JOURNAL'S NAME
Vol. XX, No. XX, XXX-XXX 200X

Size of N for Word Sense Disambiguation using N gram model for Punjabi Language

GURPREET SINGH JOSAN*, GURPREET SINGH LEHAL#
**Lecturer, Dept. of CSE, Yadwindra College of Engg., Talwandi Sabo.*
E-Mail: josangurpreet@rediffmail.com

#Professor, Dept. of Computer Science, Punjabi University Patiala.
E-Mail: gslehal@gmail.com

ABSTRACT

N-grams are consecutive overlapping N-character sequences formed from an input stream. N-gram models are extensively used in word sense disambiguation. In this paper we tried to find out whether higher order n gram models improves the word sense disambiguation in Punjabi language and whether it has any relation with entropy of the models. In our experiments statistical analysis of n gram models for n ranging from ±1 to ±6 is done. We also tried to explore the possibility of disambiguation by using future knowledge. From this experiment it became clear that lower order n gram models are sufficient for word sense disambiguation and larger n gram model gives little improvement. Disambiguation with the help of future knowledge also gives promising results.

Keywords: N Gram Model, Word Sense Disambiguation, Entropy

1. INTRODUCTION

Word sense disambiguation is widely studied and discussed area of NLP for any natural language under consideration. The potential for word sense disambiguation varies by task. Different major applications of language differ in their potential to make use of successful word sense information [13]. The potential for using word senses in machine translation seems rather more promising [13, 14]. Statistical language modeling has been widely used for such type of problems. The goal of language modeling is to predict the probability. These probability

estimations are further exploited to perform higher level tasks such as structuring and extracting the information from natural languages. The concept is widely implemented in spoken as well as written text. In Machine translation, one of the uses of statistical language model is selecting correct word sense among the possible senses, given a sequence of words in the local context of the ambiguous words. One of such statistical model is N gram model. An N-gram is simply a sequence of successive n words along with their count i.e. number of occurrences in training data [6,8]. For computational reasons, Markov assumptions are applied which states that current word does not depends on the entire history of the word but at most on the last few words [8]. The number of words in the local context of ambiguous words makes a window. The size of window i.e. number of words to be considered at ± n positions is important because while constructing n size window following factors are of main concern.

a) Larger the value of n, higher is the probability of getting correct word sense i.e. for the general domain; more training data will always improve the result. But on the other hand most of the higher order n grams do not occur in training data. This is the problem of sparseness of data.
b) As training data size increase, the size of model also increase which can lead to models that are too large for practical use. The total number of potential n grams scales exponentially with n. Computer up to present could not calculate for a large n because it require huge amount of memory space and time.
c) Does the model get much better if we use a longer word history for modeling an n-gram?
d) Do we have enough data to estimate the probabilities for the longer history?

To deal with the problem of selecting size n of the language model for word sense disambiguation, the two most widely used evaluation metrics are the entropy and sense disambiguation rate. In this study we tried to investigate the effect of size n of window by correlating them with perplexity and sense disambiguation rate. Word sense disambiguation rate is defined as percentage of words which are correctly disambiguated in the translation. The entropy on the other hand is a measure of information and it can be used as metric for how predictive a given N gram model is about what the possible sense of word could be. Given a word sequence w1,w2,.....wn to be used as a

test corpus, the quality of language model can be measured by the empirical perplexity and entropy scores on this corpus as

$$\text{Entropy} = -\frac{1}{n}\sum_{W1n}\Pr(w_1.....w_n)\log_2\Pr(w_1.....w_n)$$

$$\text{Perplexity} = \sqrt[N]{\prod_{i=1}^{N}\frac{1}{\Pr(w_i \mid w_1, w_2, ... w_{i-1})}}$$

$$= 2^{\text{Entropy}}$$

Where n = Total number of words in test set.

Pr = conditional probability

$$= \frac{Count(h_i, w_i)}{Count(h_i)}$$

For a stationary and ergodic language entropy can be measured as

$$\text{Entropy} = -\frac{1}{n}\sum_{W1n}\log_2\Pr(w_1.....w_n)$$

The goal is to obtain small values of these measures. Language model with lower perplexities and entropies tend to have higher word sense disambiguation rates. Or in other word perplexity is related inversely to the likelihood of the test sequence according to the model [8, 15]. But there have been numerous examples in the literature where language model providing a large improvements in perplexity over a base line model have yielded a little or no improvement in evaluation estimations [3]. In this research, we attempt to find out the relations between Entropy and improvement in word sense disambiguation rates by applying the concepts on Punjabi language which is an official language of Punjab state in India and ranked as 12th highest used language in world. We also aim to find out the optimum size of n for n gram model for using it in word sense disambiguation purpose.

2. PREVIOUS WORK

Claude E. Shannon [16] established the information theory in 1951. This theory included the concept that a language could be approximated

by an nth order Markov model by n to be extended to infinity. Shannon computed the per letter entropy rather than per word entropy. He gives entropy of English text as 1.3 bits per letter. Since his proposal there were many trials to calculate n grams for a big text data of a language. Brown et. al.[2] performs a test on much larger text and give an upper bound of 1.75 bits per character for English language by using trigram model. Iyer et al. [7] investigate the prediction of speech recognition performance for language model in the switchboard domain, for trigram model built on different amounts of in domain and out of domain training data. Over the ten models they constructed, they find that perplexity predicts word error rate well when only in domain training data is used, but poorly when out of domain text is added. They find that trigram coverage or the fraction of trigram in the test data present in training data is a better predictor of word error rate than perplexity.

Chen et al.[3] investigate their language model for speech recognition performance in the Broadcast news domain and concluded that perplexity correlates with word error rate remarkably well when only considering n gram model trained on in domain data.

Manin[9] performs a study on predictability of word in context and found that unpredictability of a word depends upon the word length. Marti et. al.[10] tested different vocabulary size and concluded that language models become more powerful in recognition tasks with larger vocabulary size. Resnik et. al. [13,14] made several observations about the state of the art in automatic word sense disambiguation and offer several specific proposals to the community regarding improved evaluation criteria, common training and testing resources, and the definitions of sense inventories.

No such attempt has been found in the literature for Punjabi language. In this work, we attempted to find the optimum value for window size for the problem of word sense disambiguation in Punjabi language. We are going to investigate the relation between word error rate and perplexity for the Punjabi language. We also aim to find out whether increasing the size of window will generate lower values of perplexity and word error rate or not.

3. METHODOLOGY

In this research, we investigate the improvement of machine translation system with respect to word sense disambiguation of Punjabi text.

The Training Data: We generated different n gram models where n ranges from ±1 to ±6 i.e. we generate a window of size ±1 to ±6 words around the given ambiguous word. The n grams are generated from 500K words corpus from different sources like essays, stories, novels, NEWS, articles etc.

The Test Set: Two types of test sets are created; one is with data from training set and other with data not from training set. Both sets contain approx. 1000 tokens. Sparseness data problem is dealt by smoothing the n-gram models with deleted interpolation method described in [8].

Probabilities of different n grams are found out for the two test data sets and then entropy is computed according to the formula discussed earlier. The test sets are then checked for the %age of incorrectly disambiguated words by using different n grams for disambiguating purpose.

4. RESULTS AND DISCUSSION

Table 4.1 shows the entropy and %age of Incorrect Disambiguated words for different n gram models. For the in domain data, the entropy of model decreases with the increase in size of the model. This indicates that higher the value of n, better are the chances of providing information by the n gram model for disambiguating the word. This is evident from the corresponding values of percentage of incorrectly disambiguated words. The results are looking promising because the test data is from the training data domain and so the frequency of occurrence of a particular n gram in model is higher. Consequently every n gram gets higher probability values and can give better prediction about the possible sense of words. This is shown in figure 4.1 and 4.2.

On the other hand, for the out of domain data, entropy of models decreases for bi-gram and then it increases as we increase the model size. This behavior indicates that a bi-gram model definitely has an edge over uni-gram model as far as word sense disambiguation is concerned.

Table 4.1 Entropy and %age of Incorrect Disambiguated words for different n gram models

N gram	SET 1 (Data not from training set)		SET 2 (Data from training set)	
	Entropy	%age of Incorrect Disambiguated words	Entropy	%age of Incorrect Disambiguated words
1	11.05	35.18	10.18	37.73
2	9.53	14.81	6.36	11.3
3	9.91	27.77	3.07	7.5
4	10.14	29.03	1.89	5.8
5	12.31	32.31	1.57	5.8
6	13.12	33.07	1.91	5.2
Hybrid (tri-bi-uni)	--	6.96	--	3.7

This is also indicated by the percentage of incorrectly disambiguated words the value of which decrease sharply while shifting from unigram to bigram. For trigram and higher, increase in entropy values are due to sparseness of data. In other words, we have not enough number of n grams in a particular model and consequently we have very little probability of getting a particular sequence of words in an n-gram model. Due to lower probability of getting a sequence in an n-gram model, its chances to disambiguate a word are also very few. This pattern is shown in the percentage of incorrect disambiguated words. These figures are also increased after decreasing sharply. See Fig 4.3

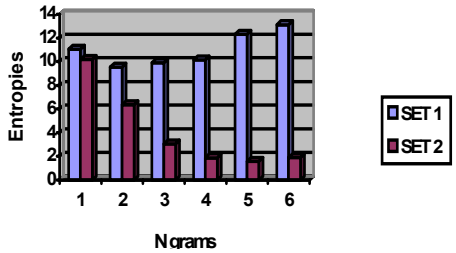

Fig 4.1 Entropies for SET 1 and SET 2

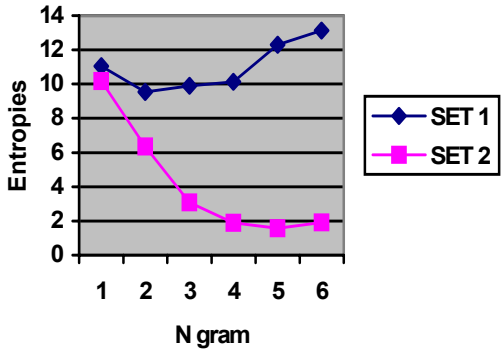

Fig 4.2 Change In Entropies With N Gram For In Domain And Out Of Domain Data

As far as relation between entropy and size of n is concerned, we can conclude that they are directly associated with each other as far as the question is of word sense disambiguation. Entropy is a reliable parameter to find out the suitability of any model for the purpose of information handling and manipulation in NLP as proved earlier in many literatures. Similar are the findings for the language under consideration and shown in fig 4.4.

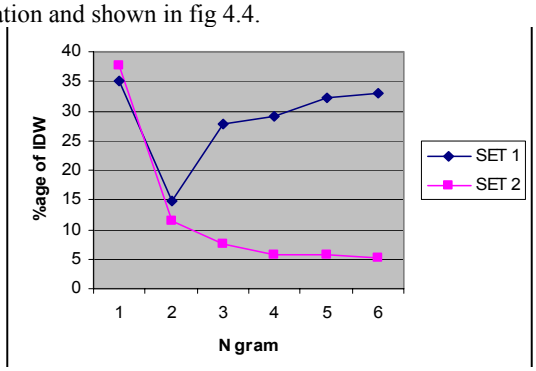

Fig 4.3 Change in %age of Incorrectly Disambiguated words with n grams

Another interesting point observed is that instead of making and using a higher order n gram models, we can improve the efficiency of the system tremendously by utilizing lower order models jointly. That is we can use tri-gram model in the first place to disambiguate a word. If it

fails to disambiguate then we move to lower order model i.e. bi-gram model for WSD. If it also fails, we can use the unigram model. With this technique we get only 7.96% and 3.7% of incorrectly disambiguated words for both SET1 and SET 2 respectively. This shows that we can effectively use this methodology for getting good results.

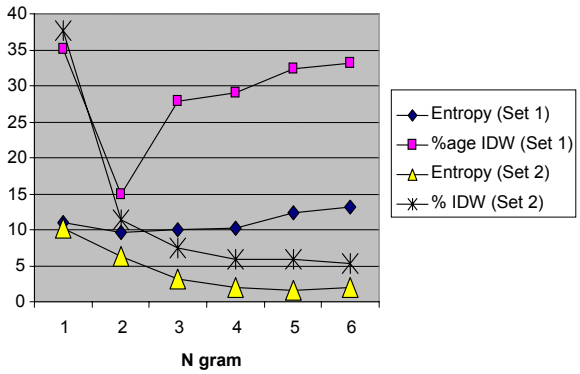

Fig 4.4 Relationship Between Entropies And Percentage Of Incorrectly Disambiguated Words For Set 1 And Set 2.

Lastly, improvement in results for percentage of incorrectly disambiguated words is noticed if we also consider the words next in sequence. Because in any n gram model, we have the possibility of looking the word sequence next to the current word and we can estimate the probabilities of such sequences easily. This information can be exploited for the word sense disambiguation process. Due to the language structure of Punjabi, there are about 7% of ambiguous words in Punjabi text that can be disambiguated by looking on the next words in sequence. Most of the cases are solved by jointly using tri-gram and bi-gram as discussed in previous paragraph.

5. CONCLUSION

In this experiment, we tried to find out whether higher order n gram models improves the word sense disambiguation in Punjabi language and whether it has any relation with entropy of the models. The most

important of our observation in this work is that we can improve the word sense disambiguation for Punjabi language by using the n gram models. In stead of generating a higher order model of n gram, which is time consuming, hard to create and maintain, and of course need a lot of data to get meaningful results, we can make use of combination of lower order n gram model. It is also observed that word sequence next to the current word can effectively be used for the word sense disambiguation purpose. Entropy is proved to be a reliable parameter to judge the suitability of n-gram models for word sense disambiguation process.

6. REFERENCES

1. Bonafonte A., & Marino J.B.,"*Language Modeling using x-grams*", Spoken Language, ICSLP 96. Proceedings., Fourth International Conference on 3-6 Oct 1996, Volume: 1, pp 394-397
2. Brown P.F. & Pietra S.A.D., "*An estimation of upper bound for the entropy of english*", Association for Computational Linguistics, Volume 18, Number 1,1992, pp 31-40
3. Chen S., Beeferman D., & Rosenfeld R.,"*Evaluation Metrics for language models*", Appeared at the Broadcast News Transcription and Understanding Workshop, February 1998.
4. Diab M.,"*Relieving the data acquisition bottleneck in word sense disambiguation*", Proceedings of the 42nd meeting of the Association for Computational Linguistics (ACL), Barcelona, Spain, 303-310.
5. Gao J, Li M., Lee k.,"*N-gram Distribution based language model adaptation*", In proceedings of ICSLP, 2000
6. Gotoh Y. & Renals S., "*Statistical Language Modeling*", Text and speech Triggered Information Access, S. Renals and G. Grefenstette(eds.), Springer,2003.
7. Iyer R., Ostendorf M., Meteer M., “Analysing and predicting language model improvements”, In proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding, 1997.
8. Jurafsky D. & Martin J.,"Speech and Language Processing: An Introduction to speech recognition, computational linguistics and natural language processing", Prentice-Hall, New Jersey, chapter 4.

9. Manin D.Y., "*Experiments on predictability of word in context and information rate in natural language*", INFORMATION PROCESSES, Electronic Scientific Journal, ISSN: 1819-5822, March,2006, pp 229-236
10. Marti U.V. & Bunke H.,"*On the influence of vocabulary size and language models in unconstrained handwritten text recognition*", Proc. 6th Int. Conference on Document Analysis and Recognition, 2001, 260 – 265.
11. Matsuoka T., Taguchi Y, Ohtsuki K., Furui S., Shirai k., "*Toward Automatic recognition of Japanese Broadcast NEWS*", In the Proceedings of DARPA, 1997, pp 181-184
12. Moradi H., Grzymala-Busse J.W., Roberts J. A., "*Entropy of english text: experiments with human and a machine learning system based on rough sets*", Information Sciences, An International Journal 104(1998), pp 31-47
13. Resnik P. & Yarowsky D., *"A Perspective on word sense disambiguation methods and their evaluation*", In proceedings of ACL-SIGLEX Workshop on "*Tagging Text with Lexical Semantics: Why, What, and How?"* April 4-5, 1997, Washington, D.C., 79-86
14. Resnik P. & Yarowsky D., "*Distinguishing Systems and distinguishing senses: New evaluation methods for word sense disambiguation*", Natural language engineering 1(1):000-000, cambridge university press, 1998.
15. Roukos S.,"*Language representation*", in R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, V. Zue (eds), *Survey of the state of the art in human language technology*, Chapter 1.6, Center for Spoken Language Understanding, 1996.
16. Shannon C.E., "*Prediction and entropy of printed english*", The bell system technical journal, january 1951, pp 50-65
17. Wang S. Schuurmans D. & Peng F.,"*Latent Maximum Entropy Approach for Semantic N-Gram Language Modeling*", In C. M. Bishop and B. J. Frey (eds), Proceedings of the 9th International Conference on Artificial Intelligence and Statistics (AISTATS-03). January 3-6, 2003, Key West, Florida, USA

GURPREET SINGH JOSAN
LECTURER, DEPT. OF COMPUTER SCIENCE
YADWINDRA COLLEGE OF ENGINEERING,
TALWANDI SABO
CONTACT: JOSANGURPREET@REDIFFMAIL.COM,
PHONE 9914347847

GURPREET SINGH LEHAL
PROFESSOR, DEPT. OF COMPUTER SCIENCE,
PUNJABI UNIVERSITY PATIALA,
CONTACT: GSLEHAL@GMAIL.COM.