Word Sense Disambiguation by Domain Specification using Babelnet

[1] Risha Ranganath Dhargalkar, [2] Kavita Asnani

[1] Department of Information Technology, Goa University,
Padre Conceicao College of Engineering,
Verna - Goa, INDIA 403722

[2] Department of Information Technology, Goa University,
Padre Conceicao College of Engineering,
Verna - Goa, INDIA 403722

Abstract - Over the decades many approaches have been carried out in order to resolve the problem of Word sense disambiguation (WSD).WSD is a process of identifying correct sense of polysemy word. The current research trends in Babelnet, which is huge knowledge base which integrates Wikipedia and WordNet along with WordNet domain. In this paper, we propose an algorithm which automatically identifies the domain of each word. Senses of a word have been annotated with Babelnet which follows state-of-art approach. Further the algorithm has been proposed for converting SemCor into gold standard corpus. The analysis has been done on this gold corpus and results shows that average precision obtained is 0.5 and recall obtained is 0.4.

***Keywords* - Natural Language Processing, Word Sense Disambiguation, GETALP system**

1. Introduction

Human language is ambiguous, so that many words can be interpreted in multiple ways depending on the context in which they occur. For instance, consider the following sentences:" She is going to the bank to withdraw money. She is going to take a walk along the river bank.

The occurrences of the word bank in the two sentences clearly denote different meanings: Financial bank and a sloping land, respectively. 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 inorder to determine the underlying meaning. The computational identification of meaning for words in context is called word sense disambiguation. Different approaches have been proposed to perform WSD.They are categorized as supervised, Unsupervised and Knowledge Based approach. The supervised approach requires training corpus and creating such sense annotated corpus is expensive and difficult. [5]. Unsupervised approach is based on text itself and senses are assigned to word from text itself but problem is it does not have external lexical resource [5]. Knowledge based WSD follows graph based state-of-art graph approach by using information from external lexical resource. In this paper, we proposed an algorithm which automatically identifies the domain of each word. The proposed algorithm uses knowledge based lexical resource Babelnet1.1.1 [14] using state-of-art graph approach. Babelnet consists of labelled directed graph where nodes represent concepts or named entities and edges represent semantic relation between them. It follows graph based approach and integrates Wikipedia and WordNet [14]. Further algorithm has been proposed for converting SemCor data into gold standard corpus. The analysis is done on our gold standard SemCor corpus by selecting health domain files by using GETALP system [21]. This paper is organized as follows. Related work on WSD and GETALP system in Section 2. Proposed method is described in section 3 followed by proposed algorithm for converting SemCor into gold standard corpus in section 4.Section 5 provides experimental results followed by section 6 which concludes paper along with direction for future work followed by acknowledgement and references in section 7 and 8.

2. Related Work

Word Sense Disambiguation is a process of finding the sense of polysemy word. Lesk 1986 [11] involves looking for overlap between the words in dictionary definitions with words from the text surrounding the word to be disambiguated. Sense whose overlap is largest is selected as the target sense. Unfortunately, the algorithm determines overlaps only among the glosses of the senses being considered. This is a significant limitation in that dictionary glosses tend to be fairly short and do not provide sufficient vocabulary to relate fine-grained sense distinctions. Banerjee and Pederson in [8] came up with an adapted Lesk algorithm in which they considered a window of context containing of a few words to the right & left of the target word and only applied the gloss overlaps method to words within this window of context.

Kolte and Bhirud [6] Proposed methodology for Word Sense Disambiguation based on domain information. Set of words where there is strong semantic relation is called as Domain of set of .The words in the sentence contributes to determine the domain of the sentence. WordNet domains help in determining the domain of words. The lexical resource used was “WordNet Domain” which is an extension of English WordNet. Wei Lee and Mit [5] Proposed WSD approach with Domain knowledge. In this approach, WordNet is adopted which serves as an external knowledge source in order to gain the information about the knowledge of domain. This approach represents a combination between knowledge source WSD approach and Unsupervised WSD approach.

Agirre and Lacalle [3] Proposed the application of knowledge based Word Sense Disambiguation systems to specific domains, based on state-of-the-art graph based WSD system that uses the information in WordNet. Knowledge-based systems exploit the information in a Lexical Knowledge Base (LKB) to perform WSD, without using any corpus evidence. Graph-based algorithm using Personalized PageRank is used which outperformed other knowledge-based WSD systems in publicly available datasets.

Mitesh Khapra [1] proposed method which is based on supervised system with far or less demand on annotation. This approach is not restricted to any specific target words. The interest of this paper is adaptation setting which is applied and tested on two domain specific corpus ie Health and Tourist and other mixed corpus SemCor. Agirre and Lacalle [4] Showed that WSD system trained on a general source corpus (BNC) and the target corpus, obtains up to 22% error reduction when compared to a system trained on the target corpus alone.

David Yarowsky [9] Proposed unsupervised learning algorithm based on two constraints: One sense per collocation: Nearby words provide strong and consistent clues to sense of target words conditional on relative distance, order and syntactic relationship .One sense per discourse: Sense of target word is highly consistent with any given document.Yee chan and Ng [2] proposed Active Learning Algorithm/Method is used which is used for selecting words and adding it to target domain. Timothy Baldwin [10] Reconsiders task of MRD-based WSD, in extending basic Lesk algorithm to investigate impact on WSD performance of different tokenization schemes, scoring mechanisms, methods of gloss extension and filtering methods.Roberto Navigli and Simone Paolo Ponzetto [14] proposed Babelnet API which can be accessed by few lines of code in order to perform multilingual WSD. Roberto Navigli, David Jurgens and Daniele Vannella [15] proposed SemEveal 2013 task for multilingual WSD.The corpus is annotated with babelnet 1.1.1. Didier Schwab, Andon Tchechmedjiev, Jérôme Goulian,Mohammad Nasiruddin, Gilles Sérasset, Hervé Blanchon proposed [20]GETALP system for performing analysis on babelnet using SemEvel-2013 task.

3. Proposed Method

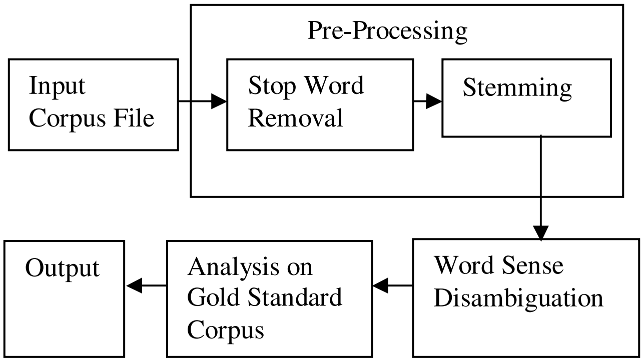


Fig 1. Proposed System Architecture

3.1 Algorithm

```
1. Input corpus file, which will be list of string of tokens.
2. Remove stop words from input file.
3. Perform stemming on processed file.
4. .For each of token from processed file(sense tag based on context)
    1. Identify the senses of word from babelnet.
    2. Find all connecting paths in whole network.
    3. .Merge identified paths inorder to obtain paths from source to destination.
    4. Score the graph nodes using out degree measure.
    5. Display all the senses of each node along with their scores, in descending order.
5. Display the domain of word, synsetID of each word is mapped with WordNet Domain database consisting of WordNet synsets annotated with domain labels.
```

6. Perform analysis of input file using GETALP system by using gold standard corpus which has been designed.

4. Algorithm for Converting SemCor into Gold Standard Corpus

Didier Schwab, Andon Tchechmedjiev [20] proposed the GETALP system to perform analysis on babelnet using SemEval corpus .This system works by adapting the Lesk measure propagated through an Ant Colony Algorithm. The main shortcoming of their system was it works only on SemEval corpus and it requires gold standard corpus to do analysis. Creating such gold corpus is challenging and difficult task.

In our work we have proposed an algorithm for converting SemCor into gold standard Corpus.

4.1 Algorithm

1. Add quote characters around tag parameters, for example <wf cmd=ignore pos=DT>The</wf> should become <wf cmd="ignore" pos="DT">The</wf>
2. Remove all <p […]> and </p> tags
3. Transform <s snum=YYY> </s> tags to <sentence id="dXXX.sYYY"> </sentence> where dXXX is a document number and YYYthe index of the sentence in that document.
4. Transform <punct> [...] </puct> to just [...]
5. <wf cmd=done pos=VB lemma=retain wnsn=0 lexsn=2:40:01 ::> retained</wf> this corresponds to an annotated word. Pos are the part of speech, wnsn is the sense number in wordnet, lexsn is the sense id from wordnet, and lemma is the root form of the word.
6. For gold standard we need an id, which can build it in the following way: dXXX.sYYY.tZZZ, XXX is a text number.YYY is the sentence number and corresponds to the index of the sentence in the current document, ZZZ is the index of the word in the sentence, so the second sentence of the first document will have as id "d001.s002.t003".
7. Furthermore the pos tag should be simplified to one letter, essentially anything that start with N should become n, anything starting by V should become v, anything starting with J will be a, Anything starting with R should become r. The rest should be ignored.
8. Finally gold standard is achieved, by concatenating lemma and lexsn like so: lemma%lexsn and associating it to the id and should put that in another file.

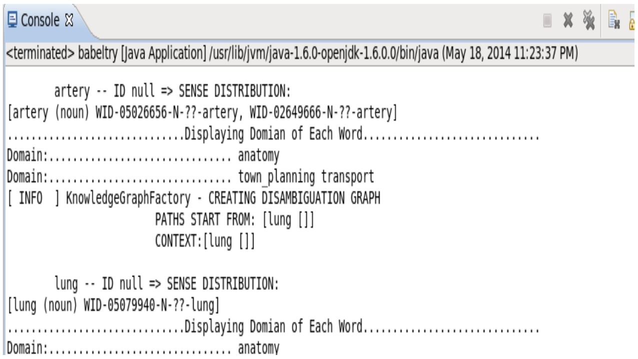
5. Experimental Result

5.1 Experimental Setup

Experiments are evaluated on Intel core duo 2 2.20GHz processor with 3GB RAM Fedora (Linux) O.S.Programs are implemented in java in eclipse IDE .Algorithm is applied on health domain files.

5.2 Observation and Results

Results are obtained for health corpus files from SemCor corpus. For each of ambiguous words it displays the domain based on context.



```
<terminated> babeltry [Java Application] /usr/lib/jvm/java-1.6.0-openjdk-1.6.0.0/bin/java (May 18, 2014 11:23:37 PM)

        artery -- ID null => SENSE DISTRIBUTION:
[artery (noun) WID-05026656-N-??-artery, WID-02649666-N-??-artery]
..............................Displaying Domian of Each Word..............................
Domain:............................... anatomy
Domain:............................... town_planning transport
[ INFO  ] KnowledgeGraphFactory - CREATING DISAMBIGUATION GRAPH
                         PATHS START FROM: [lung []]
                         CONTEXT:[lung []]

        lung -- ID null => SENSE DISTRIBUTION:
[lung (noun) WID-05079940-N-??-lung]
..............................Displaying Domian of Each Word..............................
Domain:............................... anatomy
```

Fig 2. Output of Disambiguation

Analysis is done on GETALP system by using gold standard SemCor corpus. The analysis is done by using parameters like precision, recall and F-measure.

Table 1. Analysis of health domain files on gold standard SemCor corpus

No of health domain files	Precision	Recall	F-Measure

1	0.63968	0.63968	0.63968
2	0.52857	0.43622	0.44632
3	0.56308	0.45865	0.50552
4	0.56845	0.46980	0.51444
5	0.58341	0.47005	0.52063
6	0.58410	0.44859	0.50745

5.3 Result Analysis

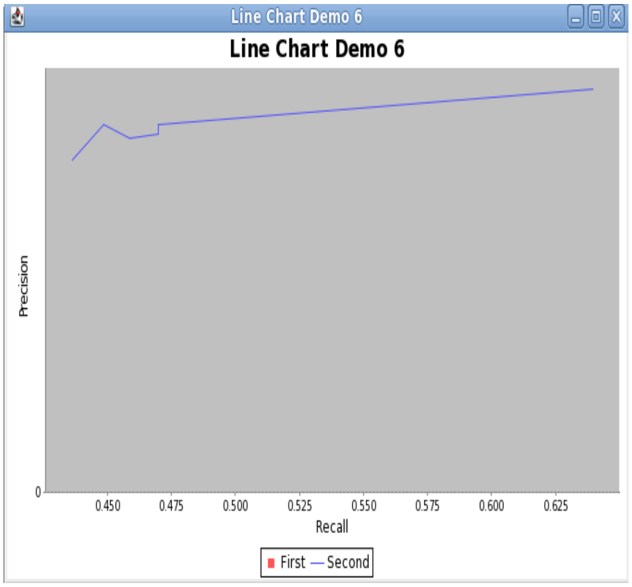


Fig 3. Recall v/s Precision

6. Conclusion and Future Work

Earlier there were no enough and proper repository for performing WSD but recently lot of progress has been done in developing knowledge bases. Supervised approach had problem because it required annotated corpus. These entire problems have been solved by using huge repository like babelnet which follows knowledge based system. The results obtained for disambiguation using this repository outperforms other supervised and unsupervised technique. Futher by creating gold standard corpus analysis was done using GETALP system and average precision obtained is 0.5 and recall is 0.4 for six domain files of health corpus.

In future, Babelnet can be used to perform multilinguality of data and analysis can be done by using any other corpus by using GETALP system by creating gold standard corpus.

Acknowledgment

This work was performed as part of a Master of Engineering thesis in project Word Sense Annotation and Disambiguation. We want to acknowledge Andon Tchechmedjiev for support and guidance for corpus creation.

References

[1] Mitesh M. Khapra ,Anup Kulkarni ,Saurabh Sohoney ,Pushpak Bhattacharyya .2010 All Words Domain Adapted WSD: Finding a Middle Ground between Supervision and Unsupervision.In Proceeding ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA.

[2] Yee Seng Chan and Hwee Tou Ng June 2007. Domain Adaptation with Active Learning for Word Sense Disambiguation In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 49–56, Prague, Czech Republic.

[3] Eneko Agirre and Oier Lopez de Lacalle and Aitor Soroa 2009. Knowledge-Based WSD on Specific Domains: Performing Better than Generic Supervised WSD. In Proceedings of the 21st international joint conference on Artifical intelligence, Pages 1501-1506 San Francisco, CA, USA.

[4] Eneko Agirre and Oier Lopez de Lacalle 2009.Supervised Domain Adaption for WSD.In EACL '09 Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pages 42–50 Stroudsburg, PA, USA, Association for Computational Linguistics.

[5] Wei Jan Lee, Edwin Mit, 2011.Word Sense Disambiguation by Using Domain Knowledge. In Proceedings of International Conference on Semantic Technology and Information Retrieval,pages 237-242, Putrajaya, Malaysia.

[6] S. G. Kolte, S. G. Bhirud, 2008.Word Sense Disambiguation using WordNet Domains. In Proceedings of Eighth Mexican International Conference on Artificial Intelligence ,pages 1187-1191 India,Mexico

[7] Sandhya Sachidanandan, Prathyush Sambaturu, and Kamalakar Karlapalem. May 13th 2013, Named Entity Recognition on tweets using Wikipedia. In Proceedings MSM2013 Workshop Concept Extraction Challenge, Rio de Janeiro, Brazil.

[8] Satanjeev Banerjee, Ted Pedersen 2002.An Adapted Lesk algorithm for word sense disambiguation using Wordnet.In Proceedings of Third International

Conference, University of Minnesota, pages 136-145,55812, Duluth, MN, USA.

[9] David Yarowsky 1995.Unsupervised word sense disambiguation rivaling supervised methods .In ACL '95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics, Pages 189-196, Stroudsburg, PA, USA.

[10] Timothy Baldwin 2007 MRD-based Word Sense Disambiguation: Further#2 Extending#1 Lesk Hikari-dai, Seika-cho Soraku-gun, Kyoto Japan.

[11] Jonas EKEDAHL, Koraljka GOLUB Word sense disambiguation using WordNet and the Lesk algorithm KnowLib, Dept. of IT, Lund Univ,Lund, Sweden.

[12] Word Sense Disambiguation: A Survey By Roberto Navigli.

[13] Word Sense Disambiguation by Esha Palta Under the guidance of Prof. Om Damani Kanwal Rekhi School of Information Technology.

[14] Roberto Navigli and Simone Paolo Ponzetto, 8-14 July 2012 “Multilingual WSD with Just a Few Lines of Code: the BabelNet API”, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 67–72, Jeju, Republic of Korea,2012 Association for Computational Linguistics.

[15] Roberto Navigli, “A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches”, M. Bielikov´a et al. (Eds.): SOFSEM 2012, LNCS 7147, pp. 115–129, Springer-Verlag Berlin Heidelberg, 2012.

[16] Roberto Navigli, “Word Sense Disambiguation: A Survey”, ACM Computing Surveys, Vol. 41, No. 2, Article 10, February 2009.

[17] Roberto Navigli, Simone Paolo Ponzetto, “BabelNet: Building a Very Large Multilingual Semantic Network”, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 216–225, Uppsala, Sweden, July 2010.

[18] Roberto Navigli, Simone Paolo Ponzetto, “BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network”, Artificial Intelligence 193 (2012) 217–250, August 2012.

[19] Roberto Navigli and Simone Paolo Ponzetto, “Joining Forces Pays Off: Multilingual Joint Word Sense Disambiguation”, Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1399–1410, Jeju Island, Korea, July 2012.

[20] Didier Schwab, Andon Tchechmedjiev, Jérôme Goulian,Mohammad Nasiruddin, Gilles Sérasset, Hervé Blanchon June 14-15, 2013 “GETALP: Propagation of a Lesk Measure through an Ant Colony Algorithm” Proceedings of Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 232–240, Atlanta, Georgia,. 2013 Association for Computational Linguistics.

[21] Navigli, Roberto, David Jurgens, and Daniele Vannella. "Semeval-2013 task 12: Multilingual word sense disambiguation." Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (* SEM 2013). 20.

Risha R Dhargalkar is a ME student of Information Technology from Goa University. She is graduated with BE in Information Technology from Goa University, India in 2008. She has around 4+ years of teaching experience. Her research interest includes data mining, natural language processing. She is currently based in Goa, India.

Kavita Asnani is Head of Department of Information Technology Department at Padre Conceicao Engineering College affiliated to Goa University, Goa. She received her Masters degree in Information Technology from Goa University, Goa, India. She has 13 years of teaching experience at College level. She has published many papers in International and National Journals, and also at International and National Conferences. Her area of research includes Data Mining, Information Retrieval and Distributed Systems.