CSE- 628

NATURAL LANGUAGE PROCESSING

RESEARCH REPORT

Submitted by:Animesh Gupta (109768122)

Problem Overview:
The research papers presents various algorithms and their variations to tackle the problem of identifying correct meaning/sense for a target word depending on the context in which it occurs. This process is termed as Word Sense Disambiguation (WSD). The input for the problem would be sentences from "WordNet", the target-ambiguous word and its position in the input sentence.The output contains the predicted definition or sense from the WordNet synset.This WSD problem is considered an AI-complete problem, that is, a task whose solution is at least as hard as the most difficult problems in artificial intelligence.Thus, the motivation for the papers is to solve the ambiguity of words by proposing techniques which achieves better accuracy than earlier implemented algorithms and baseline models.

Research Paper I: **An Adapted Lesk Algorithm for Word Sense Disambiguation**
The paper is based on the adaption of Lesk's dictionary-based word sense disambiguation algorithm.The novel idea in this paper is to incorporate the rich hierarchy of semantic relations between words using WordNet database. As WordNet is arranged semantically not like traditional alphabetically arranged dictionaries,it provides lexical database of nouns,verbs,adverbs and adjectives. Moreover, Lesk's approach is very sensitive to the exact wording of definitions, so the absence of a certain word can radically change the results.This paper uses some common relations described in WordNet like synset, hyponymy,hypernymy, meronym etc. to calculate the combination score for every candidate combination i.e. pair of words including target word in the predefined context window. Based on the scores among the various senses of the target word, maximum scored sense is predicted as the output for the given target word in the input sentence.

Results: The author claimed 32% of overall accuracy evaluated on the English lexical dataset from the Senseval-2 as compared to 16% and 23% accuracy attained by variations of Lesk Algorithm.The huge improvement claimed in the paper is mainly attributed to the fact of using Wordnet which has highly interconnected set of relations among synonyms instead of traditional Oxford Dictionary.

Research Paper II: **Using Measures of Semantic Relatedness for WSD**
This paper is the generalisation of the previous paper by the same author. It is based on the generalised adaption of Lesk's dictionary- based word sense disambiguation algorithm.It also compares the various other algorithms to determine the semantic relatedness.The paper improved Adapted Lesk version by specifying some conditions which are not specfied in the previous paper.The context window is taken as 3, that means only left and right words are considerd for determining the relations. Also,only relations between nouns is considered not other categories like verbs,adverbs and adjectives.Other algorithms includes Leacock-Chodorow measure, which is based on the shortest path between 2 noun-contexts in a "IS-A" hierarchy, scaled by the depth of the hierarchy which is fixed 16 for all nouns. Another algorithm is Resnik measure which is based on the formulation of information content of concept, which is frequency count for that concept in a large corpus and then estimating probability via maximum likelihood estimate.The author chose to vary this original algorithm by skipping the division by number of senses for that word,as otherwise it

would give higher score to the words with less senses.The other main algorithm is Jiang-Conrath Measure which is the combination of previous 2 algorithms i.e. combining Information content with the notion of path length between concepts.All these algorithms give scores which can be used for determining the semantic relatedness between pair of concepts.

Results:The author claimed to achieve best results with gloss Adapted Lesk Algorithm and Jiang-Conrath algorithm having accuracy equals to 0.391 and 0.380 respectively.The results given in the paper are in line with the author claims as other algorithms shown slightly less accuracy compare to these.But as only 29 ambiguous noun words taken, with more dataset the result might be different as differences are not much to draw any conclusions which author made.

Research Paper III: **Word Sense Disambiguation using Conceptual Density**
This paper proposes novel technique based on "Conceptual density" as a solution to the lexical ambiguity problem.It says,given a window size, the program moves the window one noun at a time from the beginning of the document towards its end, disambiguating in each step the noun in the middle of the window and considering the other nouns in the window as context. Non-noun words are not taken into account.The conceptual density is calculated for all hypernyms of all senses of the nouns in context. The highest conceptual density among all the synsets determines a set of sense choices: the senses included in its sub hierarchy are chosen as interpretations of the respective words in context. The rest of senses of those words are deleted from the hierarchy, and the procedure is then iterated for the remaining ambiguous words. The input for the system is the dataset from Semcor which is brown corpus, with words tagged by WordNet word senses And output is same as the maximum scored word sense for the target word.

Results:The author claimed to attained overall Precision of 64.5 % and Recall of 55.5% in relation with nouns at sense level. Overall results are poor in comparison with other similar algorithms but author supported his results by saying his approach makes fine-grained distinctions not coarse grained distinctions like previous algorithms ,on all the senses around more than 9000 nouns in WordNet.Further,author implemented these similar algorithms on his datasets which shows one algorithms is equal and other perform poorly which supports author claim.

Comparison:
Similarity:All papers based on the Knowledge based techniques to identify the correct meaning for the ambiguous word. Common use of some external database like WordNet,Senseval-2. The papers proposes idea on identifying the semantic relation between the neighbour words of the target word in input sentence.Paper II and Paper III,both based on determining the semantic relation using "IS-A" hierarchy relation between the words.

Differences:The Paper II improves Adapted Lesk version by specifying some conditions which are not specified in the previous Paper I.The context window is taken as 3, that means only left and right words are considered for determining the relations. Also,only relations between nouns is considered in Paper II not other categories like verbs,adverbs and adjectives as proposed by Paper I. Paper III extended the concept of Path given in Paper III by calculating density for the noun word in middle of context window.

Pros: Paper I and Paper II are practical to implement and efficient in comparison to the original Lesk Algorithm.Use of WordNet as external database make it relevant to the problems related to ambiguity and finding relations between the group of words in the sentences. Paper III uses SemCor as database which is tagged corpus from Wordnet and proposes automatic evaluation of correct sense for target word without manual tagging of data or training any process, which is also applicable for Paper I and Paper II, as algorithms proposed in all the papers are knowledge based only.

Cons:Paper I limits the solution with just one variation of the algorithm..Paper II presents only existing algorithms with minor variations and included promising ideas like combining two different algorithms in the future work only.Paper III provides overview on determining the score using Conceptual Density and failed to produce any significant improvement over existing algorithms.Moreover, all papers rely on WordNet for their knowledge but Wordnet has only limited relations between words. So the results are not fine-grained sensing particularly for Paper I and Paper II. Also, only looking relations between neighbour words does not give correct sense for the target word, sometimes we even need to look entire context or previous sentences to identify the correct sense.

Scenarios: All the papers can be used to tackle the scenarios where training dataset is limited or not available and input sentences with ambiguous word will be processed at run time only with the help of tagged database like WordNet.The methods presented in the papers can be used for the applications related to information retrieval or machine translation, but due to limited accuracy can only used as the building blocks for more sophisticated algorithms.

Relation to my project:
The Paper I and Paper II describes various practical algorithms which can be implemented and are also the part of my project to be used for WSD, which I will be comparing against baseline models like Most Frequent Sense and Random Sense. Paper III describes the term Conceptual density, and Paper II describe one algorithm based on path length, I will be using new algorithm to combine both for my project.The difficulty in using the wordnet for the path length and gloss overlap are specified by the paper, because in Wordnet there is hierarchy and concepts higher in hierarchy are more general than the concepts in lower words, so equal path length of 1, suggests different meaning depending upon the level. After analysis of various techniques, I will be experimenting by applying these techniques as a knowledge base to train my model for Supervised learning. The feature vector generated is of the form [feature-value,sense,score]. The suitable values for “feature-value” can be group of words in context window and “score” will be generated based on the various algorithms described in the research papers.