Concept-based Word Sense Disambiguation for Information Retrieval System

K Kaye, Win Thandar Aung
University of Technology (Yatanarpon Cyber City)
kkaye23@gmail.com, winthanda.monywa@gmail.com

Abstract

Word sense disambiguation (WSD) is an important technique for many NLP applications such as machine translation, content analysis and information retrieval. 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 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 each keyword in the documents 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 concept-based WSD algorithm have been used as semantics for indexing the documents to improve performance of IR system. This system is implemented by using C# programming language.

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

1. Introduction

Nowadays, 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. 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. This system also used the WordNet as the lexical resource to support semantic search. In this system, optimal concept-based word sense disambiguation has been semantically performed over the words to increase the accuracy of the IR system.

2. Related Work

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.

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.

3. Word Sense Disambiguation

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 [5].

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 such as the following:

- IS-A relation (Hyponym).
  - E.g. Apple is a fruit.
- INCLUDES relation (Hypernym).
  - E.g. Fruits include apple.

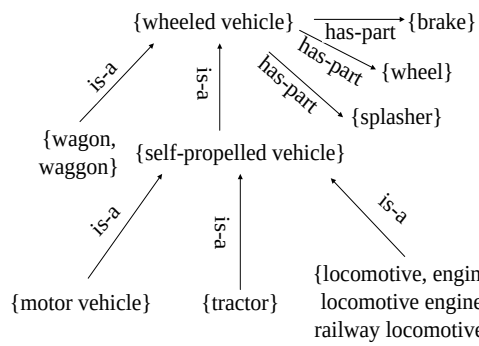

Figure 1. WordNet Semantic Network

The WordNet semantic network is shown in Figure 1. 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. 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:

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.
6. Search hypernymy and hyponomy synset of the optimal sense from the WordNet.
7. If the optimal sense for each word within SAW is searched, these words are removed from SAW and added to SDW.

3.2.1. 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:

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

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:

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

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:

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

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

4. Information Retrieval System

Information retrieval (IR) is the study of helping users to find information that matches their information needs. 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].

4.1. 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:

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

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. 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. Proposed System Architecture

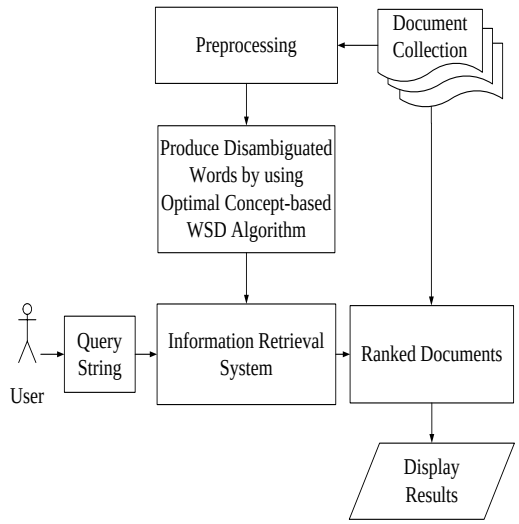

Figure 2. Proposed System Architecture

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 to produce the optimal sense of each word.

At first, this system produces various senses of the words within 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 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.

6.1. System Flow Diagram

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 searches the optimal sense for each ambiguous word within documents according to the optimal concept-based word sense disambiguation algorithm. In this algorithm, 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. System flow diagram is shown in Figure 3.

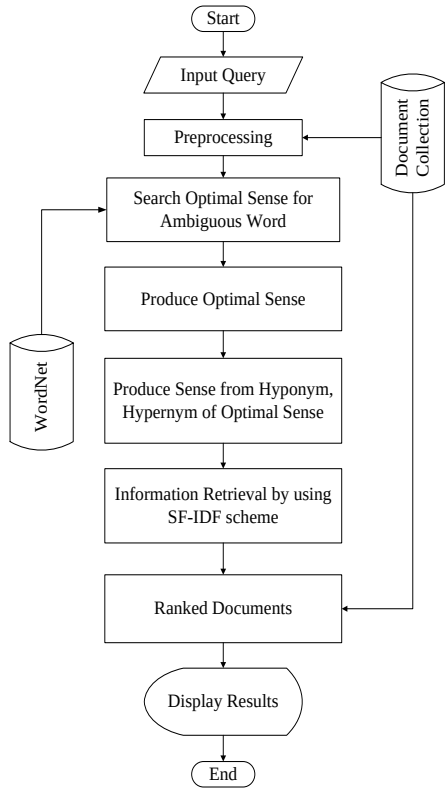

Figure 3. System Flow Diagram

6.2. Explanation of the System

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.

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. As an example, the database consists of three documents with the following content in this system.

Document1: To explore information mining on the web, it is necessary to know data mining which has been applied in many web mining tasks.

Document2: Data mining is the process of discovering useful patterns or knowledge from data sources.

Document3: Information is a collection of facts in the technology domain.

Figure 4. Documents in the Database

In this system, the user first input the query.
Query: information mining.

After receiving the user query, this system removes stop words from these documents. And then, this system defines the ambiguous words within each document. After defining, this system searches all sense of these ambiguous words by using WordNet. Table 1 shows some ambiguous words and their senses in each document.

Table 1. Ambiguous Words and their Senses

Ambiguous Word	No: of Sense	Sense 1	Sense 2	Sense 3	Sense 4
data	1	information	-	-	-
technology	2	engineering	applied science	-	-
facts	4	information	info	reality	concept
mining	2	excavation	minelaying	-	-
explore	4	search	investigate	examine	diagnose

And then, this system computes the optimal sense of the ambiguous word. As an example, the ambiguous word "Information" in the document 3 has four senses: **info**, **data**, **knowledge**, **entropy** and **accusation**.

In the Training Vector:
info = [message, received, understood],
data = [collection, facts, conclusion, drawn, statistical, data],
knowledge = [knowledge, acquired, study, experience, instruction]
entropy = [numerical, measure, uncertainty, outcome, signal, contained, thousands, bits, information]
accusation = [formal, accusation, crime]

In the Testing Vector:
[collection, facts, technology, domain]

To obtain the optimal sense, the similarities between the training vector and the testing vector are computed by using weighted K-NN classifier. And then, hyponym and hypernym of each optimal sense are extracted from the WordNet.

After searching the optimal sense for each ambiguous word, this system must perform the information retrieval process. In sense-based information retrieval, this system measures the angle between the query and document vectors. Table 2 shows weight value of some vector (sense) within each document that are calculated by using SF-IDF weighting scheme.

Table 2. Weight Value of Some Sense

Sense	Sense Frequency			Inverse Document Frequency	Weight		
	D1	D2	D3		D1	D2	D3
explore	0.33	-	-	1.585	0.523	-	-
Information (data)	0.67	1	1	0	0	0	0
mining	1	1	0	0.585	0.585	0.585	0
web	0.67	0	0	1.585	1.062	0	0
necessary	0.33	0	0	1.585	0.523	0	0

know	0.33	0	0	1.58	1.585	0.523	0
applied	0.33	0	0	1.585	0.523	0	0

After calculating sense weights in each document, this system also calculates sense weights in query. The sense weight in query is as follows:

- $W_{information,\ query} = 0$
- $W_{mining,\ query} = 0.585$

And then, this system used cosine similarity measure method to retrieve documents that is relevant to the user query. The calculation of cosine similarity measure is as follows:

- $Cosine_{(document1,\ query)} = (0 * 0) + (0.585 * 0.585) / (0.585 * 0.585) = 1$
- $Cosine_{(document2,\ query)} = 1$
- $Cosine_{(document3,\ query)} = 0$

Finally, this system retrieves document 1 and 2 those are relevant to the query by using sense-based information retrieval method.

6.3. Experimental Result

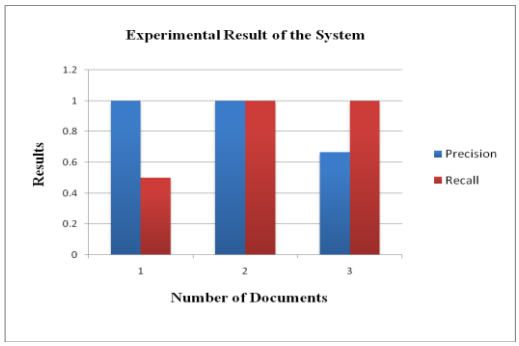

Figure 5. Experimental Result

In this paper, experimental result for sample calculation is shown in Figure 5. 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.

7. 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.

References

[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] 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.

[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.