International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue (6) NCRTCST-2013, ISSN 2249–071X

National Conference on Recent Trends in Computer Science and Technology (NCRTCST)-2013

Decision List Algorithm for Word Sense Disambiguation for TELUGU Natural Language Processing

Durga Prasad Palanati
Department of Information Technology
Sridevi Women's Engineering College, Hyderabad, India
Email: dp.cse5@gmail.com

Ramakrishna Kolikipogu
Department of Information Technology
Sridevi Women's Engineering College, Hyderabad, India
Email: krkrishna.csit@gmail.com

Abstract **– Word Sense Disambiguation plays a key role in Information Retrieval Applications. Telugu language contains more ambiguous words when compared to other European languages. Word complexity is high for Telugu and Information Retrieval in a Telugu language is a challenging task. This paper focus on assigning a particular sense for a word based on the context of use. Word cooccurance probability over the corpus is measured with the help of Decision list algorithm and assigned the sense. This algorithm is proven to be accurate for English text corpus about 90%. We adopted the same decision list algorithm to disambiguate sense of Telugu words with unsupervised learning process. The performance of Telugu Information Retrieval is measured with and without WSD. The results are more accurate with WSD when compared to word sense ambiguity.**

Keywords **– Word Sense Disambiguation, Decision List Algorithm, Natural Language Processing, Information Retrieval, Telugu, , Log-Likely Hood, Seed Word.**

I. INTRODUCTION

Word Sense Disambiguation (WSD) is a method for resolving ambiguity of words in a context. WSD is the process of distinguishing the original sense of a word in a particular context. When words have more than one sense, it is difficult for indexing the terms. WSD identify the senses of words and these senses as index terms will improve the accuracy of text processing applications, when it is used in information retrieval applications the terms are replaced by their senses in the document vector [15]. Word meanings can have different senses. Polysemy words may have multiple meanings. WSD is the problem of determining the context of using a sense in a given sentence. Word Sense Disambiguation for single term or short queries is hard because of the missing context, and it is not necessary for long queries because the other terms narrow down the search anyways. Ide and Veronis et al specified WSD is a two step process. At the first step, under the given context, every word relevant to the sentence is disambiguated by determining possible senses with help of external semantic resources like Dictionaries, Thesaurus and WordNet etc. Secondly the word sense can be assigned based on cooccurance of words in a context. We use second approach to define sense features using decision list algorithm. This paper is mainly focused on Telugu information retrieval and how Word Sense Disambiguation improves the retrieval performance.

To assign a correct word sense we adopted decision list algorithm. Decision List algorithm identifies sub-parts with higher or lower likelihood in the overall sample space. Decision lists are, in simple terms, a list of feature queries that are posed about the input data points, ranked in some well defined order to obtain the classification [15]. Basically two properties of natural languages are making use by this decision list algorithm, such as collocation and discourse. As the natural languages are more redundant these two properties over determine the word sense. Many empirical evidence to support both the above two properties are provided by Yarowsky [1]. Collocating words are ranked by probability log likely-hood based on word feature vectors defined over entire text corpus. For example, *you might look for goods that are most likely to be acceptable in a particular period of time*. The conditional probability is applicable in this scenario. For example, *more coats are consumed and provided in winter*. The Decision List gives control over the model, which allows us to edit partitions of sample space, to add specific rules and give score to each part, and customize the model in a number of ways to optimize the proportion of hits across all segments. So, the decision list algorithm works based on rules that are intended to be interpreted in a sequence of terms. For some instances they might be covered by multiple rules, which gives-raise to ambiguity. By defining ordering rules with support of decision list can solve the problem. The decision-list algorithm resolves many problems in a large set of non-independent evidence source by using only the most reliable piece of evidence rather than the whole matching collocation set. The algorithm will continue to iterate until no more reliable collocations are found. The 'One sense per discourse' property can be used here for error correction. The Sense 2 will be relabeled as Sense 1 if Sense 1 exceeds the Sense 2 above certain threshold. According to Yarowsky [2] [9], for any sense to be clearly dominant, the occurrences of the target word should not be less than 4. The remainder of the paper is organized as follows. Various approaches to Word Sense Disambiguation are discussed under Related Work in Section II. The Decision List algorithm for Telugu Word Sense Disambiguation is provided in Section III. In Section IV, experimental details with result are plotted and this Section is followed by conclusion and future work. The method adapted to Telugu IR is working similar to other languages, but the process is difficult for Telugu to

identify seed words like other European languages. Due high complexity and morphological features, finding decision list for conflated words is complex in this paper.

II. RELATED WORK

Word Sense Disambiguation (WSD) is a common problem in computational linguistic and natural language processing applications. When a word has multiple meanings, WSD identifies correct sense in a sentence or context. WSD is typically involves two main phases. In first phase possible senses are being determined and extracted for each word. I second phase corresponding words are tagged with appropriate sense for disambiguation. Many researchers worked on WSD Yarowsky stated that the decision lists are simple means to solve ambiguity problems [1]. The same authors further investigated by successfully applying same method to accent restoration, word sense disambiguation and homograph disambiguation [2][3]. It was one of the most successful systems on the senseval word sense disambiguation competition proved by Kilgarriff and Palmer in 2000[8]. Lot of work has been done in English language on word sense disambiguation. Machine readable dictionaries like LDOCE, Logman Dictionary of Contemporary English developed by Procter was frequently used as a research Lexicon [9] and for tagging word sense usages Bruce and Wiebe [10] were initially used in their work. Gally and McKeowm presented an efficient linear-time algorithm to build lexical chains, shown that one sense per-discourse can improve performance [11]. In this paper we have shown that how the separation of Word Sense Disambiguation forms the construction of the chains which enables a simplification of the task and improves running time. The evaluation of decision list algorithm used in this paper against two known lexical chaining algorithms shows that our algorithm is more accurate when it chooses the senses of nouns to include in lexical chains. The HyperLex algorithm presented in [12] is entirely corpus-based. In the context of the target word the HyperLex algorithm builds a co-occurrence graph for each pair of words co-occurring and shown that these graph fulfills the properties of small word graphs, and thus possesses highly connected components in the graph. These hubs eventually identify the main word uses senses of the target word, and can be used to perform word sense disambiguation. These hubs are used as a representation of the senses induced by the system, the same way that clusters of examples are used to represent senses in clustering approaches to WSD [13]. Tagged corpora have been used to map the induced senses, and then compare the systems over publicly available benchmarks [14], which offers the advantage of comparing to other systems, but converts the whole system into semi-supervised. Evaluating clustering solutions is not straight forward. The unsupervised evaluation seems to be sensitive to the number of senses in the gold standard, and the coarse grained sense inventory used in the gold standard had a great impact in the results. WordNet is a large lexical database for English developed by the Cognitive Science Laboratory at Princeton University under the direction of Professor George Miller. WordNet consist information about huge amount of English words. Hindi WordNet developed at IIT Bombay, which is based on an idea of English WordNet. Hindi Wordnet has many uses than a conventional Hindi dictionary. Some unique concepts are represented with different relations between synsets or synonym sets. There are no sufficient resources for Indian languages like English. There is a need of more language process tools for Telugu and other Indian languages. Our current work focuses on how to adopt the word sense disambiguation approaches used in English language processing to the Indian languages especially for Telugu.

A. What are the features?

The features are extracted from the trained data. All the extracted features are ordered according to their log-likelihood. The decision list constitutes these features. For example the feature may be like

(1) ***word*$_1$** may come before ***word*$_2$** in the context
(2) ***word*$_1$** may come after ***word*$_2$** in the context
(3) There may be a fixed number of words in between ***word*$_1$** and ***word*$_2$**.

Based on the features selected, a word is assigned a particular sense for the context. The highest weighted feature will be selected to assign the sense. The formula to calculate log-likelihood is given Equation (1).

$$weight(sense_i, feature_k) = Log\left(\frac{\Pr(sense_i \mid feature_k)}{\sum_{j \neq i} \Pr(sense_j \mid feature_k)}\right) \quad (1)$$

We insert all the features in the decision list except those features with zero or negative value.

B. Approaches to Word Sense Disambiguation (WSD)

Many researchers proposed different approaches to detect ambiguity of words in various languages. We adopted a best suitable approach for Telugu language and suggested for other Indian languages. Highest Sense Count (HSC) adopted from [3] and applied to Indian languages. HSC simply counts the occurrences of a term in the context of a particular sense and highest sense count determines the sense in that context. Some approaches [4 and 5] states that accuracy can be increased with the expansion of Nearness in the hypothesis is Words near a given Word

In a text are semantically closer to the given word". DAG Based Algorithm for WSD using WordNet [6] applied on a Hindi text gives an accuracy of around 65%.

C. Word Sense Disambiguation for Telugu Language

Telugu is one of the South Indian languages belongs to Dravidian languages family recognized by government of India. Telugu is the 2nd most spoken languages in India

after Hindi language. As per the statistics the Telugu is 15th most spoken language throughout the world [16]. Ramakrishna K and Padmaja R B stated that Telugu language is more complex in nature with high morphological features compared to other languages and Word Sense Disambiguation by word cooccurance improves the recall of the information retrieval system. Use of Synset while applying sense count will improves the robustness of the system [17]. Example 1 illustrates how the sense ambiguity problem is found for Telugu language.

Example 1:

"***గురుత్వాకర్షణ లేని ప్రాంతాన్ని శూన్యం అంటారు***" [Telugu]

"***The place in which no gravity is Emptiness***" [English]

The word '*శూన్యం*' has various senses. *sense*$_1$ is related to atmosphere with the meaning '*Emptiness*' and in *sense*$_2$ it is related to numeric value with the meaning '*zero*'. And ***"గురువు లేని విద్య విలువ శూన్యం"*** is a context of *sense*$_2$.

Example 2:

"***జంతువుల రాజు సింహం"*** [Telugu]

"***Lion is the king of animals***" [English]

The word '***రాజు***' has various senses, ***sense*$_1$** is with the meaning '*king*' and *sense*$_2$ is a name of the person. ***"రాజు బాగా చదువుతాడు. / Raju studies well"***. In this sentence the word follows *sense*$_2$.

III. Decision List Algorithm

Word Sense Disambiguation by lexical ambiguity resolution with the help of Decision Lists (DL) is a proven by many authors from last two decades. Decision lists are used in variety of applications particularly in the area of information extraction because the rule outputs they produce are easily understandable by humans easily. A decision list consists of three terms ***sense***, ***feature*** value and ***weight*** in the context of natural language processing. The rules are formulated with these three terms and such rules can be sorted based on the value of corresponding weight. The weights are ordered set of rules that are used as features in Information retrieval applications. The weighting can be calculated from the training set of data using log-likelihood proposed by Yarowsky [1].

A. Unsupervised Learning Process

Step-1: Find all possible examples for each Polysemicious word over the corpus.

Step-2: Find all possible senses of the word and identify more generalized example which covers the word senses.

Step-3: Train the seed training data using Decision List algorithm.

Step-4: Classify them using sense tagging.

Step-5: One-sense-per-discourse constraint is used to filter the entire seed set.

Step-6: Re-filter the list, when miss-classified in to sets, repeat step d & e.

Step-7: Stop, when the training parameters are constant.

B.Decision list Generation

Decision list algorithm is most efficient and supervised algorithm firstly used by Yarowsky on the SENSEVAL corpus [2] generated according to the following steps:

Step-1: Features for each word are extracted from the corpus.

Step-2: Rules of the form

<feature value, sense, score> *collected.*

Step-3: The rules obtained in step ii, are sorted based on the scores (calculated using eq.1) in decreasing order.

Step-4: The table obtained is our decision list.

In any natural language some words possess two or more meanings in various contexts. So the sense is depends on the context of use. The major problem is which sense has to be chosen to assign the meaning for that word. We used a novel approach based on conditional probability using decision lists to solve the problem. Table 1 depicts different senses for Telugu words in different context.

Table 1.Example Telugu sentences in 2 senses of a keyword ***ప్రగతి****/progress*

Sense	Examples (keyword in context)	Tag
1	కష్టపడి పని చేస్తే ***ప్రగతి*** సాధించవచ్చు / ***Progress*** can be achieved with hard work	ADJECTIVE
1	యువత పైనే దేశ ***ప్రగతి*** ఆధారపడినది. ***Progress*** is based on the youth of the country.	ADJECTIVE
2	***ప్రగతి*** బాగా పాడుతుంది. / PRAGATI [progress] sings well	NOUN
2	ఈ పుస్తకం ***ప్రగతి*** సంస్థ వారిచే ముద్రించబడినది. / The book, published by the PRAGATI [Progress] company.	NOUN

Table.2 Example words in different senses

Word	Sense$_1$	Sense$_2$
శూన్యం / Emptiness	Atmosphere	Value
ప్రగతి / Progress	Improvement	Name
రాజు / king	King	Name

పాత్ర / vessel/ role	Role	Item
కొట్టు / *beat*	Shop	Beat
రసం / *Juice*	Fluid	Expression
గాలి / air	Air	Name

C.Feature extraction

The features which can be used as rules are extracted from training data. The features are extracted with following rules.

- *Word immediately to the right of a keyword KW, (+1 KW).*
  - feature value KW_{+1}
- *Word immediately to the left of a keyword KW, (-1 KW)*
  -feature value KW_{-1}
- *A pair of words immediately to the right and immediately to the left of a keyword KW, (-1 KW +1)*
  - feature value KW_{-1+1}
- *A pair of words immediately to the right and next immediate right word of a keyword KW, (KW +1 +2).*
  -feature value KW_{-1+2}
- *A pair of words immediately to the left and next immediate left word of a keyword KW, (-2 +1 KW).*
  -feature value KW_{-1-2}
- *A pair of words immediately to the left and next immediate left word of a keyword KW, (-2 -1 KW).*

D.Formation of Rules

The rules consist of feature value, sense and score triplet. The score i.e. weight, which can be calculated using conditional probability. The features corresponding to the rules with highest score are used to disambiguation the sense of a word.

IV. RESULTS ANALYSIS

In this paper, the decision list algorithm is applied on one lakh of unique Telugu words from 1, 24,000 text documents collected from web and Wikipedia sources. For each "***word + position***" feature f_i, a smoothed log-likelihood ratio is computed for each sense s_j of a word of the corpus. The likelihood ration is measured by Function (2)

$$\left(\frac{P(f_i|s_j)}{P(f_i|\neg s_j)}\right) \quad (2)$$

with smoothing based on an empirically estimated function of feature type and relative frequency. These problems with accuracy led to the adoption of precision and recall instead of (or in addition to) accuracy for performance measurement. The combination of precision and recall has been used as the primary means of performance evaluation in the SENSEVAL exercises. A training set is used for inducing a set of features. As a result, rules of the kind <***feature-value, sense, score***> are created. The ordering of these rules, based on their decreasing score, constitutes the decision list.

Keyword used is : పాత్ర

$Sense_1$: *gives meaning like role*
$Sense_2$: *gives meaning like bowl*

The features considered in our model are KW_{+1}, KW_{-1}, KW_{-1+1}.Various features corresponding to particular sense with their scores are shown in the table.3 below.

Table 3: Scores of various features

Feature	Sense	Frequency	Max. likelihood estimate Score
KW_{+1}	$Sense_1$	50	**50/95=0.52**
	$Sense_2$	45	
KW_{-1}	$Sense_1$	80	**80/152=1.37**
	$Sense_2$	58	
KW_{-1+1}	$Sense_1$	38	**46/84=0.54**
	$Sense_2$	46	

The rules above should be ordered based on the scores in descending order.

Table 4: Ordered list (Decision list)

S.No	Rule
1	(KW_{-1}, $Sense_1$, 1.37)
2	(KW_{-1+1},$Sense_2$,0.54)
3	(KW_{+1}, $Sense_1$, 0.52)

From the Table 4 , $Sense_1$ with the meaning of *role* for the word పాత్ర is assigned. The average accuracy of assigning the senses to set of words is 97% in this paper. An Information Retrieval developed to test the accuracy of Word Sense disambiguation for Telugu languages is compared with standard Information Retrieval performance metrics *Precision* and *Recall*.

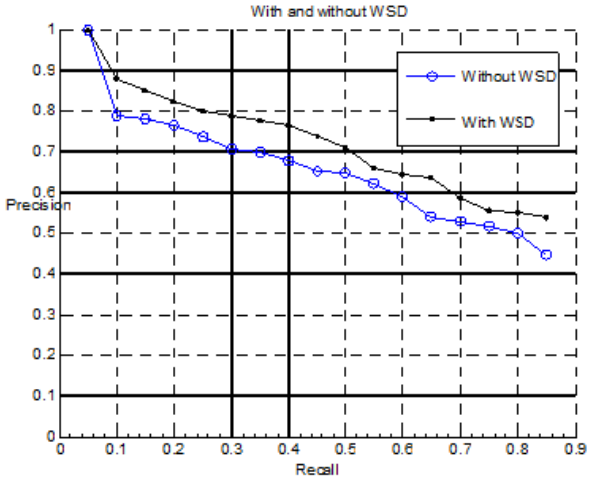

Fig.1. Recall-Precision graph with and without WSD.

The performance the Information Retrieval system with and without Word Sense Disambiguation is shown in Fig.1.

V. CONCLUSION

In this paper we considered only keywords as seeds. The performance of the system is improved and the accuracy of WSD is depends on the size of the corpus. If the corpus contain documents of various subjects with ambiguous words, then the sense of word would increase. Clustered documents with ambiguous words will improve the WSD accuracy. Word Sense Disambiguation with synonym features would give more related out come and finally improves the precision of the system.

REFERENCES

[1] Yarowsky, D. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French', in Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp. 88--95. 1994.

[2] Yarowsky,D.Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics Cambridge, MA, pp. 189-196, 1995.

[3] Yarowsky, D. Homograph Disambiguation in Text-to-speech Synthesis. J Hirschburg, R. Sproat and J. VanSanten (eds.) Progress in Speech Synthesis,Springer-Vorlag, pp. 159-175. 1996.

[4] I.Klapaftis and S. Manandhar, "Google &WordNet based Word Sense Disambiguation", in Proceedings of the Workshop on Learning and Extending Ontologies by using Machine Learning methods, Bonn, Germany, 2005.

[5] Sinha, Reddy,Bhattacharya,Pandey,Kashyap "Hindi Word Sens e Disambiguation" 2004.

[6] Agirre and Rigau G. "Word Sense Disambiguation Using Conceptual Density" in proceedings of COLING '96.

[7] Bhattacharya and Unny "Word Sense disambiguation and Text Similarity Measurement Using Word Net" 2002.

[8] Kilgarriff, A. and M. Palmer. (eds). *Special issue on SENSEVAL*. Computer and the Humanities, 34 (1-2). 2000

[9] Wilks and M. Stevenson. 1998. The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. Journal of Natural Language Engineering, 4(1):1-9.

[10] Bruce, R. and Wiebe, J. (1999) Decomposable modeling in natural language processing. Computational Linguistics 25(2): 195–207.

[11] Galley, M. & K. McKeown. 2003. "Improving Word Sense Disambiguation" in the proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003),1486-1488.Acapulco,Mexico.

[12] V´eronis. 2004. Hyperlex: lexical cartography for information retrieval. Computer Speech & Language, 18(3):223-252.

[13] Purandare and T. Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In Proc. of CoNLL-2004, pages 41-48.

[14] Niu,W. Li, R. K. Srihari, and H. Li. 2005. Word independent context pair classi_cation model for word sense disambiguation. In *Proc. of CoNLL-2005*.

[15] Voorhees, E.M.(1993), UsingWordNet to DisambiguateWord Senses for Text Retrieval, In *Proceedings of SIGIR-93*, pages 171–180, Pittsburgh, PA, USA.

[16] Kolikipogu Ramakrishna and Durga Prasad P, Text based E-Learning in Telugu Language using Information Retrieval System : A study under Information Retrieval in Indian Languages, National Conference on E-Learning and E-Learning Technologies, C-DAC, India, July 2013.

[17] Kolikipogu Ramakrishna , B.Padmaja Rani et al, Information Retrieval in Telugu Language using Synset Relationships, Proceedings of 15th IEEE International Conference on Advanced Computing Technologies-ICACT-2013,Rajampet, 21st & 22nd Sep'13, ISBN : 978-1-4673-2816-6/13, I07,2013.

AUTHOR'S PROFILE

Mr.P.Durgaprasad

is currently working as Assistant professor in Dept. of IT, Sridevi Women's Engineering College, Hyderabad. He received bachelor degree in Computer Science and Engineering from JNTUH in 2009 and Masters of Technology in Computer Science and Engineering from JNTUH in 2011. His area of research interest is natural language processing. He is renewed trainer for GATE and Other Competitive exams in the subject of Computer Science, especially in Operating systems, database management system and algorithms.

Mr.Kolikipogu Ramakrishna

was born in Peraigudem, Aswaraopet, A.P, India, in 1981. He received the B.Tech. degree in Computer Science and Information Technology from the JNT University, Hyderabad,India and the M.Tech. in CSE(Software Engineering) from JNTU, Kakinada and Pursuing Ph.D. in Computer Science & Engineering from the Jawaharlal Nehru Technological University, Hyderabad, India. At present he is working as Head of the Information Technology Department, Sridevi Women's Engineering College, Hyderabad. Since 2004, he has been working as teaching professional in the domain of computer Science and Information Technology. To the credit he has published 25 research papers in various International Journals & Conferences. He received a best paper award for his research paper presented in 15th ICACT IEEE Coonference. His current research interests include Information Retrieval Systems, Natural Language Processing, Data Mining; Information Security et al. Mr.K.Ramakrishna is a Member of different professional bodies including Association of Computing Machinery (ACM), IEEE, Computer Science Teachers Association (CSTA), International Association of Engineers (IAENG), IACSIT, Computer Society of India (CSI). He is serving as reviewer for couple international Journals and Conferences including IJCSI, IJCCT, IJECCE, IJEST, IJML, IJCL, IJSWIS and IEEE, ACM Conferences. He has been edited couple of books and conference proceedings in his career. He is a founder member president of MUNDADUGU Organization, Hyderabad.