracy" or "correctness" of the system, there are two measures of IR success, both based on the concept of relevance [to a given query or information need], are widely used: "precision" and "recall" [2].

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

$$precision = \frac{\left|\{relevantdocuments\} \cap \{retrieveddocuments\}\right|}{\left|\{retrieveddocuments\}\right|}$$

- Recall: the percentage of documents that are relevant to the query and were retrieved. It can be defined as follows:

$$recall = \frac{\left|\{relevantdocuments\} \cap \{retrieveddocuments\}\right|}{\left|\{relevantdocuments\}\right|}$$

6. 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 gloss, Hypernym synset and Hyponym 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.

7. Referencess

[1] A. R. Rezapour, S. M. Fakhrahmad and M. H. Sadreddini, "Applying Weighted KNN to Word Sense Disabiguation", *Proceedings of the World Congress on Engineering*, Vol III, U.K, July 6-8, 2011.

[2] B. Liu, *Web Data Mining*, Department of Computer Science, University of Illinois at Chicago, USA, 2007.

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

[4] 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.

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

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

[7] P. Tamilselvi and S. K. Srivatsa, "Case Based Word Sense Disambiguation Using Optimal Features", *International Conference on Information Communication and Management*, vol. 16, Singapore, 2011.

[8] 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.

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

[10] 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.