Word Sense Disambiguation using WordNet Lexical Categories

Samhith.K*1
Affine Research Associate
Affine Analytics Pvt Ltd
Email:sk29@iitbbs.ac.in

Arun Tilak.S*1
School of Electrical Sciences
IIT Bhubaneswar
Bhubaneswar, India
Email:as27@iitbbs.ac.in

Prof G.Panda
School of Electrical Sciences
IIT Bhubaneswar
Bhubaneswar, India
Email: gpanda@iitbbs.ac.in

***Abstract*—In this paper a methodology for disambiguating the word senses of polysemous words using Lexical Categories present in WordNet is presented .WordNet is a commonly used English lexical database. The algorithm is applied to the data scraped from Wikipedia articles. The representative context used in the algorithm is extracted from the Wikipedia pages of the words belonging to the category. The lexical category of the given word is determined using the words in its neighbouring context. After finding the lexical category of the word, the correct sense is found using a modified version of Lesks Algorithm. The output word sense correspond to those available in WordNet.**

***Keywords*—Polysemous,Wikipedia,Homonyms,WordNet, Word Sense Disambiguation**

I. Introduction

Word Sense Disambiguation (WSD) is the task of identifying the correct sense of a word in a given context. WSD is an important intermediate step in many Natural Language Processing (NLP) tasks especially in Information extraction, Machine translation and Question and Answering Systems. It also plays an important role in improving the relevance of a search query, providing a meaningful reply for a chatbot etc. The problem of word sense disambiguation has been described as AI-complete, that is, a problem which can be solved only by first resolving all the difficult problems in artificial intelligence(AI), such as representation of common sense and knowledge in encyclopedia. Word sense ambiguity arises when a word has more than one sense. Words which have multiple meanings are called homonyms or polysemous words. And these senses of the word vary depending on the context of their use. For example, consider the two sentences below:

1) **Click the mouse twice to execute the program**
2) **The mouse got caught in the trap**

The word mouse clearly has different senses in the two sentences above. In the first sentence it means the computer mouse used to move the cursor in computers and in the second it means the rodent. The distinction might be clear to the humans but for a computer to recognize the difference it needs a knowledge base or needs to be trained.There are four conventional approaches to Word Sense Disambiguation:

A. Unsupervised methods

Unsupervised methods make use of only raw unannotated corpora and do not exploit any sense tagged corpus to provide a sense choice for a word in context.These methods are context clustering ,word clustering and cooccurence graphs.

B. Semi-supervised methods

These make use of secondary source of knowledge such as small annotated corpus as seed data in bootstrapping process. It actually overcomes the main problems associated with building a classifier:the lack of annotated data and the data sparsity problem.

C. Supervised methods

Supervised methods make use of sense annotated corpora to train from.These use machine learning techniques to learn a classifier from labeled training sets.Some of the common techniques used are decision lists,decision trees,naive bayes,neural networks,support vector machines(SVM).

D. Knowledge based methods

Knowledge based methods rely on dictionaries, thesauri and lexical knowledge bases.

In this paper a knowledge based approach is employed, where Wikipedia articles of words belonging to different lexical categories are used as knowledge base .Representative context from these articles is extracted and they are used to distinguish the word senses.

II. Related Work

The problem of WSD is so difficult that it was one of the reasons why the Machine Translation systems were abandoned[6].But after 1980,Word Sense Disambiguation gained attention after availability of large scale lexical resources and corpora. Since then significant amount of work on Word sense disambiguation has been achieved. Our method presented in this paper relates most to that of David Yarowsky(1992).They associate the sense of word to Rogets categories. The contexts representative of the category are collected from Groliers Encyclopedia. The results were shown for 12 polysemous words. Our approach extends this method for words having multiple senses within the same category.There are words in

* The first two authors contributed equally

the language that come under the same Roget category but still have different senses.We propose steps to eliminate the disambiguity among these words using a modified version of Lesks Algorithm.Also the lexical categories available in wordnet are used instead of Roget categories.

III. About WordNet

WordNet is a manually-constructed lexical system developed by George Miller at the Cognitive Science Laboratory at Princeton University. It reflects how human beings organize their lexical memories. It can be considered as a combination of dictionary and thesaurus. It differs from traditional dictionaries and thesaurus in many ways. For example , words in WordNet are arranged semantically instead of alphabetically. The basic building block of WordNet is synset consisting of all the words that express a given concept. Synonymous words are grouped together to form synonymous sets or synsets. WordNet stores information about words that belong to four parts of speech: nouns, verbs, adjectives and adverbs. Each of these are divided into certain lexical categories. For example, Nouns contain many subdivisions like body, communication, object and more. The verbs contain sub divisions like cognition, communication and more .The lexical categories within noun ,verb , adverb and adjectives were used in our approach to recognize the sense of the word.

IV. Problem Formulation

In case of some polysemous words, a lexical category may contain more than one sense of same word as well. For example the word program has multiple meanings. It could imply a computer program or TV show or academic college program or the act of creating .The notable senses of the word program observed in WordNet are given in the table T1.In the above case the creation sense of the word can be distinguished easily as it belongs to different lexical category(verb.create) but the other sense all belong to noun.communication category and so to resolve the disambiguity among these senses a varied form of Lesks algorithm is used. The problem was broken down into two parts and the solutions were then combined. First the correct lexical category of the word is identified, then the correct sense of the word within that category is determined. We employ the method proposed in a paper by David Yarowsky for identifying lexical categories. But the lexical categories from WordNet were used instead of the Rogets categories. Once the lexical category is determined a modified form of Lesks algorithm is used to determine the appropriate sense.

V. Method

A. Finding the category

In this step the lexical category of a word is identified. The method works based on few observations: Different senses of word tend to appear in different contexts Different word senses tend to belong to different lexical categories. On basis of the above two observations, if one were able to distinguish between the lexical categories in a given context then one could thereby distinguish word senses in that context. Consider the word mercury, it has two different senses. It could either mean the planet or the element. These both belong to different lexical categories in WordNet .noun.substance implies mercury as an element and noun.object contains mercury as planet. Hence identifying the category of Mercury in a context can help us to pinpoint at its sense. To train models in natural languge processing, huge amount of data is required.Instead of using books or articles or large corpora,Wikipedia articles were used.Since Wikipedia is a huge repository of unannotated but organised text, it is used in this paper.Wikipedia is a web based free content encyclopedia project.

TABLE I: Word senses and corresponding categories of the word ”program”

Lexical Category	Meaning
verb.creation	write a computer program
noun.communication	an integrated course of academic studies
noun.communication	a radio or television show
noun.communication	(computer science) a sequence of instructions that a computer can interpret and execute

1) Collect contexts representative of the category: Wordnet has 45 lexical categories ,to which all of its words can be categorised. Contexts representing these categories is collected. The context is collected from the Wikipedia articles of the words in each of the lexical categories. From these article concordances of 100 surrounding words for each occurrence of the polysemous word in the article is extracted. Two important preprocessing must be done before including these words.They are lemmatization and removing stop words. Lemmatization-It is the technique of reducing words to their grammatical roots. A word in its base form or dictionary form is called a lemma. This however differs from stemming which does not take into account the context. Stemming reduces word forms to (pseudo)stems and lemmatization reduces the word forms to linguistically valid lemmas. Also lemmatization requires part of speech tag of the word as it is dependant on the context. For example, the word ”better” on stemming remains ”better” whereas the same word on lemmatization becomes ”good”. Removing stop words-Words which occur too frequent in articles are not useful for the purpose of classification. Such words frequently used as stop words and should be filtered out.These mainly consist of articles, prepositions and conjunctions. For example an,the,against,and.

For optimal results the concordances should only include references to the given category.But due to occurrence of polysemous words inside the article ,there may be some spurious examples. While the level of noise introduced through polysemy is substantial, it can usually be tolerated because the spurious senses are distributed through the 44 other categories, whereas the signal is concentrated in just one. Only if several words had secondary senses in the same category would context typical for the other category appear signicant in this context.

TABLE II: Results

Word	Sense	Category	Frequency	Accuracy per sense
sack(verb)	discharge from position	verb social	117	97.4 %
	plundered	verb possession	54	98.1 %
behaviour	psychology	noun act	119	97.4 %
	reaction to something	noun state	95	97.8 %
	behavioral attributes	noun attribute	76	96.5 %
band	insrumentalists	noun group	32	87.5 %
	bind or tie together	verb contact	116	95.6 %
	jewelry	noun artifact	72	95.8 %
	range of frequencies	noun communication	136	98.5 %
amaze	bewildering effect	verb cognition	51	94.11 %

2) Identify the salient words in the collective context: Next salient words in the collective context are identified and their weights are calculated .Salient words are those that appear more often in the context of a category and hence is a better than average indicator for the category. The mathematical estimate used to denote salience of word :

$$\frac{Pr(w/LCat)}{Pr(w)}$$

,the probability of a word appearing in the context of a Lexical Category divided by its overall probability in the collected articles.The log of this salience

$$\log(\frac{Pr(w/LCat)}{Pr(w)})$$

, is considered as the words weight in the statistical model of the category.

3) Use the weights to predict category for word in novel text: Next the resulting weights were used to decide the category for a word in text. The occurrence of salient words in the context of an ambiguous word supports the fact that the word belong to that category. If several such words occur then the evidence is further supported. Using Bayes rule ,the sum of the weights of all the words in a context window is calculated and the category for which the sum is greatest is considered as the correct one.

$$ARGMAX_{Lcat} \sum_{w} \frac{Pr(w|LCat) * Pr(LCat)}{Pr(w)}$$

We have taken the context as 20 words to the right and 20 words to the left of the polysemous word.

B. Identifying the correct sense

After narrowing down on the category, the correct sense of the word is found out using modified Lesks algorithm. According to Lesk Algorithm ,the dictionary definition or gloss of a words senses are compared to the dictionary definition of every word in the context of the word in the given text. The word is assigned the sense whose gloss shares the largest number of words in common with the glosses of the other words. The Lesk Algorithm restricts its comparison to just the dictionary meanings of the word being disambiguated, in our approach we also compare the meanings of the word that are present in the dictionary definition of disambiguated word with its context .The sense that has the maximum overlap is selected.

VI. Results

The model was tested against the tagged corpus obtained from SENSEVAL . The tags of senseval were not tagged using WordNet Lexical categories. Hence they were hand tagged and the tests were run. The data from the test runs are depicted on the table .The table contains results for 4 corpora containing different senses for the words amaze, band, behaviour, sack.

VII. Conclusion

This paper has described an approach to perform word sense disambiguation using lexical categories in WordNet using a modified method proposed by David Yarowsky and a modified Lesk algorithm. The accuracy rates are very good and are much more than conventional WSD methods. We can hereby state that the rise of structured data in internet is creating more opportunity for increasing the accuracy of WSD engines.

Acknowledgment

This work greatly benefited from the technical and financial assistance provided by Affine Analytics Pvt Limited.

References

[1] George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[2] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[3] David Yarowsky. Word Sense Disambiguation using Statistical Models of Rogets Categories trained on large corpora

[4] William A. Gale.Kenneth W. Church.David Yarowsky .One Sense Per Discourse

[5] David Yarowsky . Unsupervised Word Sense Disambiguation rivaling supervised methods.

[6] Roberto Navigli. Word Sense Disambiguation :A Survey

[7] M.Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of SIGDOC 86,1986.

[8] S.G.Kolte, S.G Bhirud .Word Sense Disambiguation using WordNet Domains

[9] Nancy Ide and Jean Veronis. Introduction to the special issue on word sense disambiguation: The state of the art. Association for Computational Linguistics, 24(1):

[10] Ying Liu,Peter Scheuermann,Xingsen Li and Xingquan Zhu.Using WordNet to Disambiguate Word Senses for Text Classification

[11] Sruthi Sankar K P, P C Reghu Raj, Jayan V. Unsupervised Approach to Word Sense Disambiguation in Malayalam . In Proceedings of International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015)

[12] Jean Veronis,Nancy M.IDE .Word Sense Disambiguation with Very Large Neural Networks extracted from Machine Readable Dictionaries

[13] William A. Gale,Kenneth W. Church,David Yarowsky.A Method For Disambiguating Word Senses in a Large Corpus