Maximum Entropy Approach based Named Entity Recognition in Punjabi Language

Arshdeep Singh
Computer Sc. & Engg. Dept
PTUGZS Campus
Bathinda, India

Jyoti Rani
Computer Sc. & Engg. Dept
PTUGZS Campus
Bathinda, India

Amandeep Kaur
UCOE
Punjabi University
Patiala, India

ABSTRACT
Named Entity Recognition is the task of identifying and classifying named entities into some predefine categories like person, location, organization etc. NER is used in many applications like text summarization, text classification, question answering and machine translation systems etc. For English a lot of work has already been done in the field of NER, where capitalization is a major key for rules, whereas Indian languages do not have such feature. This makes the task difficult for Indian Languages. This work reports about the evaluation of a Named Entity Recognition (NER) system for Punjabi language using the Maximum Entropy Approach (MAXENT). A manually tagged Punjabi news corpus is used for the evaluation which was developed from Punjabi newspaper available online. The training set annotated with a NE tagset of 12 tags is used. A MAXENT based NER system for Punjabi has reported an overall Precision, Recall and F-Score values of 90.92%, 72.30% and 80.55% respectively with feature set context word, Part of Speech (POS) information, NE tag of previous word and First name Gazetteer list.

Keywords
Named Entity Recognition, Named Entity, Maximum Entropy, NLP, Punjabi.

1. INTRODUCTION
Named Entities (NE) are phrases that contain person, organization, location, number, time, measure etc. Named Entity Recognition (NER) is the task of identifying and classifying the Named Entities into predefine categories such as person, organization, location, etc in the text. NER is an Information Extraction (IE) task which aims to locate named entities in a particular document. It is also considered as the core of natural language processing (NLP) system. NER is widely used in Natural Language Processing. The NER task has important significance in the in the Internet search engines and is an important task in many of the Language Engineering applications such as Machine Translation, Question-Answering systems, Information Retrieval and Automatic Summarization.

Various conferences being held that explores the work in various languages in NER. The sixth series of Message Understanding Conference (MUC-6) held in Columbia, 1995 involved the evaluation of information extraction. A Tagset of three tags (Enamex, Numex and Timex) was used in this conference. The topic domain for this conference was "Negotiation of Labor Disputes and Corporate Management Succession". Conference on Computational Natural Language Learning (CONNL-2002) was held in Taiwan,2002 that concerns with language independent Named Entity Recognition and concentrated on two European languages i.e Spanish and Dutch. A tagset containing four tags (Person, Location, Organization and Miscellaneous) were used in this conference. A boosted decision tree obtained the best performance for both data sets [3]. A workshop on NER for South and South East Asian Languages (NERSEAL-2008) organized by IJCNLP was held in IIT, Hyderabad in 2008 that dealt with the South and South East Asian Languages for the NER task.

NER approaches can be classified under three categories. The Linguistic/Rule based Approach that needs skilled linguistics to handcraft rules. It requires advanced knowledge of grammar and other related rules [8]. The Machine Learning based approach which is the current trend in NER as it is trainable, adaptable and its maintenance is much cheaper than rule based. The representative machine learning approaches used in NER are Hidden Markov Model (HMM) [1], Maximum Entropy Model [2], Decision Tree [19], Conditional Random Field (CRF) [13] and Support Vector Machine (SVM) [21]. Final category is Hybrid NER that combines both the rule based and the machine based approaches for more accuracy to identify named entities [20].

We have evaluated a Maximum Entropy Markov Model, a machine learning approach for named entity recognition in punjabi language and identified suitable features for this NER task.

The paper is organized as follows. Previous work done in NER in various languages in Section II, followed by introduction to Maximum Entropy model in Section III, in Section IV the architecture of the MAXENT based Punjabi NER system is discussed and finally in Section V we conclude the paper.

2. PREVIOUS WORK
A lot of work has been done in Indian languages using various machine learning approaches. Some of them have been summarized as follows:

The work regarding Telugu language is mentioned in [18]. The evaluation has reported precision, recall and F-Score of 64.07%, 34.57% and 44.97% respectively.

[7] has reported Lexical F-Score of 40.63%, 50.06%, 39.04%, 40.94%, and 43.46% for Bengali, Hindi, Oriya, Telugu and Urdu respectively. They have used a two stage hybrid approach using CRF based machine learning followed by some heuristics or rules for the task of named entity recognition for South and South East Asian languages.

The work of [6] reports about the development of a NER system for Bengali language using Support Vector Machine. Precision, Recall and F-Score are claimed to be 89.4%, 94.3%, and 91.8% respectively. He has experimented with 150K words data that was manually tagged with 16 NE tags.

In [11] a comparative study of Conditional Random Field and Support Vector Machines for recognizing named entities in

Hindi language is done. They have claimed F-Score of 47%, and 37% for CRF and SVM respectively.

The work using Maximum Entropy Approach for Bengali and Hindi languages was demonstrated in [9]. They have used named entity tagset with four tags namely Person name, Location name, Organization name and Miscellaneous name. They have achieved overall average recall, precision and f-score values of 88.01%, 82.63% and 85.22% respectively for Bengali and 86.4%, 79.23% and 82.66%, respectively for Hindi.

[5] demonstrates the highest maximal F-measure, nested F-measure and lexical F-measure of 53.36%, 53.46% and 59.39 respectively for Bengali language for recognizing the named entities using 122K, 502K, 64K, 93K and 35K tokens for Bengali, Hindi, Telugu, Oriya and Urdu respectively .

The work in Telugu language using Maximum Entropy is mentioned in [15]. They have experimented with four tags and reported f-measures of 72.07%, 60.76%, 68.40% and 45.28% for Person, Organization, Location and Other tags respectively.

[12] reports the development of a MAXENT in hindi language for NER task and claimed to have an average F measure of 71.9% for a tagset of four named entity tags namely person, organization, location and others.

The development of NER system in Hindi using MAXENT and transliteration by [17] reported the maximum F-Value to be 81.12% by incorporating the gazetteer lists with the best feature set.

NER system in Punjabi Language claimed by [10] reports overall Precision, Recall and F-Value to be 88.05%, 74.85% and 80.92% respectively. They have experimented with 25k manually tagged data with 12 different tags using CRF machine learning approach.

3. MAXIMUM ENTROPY MARKOV MODEL

The Maximum Entropy Framework estimates probabilities based on the principle of making as few assumptions as possible, other than the constraints imposed. Such constraints are derived from training data, expressing some relationship between features and outcome. MAXENT computes the probability p(h|t) of history h together with a tag t is defined as:

$$P(h,o) = \frac{1}{Z(h)} \prod_{j=1}^{k} \alpha_j^{f_j(h,o)}$$

where o refers to the outcomes (or tags), h is the history (or context), and Z(h) is the normalization function [14]. The features used in the maximum entropy framework are binary. An example of a feature function is:

$$f_j(h,o) = \begin{cases} 1, & if\ 0 = org - B, word = PETER \\ 0, & otherwise \end{cases}$$

The parameters α_j are estimated by a procedure called Generalized Iterative Scaling (GIS) [4]. This is an iterative procedure that improves the estimation of the parameters at each iteration. Each parameter α_j corresponds to a feature f_j.

For a given sequence of words $\{w_1,\ldots w_n\}$ and tags $\{o_1,\ldots o_n\}$ as training data, define h_i as the history available when predicting o_i. The parameters$(\alpha_1,\ldots\ \alpha_k\}$ are then chosen to maximize the likelihood of the training data using p:

$$L(p) = \prod_{i=1}^{n} p(h_i, o_i) = \prod_{i=1}^{n} \frac{1}{Z(h)} \prod_{j=1}^{k} \alpha_j^{f_j(h_i,o_i)}$$

For our development we have used a java based open NLP MAXENT toolkit to get the probability values of a word belonging to each class. That is, given a sequence of words, the probability of each class is obtained for each word. To find the most probable tag corresponding to each word of a sequence, we can choose the tag having the highest class conditional probability.

4. MAXENT BASED PUNJABI NER SYSTEM

Punjabi is the official language of the Indian state of Punjab and also one of the official languages of Delhi. According Ethnologue3 2005 estimate there are 88 million native speakers of Punjabi language and ranked 20th among the languages spoken in world.

Capitalization is a very important feature for English as most of the names are capitalized and it becomes easier to recognize named entities. Due to absence of the capitalization feature, Punjabi NER task is difficult. Also, person names are more diverse in Indian languages; many common words are used as names.

Punjabi like other Indian languages is a resource poor language – annotated corpora, name dictionaries etc. are not yet available in the required measure. MAXENT model has the capability to use different features to compute the conditional probabilities. Selection of an appropriate feature set is very important to train a ML based classifier.

Feature selection plays a crucial role in identification of Named Entities. Various Language independent and Language dependent features are used to effectively identify and classify Named Entities in the text. Different possible combinations of features are also used to enhance the accuracy of NER systems. Some of the features used to identify named entities are:

4.1 Feature Description

Feature selection plays a crucial role in MAXENT framework. Experiments carried out to find out the most suitable features for NER in Punjabi.

4.1.1 Surrounding words

As the surrounding words are very important to recognize a NE, previous and next words of a particular word are used as features. As a feature, previous *m* words ($w_{i-m}...w_{i-1}$) to next *n* words ($w_{i+1}...w_{i+n}$) can be treated depending on the training data size, total number of candidate features etc. In our work we have experimented on different combinations upto previous three words to next three words are used as features i.e a word window of 3, 5 and 7 Previous and next words of a particular word have been used as a feature. In our work we have experimented on word window 5and 7.

4.1.2 Word suffix and prefix

Word suffix and prefix information is helpful to identify the NEs. We have experimented with the word suffix and prefix of length 1 to 4 characters of the current word as a feature.

4.1.3 Part of speech information

For our evaluation, we have used a POS information of the current word, previous word and the next word.

4.1.4 Named entity information

The NE tag of the previous word is also taken as a feature. This is the only dynamic feature in the experiment.

4.1.5 Gazetteer lists

As there is scarcity of resources in electronic format for punjabi language, seven different lists namely Location, First name, Middle name, Last name, Day , Month and Person prefix lists prepared from corpus itself have been used.

4.2 Training Data

The training data used for this task contains of about 25 k words which was developed for the work [10] and was collected from the popular Punjabi newspaper available online http://www.ajitweekly.com. The training data is of an Open domain and contains variety of news like history, bollywood, international etc. 12 types of entities were used to be recognized. These are Person, Location, Organization, Brand, Measure, Term, Designation, Title object, Title Person, Number, Time and Abbreviation. The training data was formatted into *IOB* format [16] in which a B-XXX tag stands for the first word of an entity type XXX and I-XXX tag is used for the subsequent words of an entity. The tag O indicates the word that is not NE.

4.3 Evaluation

About 50 different experiments were conducted taking several combinations from the mentioned features to identify the best feature set for the NER task. We have evaluated the system using a test file of size 7 K words of an open domain. The accuracies are measured in terms of F-measure, which is the weighted harmonic mean of precision and recall. Where Precision is the percentage of the correct annotations and recall is the percentage of the total named entities that are successfully annotated and are calculated as:

$$Precision = \frac{Number\ of\ correct\ responses}{Number\ of\ responses}$$

$$Recall = \frac{Number\ of\ correct\ responses}{Number\ of\ responses\ in\ the\ key}$$

$$F\ Score = 2 * \frac{Precision + Recall}{Precision * Recall}$$

First of all, we have used only the current and the surrounding words i.e previous word and the next word. We have experimented with several combinations of previous 3 to next 3 words (w_{i-3} ...w_{i+3}) to identify the best word-window. The results are shown in Table I

Table 1. Results of MAXENT using word as a feature

Features	Precision	Recall	F-value
w_{i-1}, w, w_{i+1}	77.8	57.2	65.9
w_{i-2}, w_{i-1}, w, w_{i+1}, w_{i+2}	79.84	44.82	57.42
w_{i-3},w_{i-2}, w_{i-1}, w, w_{i+1},	82.65	36.50	50.64
w_{i+2},w_{i+3}			

From Table 1 we can observe that word window (w_{i-1}, w, w_{i+1}) i.e one previous word, current word and one next word gives the best result for tagging the named entities. When the window size is increased, the performance starts degrading. So we have chosen three word window for our NER task.

More experiments were conducted to find the best feature set for the Punjabi NER task. The features described earlier are applied separately or in a combination to build up the MAXENT based NER model. In Table II we have summarized the results using various features and gazetteer lists.

Table 2. Results of MAXENT using Different features

Features	Precision	Recall	F-value
Ww3, NE Tag	88.31	57.85	69.91
Ww3, CPOS	82.56	67.15	74.07
Ww3, CPOS, NPOS	84.40	62.88	72.10
Ww3, CPOS, PPOS	83.24	59.93	69.70
Ww3, NE Tag, CPOS	89.20	70.89	78.99
Ww3, NE Tag, CPOS, 0<|prefix|<4, 0<|suffix|<4	87.87	71.47	78.82
Ww3, NE Tag, CPOS, 1<|prefix|<5, 1<|suffix|<5	87.03	70.94	78.17
Ww3, NE Tag, CPOS, |suffix|<=4	86.62	72.40	78.87
Ww3, NE Tag, CPOS, |prefix|<=4	86.37	69.48	77.02
Ww3, NE Tag, CPOS, |prefix|<=4, |Suffix|<=4	87.40	69.97	77.72
Ww3, NE Tag, CPOS, LocationList	91.18	71.90	80.41
Ww3, NE Tag, CPOS, FirstName	**90.92**	**72.30**	**80.55**
Ww3, NE Tag, CPOS, MiddleName	91.53	71.67	80.30
Ww3, NE Tag, CPOS, LastName	91.42	71.77	80.41
Ww3, NE Tag, CPOS, PersonPrefix	91.08	71.91	80.36
Ww3, NE Tag, CPOS, DayList	91.19	72.04	80.49
Ww3, NE Tag, CPOS, DayList, MonthList	91.45	70.72	79.76
Ww3, NE Tag, CPOS,	91.00	70.32	79.33

FirstName, MiddleName			
Ww3, NE Tag, CPOS, FirstName, MiddleName, LastName.	91.74	69.0	78.76
Ww3, NE Tag, CPOS, All Gazzetteers	92.01	65.92	76.81
Ww3= Previous word, Current word and Next word NE Tag= Named entity information of the previous tag CPOS= Current word's part of speech NPOS= Next word's part of speech 0<|Prefix|<4= prefix's of length 1, 2 and 3 0<|Suffix|<4= suffix's of length 1, 2 and 3 1<|Prefix|<5= prefix's of length 2, 3 and 4 1<|Suffix|<5= suffix's of length 2, 3 and 4 |Prefix|<=4 = prefix's of length 1, 2, 3 and 4 |Suffix|<=4 = suffix's of length 1, 2, 3 and 4			

From Table II we can observe that some of the features are able to improve the system accuracy separately, but when applied in combination with other features, they cause the accuracy to decrease. From our observations the word window 3, NE tag and the CPOS gives the best result without Gazetteer information.

After adding the gazetteer lists, we have observed that the system gives the maximum f-value of 80.55 when only the First Name list is incorporated with the baseline result. It was also observed that when we go on adding the gazetteer lists the precision value goes on increasing i.e the number of entities tagged by the system increases. This means that when we add gazetteer lists the system's ability to find the named entities increases. Though the difference in the F-Value with incorporating various gazetteer lists is marginal. The highest F-value achieved by the Punjabi NER is 80.55.

5. CONCLUSION

ML based approach requires annotated data and other resources to build NER. In this paper, we have evaluated the MAXENT based system for NER task in Punjabi language. Even though we have limited resources, the system has given decent results. We have identified the suitable features for the Punjabi NER task and it has been shown that the surrounding word window [-1,0,1], POS information of the current word, NE information of the previous word and the First name gazetteer list are the best suited features of NER task in Punjabi language.

We have achieved an overall Precision, Recall and F-score of 91.21%, 72.21% and 80.61% respectively for the Punjabi NER. The performance can further be improved by improving the gazetteer lists and using more training data.

6. REFERENCES

[1] Bikel, D.M, Miller, S., Schwartz, R. and Weischedel, R. (1997). Nymble: a high performance learning name-finder. In proceedings of the fifth conference on Applied natural language processing, pp 194-201, San Francisco, CA, USA.

[2] Borthwich, A. (1999). Maximum Entropy Approach to Named Entity Recognition. Ph.D. Thesis, New York University.

[3] Carreras, X., Marques, L. and Padro, L. (2002). Named Entity Extraction using adaboost. In proceedings of the Conference on Computational Natural Language Learning, pp. 167-170, Taipei, Taiwan.

[4] Darroch, J. N. and Ratcliff D. (1972). Generalized iterative scaling for loglinear models. Annals of Mathematical Statistics, pp 1470-1480.

[5] Ekbal, A., Haque, R., Das, A., Poka, V. and Badyopadhyay, S. (2008 a). Language Independent Named Entity Recognition in Indian Languages. In the Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pp 33-40, Hyderabad, India.

[6] Ekbal, A., Badyopadhyay, S. (2008 b). Bengali Named Entity Recognition using Support Vector Machine. In the Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pp 51-58, Hyderabad, India.

[7] Gali, K., Surana, H., Vaidya, A., Shishtla, P. and Sharma, D.M. (2008). Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition. In the Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pp 25-31, Hyderabad, India.

[8] Grishman, R. (1995).The NYU System for MUC-6 or Where's the Syntax. In the Procedings of Sixth Message Understanding Conference (MUC-6), pp 167-195, Fairfax, Viginia.

[9] Hasanuzzaman, M., Ekbal, A. and Bandyopadhyay, S. (2009). "Maximum Entropy Approach for Named Entity Recognition in Bengali and Hindi," International Journal of Recent Trends in Engineering, vol. 1, Number 1, pp 408-412.

[10] Kaur, A., Josan, G.S. and Kaur, J. (2009). Named Entity Recognition in Punjabi Language: A Conditional Random Field Approach, In the Proceedings of International Conference on Natural Language Processing, pp 277-282.

[11] Krishnarao, A.A., Gahlot, H., Srinet, A. and Kushwaha, D.S. (2009). A Comparison of Performance of sequential Learning Algorithms on the task of Named Entity Recognition for Indian Languages. In the Proceedings of 9th International Conference on Computer Science, pp 123-132, Bton Rouge, LA, USA.

[12] Kumar, N. and Pushpak, B. (2006). Named Entity Recognition in Hindi using MEMM. In Technical Report, IIT Bombay.

[13] Lafferty, J.D., McCallum, A., Pereira, F.C.N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and LAbelling Sequence Data. In the proceedings of International Conference on Machine Learning, pp 282-289, Williams College, Williamstown, MA, USA.

[14] Pietra S. D., Pietra V. D. and Lafferty J. (1997). Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 19(4): 380-393.

[15] Raju, G.V.S., Srinivasu, B., Raju, S.V. and Kumar, K.S.M.V. (2008). Named Entity Recognition for Telugu using Maximum Entropy Model. In the Proceedings of Journal of Theoretical and Applied Information Technology, pp 125-130.

[16] Ramshaw, L. and Marcus, M.(1995). Text chunking using transformation-based learning. In Proceedings of the Third Workshop on Very Large Corpora,, Somerset, New Jersey Association for Computational Linguistics, pp82-94, Somerset, New Jersey.

[17] Saha, S.K., Ghosh, P.S., Sarkar, S. and Mitra, P. (2008). Named Entity Recognition in Hindi using Maximum Entropy and Transliteration. Polibits 38, pp 33-42.

[18] Shishtla, P.M., Gali, K., Pingali, P. and Varma, V. (2008). Experiments in Telugu NER: A Conditional Random Field Approach. In the Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pp 105-110, Hyderabad, India.

[19] Sekine, S., and Eriguchi, Y. (2000). Japanese named entity extraction evaluation: analysis of results. In Proc. of the 18th COLING, pp 1106–1110.

[20] Srihari R., Niu C. and Li W. (2000). A Hybrid Approach for Named Entity and Sub-Type Tagging. In: Proceedings of the sixth conference on applied natural language processing, pp 247-254, Washington, USA.

[21] Yamada, H., Kudo, T. and Matsumoto, Y. (2002). Japanese named entity extraction using support vector machine. Transactions of Information Processing Society of Japan, pp 44–53.