Advanced Science and Technology Letters
Vol.86 (Ubiquitous Science and Engineering 2015), pp.77-80
http://dx.doi.org/10.14257/astl.2015.86.16

A Graph-based Word Sense Disambiguation Using Collocation

JungGil Cho

Division of Computer Engineering, SungKyul University

Abstract. Word sense disambiguation (WSD) is an essential part of Natural Language Processing (NLP) task to identify the senses of words in context. A new WSD approach based on collocation is proposed in this paper. In this approach, collocation is used to characterize the senses of words and the graph connectivity is measured by using various meanings which exist in the WordNet, when no collocation exists among the words.

Keywords: Collocation, Word Sense Disambiguation (WSD), Tree-based Association Rule (TAR), WordNet.

1 Introduction

The precise sense of an ambiguous word, which is general feature of all natural languages, can be selected from the context where words are located, and can be defined as the task in which inherent sense of polysemy word is allocated automatically in context. This task is called WSD, which figures out the precise senses of words in context.

The unsupervised WSD method is used in case of performing WSD without data learning, and is classified into graph-based ones and the similarity-based ones. In graph-based methods, a sense graph is built from the words in context, and it is processed to select the most appropriate meaning for each word from the created sense graph. Experimental comparisons between the two algorithm types indicate that graph-based algorithms outperform similarity-based ones, often by a significant margin [3].

In this paper, a new WSD approach is proposed which shows better performances over the conventional unsupervised WSD approaches. The suggested method establishes the framework to measure the best connectivity by increasing efficiency with the use of collocation before graph formulation. Moreover, the connectivity is evaluated suing the degree centrality algorithm in which connectivity measurement is proved as best. When using the proposed graph-based measurement approach combined with collocation, better accuracy with better precise meaning is resulted from the performance evaluation over the conventional approaches.

ISSN: 2287-1233 ASTL

2 Related Researches

Collocation refers to a group of practical words that habitually go together, whereas the sense of a word is figured out by accompanying words. In this case words are to be classified in terms of co-occurrence relation as well as sense. The co-occurrence relation means the constraints shown in the sense combination relation, which is called collocation constraint or selection constraint. Therefore, a collocation is a combination of words that occur more often and naturally than would be expected by random, and is divided into strong collocation in which closely-coupled words are used like nearly one word and weak collocation[4].

3 New Criteria of WSD

3.1 WSD Using Collocations

In this paper, WSD algorithms measure sense accuracy based on co-occurrences and collocation information. In case of using collocation information, the tendency to have one meaning for one collocation can be utilized since a word with ambiguity owns one sense in one literature. Hence, if a word with ambiguity includes collocation, WSD process can be handled more efficiently by having priority to use collocation information.

Collocation is classified into three types, idiom, phrasal verbs and collocation in terms of its characteristics. Especially, the syntactic patterns of collocation are described in details in Table 1.

Table 1. Syntactic Pattern Types of Collocation

Type	Grammar Type	Example
Type 1	adjective + noun	a huge profit
Type 2	noun + noun	a pocket calculator
Type 3	verb + adjective + noun	learn a foreign language
Type 4	verb + adverb	live dangerously
Type 5	adverb + verb	half understand
Type 6	adverb + adjective	completely soaked
Type 7	verb + preposition + noun	speak through an interpreter

3.2 Collocation-based WSD Algorithms

The WSD algorithms proposed in this paper are executed basically by the unit of sentence. Step-wise execution process of the algorithms, shown in Figure 1, is explained as below. At the 1st step, if a word with ambiguity includes collocation, WSD process can be handled more efficiently by having priority to use collocation information. Secondly, a graph G is built by using set G, and then a score of each vertex is counted. Finally, the sense of the word is determined from the label with the highest score out of sense of each word.

Input: Sequence : $W = (w_i \,|i = 1 \ldots N)$
Output: Sequence :$(L = (L_{w_i}|i = 1 \ldots N)$

1: Check the collocation and if it matches allocate sense
  1.1: Analyze morpheme
  1.2: Analyze by word class
  1.3: Test by category
  1.4: Collocation matching
2: If the collocation matches, go to stage 7
3: Use group G and create graph G in relation with WordNet
4: Mark peak score in graph G
5: Allocate sense with peak score
6: If there is no sense in graph G, allocate the original sense
7: Generate word sense

Fig. 1. WSD Algorithm based Collocations

4 Experiments and Results

This research used the NLTK (the Natural Language Toolkit), which is a strong natural language processing package of Python, to analyze the morpheme and process collocations. The data used in the experiment was conducted within data sets of English words in the Senseval-2, which analyzes the recent functions of the WSD system. When using the contextual words, which appears within the context surrounding ambiguous words, the window size of the context has to be considered. In this paper, considering the size of data sets, Windows 5 was selected as the default value.

Before constructing graph, if the collocation in the sentence is comprehended, it increases the connectivity measurement value. The WSD method using collocation can be considered as a good method regarding its processing speed and performance, if it is possible to extract sufficient amount of collocations. However, if the collocation rarely appears in literature, it is difficult to use WSD. Therefore, considering two dimensions, efficiency and performance; the word sense disambiguation was made using collocations, and for the words that could not use collocations, graph centrality algorithm was used for word sense disambiguation. By using this mixed method, this research was able to achieve good experimental results.

In Table 2, the precision of the methods suggested in this paper and that of the contrasting graph-based methods are compared.

Table 2. Comparison with related work

	[5]	[6]	This paper
Senseval-2 precision	59.54	62.2	66.5

5 Conclusion

In this paper, a new unsupervised WSD method was suggested that does not require tagged Corpus and shows better performance than the existing unsupervised WSD method. The algorithm of this paper increased efficiency and constructed graphs by using collocations to measure the best graph connectivity. Also, it measured the connectivity using degree centrality algorithms. Moreover, the assessment results imply that the method of measurement suggested in this paper, which is based on graph using collocations, yield a more accurate graph connectivity measurement value than the conventional methods.

References

1. Cho J. K., Shin K. C.: A Graph-based Word Sense Disambiguation using Measures of Graph Connectivity. In: KIIT, Vol. 12, No. 6, pp. 143-152 (2014)
2. Sinha R., Mihalcea R.: Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity. In: In IEEE International Conference on Semantic Computing (ICSC) (2007)
3. Navigli R., Lapata M.: An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. In: Pattern Analysis and Machine Intelligence, IEEE Transactions on, IEEE Computer Society (2010)
4. Lewis M.: Language in the Lexical approach. In: M. Lewis (Ed.), Teaching collocation: Further developments in lexical approach, pp. 126-153 (2000)
5. Agirre E., Soroa A.: Personalizing pagerank for word sense disambiguation. In: Proc. of EACL, pp.33-41 (2009)
6. Hessami E., Mahmoudi F., Jadidinejad H.: Unsupervised Graph-based Word Sense Disambiguation Using Lexical relation. In: International Journal of Computer Issues (IJCSI), Vol. 8, Issue 6, No 3, pp. 225-230 (2011)

