AN HYBRID APPROACH TO WORD SENSE DISAMBIGUATION WITH AND WITHOUT LEARNED KNOWLEDGE

Roshan Karwa[1] and Manoj Chandak[2]

[1] M.Tech Scholar, CSE Department, SRCOEM, Nagpur, India
[2] Associate Professor and Head, CSE Department, SRCOEM, Nagpur, India

ABSTRACT

Word Sense Disambiguation is a classification of meaning of word in a precise context which is a tricky task to perform in Natural Language Processing which is used in application like machine translation, information extraction and retrieval, automatic or closed domain question answering system for the reason that of its semantics perceptive. Researchers tried for unsupervised and knowledge based learning approaches however such approaches have not proved more helpful. Various supervised learning algorithms have been made, but in vain as the attempt of creating the training corpus which is a tagged sense marked corpora is tricky. This paper presents a hybrid approach for resolving ambiguity in a sentence which is based on integrating lexical knowledge and world knowledge. English Wordnet developed at Princeton University, SemCor corpus and the JAWS library (Java API for WordNet searching) has been used for this purpose.

KEYWORDS

Natural language processing, world knowledge, knowledge sources, JAWS & Wordnet

1.INTRODUCTION

In current era, almost the searching of any kind is relying on internet and it has been found that the result of search is not what actually needed. The search engines sometimes give data relevant to the context and sometimes give irrelevant data. Such thing happens because, while querying for certain data, there is likelihood that the query may contain ambiguous words or the words having multiple meaning. For example, Play in English can either have a drama or dramatic play meaning or a sport meaning. Identifying the relevant sense of a polysemous word i.e. word having multiple meaning is involuntary in human being but for a machine, it is difficult as computer do not have a basis of commonsense knowledge. The task of identifying the appropriate sense of an ambiguous word in a sentence is known as Word Sense Disambiguation [13] Disambiguating a word needs two things: Dictionary having a list of senses of ambiguous word i.e. semantic relations of a polysemous word and Corpus (Real World Text) consisting real world knowledge. It is difficult for system or even to human being to identify the correct sense without a sense repository i.e. one or more type of knowledge sources. There are two types of a knowledge sources, first is corpus which is tagged or untagged with the word sense, and former is dictionaries like a Wordnet related to machine readable dictionaries etc. Steps in resolving ambiguous word is to first identifying the set of ambiguous word in a sentence, then a algorithm is to be applied which will make a use of knowledge sources like syntactic, semantic etc to find out a relevant sense. Several WSD techniques have been proposed in the past ranging from

DOI : 10.5121/ijnlc.2015.4203

knowledge based to supervise to unsupervised methods. Supervised and unsupervised rely on corpus evidence. Researcher proposed many WSD techniques ranging like knowledge based, supervised, unsupervised and semi supervised learning, but all are limited and have some disadvantages.

Example of word ‘bank’ ambiguity in sentence:

1.John is going to rob bank.
Ambiguous words are Go, Rob, Bank in this sentence. Here the word “Bank” sense is financial institution.

2.John is planning to take a toddle beside the river bank.
Ambiguous words are Plan, Bank, Toddle in this sentence. Here the word “Bank” sense is not a financial but a sloping land.

Example of word ‘tree’ ambiguity in sentence

1.This tree is from ancient times.
Ambiguous word is “tree” in this sentence and the sense is of trunks and branches.

2.Tree diagram should be present in thesis.
Ambiguous words are tree, present and diagram and the correct sense of tree is a logical programming flowchart.

This paper is organized as follows: Section 2 comprises related works in WSD; section 3 comprises knowledge sources; section 4 comprises open source tools used; section 5 is of proposed approach and implementation details, section 6 comprises experimental details and last section is of conclusion and future work is in last section.

.

2. PAST ACCOUNTS OF WSD

After facing problems related to natural language processing, Researchers proposed a various approaches to overrule the problems, some approaches are based on dictionaries and some are on the corpus evidence.

Knowledge based approach are depend on the Knowledge resources like Wordnet which is a dictionary, different thesaurus. They are also referred as Dictionary based approach. To get the correct sense, knowledge based depends on the dictionaries. Agirre et al. [1], 1996 proposed Word Sense Disambiguation with Conceptual Density method. This method’s basic idea is to select a sense based on the conceptual distance i.e. how the ambiguous word and its context word are related. This result is later extended by the same researcher i.e. Agirre et al.[2], 2001 with the change approach to find the correct sense, they called the approach as the Selectional preference method. This method look for the probable associations between word categories, simplest measure for this word to word relation is frequency count. Overlap based approaches like Lesk, Extended lesk are purely a based on the matching of word and contexts words. This approach is suggested by Satanjeev Banerjee, Ted Pedersen [14], 2002. Basic problems with this approach is it is heavily depends on dictionaries, which is also have some restrictions over acquiring the common sense knowledge.

Machine learning approaches are purely based on the corpus which is tagged or tagged, Supervised and unsupervised are come under the machine learning. Supervised WSD learning uses tagged corpus which includes training and testing module, while training, preprocessing has to be done first and then applying some trained algorithm and to test an unknown sample based on trained data. Naïve baye's, Decision list, Support vector machine are some of the supervised approaches. Naive baye's learning approach is a mathematical as to find the correct sense; it depends on the simple conditional probability calculation, consisting of feature as collocation, co-occurrence, part of speech (Gerard Escudero et al.[7], 2000). Decision list algorithm is simple if else then approach; most appropriate feature in decision list is one sense per collocation (Agirre, E. and Martinez, d. 2000). Algorithm which is mostly depend on the examples is Exemplar-based learning researched by a same researcher of Naïve baye's (Gerard Escudero et al.[7], 2000) . Then the latest algorithm is Support Vector Machines (SVM) which is based on the binary classes, based on the irrelevant and relevant senses, it separates the classes (Navigli and roberto, 2009[10]) are some of supervised approach. Main problem is with supervised is the effort of generating the manually tagged corpus.

To overcome the disadvantage of supervised that is generating a manually creating corpus which is tagged one, researcher proposed the unsupervised methods. Mihalcea and Moldovan, 2001[9], used this unsupervised approach i.e. corpus which is untagged called feature selection method. This method is automatic in nature, researcher then tried for another unsupervised approach i.e. based on the rank system which is Personalized PageRank algorithm as in (E. Agirre and A. Soroa [4], 2009)**,** Similarity-based algorithms as in (R. Navigli and M. Lapata [10], 2010) are some of unsupervised approach. A clear disadvantage is that, so far, the performance of unsupervised systems lies a lot lower than that of supervised systems due to a cluster issues.

3.KNOWLEDGE SOURCES: WORLD KNOWLEDGE

Knowledge sources employed for Word Sense Disambiguation are two: one is lexical knowledge in form of frequency or other measures and the other is world knowledge, also refer as Common sense knowledge can be acquired through a training corpus.

3.1 Lexical Knowledge

Lexical knowledge is associated with a dictionary**.** It is used in all approaches i.e. knowledge based, supervised and mainly in unsupervised one. Sense Frequency, Sense Gloss, Concept trees, selectional restrictions is some of the components of lexical knowledge

3.2 World Knowledge

World Knowledge is all possible associative links, which mined out of corpora are considered to be a part of Lexical meaning. For instance, word Airplane can be associated to vehicle, fly, pilot, plane crash, aerobus, Wright brothers etc. World knowledge, also we can refer it as Common sense knowledge is complex to understand as a result we can attain that knowledge from the dictionary like a WordNet by incorporating the JAWS i.e. JAVA API for Wordnet searching library. We can also get world knowledge through the training corpus by using machine learning algorithms. With the exact ambiguous word, along with its commonsense knowledge and the words in the environment of target ambiguous word can provide as the clue of target senses.

4.OPEN SOURCE TOOLS

For this project, we use a number of open source tools that are referenced all the way through the project. The tools and resources include WordNet, a java interface of WordNet, a part of speech tagger, SemCor. Some of these tools, WordNet for example, provide the definitions and relations. Some resources, like SemCor, provide examples of correctly translated text. The sections below explain what each tool/resource is and how this project uses them.

4.1 WordNet

WordNet is a publicly available lexical database developed by Princeton University (Miller). There are 206941 words across 117659 SynSets, which are groups of synonyms, in WordNet 3.0. This means that there are 117659 unique definitions available. This project uses wordnet for getting the correct definition of sense. Fellbaum.1998 [5], all of semantic relations about words is why WordNet is a lexical database. WordNet is so rich in information and so well executed that it is one of the most familiar tools for word sense disambiguation.

4.2 WordNet Interface

Wordnet interface is interface to WordNet using the Java API to WordNet. English nouns, verbs, adjectives and adverbs are prepared into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. Different methods like getDict, getLemma, getIndexTerms, getSynonym, getSynset etc. are used for different purposes. WordNet is helpful while identifying whether the word is ambiguous or not.

4.3 SemCor Corpus

Princeton University developed Semcor, which originates from the Brown Corpus (Princeton University, 2011). The SemCor files contain over 20,000 tagged words across 352 files. Every tag contains the part of speech, the lemma, and the correct WordNet sense. This makes SemCor extremely useful for researchers using WordNet. SemCor provides a professionally tagged resource to compare the accuracies of word sense disambiguation algorithms. As data is unstructured, to make it suitable, we have used JAVA's DOM parser and transformed the corpus XML Format to 4 forms Word Form, POS, Lemma, and Word Sense Number

4.4 Part Of Speech Tagger

We have used Stanford's MaxentTagger for the purpose of part of speech. For that, we use it through JAVA library. There are two taggers in this package; one is bi-directional dependency network tagger whose accuracy was calculated 97.32% and second tagger given by Maxent tagger is by using only left second-order sequence information whose accuracy mentioned was 96.92%. We have used Java API: A MaxentTagger can be made with a constructor taking as argument the location of parameter files for a trained tagger. It is giving a proper part of speech, later they plays very important role in disambiguation.

4.5 JAWS Library

Java API for WordNet searching (JAWS) is a library that we use for retrieving relations from wordnet lexical database. It is maintained by CSE department at Southern Methodist University. We can provide knowledge of ambiguous terms by retrieving information associated with it i.e. its word form, phrases etc.

5. PROPOSED APPROACH AND IMPLEMENTATION

We proposed Hybrid Approach which will integrate all knowledge sources i.e. to use corpus evidence as well as semantic relations from Wordnet in terms of World Knowledge. We termed our approach as Hybrid as we are integrating Knowledge sources Syntactic (POS, Morphology, and Collocations), Semantic (Word associations), Features and lexical resource (Target Word specific feature), Verb object syntactic relation. The steps for our purpose of disambiguation is

- Preprocessing of Corpus
- Feature Extraction
- Preprocessing of Target Sentence
- Classification Task
- Finding correct sense(Removing ambiguity)

5.1 Preprocessing of Corpus

As text is an unstructured source of information to make it suitable to an automatic method it is transformed into structured format. We have used the SemCor corpus which is of XML format; its tag contains the part of speech, the lemma, and the correct WordNet sense.

```
<wf cmd=done pos=NN lemma=committee wnsn=1
lexsn=1:14:00::>Committee</wf>

<wf cmd=done pos=NN lemma=approval wnsn=1 lexsn=1:0
```

Figure 1. Corpus file

For that purpose, we use DOM parser to parse XML files.

```
committee NN committee 1
approval NN approval 1
gov._price_daniel NN person 1
certain ADJ certain 4
```

Figure 2. Structured Format (Preprocessed)

5.2 Feature Extraction

After pre-processing of corpus, we get sentence which does not contain any irrelevant data in it and the data in its root form. From that data features are extracted. The window size of feature vector is of [-2, +2]. Here the features are the words itself and the part-of-speech of that words. This feature helps to train the classifier. Then these feature set is directly used to compare with the feature of target sentence for disambiguation.

5.3 Preprocessing of Target sentence

For the better result of disambiguation, in the pre-processing step firstly we are removing the stop words and stop symbols from the target sentences. For checking the co-occurrence of the words, these words are to be brought to their original form or root form i.e. stemming is done.

5.4 Task of Classification

Most of the approaches to the removal of ambiguity of word are from the machine learning field. We use Naïve baye's algorithm which is a probabilistic based approach, with that we use knowledge of ambiguous word with the exact ambiguous word i.e. associated word knowledge for the purpose of comparison. In classification procedure, after removing function words, only considering content words, first objective is to find the ambiguous word in the sentence. There is chance of more than one ambiguous word in the target sentence. We have to look for ambiguous word with the help of wordnet. Next step is to perform the process of pre-processing of target sentence steps. After that JAWS library methods, then these features are compared with the features of training data. The count is increase for each feature as the data is found same as the data of target feature and the training feature.

5.5 Finding correct sense

As approach is hybrid one and we are using probabilistic measure. Parameters in the probabilistic WSD are: Pr(s) i.e probability of sense and $Pr(V_w^i|s)$ i.e. probability of feature w.r.t. particular sense

$$Pr(s)= count(s,w) / count(w) \text{ and } Pr(V_w^i|s)= count(V_w^i,s,w)/count(s,w)$$

The sense with the highest probability is return and along with its sense id. This sense id is used to map with definition associated with that sense for the target word in the sentence.

6.EXPERIMENTAL DETAILS AND DISCUSSION

6.1 Resolution of ambiguity

Enter the target sentence and load the tagger to tag the sentence as shown in figure

Enter Sentence: John is playing cricket and then he will go to bank

Loading default properties from tagger ./english-bidirectional-distsim.tagger
Reading POS tagger model from ./english-bidirectional-distsim.tagger ... done [8.3 sec].

John_NNP is_VBZ playing_VBG cricket_NN and_CC then_RB he_PRP will_MD go_VB to_TO bank_NN

Figure 3: Part Of Speech Tagger output

For feature selection, we have used the Naïve baye's which is a Probabilistic Approach. Features used POS, Word (Lemma), Collocation i.e neighbor contextual information (+2,-2).

John is playing cricket and then he will go to bank.
John
play
cricket
go
bank

Figure 4: Removal of stop words (preprocessing)

play 35
play 157

Figure 5: Ambiguous word without knowledge

Without knowledge is our first approach while also we use the JAWS library for retrieving more information of ambiguous word so that we can use information with the learned world knowledge.

play 35
make for 27, flirt 2,toy 16, trifle 21 ,meet 34 ,recreate 11 ,fiddle 19,wager 30,bring 27,act as 8,spiel 6, run 18, dally 21, bet 30, play 1, represent 4, encounter 34, diddle 19, playact 25, take on 34, role play 25, act 4, work 27 ,wreak 27
play 157

Figure 6: Knowledge of Ambiguous Word

After the step of feature extraction, classification and finding correct sense, we get following correct sense.

1
participate in games or sport

Figure 7. Winner sense.

As our example sentence is John is playing cricket and then he will go to bank and ambiguous word in sentences are play, cricket, go and bank. Step by step we discussed the particulars how ambiguity of example word ‘play’ is resolved; same procedure can be applied to other ambiguous word present in the same sentence. Another example of resolving an ambiguity:

Enter target sentence consisting a set of ambiguous word: Ambiguity is the possibility of interpreting sense of the word used in the sentence

Stanford's POS tagger:
ambiguity_NN is_VBZ the_DT possibility_NN of_IN interpreting_VBG sense_NN of_IN the_DT word_NN used_VBN in_IN the_DT sentence_NN

Elimination of unnecessary words and then lemmatization:
Ambiguity possibility interpret sense word us sentence

Ambiguous word 1: possibility

World Knowledge:
possibility 4, possible action 4, possibility 1, theory 3, opening 4, possibleness 2, hypothesis 3, possibility 53

Without World Knowledge:
Possibility 4
Possibility 53

Possible meanings:
250 NN 1 null
1 NN 17 a future prospect or potential
2 NN 17 capability of existing or happening or being true
3 NN 11 a tentative theory about the natural world; a concept that is not yet verified but that if true would explain certain facts or phenomena
4 NN 7 a possible alternative

After feature extraction and classification
Winner sense: a possible alternative

Ambiguous word 2: interpret

World knowledge:
interpret 6, see 1, understand 6, interpret 1, represent 4, render 3, construe 1,rede 2
, translate 5, read 6, interpret 22

Without World knowledge:
interpret 6
interpret 14

Possible meanings:
250 VB 1 null
1 VB 12 make sense of; assign a meaning to
2 VB 6 give an interpretation or explanation to
4 VB 2 create an image or likeness of
6 VB 1 make sense of a language

After feature extraction and classification
Winner sense: make sense of a language

Figure 8. Resolution of ambiguity in target sentence

6.2 Performance measure

Domain wise testing for Precision, Recall and F-measure is done. Precision is proportion of correctly classified instances of total classified. Recall is proportion of correctly classified of total to be classified. F-measure is harmonic mean for Precision and Recall. Some ambiguous word is taken and check for ambiguity resolution. Astrology, Banking, Biology, Medicine and sport are considered randomly for testing the accuracy.

Domain	Ambiguous words tested
Astrology	study, position, sun, moon, birth, nature, time
Banking	bank, man, state, go, account, transfer, side
Biology	generation, branch, relation, soil, type, study, living
Medicine	aid, minor, body, study, collect, pay, carry
Sport	play, start, mind, cricket, go, together

Table 1: Domain & Ambiguous Words

Table 2 shows the precision, recall and F-measure calculated.

Domain	Precision	Recall	F-measure
Astrology	0.875	0.7	0.778
Banking	0.7857	0.611	0.687
Biology	0.667	0.5455	0.6
Medicine	0.6	0.5455	0.570
Sport	0.667	0.5	0.571
Average	0.71894	0.5804	0.6412

Table 2: Domain wise Performance measure

Random sentences are tested with and without World Knowledge. Without knowledge, only exact ambiguous word is considered for comparison. With knowledge, ambiguous word with the associate information of ambiguous word is considered for comparison with corpus.

Source	Precision	Recall
Without World Knowledge	**70.96**	**68.75**
With World Knowledge	**67.28**	**65.51**

Table 3: Random sentence wise Performance measure (in percentage)

While checking for ambiguous word such as bank, play, state, side, observe etc, it is found that World knowledge consist many unnecessary information too which results in excess comparison, also affects the time factor.

Performance measures i.e. precision and recall get better as we increase the total number of documents. If we consider only 1 corpus document, recall is not good and as we gradually increase the number of documents, performance measure also get better. Total six cases are performed and result is stated as below.

Test	Number of documents	Precision	Recall
#1	**2**	**0.25**	**0.20**
#2	**20**	**0.33**	**0.25**
#3	**80**	**0.45**	**0.33**
#4	**116**	**0.75**	**0.75**
#5	**145**	**0.83**	**0.72**
#6	**186**	**0.88**	**0.7142**

Table 4: Performance measure over number of documents

6. COMPARISON WITH EXISTING SYSTEM

Our hybrid disambiguating system is compared with the already existed word sense disambiguation system, and the performance measure of those existing system is taken from Judita Preiss, 2006 [6] Probabilistic word sense disambiguation: Analysis and techniques for combining knowledge sources, Technical report, University of Cambridge.

System	Description
SMUls	**A system which combines a pattern learning module with active feature selection. The sense tagged corpora semcor, WordNet definitions and gencor (this is a sense tagged corpus created automatically by the authors of the SMUls system (Mihalcea and Moldovan)** **Precision= 63.8%** **Recall= 63.8%**
BCU-ehu-dlist-all	**semcor, Based on Yarowsky's decision lists, learns lemmas, word forms and PoS from training data.** **Precison= 57.2%** **Recall= 29.1%**
CNTS-Antwerp	**semcor, A number of machine-learning word experts are trained, and the best one (based on training data) is individually selected for each word-PoS combination.** **Precision=63.5%** **Recall= 63.5%**
Hybrid(Without World Knowledge)	**Semcor , Hybrid which comprises combination of all knowledge resources mainly collocation and PoS, Probablistic approach, using only exact ambiguous word** **Precision= 70.96%** **Recall= 68.76%**
Hybrid(With World Knowledge)	**Semcor, Hybrid which comprises combination of all knowledge resources mainly collocation and PoS, Probablistic approach, ambiguous word plus information associated with the ambiguous word** **Precison= 67.28%** **Recall= 65.51%**

Table 5: Comparison with existing system

7. CONCLUSION & FUTURE WORK

Based on our study of WSD scenarios, we make the following conclusions:

1.Considering the disadvantages of all existing approaches i.e. knowledge based requires exhaustive enumeration search and knowledge resources, supervised has a problem of data sparseness, also huge number of parameters require to be trained and the unsupervised algorithm fails to distinguish between finer sense of a ambiguous word so effort has been made to resolve the issue by suggesting the hybrid approach.

2.Integration of various knowledge resources for a feature set such as Part of speech, morphological form(Lemma) of word, Neighboring words(in form of collocation vector), verb noun syntactic relation are helping us to obtain a good accuracy for classification.

3.System is working with high accuracy when the inappropriate information is detached from the sentences and also when the training data is increased.
In future, we would like to test our hybrid wsd work on the regional language of an India such as Hindi, Sanskrit, Guajarati, Bhojpuri etc.

REFERENCES

[1] Agirre, Eneko & German Rigau. 1996. "Word sense disambiguation using conceptual density", in Proceedings of the 16th International Conference on Computational Linguistics (COLING), Copenhagen, Denmark, 1996.

[2] Agirre, Eneko & David Martínez. 2001. Learning class-to-class selectional preferences. Proceedings of the Conference on Natural Language Learning, Toulouse, France, 15–22.

[3] Agirre, E. and Martinez, d. 2000. Exploring automatic word sense disambiguation with decision lists and the web. In Proceedings of the 18th International Conference on Computational Linguistics (COLING, Saarbr ¨ ucken, Germany). 11–19.

[4] E. Agirre and A. Soroa, “Personalizing PageRank for Word Sense Disambiguation,” Proc. 12th Conf. European Chapter of the Assoc. for Computational Linguistics (EACL 09), Assoc. for Computational Linguistics, 2009, pp. 33–4.

[5] Fellbaum. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Massachusetts, 1998.

[6] Judita Preiss, 2006: Probabilistic word sense disambiguation: Analysis and techniques for combining knowledge sources, Technical report, University of Cambridge, UCAM-CL-TR-673 ISSN 1476-2986, Number 673

[7] Gerard Escudero, Llu´ıs M`arquez and German Rigau, “Naïve Bayes and Exemplar-based approaches to Word Sense Disambiguation Revisited”, arXiv:CS/0007011v1, 2000.

[8] Mitesh M. Khapra, Anup Kulkarni, Saurabh Sohoney, and Pushpak Bhattacharyya. 2010. All words domain adapted WSD: Finding middle ground between Supervision and unsupervision. In Jan Hajic, Sandra Carberry, and Stephen Clark, editors, ACL, pages 1532-1541.

[9] Mihalcea and D.I. Moldovan. Pattern Learning and Automatic Feature Selection for Word Sense Disambiguation. In Proceedings of the Second international Workshop on Evaluating Word Sense Disambiguation Systems(SENSEVAL-2), 2001.

[10] Navigli, roberto, “word sense disambiguation: a survey”, ACM computing surveys, 41(2), ACM press, pp. 1-69, 2009.

[11] Ping Chen and Chris Bowes, University of Houston-Downtown and Wei Ding and Max Choly, University of Massachusetts, Boston Word Sense Disambiguation with Automatically Acquired Knowledge, 2012 IEEE INTELLIGENT SYSTEMS published by the IEEE Computer Society.

[12] R. Navigli and M. Lapata, “An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 32,no. 4, 2010, pp. 678–692.

[13] Roshan R. Karwa , M.B.Chandak "Word Sense Disambiguation: Hybrid Approach with Annotation Up To Certain Level – A Review", International Journal of Engineering Trends and Technology (IJETT), V18(7),328-330 Dec 2014. ISSN:2231-5381. www.ijettjournal.org. published by seventh sense research group.

[14] Satanjeev Banerjee, Ted Pedersen, “An adaptive Lesk Algorithm for Word Sense Disambiguation Using WordNet”, Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, page no: 136-145, 2002.