SYNOPSIS OF

POLYWORDNET: A WORD SENSE DISAMBIGUATION SPECIFIC WORDNET OF POLYSEMY WORDS

A DISSERTATION
to be submitted by

UDAYA RAJ DHUNGANA

under the supervision of
PROF. DR. SUBARNA SHAKYA

for the award of the degree of

DOCTOR OF PHILOSOPHY IN
COMPUTER ENGINEERING

Department of Electronics and Computer Engineering
Pulchowk Campus, Institute of Engineering
Tribhuvan University

AUG 2017

1.0 Introduction

1.1 Ambiguity and Word Sense Disambiguation

This research focuses only on the lexical ambiguity in natural languages. The lexical ambiguity arises when a word has multiple meanings according to the different context. For example, "I go to bank". Here, the word "bank" has multiple meanings. The word "bank" may mean a financial institution or it may mean a bank of a river. Here, the two same spelled words "bank" and "bank" have the two different meanings. Such polysemy words called polysemy words create ambiguity for machine. The process of finding the accurate meaning of the polysemy word by machine by analyzing the context is referred as Word Sense Disambiguation (WSD). The two main WSD approaches are: corpus-based and knowledge-based. The WSD approaches, which are **corpus based**, utilize the information evidence obtained from corpus to disambiguate the ambiguous word. The **knowledge-based WSD** approaches use the resources such as dictionaries, thesauri, ontology, collocation etc to disambiguate the meaning of a polysemy word in a given context. **This research work focuses its study on the knowledge-based approach.**

1.2 Motivation to Research

The information in the dictionary were found to be insufficient to represent the context of a target word and to resolve the ambiguity since the dictionary only contains the short definitions of the words. With the development of the lexical database called the WordNet resolved the problem of lack of information. The WordNet provided more information using their semantic relations such as synonym sets, hypernymy/hyponymy, meronymy/holonymy relations etc. Pedersen et al adapted the Lesk algorithm by using the WordNet for and the accuracy of their WSD approach is increased from 16% to 32%. However, the accuracy at satisfactory level was not obtained till now. ***The motivation for this research work is originated from right this point.***

1.3 Hypothesis

It is strongly believed that if the senses of polysemy word and their corresponding contextual related words are interlinked, it resolves the problem of noise information that is introduced due to the use of common information from WordNet and increases the accuracy of the WSD approaches.

1.4 Research Objectives

1. To study the structure of existing WordNets and organization of words in these WordNets.
2. To investigate the factors/issues on these WordNets due to which WSD methods are lacking to obtain a satisfactory level of accuracy on WSD.
3. To design a new logical model to organize the senses of a polysemy word and the related words with each sense of the polysemy word so that by using this model, the accuracy level of WSD methods could be increased than using the WordNet.

1.5 Scope and Limitations

This research work limits its study in the contextual overlap-count knowledge-based WSD approaches and the use of the information from the WordNet for word sense disambiguation. The main limitation of this research is that small amount of information is used to build the sample WordNet and PolyWordNet and the size of the Test Data which contains 180 English sentences.

2.0 Literature Review

Lesk Michael in 1986 used the overlap of word definition from the *Oxford Advanced Learner's Dictionary of Current English* (OALD) to disambiguate the word senses [**1**]. He used only the definitions of the words that need to be disambiguated from the dictionary. After this, Guthrie et al. in [**2**] and Yarowsky in [**3**] also used this concept of dictionary based approach in their methods. The dictionary based approaches were promising. However, due to the lack of sufficient information in the dictionary, they were not giving better accuracy.

Banerjee and Pedersen in 2002 adapted the original Lesk algorithm using the information from WordNet and found an overall accuracy of 32% which was double of the original Lesk algorithm. In [**4**], authors used Hindi WordNet for word sense disambiguation in Hindi language. The accuracy of their algorithm was found in the range from 40% to 70%. Kostas et al. (2003) in [**5**] formed their sense bag and context bag by using the WordNet definition and the definition of all the hypernyms associated with the nouns and verbs in the sense definition. They also tested their system including the hyponymy relation but the experiment showed that there was no improvement in accuracy using the hyponymy. Shuang Liu et al. in [**6**] had also used the information from the synonyms, hyponyms, hypernyms to determine the senses of words by using the WordNet. Similarly, in other WSD approaches such as [**7**], [**8**], [**9**], [**10**], [**11**], [**12**], [**13**], [**14**] has used the information from synonyms, hyponyms, hypernyms relationship in their algorithms to disambiguate the senses of words.

3.0 Problem Statement

This research investigated three problems on the contextual overlap count WSD approaches:

1. **Insufficient Information in WordNet for Disambiguation:** The definitions in the WordNet still don't contain sufficient information for sense disambiguation.
2. **Noise Information and Wrong Disambiguation:** The use of common hypernym induces the noise information and this noise information causes the wrong sense disambiguation.
3. **Disambiguation depends on the gloss's words:** While defining the gloss of a word, the words, which are used to define their meaning, does not have any fixed rule. These words determine the number of overlaps with given context. This is not fair for all contexts.

4.0 Solution Approach

4.1 Related Words and Their Generation

This research believes that each sense of polysemy word has some distinct words related only to that sense and describe the sense. These words are called as related words for the sense. For example, the words "copy", "write" etc. are the related words for the sense "writing implement with a point from which ink flows" of polysemy word "pen". To find out the related words for a sense of a polysemy word, take the sense as an entity and find out the following information (if applicable) for that sense:

1) All possible **a**ttributes (A)
2) All possible **f**unctions (F)
3) All entities it contains or all its constituent **p**art (P)
4) All entities with which its functions are **r**elated (R)
5) All entities with which it is **u**sed for (U)
6) All entities that **d**escribe it or its function or the way of doing its function (D)
7) All entities along with it **o**ccurs (O)

The related words are generated using the above AFPRUDO rules.

4.2 Organization of words: PolyWordNet- A New Lexical Database

These related words are inter-linked with their corresponding senses of polysemy words. The resulted new lexical database is named as 'PolyWordNet'. In PolyWordNet, each related word is linked only with a sense of a polysemy word. If a word is equally semantically related with more than one sense of the same polysemy words, it is just ignored and is not included in the related words of either sense.

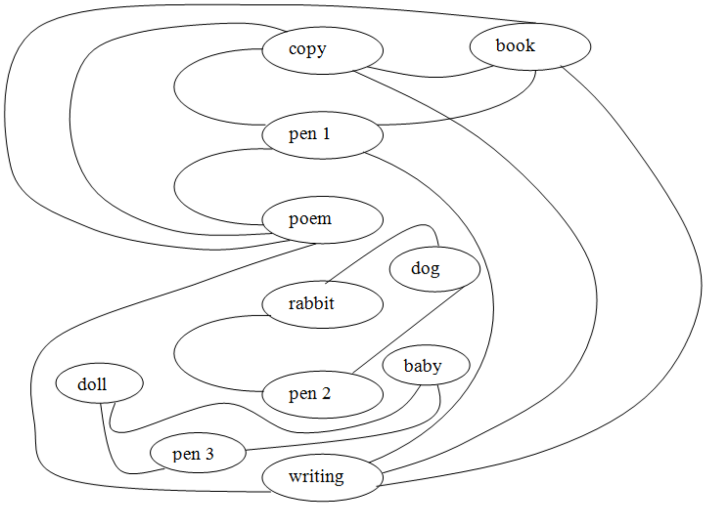

Fig. 3. 1: A sample of organization of words in PolyWordNet

The words in WordNet are connected to the multiple senses of a polysemy word. This condition introduces the ambiguity in ambiguity and hence produce noise information. These are resolved in PolyWordNet by linking one related word only with a single sense of a polysemy word.

4.3 Disambiguation Process: A New WSD Algorithm

The developed new WSD algorithm searches the paths or links of context words with a sense of the target word. If the paths thus obtained connect the context words only with one sense of the target word, the algorithm outputs the linked sense as the correct sense of the target word. If there are paths that link more than one senses, then the algorithm counts the number of paths or links or connection for each linked sense. Then the sense for which the number of connection paths is maximum is selected as a correct sense. If the two or more senses have the equal number of connections, the first sense in the array maintained by algorithm is selected as correct sense. If no connection is found, the algorithm displays an information indicating the failure of disambiguation.

5.0 Research Design: Experimental

The hypothesis of this research has formulated a relation between the independent and dependent variables. It is tested using experiments to observe the effect of the independent variables on dependent variable. Therefore, **experimental research design** is chosen to carry out appropriate experiments to confirm or refute the stated hypothesis. All together four experiments are designed and are described in the following subsections.

a) Experiment 1: Experimental Setting

The points to be noted in this setting are:- 1) use of the simplified Lesk algorithm, 2) use of information only from relationships:- synset and gloss in the WordNet to form the sense and context bags and 3) for second run, the information in sense bags is increased by including the hypernyms of words in each sense bag.

b) Experiment 2: Experimental Setting

The points to be noted in this setting are:- 1) use of the simplified Lesk algorithm, 2) use of information from relationships:- synset, gloss and hypernyms in the WordNet to form the sense and context bags and 3) for second run, the information in sense bags are increased by including the hypernyms of words in each sense bag.

c) Experiment 3: Experimental Setting

The points to be noted in this setting are:- 1) use of the simplified Lesk algorithm, 2) use of information from relationships:- synset, gloss, hypernyms, hyponyms and meronyms in the WordNet to form the sense and context bags and 3) for second run, the information in sense are increased by including the hypernyms of words in each sense bag.

d) Experiment 4: Experimental Setting

The points to be noted in this setting are 1) use of new WSD algorithm, 2) use of PolyWordNet instead of WordNet and 3) use of direct inter-linked relations between senses of polysemy word and the corresponding related words.

5.0 Data Collection

To collect the required data, ***telephone interview, questionnaire*** *and* ***documents*** methods of data collections are used in this research. *To build sample WordNet and PolyWordNet,* 20 polysemy words are selected from internet, research papers and by telephone interview with a English lecturer. Then, the different senses and other related relationships- synset, gloss, hypernyms, hyponym and meronyms are collected from the WordNet 2.1. To collect the test data, 4 set of questionnaire each set containing 5 polysemy words are prepared. Then, the respondents are asked to write a simple sentence for each sense listed in the questionnaire. They are also asked to write the possible words which can be used with each sense to collect the related words.

6.0 Result and Analysis

The Fig. 5.1 shows the accuracies obtained in all four experiments. The first three experiments has two runs A and B. The result of Experiment 4, which uses the PolyWordNet, shows the highest accuracy. The 173 test sentences out of 180 are correctly disambiguated giving the accuracy of 96.11% which is significantly higher than the accuracies found in the first three experiments.

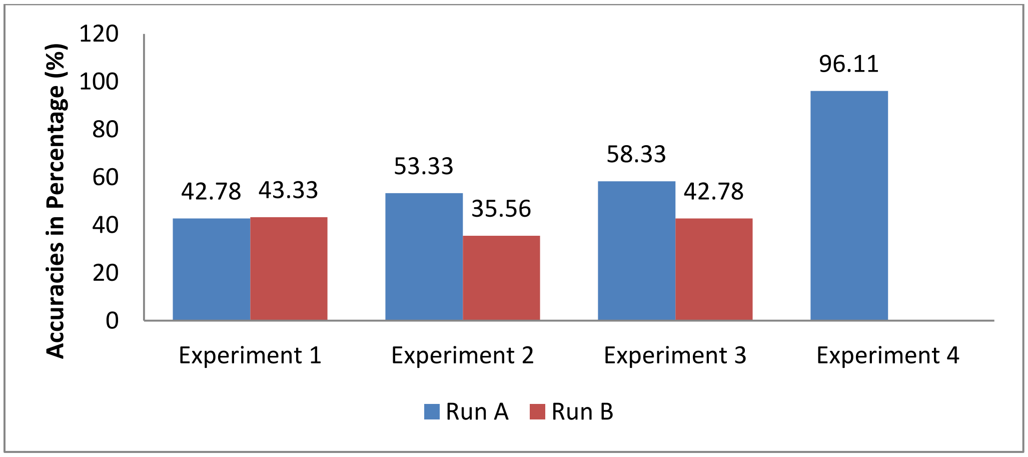

Fig. 5. 1: Accuracies obtained in experiments

From the observations of results obtained in the experiments, it is found that when only the synset and gloss are used, it contains less information for sense disambiguation. When the information in both context and sense bags are increased, the accuracy is found to be increased. However, at the same time, it is observed that correctly disambiguated test sentences are incorrectly disambiguated by significantly large number when the information from the WordNet are increased for sense disambiguation. This proves that the knowledge-based contextual overlap count

WSD methods that use the information from WordNet for sense disambiguation suffer from the noise information. The reason is that in WordNet, the words that are used as context words, are connected to the multiple senses of a polysemy word. This noise information increase the overlaps between the incorrect sense of a target word with the context resulting in the wrong disambiguation. This is resolved in PolyWordNet by linking one related word only with a single sense of a polysemy word. Therefore, the accuracy of the new WSD algorithm which uses the PolyWordNet is significantly found higher than that of contextual overlap count WSD algorithms.

7.0 Conclusions

The purpose of this research and all the stated objectives are successfully achieved. This research concludes that the contextual overlap count WSD approaches which use the information from WordNet for sense disambiguation suffer from noise information. The noise information causes the wrong disambiguation. This is resolved by organizing the senses of polysemy words and their corresponding related words in the new lexical database called PolyWordNet. The results of experiments show the significantly higher accuracy (96.11%) of this new WSD algorithm than that of the accuracy (58.33%) of the contextual overlap count WSD approaches which uses the information from WordNet. Hence, the stated hypothesis is proved.

8.0 Recommendation

1. A detail research in relation between the senses of a polysemy words and prepositions is recommended so that the prepositions can also be included in PolyWordNet as related words.
2. It is highly recommended to research for automatic generation of related words.

9.0 Significance of Research

1. The organization of the words in **PolyWordNet resolves the problem of noise information** which is produced with the use of information from WordNet.
2. Using the PolyWordNet, **the new WSD algorithm gives higher accuracy** for sense disambiguation in NLP tasks such as Machine Translation, Information Extraction, Text Summarization etc with a higher accuracy.

10.0 Contributions of Research

1. PolyWordNet is a completely **new** lexical database.
2. The organization of the words in **PolyWordNet** is **completely different**. The dictionary organizes the words based on the alphabetical order and the WordNet organizes the words based on synonym set. PolyWordNet organizes the words based on the relation between a sense of polysemy word and its related words.
3. The WSD algorithm which uses information from PolyWordNet for sense disambiguation is a completely **new** WSD algorithm.

References

[1] Michael Lesk, "Automatic Sense Disambiguation Using Machine Readable Dictionries: How to Tell a Pine Cone from an Ice Cream Cone," in *Proceedings of the 5th Annual International Conference on Systems Documentation*, Toronto, Ontario, Canada, 1986, pp. 24-26.

[2] Joe A. Guthrie, Louise Guthrie, Yorick Wilks, and Homa Aidinejad, "Subject dependent co-occurrence and word sense disambiguation," in *The 29th Annual Meeting of the Association for Computational Linguistics*, 1991, pp. 146-152.

[3] David Yarowsky, "Word-sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora," , vol. 2, Nantes, France, 1992, pp. 454-460.

[4] Manish Sinha, Mahesh Kumar, Prabhakar Pande, Lakshmi Kashyap, and Pushpak Bhattacharyya, "Hindi word sense disambiguation," in *Proceedings of International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems*, Delhi, India, 2003.

[5] Kostas Fragos, Yannis Maistros, and Christos Skourlas, "Word sense disambiguation using wordnet relations," in *First Balkan Conference in Informatics*, Thessaloniki, Greece, 2003.

[6] Shuang Liu, Clement Yu, and Weiyi Meng, "Word Sense Disambiguation in Queries," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, Bremen, Germany, 2005, pp. 525-532.

[7] Kiyoaki Shirai and Tsunekazu Yagi, "Learning a Robust Word Sense Disambiguation Model Using Hypernyms in Definition Sentences," in *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, 2004.

[8] S G Kolte and S G Bhirud, "WordNet: a knowledge source for word sense disambiguation," *International Journal of Recent Trends in Engineering*, vol. 2, no. 4, 2009.

[9] S. G. Kolte and S. G. Bhirud, "Exploiting Links in WordNet Hierarchy for Word Sense Disambiguation of Nouns," in *Proceedings of the International Conference on Advances in Computing, Communication and Control*, Mumbai, India, 2009.

[10] Arindam Roy, Sunita Sarkar, and Bipul Syam Purkayastha, "Knowledge Based Approaches to Nepali Word Sense Disambiguation," *International Journal on Natural Language Computing (IJNLC)*, vol. 3, no. 3, pp. 51-63, 2014.

[11] Hee-Cheol Seo, Hoojung Chung, Hae-Chang Rim, and Sung Hyon Myaeng, "Unsupervised word sense disambiguation using WordNet relatives," *Computer Speech & Language*, vol. 18, no. 3, pp. 253-273, July 2004.

[12] Andres Montoyo, Armando Suárez, German Rigau, and Manuel Palomar, "Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods," *Journal Of Artificial Intelligence Research*, vol. 23, no. 1, pp. 299-330, March 2005.

[13] Udaya Raj Dhungana and Subarna Shakya, "Word sense disambiguation in Nepali language," in *Digital Information and Communication Technology and it's Applications (DICTAP), 2014 Fourth International Conference on*, Bangkok, Thailand, May 2014, pp. 46-50.

[14] Udaya Raj Dhungana, Subarna Shakya, Kabita Baral, and Bharat Sharma, "Word Sense Disambiguation using WSD specific WordNet of polysemy words," in *Semantic Computing (ICSC), 2015 IEEE International Conference on*, California, USA, Feb, 2015, pp. 148-152.

List of Publications

1) Papers Published in **International Journal**

1. Udaya Raj Dhungana and Subarna Shakya, "Hypernymy in WordNet, Its Role in WSD and Its Limitations," *International Journal of Simulation, Systems, Science & Technology*, vol. 16, no. 6, December 2015. **Link:** http://ijssst.info/Vol-16/No-6/cover-16-6.htm

   Award: Selected as one of the **BEST PAPER** from *CICSyN*, Riga, Lativa 2015
2. Udaya Raj Dhungana, Subarna Shakya, K. Baral, and Bharat Sharma, "Word Sense Disambiguation using WSD specific WordNet of polysemy words," *International Journal on Natural Language Computing (IJNLC)*, Vol. 3, No. 4, Aug, 2014, pp. 148-152.

   Link: http://airccse.org/journal/ijnlc/vol3.html

2) Papers published in **IEEE Xplore Digital Library**

3. Udaya Raj Dhungana and Subarna Shakya, "Polywordnet: A lexical database," in *Inventive Computation Technologies (ICICT), International Conference on*, vol. 2, 2017, pp. 1-7.
4. Udaya Raj Dhungana and Subarna Shakya, "Word sense disambiguation using PolyWordNet," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, vol. 2, 2017, pp. 1-6.
5. Udaya Raj Dhungana and Subarna Shakya, "Hypernymy in WordNet, Its Role in WSD, and Its Limitations," in *The 7th International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN)*, Riga, 2015, pp. 15-19.
6. Udaya Raj Dhungana, Subarna Shakya, Kabita Baral, and Bharat Sharma, "Word Sense Disambiguation using WSD specific WordNet of polysemy words," in *Semantic Computing (ICSC), 2015 IEEE International Conference on*, California, USA, Feb, 2015, pp. 148-152.
7. Udaya Raj Dhungana and Subarna Shakya, "Word sense disambiguation in Nepali language," in *Digital Information and Communication Technology and it's Applications (DICTAP), 2014 Fourth International Conference on*, Bangkok, Thailand, May 2014, pp. 46-50.

   Link: http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&
   queryText=Udaya%20Raj%20Dhungana