An Investigation to Semi supervised approach for HINDI Word sense disambiguation

Neetu Mishra
Indian Institute for Information Technology, Allahabad
Allahabad, India
Neetumishra1508@gmail.com

Tanveer J. Siddiqui
JK Institute Of Applied Physics
Allahabad University
Allahabad, India
jktanveer@yahoo.com

Abstract— This **paper investigates yarowsky algorithm for Hindi word sense disambiguation. The evaluation has been developed on a manually created sense tagged corpus consisting of Hindi words (nouns). The sense definition has been obtained from Hindi Word Net, which is developed at IIT Bombay. The maximum observed precision of 61.7 on 605 test instances corresponds to the case when both stemming and stop words elimination has been performed.**

Index Terms*— Seed, *Word sense disambiguation, Hindi word sense disambiguation

I. Introduction

Natural Languages contain words that have multiple senses or meaning .Human beings easily recognize the correct meaning of a word without even considering all of its senses[11]. However it creates problem during automatic processing of text .In Order to get the correct meaning of a word disambiguation has to be performed. The task of disambiguation is concerned with identifying correct meaning of an ambiguous word in a specific use. For example consider the word '**फल**' in the following sentences

"बाज स फल आर फल स बाज का चक्राय गात क साय ाबदु रुप बाज म एक सरल- रखाय गात भा ह, और वह है आनुवाशकता का ाकास"

"जा जसा शुभ व अशुभ काय करता ह, वा वसा हा फल भोगता है"

"बरछा का फल नुकाला हाता ह"

The occurrences of the word '**फल**' In these three sentences clearly denote different meanings, the identification of the specific meaning that a word assumes in context is obviously simple. The word '**फल**' has three senses as listed in HindiWordNet .A simple dictionary lookup operation will not get the intended meaning.

फल, **फर, प्रसून** - वनस्पात म हान वाला गूद या बाज स भरपूर बाजकाश जा किसा ाविशष्ट ऋतु म फूल आन क बाद उत्पन्न हाता है *"उसन फल का दुकान स एक किला आम खराद।"*

पारणाम, अंजाम, अन्जाम, अंत, अन्त, नतीजा, प्रातफल, फल, ताबीर, ख़ामयाज़ा, खामयाजा, ख़ाामयाज़ा, खाामयाजा, पारणात, व्युष्ट, ाविपाक - किसा काय क अत म उसक फलस्वरूप हानवाला काय या काइ बात *"उसक काय का पारणाम बहुत हा बुरा निकला"*

गाँस, फल, गाँसी, गांस, गासी, गांसी, अंकुड़ा, अँकुड़ा, अँकड़ा, अंकड़ा, आँकुड़ा - तार या बरछा आाद क आगे का धारदार भाग *"इस तार का गास बहुत नुकाला है"*

Fig 1. Senses of फल **obtained from the Hindi WordNet**

Word Sense Disambiguation is an open problem in Natural Language Processing(NLP).There are some other areas ,like Information retrieval, text classification, Machine Translation ,which can benefit from an accurate WSD method. Most of the work on word sense disambiguation (WSD) focuses on English. Very few amount of research has been done on automatic word sense disambiguation for HINDI [2].The main work in

this a classifier is learned from Unlabelled Hindi corpus. The approach is inspired from Yarowsky's work [5] but differs from it in the following points:

(i) Unlike the work in [5], we create seed instances automatically.

(ii) After stop word removal from the raw context we perform stemming on remaining words. It takes care of the case when the word appearing in the context is morphological variant of the word in the decision list[11].

(iii) Unlike Yarowsky's algo we are handling senses with homonyms from WordNet.

2 Related Work

A number of studies have been conducted in this area of NLP in English Language rather Asian languages History of WSD is presented in this paper [1], starting from the 1950's. It covers the major areas of work and outlines the broad lines of development in this field. The automatic disambiguation of word senses has been a matter of interest since the earliest days of computer understanding of natural language in the 1950's.

The first most Lesk algorithm (Lesk 1986) which is developed for the semantic disambiguation of all words in unrestricted text. In this work they were considering all the sense definitions for the word in the dictionary and knowledge about the immediate context, where the sense disambiguation is performed. Second one is Measures of semantic similarity computed over semantic network. Words that share a common context are usually closely related in meaning, and therefore the appropriate sense can be selected by choosing those meaning found within the smallest semantic distance.(Rada et al.1989)This category includes methods for finding the semantic density/distance between concepts. Depending on the size of the context these measures are in turn into two main categories local context and global context. There is a lot of work which has considered the use of local context. (Patwardhan et al. 2003) applied the first five similarity measures to decide upon the context sense of 1,723 instances of ambiguous nouns from the Senseeval-2 English lexical sample data.The work in (Walker, 1987) uses subject categories provided by some dictionaries e.g. Longmans Dictionary of Contemporary English (LDOCE) (Procter, 1978), Roget's thesaurus, etc. in disambiguation whereas the work in (Wilks et al., 1990) attempted to expand dictionary definition with words that commonly co-occur with that definition. Yarowsky (1992) extended this idea by including additional evidences from corpora and training machine learning algorithms. Hand coded knowledge may also be used. The work has been carried out using existing lexical knowledge sources such as WordNet (Aggire & Rigau,1996; Resnick, 1995; Voorhees, 1995), LDOCE (Guthrei et al.,1991) and Roget's International Thesaurus (Yarowsky, 1992).Sense disambiguation [3] is an "intermediate task" (Wilks and Stevenson, 1996) which is necessary to achieve most natural language processing tasks. It is indispensable for language understanding applications such as content understanding, machine-man communication, etc.; it is also helpful, and in some cases required, for applications whose aim is not language understanding: Early works on disambiguation were dictionary-based. They make use of lexical resources, e.g. dictionary, thesaurus, ontology, etc., for disambiguation. Yarowsky extended the idea by combining evidences from thesaurus and supervised learning [4]. Other works in supervised disambiguation include (Gale, 1992 & Mooney, 1996). Creating sense tagged corpus required by supervised approaches is quite time consuming. Yarowsky[5] proposed an unsupervised approach for disambiguation which uses unlabelled text for training. It can be easily extended for languages for which sense tagged corpus is not available. Unsupervised algorithm broadly fall in two categories: similarity based &graph based. Similarity based algorithm utilize surrounding context to disambiguate a word whereas graph based algorithm work by build a graph and identifying the most important node for each word. Nodes in the graph correspond to word senses and edges correspond to dependencies between them. A comparative study of these two types of algorithm has been made in (Mihalcea, 2005) and Brody et a., 2006).

3 Proposed Method

In general in an unsupervised technique, the ambiguous word (without labeled instances) are given as input and are then grouped into clusters according to a similarity metric. These clusters are then labeled by hand with known word senses. We will implement this method in two steps:

1. Finding similarity metric: for this we will use Yarowsky's algorithm.
2. Labeling the clusters with known word senses: Instead of labeling by hand, we will use Word Net to label the cluster. This will save the time and cost.

Yarowsky's algorithm:

It's an unsupervised algorithm which can accurately disambiguate word senses in a completely large untagged corpus[5]. The basis of this algorithm is that there are constraints between different occurrences of the ambiguous word within the corpus.

1. One sense per discourse: The sense of a target word is highly consistent within any given document.
2. One sense per collocation: Nearby words provide consistent and strong information about the sense of the target word.

Word not only tends to occur in collocation that indicates their sense but they tend to occur multiple times in collocation. This provides a mechanism for boot strapping, a sense tagger. If we start with a small set of seed examples representative of two senses of a word, by fallowing the combination of the constraints one sense per discourse and one sense per collocation, we can incrementally augment these seed examples with additional examples.

Step 1.

Identify all examples of the given polysemous word in the corpus, storing their contexts as lines in an initially untagged training set. For example:

Sense	Training examples (keyword in context)
?	… फल या सब्जा एसा जिन्स ……..
?	……..मैदानी फला म आम का प्रथम ……
?	……अवज्ञानक विाध स फल-वृक्ष लगात ……

Fig 2: Example of untagged instances

Step 2:

For each possible sense of the word, identify a relatively small number of training examples representative of the sense. This can be done by selecting small no of seed collocation representative of each sense and then tagging all training examples containing the seed collocation with the seed's sense level.

Sense	Training examples (seed word in context)
A	……लाभ उठाकर याद फला का बागवाना……..
?	……..मैदानी फला म आम का प्रथम ……
B	… अवज्ञानक विाध स फल-वृक्ष लगात ……
B	पता चलगा कि फल वृक्षा का रापत समय …
B	…जितन काट-रोग फल-वृक्षा पर आक्रमण …
C	… इस तीर का फल बहुत नुकाला है……..

Fig 3: Seed words in Context

Step 3a:

Train the supervised classification algorithm on the sense – A/sense-B/sense-C seed sets. To identify other collocation, the decision list algorithms are used, that partition the seed training data, ranked by the purity of the distribution. This step is used in my previous work [11] for training the data by manual seed selection.

Step 3b:

Apply the classifier to entire sample set. Take those members in the residual that are tagged as sense-A or sense-B with probability greater than some threshold. Add those examples to the growing seed sets.

Step 3C:

The one sense per discourse constraint is then used optionally to filter and augment this addition.

Step 3D:

Repeat step 3 iteratively.

Step 4:

Stop when the training parameters are held constant. It will converge on a stable residual set.

Step 5:

This classification procedure is now applied to new data.

4 Experiments

4.1 Data Set

We have developed our own training and test corpus which includes the Hindi corpus from IIT Bombay and news articles from the India info Dainik Jagran, Khoj, Hindi Wikipedia and webdunia websites. Some documents have been taken from EMILLE Corpus. All data are encoded in two byte Unicode text. Fig 4 shows the statistics of our dataset.

Total Number of Words	40
Total Number of Instances	1470
Total Number of Training Instances	865
Total Number of Testing Instances	605
Average Number of Instances Per Word	36.75
Average Number of Sense Per Word	2.42
Maximum Number of Senses	4
Minimum Number of Senses	2

Fig 4: Statistics of the Dataset

In Our training dataset we have on average of 2.42 senses per word, the maximum number of senses which we are considering in this work is 4 and the minimum number of

senses are 2.and the total number of instances for 40 words are 1470.

4.2 Evaluation

We have two test runs for our experiment in which one is for with stop word removal and manual seed selection and second one is withstopword removal and automatic seed selection.

4.3 Experiment

We have tested our data on average 15 instances per word, therefore total testing instances are 605 and the training instances are 865of all 40 words. The tagged instances are shown in the following

इनका झापड़ा निकटवता जगला म पाय जानवाल वृक्षा का छाटा-छाटा शाखाओं
एव उनका पात्तया का बना हाता ह [0]
सवप्रथम य वृक्ष का पतला-पतला शाखाआ का बाधकर झा पड़ा का ढ़ाचा तैयार
कर लत ह आर इसक बाद इस ढ़ाच पर पत्ता का छाना ढ़ाल दत ह [0]
उराल नामक एक छाट वृक्ष (शाल क सदृश्य) का शाखाआ स भा निकाला जाताहै [0]

सगर वंश शाखा क सम्नबध म भा एसा हा बात ह [1]
परन्तु सगर, हारश्चन्द्र हार वश आर दाक्षण काशल शाखाए सवमान्य घटनाआ स पृथक प्रमाणत ह [1]
ावदह क तान सूय वश, मायलशाखा, साकाश्यशाखा आर ऋतुाजत शाखाआ मऋतुाजत शाखा क सारध्वज-जनक [1]

ऋण आवदन का स्वाकृात क लए बक शाखा का १५ दन
का वतमान समय सामा जारा रहगा [2]
ब्लाक कायालय स बका म समान गात स आवदन जात रहन चाहए ाजसस आवदना क ढर न लग आर बक शाखाआ पर भार न बढ [2]
ऋण आवदन का स्वाकृात क लए बक शाखा का १५ दन का वतमान समय सीमा जारा रहगा [2]

Fig 6: Manually tagged Instances of word "शाखा"

5 Results and Discussion

These are results of the ten words and their three senses. When we are extending the number of words their average precision is varies between 50 to 60%.When we are considering these type of instances

"माँग को सामान्यतः एक तालिका या ग्राफ़ के रुप म प्रदाशत करते ह ाजसम कामत और इाच्छत मात्रा का सबन्ध दखाया जाता है"

In the above sentence there are three ambiguous words in one instance, and for this case our algorithm did not give better result.

Word	No Noof Senes	Precision			
		Sense number			
		First	Second	Third	Fourth
I	3	0.634	0.567	0.588	-
II	2	0.579	0.372	-	-
III	2	0.420	0.479	0.530	-
IV	2	0.704	0.590	-	-
V	3	0.517	0.501	0.500	-
VI	2	0.608	0.68	0.675	-
VII	2	0.525	0.534	-	-
VIII	4	0.294	0.394	0.411	0.435
IX	3	0.438	0.463	0.396	-
X	2	0.362	0.432	-	-

Fig 7 : Some set of Words and their precisions for three senses

In the above figure the set of ten words and their respective precisions for possible senses (maximum 4).

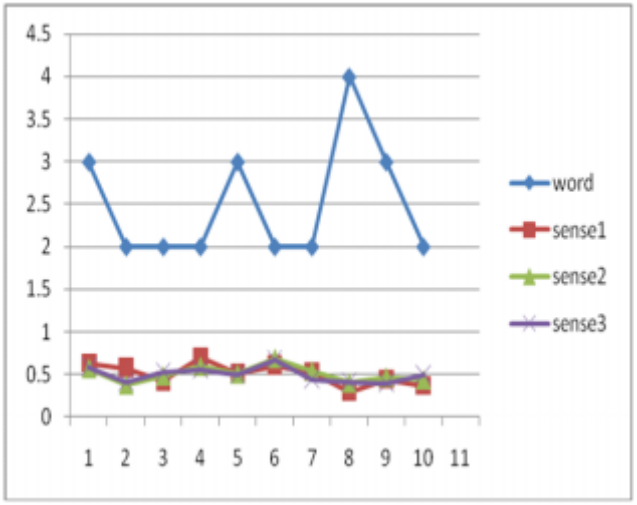

Fig: Performance analysis graph words

6 Conclusion and Future Work

In this research paper, we investigate the yarowsky [5] algorithm, in which we did some modifications like stop word removal and stemming. We observed the best overall performance for the case when both operations had jointly performed on them.
We also extend our dataset including their synonyms and homonyms'.

References

1. Ide N.Veronis J. Word Sense Disambiguation.The state of the art (1998)
2. Sinha M.Kumar M.,Pande P.,Kashyap L.,:Hindi Word Sense Disambiguation. International Symposium on Machine Translation,Natural Language Processing and Translation Support System,Delhi, India,(2004)
3. Stevenson M.,Yorick,W.: The Interaction of Knowledge Sourses inWordSenseDisambiguation,ComputationalLinguistics,321-349(2001)
4. Yarowsky ,D. :Word Sense Disambiguation using statistical models of Rogets Categories trained on larged corpora". In Proceedings of the 14th International conference on computational Linguistics(COLING-92),pp.454-460,Nantes,France(1992)
5. Yarowsky D.: " Unsuperwised word Sense disambiguation Rivaling supervised Methods" ,Proceedings of the 33rd Annual Meeting of the association for Computational Linguistics,Cambridge,MA, pp.189-196.(1995)
6. Lesk, Michel : Automated sense Disambiguation Using Machine-readable dictionaries: How to tell a pine cone from an Ice cream Cone." Proceedings of the 1986 SIGDOC conference, Toronto, Canada, June 1986, pp 24-26.(1986)
7. Fellbaum C., Alkhalifa M., Black W., Elkateb S., Pease A., Rodriguez H., Vosesen P.: Domain-specific Word Sense Disambiguation. In Eneko Agirre and Philip Edmonds (eds) Word Sense Disambiguation – Algorithms and Applications, Springer, June 2006
8. W.Gale, K. Church, and D.Yarowsky. : One sense Per discourse. In Proceedings of the DARPA speech and Natural LanguageWorkshop,pp.233-237,Harriman,NY,Feburary.(1992)
9. Mihalcea R.: Unsupervised Large-Vocabulary word sense disambiguation with graph based algorithms for sequence data labeling. In Proceedings of HLT/EMNLP, pp.411-418, Vancouver, BC (2005)
10. Broody et al, Broody S., Navigli R. Lapata M.: Ensemble methods for Unsupervised WSD .In Proceedings of the ACL/COOLING, Sydney, Australia,(2006).
11. Mishra et al.,: An Unsupervised approach to Hindi Word Sense Disambiguation in proceedings of the IHCI, pp.327-336 Allahabad, India(2009).
12.