EMPIRICAL IMPLEMENTATION DECISION TREE CLASSIFIER TO WSD PROBLEM

Boshra F. Zopon Al_Bayaty[1], Dr. Shashank Joshi[2]

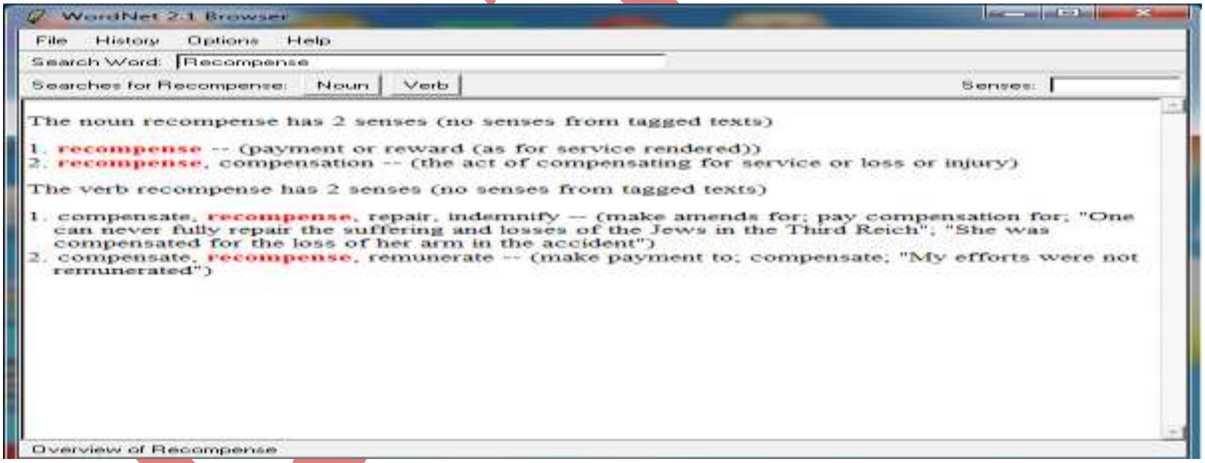
[1]*Department Of Computer Science,Yashwantrao Mohite College, Bharati Vidyapeeth University*
AL-Mustansiriyah University, Baghdad, (Iraq)
[2] *Department Of Computer Engineering, Engineering College, Bharati Vidyapeeth University, (India)*

ABSTRACT

We have applied on of most successful supervised learning approach to word sense disambiguation. We select the popular algorithm called decision tree. Empirically, we used senseval-3 to evaluation the word sense disambiguation in training and test data. We designed the experiment to fined the accuracy of decision tree, in this paper our study achieved (45.14 %) accuracy.

Keyword: Decision Tree, Naïve Bayes, Supervised Learning Approaches, WSD, Wordnet

I. INTRODUCTION



The noun recompense has 2 senses (no senses from tagged texts)

1. recompense -- (payment or reward (as for service rendered))
2. recompense, compensation -- (the act of compensating for service or loss or injury)

The verb recompense has 2 senses (no senses from tagged texts)

1. compensate, recompense, repair, indemnify -- (make amends for; pay compensation for; "One can never fully repair the suffering and losses of the Jews in the Third Reich"; "She was compensated for the loss of her arm in the accident")
2. compensate, recompense, remunerate -- (make payment to; compensate; "My efforts were not remunerated")

Fig. 1: The Screenshot Shows the Multiple of Recompense Word

There are many words have multiple meaning according to the context of speech. For example the word (Recompense) has different meaning in context, as in screenshot below:

Decision tree is one of prominent method to address sense disambiguation. In this approach meaning of word are mapped with leafs of tree. One with high value (accuracy) is considered and other leafs are rejected. This process requires calculation of "entropy", which provides base for the meaning or sense to be accepted or rejected. We applied the Entropy calculation and information Gain as in formulas below:

- Entropy(S)= $-P+\log_2 P+ - P-\log_2 P-$
- Gain (S,A)= Entropy(S) - $\sum_{v \in DA} \frac{|Sv|}{|S|}$ Entropy (Sv)

II. ADVANTAGES and DISADVANTAGES of DECISION TREE

2.1 Advantages

1. Efficient technique to mine a data by efficiently classifying values as per the attributes of.
2. Robust approach to filter exact data if size of value tree is small.
3. Non linear structure helps to reduce the required for traversal.

2.2 Disadvantages

1. Data is classified or mined at the cost of over fitting.
2. Maintenance of data becomes difficult, in case more number of sub-trees.

III. ALGORITHM APPLIED IN THIS WORK

For implementing WordNet data source is used this is repository which provides the mapping of word and different sense associated with that word. For performing on experiment we referred a data set 10 nouns and 5 verbs which contains following words:

Data set of pos (n) = {Praise, Name, Lord, Worlds, Owner, Recompense, Straight, Path, Anger, Day}.

Data set of pos (v) = {Worship, Rely, Guide, Favored, Help}.

To use WordNet repository senseval XML mapping technique is used, where the given data set and senses are expressed with XML. And to ensure effective working of decision tree training and testing file is used. Job of file is to provide the context which will be extremely useful exactly know meaning of particular word. For implementing C4.5 algorithm eclipse ID2, is used, while implementing it equations related with entropy are implemented. Below the algorithm we applied:

Box (1) C4.5 Algorithm implemented

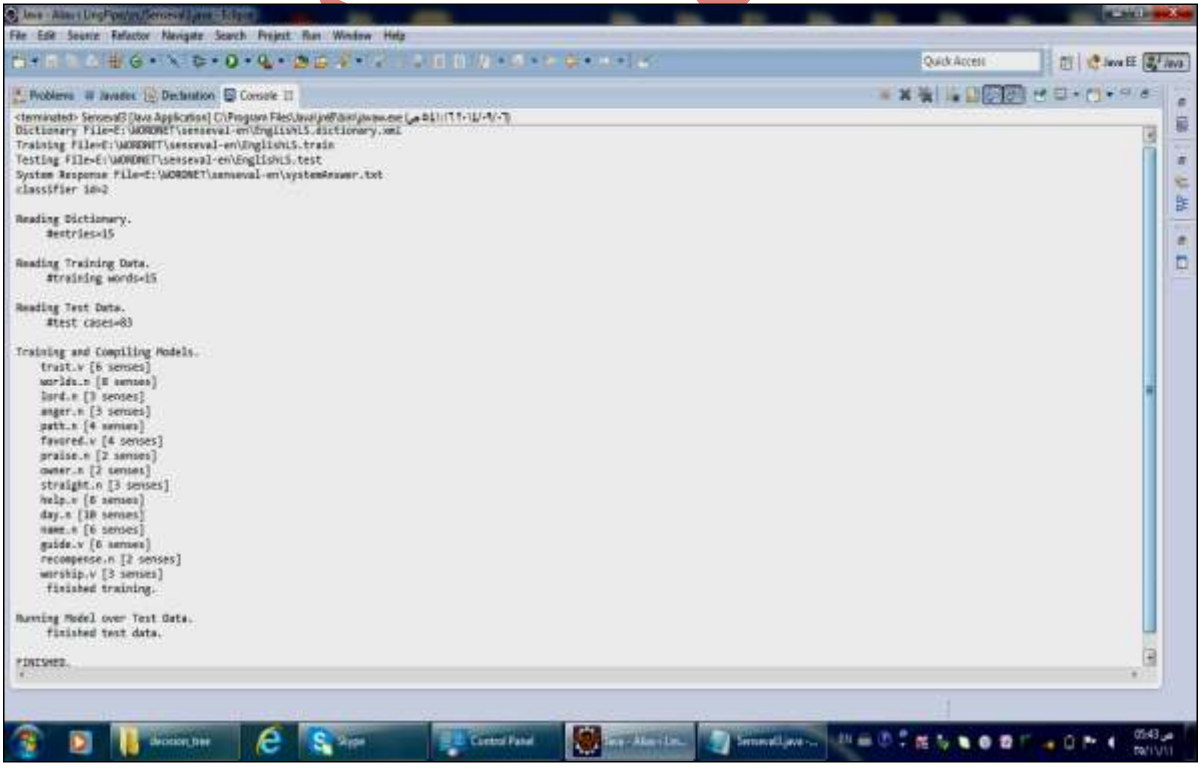
```
1. Read data set and calculation POS (e.g. recompense.)
2. Prepare context containing various senses of word (e.g. Recompense- reward)
3. Calculate frequency at context (i.e. - p- and +P+)
        -P- Negative
        -P+ Positive
4. Calculate information gain for calculating entropy (S) = $-P+\log_2 P+-P-\log_2 P-$
5. Gain (S,A)= Entropy(S) - $\sum_{v \in DA} \frac{|Sv|}{|S|}$ Entropy (Sv)
6. Select highest (Entropy, Attribute ratio)
7. E.g. (S,A) for recompense = 0.593
        For = reward
```

IV. RESULT

From the results that we acquired by implementing decision tree by using C4.5 algorithm, we can derive conclusion that this algorithm is useful in extracting meaning or sense of few words like {worlds-1000, name-1000, praise-593, owner-595, recompense-595}. But simultaneously there are some words for which this algorithm fail shoot by providing low accuracy, for example, {help-125, day-109}. As summary this is very helpful approach solve disambiguation. The results for our dataset shown in table (1) and Fig. 2 belwo shows the Screenshot Shows Taraining and Compilation Model

Table (1) Data Set of Words and Results of Decision Tree Classifier

Word	POS	# Senses	Score	Accuracy
Praise	n	2	405	593
Name	n	6	184	1000
Worship	v	3	308	425
Worlds	n	8	1000	1000
Lord	n	3	187	426
Owner	n	2	405	595
Recompense	n	2	405	595
Trust	v	6	167	167
Guide	v	5	199	247
Straight	n	3	462	462
Path	n	4	316	316
anger	n	3	462	462
Day	n	10	109	109
Favored	v	4	250	250
Help	v	8	125	125


File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Problems Javadoc Declaration Console
decision_tree
Skype
Control Panel

```
<terminated> Senseval3 [Java Application] C:\Program Files\Java\jre8\bin\javaw.exe [illegible]
Dictionary File=E:\WORDNET\senseval-en\English15.dictionary.xml
Training File=E:\WORDNET\senseval-en\English15.train
Testing File=E:\WORDNET\senseval-en\English15.test
System Response File=E:\WORDNET\senseval-en\systemAnswer.txt
classifier id=2

Reading Dictionary.
     #entries=15

Reading Training Data.
     #training words=15

Reading Test Data.
     #test cases=83

Training and Compiling Models.
    trust.v [6 senses]
    worlds.n [8 senses]
    lord.n [3 senses]
    anger.n [3 senses]
    path.n [4 senses]
    favored.v [4 senses]
    praise.n [2 senses]
    owner.n [2 senses]
    straight.n [3 senses]
    help.v [8 senses]
    day.n [10 senses]
    name.n [6 senses]
    guide.v [6 senses]
    recompense.n [2 senses]
    worship.v [3 senses]
     finished training.

Running Model over Test Data.
     finished test data.

FINISHED.
```

Fig. 2: The Screenshot Shows Training and compilation Model

V. CONCLUSIONS

By looking at the results that we came across there are few words which are providing accurate results. Overall accuracy of this approach is (45.14%), so there is definitely scope for modifying the accuracy of this approach.

VI. ACKNOWLEDGMENT

I would like to thank my research guide respected Dr. Shashank Joshi (Professor at Bharati Vidyapeeth University, College of Engineering) to inspire me always, pushed me, and helped me be the best I can.

REFERENCES

Journal Papers:

[1] Boshra F. Zopon AL_Bayaty, Dr. Shashank Joshi, Conceptualisation of Knowledge Discovery from Web Search, Bharati Vidyapeeth University, International Journal of Scientific & Engineering Research, Volume 5, Issue 2, February-2014, pages 1246- 1248.

[2] Boshra F. Zopon AL_Bayaty, Shashank Joshi, Empirical Implementation Naive Bayes Classifier for WSD Using WordNet., Bharati Vidyapeeth University, international journal of computer engineering & technology (IJCET), ISSN 0976 – 6367(Print), ISSN 0976 – 6375(Online), Volume 5, Issue 8, August (2014), pp. 25-31,© IAEME: ww.iaeme.com/IJCET.asp, Journal Impact Factor (2014): 8.5328 (Calculated by GISI), www.jifactor.com.

[3] Ted Pedersen, A Decision Tree of Bigrams is an Accurate Predictor of Word Sense, department of computer science, university of Minnesota Duluth, Duluth, MN 55812 USA,2004.

[4] Approaches for Word Sense Disambiguation – A Survey, Pranjal Protim Borah, Gitimoni Talukdar, Arup Baruah, International Journal of Recent Technology and Engineering (IJRTE), ISSN:2277-3878, Volume-3, Issue-1, March2014.

[5] Mark Alan Finlayson, Java Libraries for Accessing the Princeton WordNet: Comparison and Evaluation, Computer Science ad Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 2007.

[6] Oi Yee Kwong, Psycholinguistics, Lexicography, and Word Sense Disambiguation, Department of Chinese, Translation and Linguistics, copyright 2012 by Oi Yee Kwong, 26th Pacific Asia Conference on Languge, Information and Computation pages 408-417, 2012.

[7] Ted Pedersen, Evaluation the Effectiveness of Ensembles of Decision Trees in Disambiguation Senseval Lexical Samples, department of computer science, university of Minnesota Duluth, Duluth, MN 55812 USA.

[8] Miller, G. et al., 1993, Introduction to WordNet: An On-line Lexical Database, ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.pdf, Princeton University.

[9] Navigli, R. 2009.Word sense disambiguation: A survey. ACM Compute. Survey. 41, 2, Article 10 (February 2009), 69 pages DOI = 10.1145/1459352.1459355 http://doi.acm.org/10.1145/1459352.1459355.

Books:

[10] Nitin Indurkhya and Fred J. Damerau “HANDBOOK OF NATURAL LANGUAGE PROCESSING” SECOND EDITION. Chapman & Hall/CRC, USA, 2010.

[11] Daniel Jurafsky and James H. Martin, Naïve Bayes Classifier Approach to Word Sense Disambiguation, chapter 20, Computational Lexical Semantics, Sections 1 to 2, University of Groningen, 2009.

Theses:

[12] Ahmed H. Aliwy. Arabic Morphosyntactic Raw Text part of Speech Tagging System. PhD dissertation, University of Warsaw, 2013.

Links:

[13] http://www.senseval.org/senseval3.

[14] http://www.e-quran.com/language/english.

[15] http://wordnet.princeton.edu.