INTERNATIONAL JOURNAL OF ADVANCE RESEARCH AND DEVELOPMENT

(Volume 4, Issue 7)

Available online at: www.ijarnd.com

Tanimoto coefficient based Word Sense Disambiguation

Hnin Yu Yu Win[1], Htwe Htwe Pyone[2]

[1]University of Computer Studies, Taungoo, Myanmar

[2]University of Computer Studies, Myitkyina, Myanmar

ABSTRACT

In many NLP applications such as machine translation, content analysis, and information retrieval, Word Sense Disambiguation (WSD) is an important technique. Word sense disambiguation is the essence of communication in a natural language. WSD process is useful for automatically identifying the correct meaning of an ambiguous word in the sentence or the query when it has multiple meanings. So, this system proposes as the Tanimoto coefficient based word sense disambiguation system to increase the precision of the NLP application. This system provides additional semantic as conceptually related words with the help of glosses to each keyword in the inputted sentence by disambiguating their meanings. This system uses the WordNet as the lexical resource that encodes concepts of each term.

Keywords— *Ambiguous Word, WSD, WordNet*

1. INTRODUCTION

Human language is ambiguous, so that many words can be interpreted in multiple ways depending on the context in which they occur. Unfortunately, the identification of the specific meaning that a word assumes in context is only apparently simple. While most of the time humans do not even think about the ambiguities of language, machines need to process unstructured textual information and transform them into data structures which must be analyzed in order to determine the underlying meaning. The computational identification of meaning for words in context is called Word Sense Disambiguation (WSD).

WSD involves the association of a given word in a text or discourse with a definition or meaning which is distinguishable from other meanings potentially attributable to that word. Therefore, the WSD task necessarily involves three steps. The first step is to determine all the different senses for every word relevant to the text or discourse under consideration, i.e., to choose a sense inventory, e.g., from the lists of senses in everyday dictionaries, from the synonyms in a thesaurus. The second step involves a means to assign the appropriate sense to each occurrence of a word in context. In the third step, the computer needs to learn how to associate a word sense with a word in context using either machine learning or manual creation of rules or metrics. A rich variety of techniques have been researched from dictionary-based methods that use knowledge encoded in lexical resources, supervised machine learning works on classifiers, and unsupervised learning method supports clusters.

Among them, the proposed system uses the knowledge-based (dictionary-based) WSD method that is based on Tanimoto coefficient similarity and WordNet. Using this knowledge-based method, this system can convert ambiguous sentence into disambiguated sentence. By disambiguating user inputted sentence, this system provides to more accurate the performance of the natural language processing applications.

2. RELATED WORK

In 2012, J. Hui and Z. Yangsen [3] considered not only the sentence collocation but also the syntax and semantic to obtain more knowledge in line with human cognitive behavioral models. Co-occurrence words, collocation words and demonstratives have different degrees of constraint on determining the polysemy sense. Therefore, they can be extracted from the corpus, dictionaries, and knowledge source to construct the unified disambiguation knowledge base, and then use them for disambiguation. But, there are still some shortcomings that lead to disambiguation correct rate is not high for the use of the knowledge base.

In 2015, D. Jianyong and L. Xia [2] introduced the attribute knowledge into word sense disambiguation task. Every sense of the polysemous words can be described by the different attribute sets. These attributes can be viewed as a kind of context features. The attribute knowledge bases are built for every polysemous word, and employed into the Naive Bayes classifier and Maximum Entropy classifier as a dimension feature to judge the specific semantic of polysemous words in the specific context. Experimental results show that this method can effectively improve the accuracy of Chinese word sense disambiguation.

In 2016, G. Sreelakshmi and P. H. Rosna [5] proposed a Supervised Malayalam word sense disambiguation system using Naive Bayes classifier. Word sense disambiguation is a complex problem in NLP because a particular word may have different meanings in different situations. For all human beings it is very easy to find out the accurate sense in a particular context but for machines it is very difficult to predict. Some extent of intelligence may add to the machine for an accurate prediction. Here, this system provided 95% reliability using corpora of 1 lakh words.

3. LEXICAL AMBIGUITY

For lexical items, two types of ambiguity are traditionally known as polysemy and homonymy. In a piece of context, an individual can come across a polysemous term or a homonymous one. In polysemy, a word is associated with more than one meaning

which is traditionally called as senses and is distinct but related in some semantic way. In homonymy, a word has meanings which are distinct but not related in a manner. Like the term "bank" signifies river edge and financial institution is a common example of homonymy. Whereas the term "mouth" signifies mouth of the river and mouth of the man is an example of polysemy [4].

4. WORD SENSE DISAMBIGUATION

Word Sense Disambiguation (WSD) is the process of finding the correct meaning or senses for words that have two or more different meanings in a sentence [8]. WSD determines correct sense of an ambiguous word in a definite context. It has been an important search issue in natural language processing, and it has many applications containing text analysis, data mining, information retrieval, machine translation and so on [6].

WSD technology is developing from matching level to syntactic and semantic level. With the continuous improvement of syntactic analysis, more and more grammatical attributes can be used in WSD, and closer to the way by which humans understand the natural language [7].

4.1 Approaches of WSD

Word sense disambiguation is the problem of selecting a sense for a particular word from a set of predefined possibilities. Many polysemic words are there in any natural languages. Polysemic words are nothing but the words with different meanings. Ambiguity is really a problem in the field of natural language processing (NLP). WSD is important for all NLP tasks which require semantic interpretation [9]. There are three WSD approaches. These are as follows:

- Knowledge based WSD approach involves the use of dictionaries, thesauri, ontologism, etc to understand the sense of words in context. Even though these methods have a comparatively lower performance than some other forms of approaches, but they do have large-scale knowledge resources. Knowledge based approach involves using an external dictionary source like WordNet or some other machine language dictionary. Knowledge based approach either uses grammar rules for disambiguation or use hand coded rules for disambiguation.
- Supervised based WSD approach uses machine learning techniques to set a classifier from manually sense-annotated data sets. The classification task for assigning the correct sense to each instance of that word is done by a word expert known as the classifier. The training set usually contains some examples with the target word manually tagged with a sense from the sense inventory of a dictionary.
- Unsupervised based WSD approach has the potential to overthrow the knowledge acquisition bottleneck which is the huge amount of resources manually marked with word senses. This approach is based on the thinking that same sense of a particular word will have alike neighbor words. Clustering word occurrences and classifying new occurrences into induced clusters is how they induce word sense, that is, they are independent of training sets and do not require machine readable resources like dictionaries, etc [10].

4.2 Applications of WSD

Word sense disambiguation hasn't demonstrated any clear cut benefits in human language technology applications till date but this failure is more due to the current deficiencies in its accuracy, and it will start becoming a very resourceful tool. Real world applications of WSD are as follows:

- Machine translation: It is an extremely difficult task to automatically identify the precise translation of a word in context. WSD has been considered as a major task which needs to be solved to enable an accurate machine translation, this is because it is widely known that disambiguation of words in a sentence can help choose better candidates as depending on the context words can have totally different translations.
- Information Retrieval: Explicit semantics are not used to narrow down documents which are not relevant to the user by even the most advanced search engines. An accurate disambiguation of the document database along with the disambiguation of the queried words will facilitate the selection of only those documents which are actually required.
- Content analysis: Analysis of text with respect to ideas, themes, tones, etc can benefit from WSD used in content analysis domain. Content analysis using WSD can help in classification of data with as per user requirements.
- Semantic Web: Semantic Web is nothing but a collaborative movement by World Wide Web consortium to encourage web pages to include semantic content into their web pages to convert the currently existing unstructured or semi structured documents into a web of data.

5. KNOWLEDGE BASED WSD

In the knowledge based word sense disambiguation (WSD), WordNet is used as the lexical knowledge resource and Tanimoto coefficient similarity method is used to search the relevant sense for the ambiguous word.

5.1 WordNet

WordNet is a manually-constructed lexical system developed by George Miller at the Cognitive Science Laboratory at Princeton University. It reflects how human beings organize their lexical memories. It can be considered as a combination of dictionary and thesaurus. It differs from traditional dictionaries and thesaurus in many ways. For example, words in WordNet are arranged semantically instead of alphabetically. The basic building block of WordNet is synset consisting of all the words that express a given concept. Synonymous words are grouped together to form synonymous sets or synsets. WordNet stores information about words that belong to four parts of speech: nouns, verbs, adjectives and adverbs. Each of these are divided into certain lexical categories. For example, Nouns contain many subdivisions like body, communication, object and etc. Verbs contain subdivisions like cognition, communication and etc. Lexical categories within noun, verb, adverb and adjectives were used to recognize sense of the word [1].

5.2 Tanimoto Coefficient Similarity Method

In TF-IDF (term frequency- inverse document frequency) scheme, the inputted sentence is represented as a weight vector (testing vector), in which each component weight is computed based on some variation of TF or TF-IDF scheme. In this scheme, N is total number of training vectors. The df_i is number of training vector in which term t_i appears at least once. The f_{ij} is the raw frequency count of term t_i in training vector d_j. Then, the normalized term frequency (denoted by tf_{ij}) is given as follows:

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

The inverse training vector frequency (denoted by idf_i) of term t_i is given as follows:

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

The final TF-IDF term weight is given as follows:

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

The Tanimoto coefficient method is used to compute the degree of relevance between training vector and testing vector.

$$sim(d_j, q) = \frac{\sum_{i=1}^{|v|} w_{ij} \times w_{iq}}{\sum_{i=1}^{|v|} (w_{ij})^2 + \sum_{i=1}^{|v|} (w_{iq})^2 - \left(\sum_{i=1}^{|v|} w_{ij} \times w_{iq} \right)} \quad (4)$$

In this method, $sim(d_j, q)$ is the similarity between training vector d_j and testing vector q. The w_{ij} is weight of the term t_i within training vector d_j and the w_{iq} is weight of the term t_i within testing vector q.

6. PROPOSED SYSTEM DESIGN

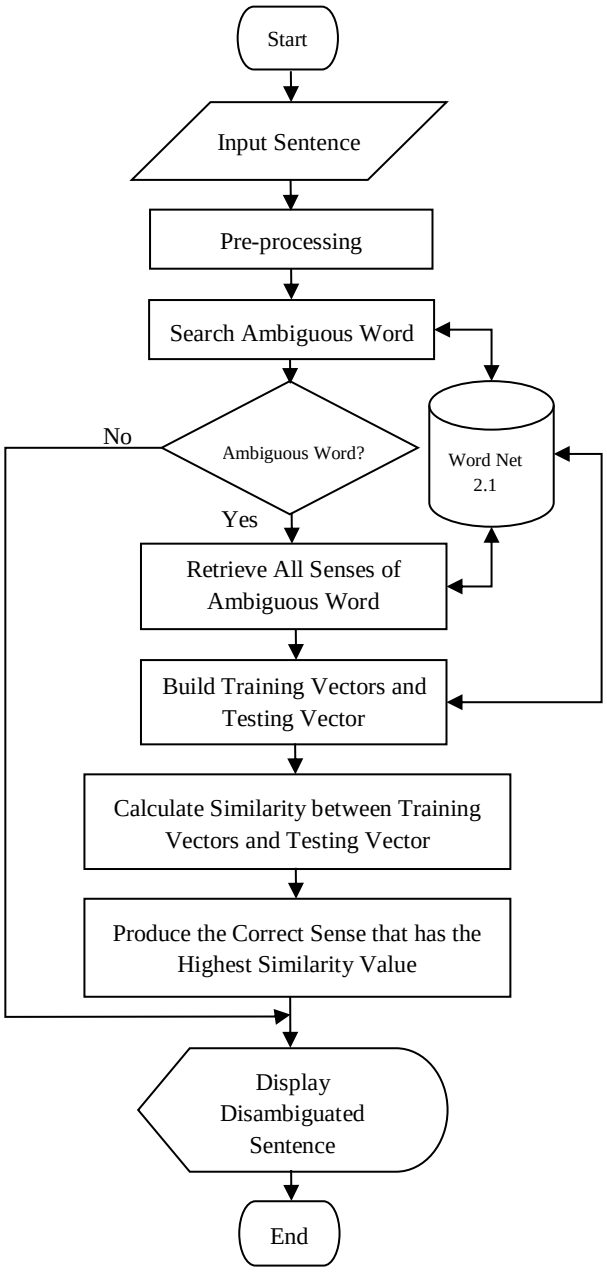

Fig. 1: Proposed system design

Proposed system design is shown in figure 1. At first of the system, the user must input the English sentence. Then, this system performs the preprocessing about the inputted sentence. Preprocessing includes tokenization, stopword removal and stemming.

After performing the preprocessing, this system checks the keywords from the inputted sentence that is ambiguous word or disambiguous word. If the keyword is ambiguous word, this system retrieves all relevance senses about ambiguous word.

For relevance senses retrieval, this system uses the WordNet lexical resource. Then, this system builds the training vectors and testing vector. By using Tanimoto coefficient similarity method, this system calculates the similarity between training vectors and testing vector. After finishing similarity calculation, this system produces the most relevant sense (correct sense) that has the highest similarity value. Finally, this system produces the disambiguated sentence that is useful for NLP application.

6.1 Explanation of the system

In this system, the user must first input the English sentence.

Input Sentence: It is about information measurement.

After performing the preprocessing step, this system checks "information" and "measurement" keyword to denote this word that is ambiguous word or disambiguous word. By using WordNet lexicon, this system extracts the relevant sense for the ambiguous word.

Table 1: Relevant Sense and its gloss

Sense ID	Sense Name	Gloss
Sense 1	Info	message received understood communication something communicated people groups message received understood written evaluation students scholarship deportment father signed report card
Sense 2	Data	collection facts conclusions drawn statistical data several things grouped together considered number entities members considered unit collection facts conclusions drawn statistical data unanalyzed data data subjected analysis
Sense 3	Entropy	communication theory numerical measure uncertainty outcome signal contained thousands bits information system measurement information based probabilities events convey information system related measures facilitates quantification particular characteristic something quantify

Then, this system builds the training vectors and testing vector. Using these vectors, this system calculates the TF, IDF and weight for similarity calculation.

Table 2: TF and IDF calculation results

S no.	Terms (Words)	"TF" about Training Vectors			"IDF"
		Infoo	Data	Entropy	
1	Message	1	-	-	1.585
2	Received	1	-	-	1.585
3	understood	1	-	-	1.585
4	communication	0.5	-	0.33	0
5	Something	1	-	0.33	0
6	communicated	0.5	-	-	1.585
7	People	0.5	-	-	1.585
. . .	. . .	. . .	. . .	. . .	. . .
58	Quantify	-	-	0.33	1.585

Table 3: Weight calculation result

S no.	Terms (Words)	"TF" about Training Vectors			"IDF"
		Infoo	Data	Entropy	
1	Message	1.585	-	-	-
2	Received	1.585	-	-	-
3	understood	1.585	-	-	-
4	communication	0	-	0	-
5	Something	0	-	0	-
6	communicated	0.7925	-	-	-
7	People	0.7925	-	-	-
. . .	. . .	. . .	. . .	. . .	
58	Quantify	-	-	0.523	-
	measurement	-	-	-	1.585

By using TF, IDF and weight results, this system calculates the similarity using Tanimoto coefficient similarity method.

Table 4: Similarity result

S no.	Sense Name	Similarity Value
1	Info	0
2	Data	0
3	Entropy	1

In this sample, "entropy" is the most relevant sense (correct sense) of "Information" ambiguous word. After disambiguating the user inputted sentence, this system produces the disambiguated sentence. The disambiguated sentence is "information entropy measurement".

6.2 Experimental Result of the system

To access the "accuracy" or "correctness" of the system, this system uses the following equation.

$$A_i = t/n * 100 \quad (5)$$

A_i is the accuracy of the system. The "t" is the number of corrected English sentence. The "n" is the number of tested English sentence. For performance measurement, this system is tested 200 English sentences. This system obtains the correct rate that is 85% and the error rate that is 15% respectively. Experimental result of the system is shown in figure 2.

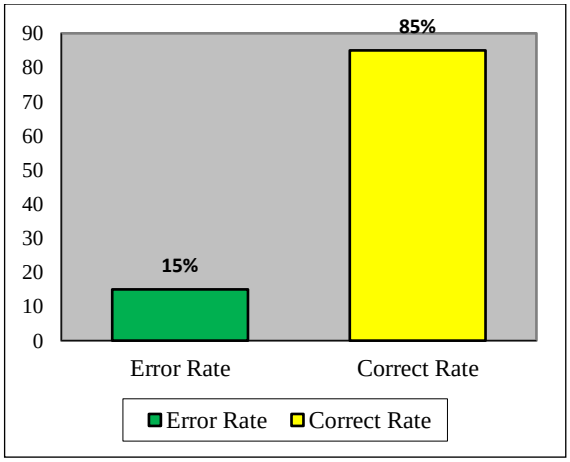

Fig. 2: Experimental Result

7. CONCLUSION

In conclusion, this system is developed based on the semantic oriented methodology. Thus, this system is useful not only to improve the performance of natural language processing (NLP) applications but also to find the correct sense of the word by using Tanimoto coefficient based WSD method. 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 word sense disambiguation system.

8. REFERENCES

[1] K. Samhith, S. A. Tilak and G. Panda, "Word Sense Disambiguation using WordNet Lexical Categories", International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), IEEE, pp. 1664-1666, 2016.

[2] D. Jianyong and L. Xia, "Attribute Knowledge Mining for Chinese Word Sense Disambiguation", IEEE, pp. 73-77, 2015.

[3] J. Hui and Z. Yangsen, "Study of Chinses Word Sense Disambiguation Based on Scenario Words", International Conference on Fuzzy Systems and Knowledge Discovery, IEEE, pp.2804-2808, 2012.

[4] S. Jumi and K. S. Shikhar, "Survey on Word Sense Disambiguation: An Initiative towards an Indo-Aryan Language", I. J. Engineering and Manufacturing, pp. 37-52, 2016.

[5] G. Sreelakshmi and P. H. Rosna, "Malayalam Word Sense Disambiguation using Naïve Bayes Classifier", International Conference on Advances in Human Machine Interaction, IEEE, 2016.

[6] Z. Chunxiang and H. Shan, "A Word Sense Disambiguation System based on Bayesian Model", International Conference on Computer Science and Network Technology, IEEE, pp. 124-127, 2015.

[7] C. Ling and Z. Yangsen, "Study on Word Sense Disambiguation Knowledge Base Based on Multi-Sources", IEEE, 2011.

[8] C. Ganesh and K. D. Sanjay, "A Literature Survey on Various Approaches of Word Sense Disambiguation", International Symposium on Computational and Business Intelligence, IEEE, pp. 106-109, 2014.

[9] P. H. Rosna, "Word Sense Disambiguation- A Survey", Proceedings of International Colloquiums on Computer Electronics Electrical Mechanical and Civil, 2011.

[10] G. Rohit, "A Survey on Word Sense Disambiguation", IOSR Journal of Computer Engineering, vol. 14, no. 6, pp. 30-33, 2013.

