TUTA/IOE/PCU

Journal of the Institute of Engineering, Vol. 10, No. 1, pp. 192–198

Fax: 977-1-5525830

Word Sense Disambiguation using Clue Words

Udaya Raj Dhungana, Subarna Shakya

Department of Electronics and Computer Engineering, Central Campus, Pulchowk, IOE, TU, Lalitpur, Nepal

Corresponding Email: udayas.epost@gmail.com, drss@ioe.edu.np

Abstract: This paper presents a new model to disambiguate the correct sense of polysemy word based on the related context words for each different sense of the polysemy word. The related context words for each sense are referred to as clue words for the sense. The WordNet organises nouns, verbs, adjectives and adverbs together into sets of synonyms called synsets each expressing a different concept. In contrast to the structure of WordNet, we developed a model that organizes the different senses of polysemy words based on the clue words. These clue words for each sense of a polysemy word are used to disambiguate the correct meaning of the polysemy word in the given context using any WSD algorithm. The clue word for a sense of a polysemy word may be a noun, verb, adjective or adverb.

Keywords: Word Sense Disambiguation, WordNet, Polysemy Words, Synset, Hypernymy, Clue Words

1. Introduction

Words that express two or more different meanings when used in different contexts are referred to as polysemy or multi-sense words. Every natural language contains such polysemy words in its vocabulary. Such polysemy words create a big problem during the translation of one natural language to another. To translate the correct meaning of the polysemy word, the machine must first know the context in which the polysemy word has been used. Only after this, the machine can find out the correct meaning of the word in the particular context and can translate the meaning of that word into the correct word in another language. The process of finding the correct meaning of the polysemy word using machine by analyzing the context in which the polysemy word has been used is referred as Word Sense Disambiguation (WSD).

Although different methods have been tested to find the correct sense of the polysemy word at the given context in which the word is used, accuracy at satisfactory level has not been obtained yet. Among the different methods used for WSD, this research focused its study on the knowledge-based approach. The knowledge-based approaches use the resources such as dictionaries thesauri, ontology, collocation etc to disambiguate a word in a given context [6].

These days, the WordNet is becoming popular as a resource to be used in knowledge-based approach to disambiguate the meanings of polysemy words. WordNet is a lexical database developed at Princeton University [3] for English language. The WordNet organizes nouns, verbs, adjectives and adverbs into the groups of synonyms and describes the relationships between these synonym groups forming a semantic network among the words. After the development of English WordNet, many other WordNet on other languages Spanish WordNet, Italian WordNet, Hindi WordNet etc were built. These WordNet are used as one important resource to disambiguate the different meanings of a polysemy words in different languages.

To disambiguate the meaning of a polysemy word using the WordNet, the related words from synset, gloss and different levels of hypernym are collected from the WordNet database and these related words are compared to find the overlaps using different WSD algorithm. However, the collection of related words from synset, gloss and different level of hypernym that are taken from the WordNet contains only very few words that can be used to disambiguate the correct sense of a polysemy word in the given context (Dhungana U.R., 2008). This increases only the computational overhead for the WSD algorithm and for the system. Moreover, according to (Dhungana U.R., 2008), as we go increasing the levels of hypernym to collect the related words, the sense of the words becomes more general and the two different sense of a polysemy word are found to have the same hypernym. This therefore creates another ambiguity in ambiguity. To overcome this problem, we have developed a new model to organize both the single sense and multi-sense words.

2. Related Tasks

In [7], they have adapted original Lesk algorithm which was based on dictionary to use the lexical database WordNet. They have used Senseval-2 word sense disambiguation exercise to evaluate their system and found an overall accuracy of 32%. They have used a different counting method for the words overlapping in their adapted algorithm. To compare two glosses, they have used the longest sequence of one or more consecutive words that occurs in both glosses and for each overlap, a square of the number of words in the overlap is calculated and taken the sum of all overlaps.

In [5], authors had claimed that their work on automatic WSD using Hindi WordNet developed at IIT Bombay was the first attempt for Hindi language. They used statistical method for determining the senses and their system can disambiguate the nouns only. They have compared the context of the word in a sentence with the contexts constructed from the WordNet and choose the winner and evaluated their system on the Hindi corpora provided by the Central Institute of Indian Languages (CIIL). The accuracy of their algorithm was found in the range from 40% to 70%. They have used simple overlap method to determine the winner sense that is the sense with highest overlaps in count is the sense determined by the system in the given context.

In [9], authors have used overlap selection approach-the Lesk algorithm to disambiguate the Nepali ambiguous words. They had modified the Lesk algorithm as: only the words in the sentence as context words are compared with the glosses, examples and hypernym of the target word whose meaning is to be disambiguated. They did not use the glosses, example and hypernym of the context words to find the overlaps.

After the work of [9], Shrestha and Dhungana in [1], used the adapted Lesk algorithm with little modification using the sample Nepali WordNet developed by themselves. The experiments performed on 348 words (including the different senses of 59 polysemy words and context words) with the test data containing 201 Nepali sentences shows the accuracy level is 88.05%, which is found to be increased by 16.41%, in compared to the accuracy level of the earlier research [9].

3. Statement of Research Problem

3.1 WSD and WordNet

From analyzing many research works on WSD using WordNet such as [4], [7], [5], [9] and [1], it was noticed that the WordNet is not exactly suitable to use with knowledge-based, overlap selection WSD approaches. The reason is that the WordNet is built for general purpose in NLP tasks but not focused in WSD. In all WSD methods that used WordNet, the WordNet is used to take a large number of words to disambiguate the meaning of multi-sense word. However, only very few words taken from the WordNet are used to disambiguate the different senses of a multi-sense word. In this sense, all the efforts such as processing time for CPU and memory to store large number of unused words are wasted. Furthermore, it is noticed that the words taken from the WordNet to disambiguate the multiple meaning of the multi-sense word itself creates the ambiguity resulting in decrease in accuracy.

Another important point noticed from [1] is that when deeper levels of the hypernymy from the WordNet are used, the correctly disambiguated polysemy words are also incorrectly disambiguated.

Now let us take an example which explains the problems that can be found in WordNet to use in WSD methods. For this purpose, Hindi WordNet, which is same as the English WordNet in structure with some modification to adapt the need of Hindi language, is taken.

3.1.1 Example: WSD and Hindi WordNet

Let us consider a polysemy word ताल (read as Taal) in Hindi language. In Hindi WordNet, the word has eleven different meanings. The six meanings of the word ताल with the different levels of hypernymy of each meaning are shown in figure 1. First look at the figure 1 (a) and (b). Here we can see that the only the meaning of two different words are different. The hypernymy of the two senses are the same except the meaning 2 have one more hypernymy. In such case, there is no meaning of the use of the hypernymy of the two different senses of the polysemy word to disambiguate their meanings. Inclusion of these hypernymy is just waste of computational effort and waste of memory.

ताल Meaning 1	ताल गाने-नाचने -, बजाने आदि में उसके समय और क्रिया का परिमाण ठीक रखने का एक साधन “नर्तकी वादक को नृत्य की ताल समझा रही है यह राग / तीन ताल का है”
Hypernymy 1	मानव कृति, मानवकृति, मानवकृति-, मानव निर्मित वस्तु, मानवमानव - कृत वस्तु- बनाई या तैयार की हुई वस्तु द्वारा“यह मुगलकालीन मानव कृति है”
	वस्तु, चीज़, चीज "हवा एक अमूर्त वस्तु है" वास्तविक या कल्पित सत्ता -
	अस्तित्व, मौजूदगी, वज़ूद, वजूद, संभूति, विद्यमानता, सत्ता, हस्ती, भव, अस्ति, नमोंनिशान सत्ता का भाव -“कभीमें यह प्रश्न उठता है कि कभी हमारे मन- क्या ईश्वर का अस्तित्व है”

	भाव वह जिसमें होने की क्रिया निहित हो -“सुंदरता में सुंदर होने का भाव है”
Hypernymy 2	बोध, संज्ञान, ज्ञान, भान, संज्ञा, बोधि, अवबोध, अवगति, अवगम, अवभास वस्तुओं - मन या विवेक को होती है और विषयों की वह पूर्ण जानकारी जो“कन्याकुमारी में आत्मचिंतन करते समय स्वामी विवेकानंद को आत्म बोध हुआ”

(a) Meaning 1 of Word ताल

ताल Meaning 2	ताल - चश्मे के काँच का एक पल्ला "फ्रेम में ताल ठीक से नहीं बैठा है"
Hypernymy 1 (is a kind of …..)	मानव कृति, मानवकृति, मानवकृति-, मानव निर्मित वस्तु, मानवकृत वस्तु- - मानव द्वारा बनाई या तैयार की हुई वस्तु "यह मुगलकालीन मानव कृति है"
	वस्तु, चीज़, चीज - वास्तविक या कल्पित सत्ता "हवा एक अमूर्त वस्तु है"
	अस्तित्व, मौजूदगी, वज़ूद, वजूद, संभूति, विद्यमानता, सत्ता, हस्ती, भव, अस्ति, नमोंनिशान - सत्ता का भाव "कभीकभी- हमारे मन में यह प्रश्न उठता है कि क्या ईश्वर का अस्तित्व है"
	भाव - वह जिसमें होने की क्रिया निहित हो "सुंदरता में सुंदर होने का भाव है"

(b) Meaning 2 of Word ताल

What we need is that the organization of the polysemy words must be in such a way that it must only contains only such words in a efficiently organized way that these words are sufficient to disambiguate the different senses of the polysemy word.

Table 3. 1: The different three senses of Hindi word ताल with their hypernymy.

ताल Meaning 3	तालाब, ताल, पोखरा, सरोवर, सर, तड़ाग, पोखर, तलाब, पुष्कर, अरकासार, जल्ला, तोयाधार, पुष्करिणी, पुखर - पानी का बड़ा कुंड "अधिक गरमी के कारण इस तालाब का पानी सूख रहा हैतालाब में रंगीन कमल खिले हुए / हैं"
Hypernymy 1 (is a kind of …..)	जलाशय, अखात, झषनिकेत, जलाकर, मीनगोधिका, ह्रद, पूत, आबगीर, पर्परीक - वह स्थान जहाँ पानी जमा होकर ठहरा या बना रहता हो "जलाशय में कमल खिले हुए हैं"
	स्थान, जगह, स्थल, प्रदेश, आगार, केतन, निक्रमण, गाध, आस्थान, आस्पद, इलाका, इलाक़ा, प्रतिष्ठान - निश्चित और परिमित स्थितिवाला वह भूभाग जिसमें कोई - बस्ती, प्राकृतिक रचना या कोई विशेष बात हो "काशी हिन्दुओं का धार्मिक स्थान है
	भूभाग-, भूक्षेत्र-, भौगोलिक क्षेत्र, भूभाग - पृथ्वी का कोई बड़ा भाग या क्षेत्र

	"भारत एक ऐसा भूभाग है-, जहाँ नाना प्रकार की भाषाएँ बोली जाती हैं"
	क्षेत्र, इलाक़ा, इलाका, प्रदेश, प्रांत, प्रान्त, भूमि, दयार, फील्ड - जमीन का एक भाग "ग्रामीण क्षेत्रों में अभी भी बिजली की समस्या बनी हुई है"
H1	वस्तु, चीज़, चीज - वास्तविक या कल्पित सत्ता "हवा एक अमूर्त वस्तु है"
	अस्तित्व, मौजूदगी, वज़ूद, वजूद, संभूति, विद्यमानता, सत्ता, हस्ती, भव, अस्ति, नमोंनिशान - सत्ता का भाव "कभीकभी- हमारे मन में यह प्रश्न उठता है कि क्या ईश्वर का अस्तित्व है"
	भाव - वह जिसमें होने की क्रिया निहित हो "सुंदरता में सुंदर होने का भाव है"
H2	खंड, खण्ड, अंश, टुकड़ा, भाग, हिस्सा, अंग, विभाग, पुरज़ा, पुरजा, पुर्ज़ा, पुर्जा, कल, चरण, अंशक, भंग - उन अंगों या अवयवों में से कोई एक, जिनके योग से कोई वस्तु बनी हो "इस यंत्र के सभी खंड एक ही यंत्रालय में बने हैं इसके अगले / चरण में हम आपको एक नाटक दिखलाएँगे"
	भाग, हिस्सा, अंश, विधा, प्रभाग - समष्टि अथवा समूह का कोई अंश "इसका मध्य भाग कुछ मोटा है"
	अस्तित्व, मौजूदगी, वज़ूद, वजूद, संभूति, विद्यमानता, सत्ता, हस्ती, भव, अस्ति, नमोंनिशान - सत्ता का भाव "कभीकभी- हमारे मन में यह प्रश्न उठता है कि क्या ईश्वर का अस्तित्व है"
	भाव - वह जिसमें होने की क्रिया निहित हो "सुंदरता में सुंदर होने का भाव है"

(c) Meaning 3 of Word ताल

4. Problem Statement

Although WordNet is very useful lexical resource that can be used to disambiguate the different meanings of a polysemy word, it has still some limitations discussed on previous section. Therefore, to improve the accuracy of the knowledge-based overlap selection WSD methods dramatically, we have extremely need of a new logical model that could organize the words in such a way that it could disambiguate the meanings of a polysemy word correctly and efficiently resulting in higher accuracy than that can be obtained from the use of WordNet.

5. Solution

Unlike in WordNet, we organized the different senses of polysemy word and grouped them based on to which verb, noun, adverb, adjective, article, preposition etc each sense of a polysemy word generally used with, the resulting model (a new WordNet like structure but focused on WSD) contains all the necessary words/information that can sufficiently disambiguate each meaning of a polysemy word. The new model does not contain any unnecessary words/information which could itself again create ambiguity like WordNet.

For instance, a sample organization of different meanings of a polysemy word कडा in Nepali language can be as shown in figure 2. The polysemy word कडा has five different meanings [8] as shown in figure 3.

Table 3. 2: A sample organization of different meanings of a polysemy word कडा

Polysemy Word	Used with verb	Used with noun	Used with adjectives
कडा १	लगाउनु, पहिरनु	नाडी, खुट्टा, स्त्री, नारी, चाड, पर्व, तीज,	पहेँलो, सेतो,
कडा २	झुन्ड्याउनु, समाउनु	कराई, ताउलो, ताउली, भाँडा, बर्तन, भान्छा	

Experiment Settings

To check our new model, we set up the two experiments. In first experiment, we used the system developed by [2] using the normal sample Nepali WordNet. In second experiment, we only replaced the normal sample Nepali WordNet by our new model WordNet organized with clue words keeping all the settings constant.

6. Result and Discussion

After running the experiments, we found that the system with normal sample Nepali WordNet correctly disambiguated the polysemy words for 177 out of 201 test sentences. This shows the accuracy of the system with normal sample Nepali WordNet to be 88.059%. In the second experiment, the same system with our new model WordNet correctly disambiguated the polysemy words for 184 out of 201 test sentences. At this time, the accuracy of the system is found to be 91.543%. Using our new model WordNet that is specific to WSD is found to be increased by 3.484%.

Table1: Experiments Results

Experiment No.	No of polysemy words that are correctly Disambiguated	No of polysemy words that are not correctly Disambiguated	Accuracy in percentage
1	177	24	88.059
2	184	17	91.543

7. Conclusions

The WordNet organizes the words in the lexical database based on their meanings of the words instead of their forms as in dictionaries. It groups the nouns, verbs, adjectives and adverbs together into synonym sets, each expressing a distinct concept [10]. The words in a synonym set can be interchangeably used in many contexts. The main relationship among the words in WordNet is the synonym.

From analyzing many research works on WSD using WordNet such as [4], [7], [5], [9] and [1], we have noticed that the WordNet is not exactly suitable to use with knowledge-based, overlap selection WSD approaches. The reason is that the WordNet is built for general purpose in NLP tasks but not focused in WSD. In all WSD methods that used WordNet, the WordNet is used to take a large number of words to disambiguate the meaning of multi-sense word. But it is noticed that only very few words taken from the WordNet are used to disambiguate the different senses of a multi-sense word. In this sense, all the efforts such as processing time for CPU and memory to store large number of unused words are wasted. To defeat with these problems of WordNet, we developed a new model to mainly to organize the different senses of polysemy words using the clue word rather than using the hypernym relations or any other relation from the WordNet to disambiguate the sense of polysemy words. The experiments show that the accuracy of the system using our new model WordNet is found to be 91.543%.

Acknowledgements

We would like to kindly express our sincere thanks to Madan Puraskar Pustakalaya (MPP) for providing Nepali corpus. We are equally thankful to Mr. Balram Prasain, Lecturer, Tribhuwan University for his valuable help in this research.

References

[1] Dhungana, U. R. (2011). Nepali Word Sense Disambigution using Adapted Lesk Algorithm.

[2] Dhungana, U. R. (2012). Nepali WSD Specific WordNet. Pokhara: Pokhara Engineering College.

[3] George Armitage Miller, C. F. (n.d.). Wordnet. Retrieved June 12, 2012, from Princeton University: http//:wordnet.princeton.edu/wordnet/

[4] Lesk, M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionries: How to Tell a Pine Cone from an Ice Cream Cone.

[5] Manish Sinha, M. K. Hindi Word Sense Disambiguation.

[6] Nancy Ide, J. V. (1998). Word Sense Disambiguation: The State of the Art.

[7] Pedersen, S. B. An Adapted Lesk Algorithm for Word Sense Disambiguation using WordNet.

[8] Sharma, B. K. (BS 2064). Samchhipta Nepali Sabdasagar. Sabdartha Prakashan .

[9] Shrestha N., H. P. (2008). Nepali Word Sense Disambiguation Using Lesk Algorithm.

[10] WordNet A Lexical Database for English. (n.d.). Retrieved Dec 4, 2013, from What is WordNet?: https://wordnet.princeton.edu/wordnet/