Named Entity Recognition in Bengali: A Conditional Random Field Approach

Asif Ekbal
Department of CSE
Jadavpur University
Kolkata-700032, India
asif.ekbal@gmail.com

Rejwanul Haque
Department of CSE
Jadavpur University
Kolkata-700032, India
rejwanul@gmail.com

Sivaji Bandyopadhyay
Department of CSE
Jadavpur University
Kolkata-700032, India
sivaji_cse_ju@yahoo.com

Abstract

This paper reports about the development of a Named Entity Recognition (NER) system for Bengali using the statistical Conditional Random Fields (CRFs). The system makes use of the different contextual information of the words along with the variety of features that are helpful in predicting the various named entity (NE) classes. A portion of the partially NE tagged Bengali news corpus, developed from the archive of a leading Bengali newspaper available in the web, has been used to develop the system. The training set consists of 150K words and has been manually annotated with a NE tagset of seventeen tags. Experimental results of the 10-fold cross validation test show the effectiveness of the proposed CRF based NER system with an overall average Recall, Precision and F-Score values of 93.8%, 87.8% and 90.7%, respectively.

1 Introduction

Named Entity Recognition (NER) is an important tool in almost all Natural Language Processing (NLP) application areas. Proper identification and classification of named entities (NEs) are very crucial and pose a very big challenge to the NLP researchers. The level of ambiguity in NER makes it difficult to attain human performance. NER has applications in several domains including information extraction, information retrieval, question-answering, automatic summarization, machine translation etc.

The current trend in NER is to use the machine-learning approach, which is more attractive in that it is trainable and adoptable and the maintenance of a machine-learning system is much cheaper than that of a rule-based one. The representative machine-learning approaches used in NER are Hidden Markov Model (HMM) (BBN's IdentiFinder in (Bikel et al., 1999)), Maximum Entropy (New York University's MENE in (Borthwick, 1999)) and Conditional Random Fields (CRFs) (Lafferty et al., 2001; McCallum and Li, 2003).

There is no concept of capitalization in Indian languages (ILs) like English and this fact makes the NER task more difficult and challenging in ILs. There has been very little work in the area of NER in ILs. In Indian languages particularly in Bengali, the work in NER can be found in (Ekbal and Bandyopadhyay, 2007a; Ekbal and Bandyopadhyay, 2007b) with pattern directed shallow parsing approach and in (Ekbal et al., 2007c) with HMM. Other than Bengali, a CRF based NER system can be found in (Li and McCallum, 2004) for Hindi.

2 Conditional Random Fields

Conditional Random Fields (CRFs) (Lafferty et al., 2001) are used to calculate the conditional probability of values on designated output nodes given values on other designated input nodes. The conditional probability of a state sequence $S = < s_1, s_2, \ldots, s_T >$ given an observation sequence $O = < o_1, o_2, \ldots, o_T >$ is calculated as:

$$P_\wedge(s|o) = \frac{1}{Z_0} \exp(\sum_{t=1}^{T} \sum_{k} \lambda_k \times f_k(s_{t-1}, s_t, o, t)),$$

where, $f_k(s_{t-1}, s_t, o, t)$ is a feature function whose weight λ_k, is to be learned via training. The values of the feature functions may range between $-\infty, \ldots + \infty$, but typically they are binary. To make all conditional probabilities sum up to 1, we must calculate the normalization factor,

$$Z_0 = \sum_s \exp(\sum_{t=1}^{T} \sum_k \lambda_k \times f_k(s_{t-1}, s_t, o, t)),$$

which as in HMMs, can be obtained efficiently by dynamic programming.

To train a CRF, the objective function to be maximized is the penalized log-likelihood of the state sequences given the observation sequences:

$$L_\wedge = \sum_{i=1}^{N} \log(P_\wedge(s^{(i)}|o^{(i)})) - \sum_k \frac{\lambda_k^2}{2\sigma^2},$$

where $\{< o^{(i)}, s^{(i)} >\}$ is the labeled training data. The second sum corresponds to a zero-mean, σ^2 -variance Gaussian prior over parameters, which facilitates optimization by making the likelihood surface strictly convex. Here, we set parameters λ to maximize the penalized log-likelihood using Limited-memory BFGS (Sha and Pereira, 2003), a quasi-Newton method that is significantly more efficient, and which results in only minor changes in accuracy due to changes in λ.

When applying CRFs to the NER problem, an observation sequence is a token of a sentence or document of text and the state sequence is its corresponding label sequence. While CRFs generally can use real-valued functions, in our experiments maximum of the features are binary valued. A feature function $f_k(s_{t-1}, s_t, o, t)$ has a value of 0 for most cases and is only set to be 1, when s_{t-1}, s_t are certain states and the observation has certain properties. We have used the C++ based OpenNLP CRF++ package [1].

3 Named Entity Recognition in Bengali

Bengali is one of the widely used languages all over the world. It is the seventh popular language in the world, second in India and the national language of Bangladesh. A partially NE tagged Bengali news corpus (Ekbal and Bandyopadhyay, 2007d), developed from the archive of a widely read Bengali news paper available in the web, has been used in this work to identify and classify NEs. The corpus contains around 34 million word forms in ISCII (Indian Script Code for Information Interchange) and UTF-8 format. The *location*, *reporter*, *agency* and different *date* tags (*date*, *ed*, *bd*, *day*) in the partially NE tagged corpus help to identify some of the location, person, organization and miscellaneous names, respectively, that appear in some fixed places of the newspaper. These tags cannot detect the NEs within the actual news body. The date information obtained from the news corpus provides example of miscellaneous names. A portion of this partially NE tagged corpus has been manually annotated with the seventeen tags as described in Table 1.

NE tag	Meaning	Example
PER	Single-word person name	*sachin*/ PER
LOC	Single-word location name	*jadavpur*/LOC
ORG	Single-word organization name	*infosys*/ ORG
MISC	Single-word miscellaneous name	100%/ MISC
B-PER I-PER E-PER	Beginning, Internal or End of a multiword person name	*sachin*/B-PER *ramesh*/I-PER *tendulkar*/E-PER
B-LOC I-LOC E-LOC	Beginning, Internal or End of a multiword location name	*mahatma*/B-LOC *gandhi*/I-LOC *road*/E-LOC
B-ORG I-ORG E-ORG	Beginning, Internal or End of a multiword organization name	*bhaba*/B-ORG *atomic*/I-ORG *research*/I-ORG *center*/E-ORG
B-MISC I-MISC E-MISC	Beginning, Internal or End of a multiword miscellaneous name	*10e*/B-MISC *magh*/ I-MISC *1402*/E-MISC
NNE	Words that are not NEs	*neta*/NNE

Table 1: Named Entity Tagset

3.1 Named Entity Tagset

A CRF based NER system has been developed in this work to identify NEs in Bengali and classify them into the predefined four major categories, namely, 'Person name', 'Location name', 'Organization name' and 'Miscellaneous name'. In order to

[1] http://crfpp.sourceforge.net

properly denote the boundaries of NEs and to apply CRF in NER task, sixteen NE and one non-NE tags have been defined as shown in Table 1. In the output, sixteen NE tags are replaced appropriately with the four major NE tags by some simple heuristics.

3.2 Named Entity Features

Feature selection plays a crucial role in CRF framework. Experiments were carried out to find out the most suitable features for NER in Bengali. The main features for the NER task have been identified based on the different possible combination of available word and tag context. The features also include prefix and suffix for all words. The term prefix/suffix is a sequence of first/last few characters of a word, which may not be a linguistically meaningful prefix/suffix. The use of prefix/suffix information works well for highly inflected languages like the Indian languages. In addition, various gazetteer lists have been developed for use in the NER task. We have considered different combination from the following set for inspecting the best feature set for NER task: F={$w_{i-m}, \ldots, w_{i-1}, w_i, w_{i+1}, \ldots w_{i+n}$, |prefix| $\leq n$, |suffix| $\leq n$, previous NE tag, POS tags, First word, Digit information, Gazetteer lists}.

Following are the details of the set of features that were applied to the NER task:

• Context word feature: Previous and next words of a particular word might be used as a feature.

• Word suffix: Word suffix information is helpful to identify NEs. This feature can be used in two different ways. The first and the naïve one is, a fixed length word suffix of the current and/or the surrounding word(s) might be treated as feature. The second and the more helpful approach is to modify the feature as binary valued. Variable length suffixes of a word can be matched with predefined lists of useful suffixes for different classes of NEs. The different suffixes that may be particularly helpful in detecting person (e.g., *-babu*, *-da*, *-di* etc.) and location names (e.g., *-land*, *-pur*, *-lia* etc.) have been considered also. Here, both types of suffixes have been used.

• Word prefix: Prefix information of a word is also helpful. A fixed length prefix of the current and/or the surrounding word(s) might be treated as features.

• Part of Speech (POS) Information: The POS of the current and/or the surrounding word(s) can be used as features. Multiple POS information of the words can be a feature but it has not been used in the present work. The alternative and the better way is to use a coarse-grained POS tagger.

Here, we have used a CRF-based POS tagger, which was originally developed with the help of 26 different POS tags[2], defined for Indian languages. For NER, we have considered a coarse-grained POS tagger that has only the following POS tags:

NNC (Compound common noun), NN (Common noun), NNPC (Compound proper noun), NNP (Proper noun), PREP (Postpositions), QFNUM (Number quantifier) and Other (Other than the above).

The POS tagger is further modified with two POS tags (Nominal and Other) for incorporating the nominal POS information. Now, a binary valued feature 'nominalPOS' is defined as: If the current/previous/next word is 'Nominal' then the 'nominalPOS' feature of the corresponding word is set to 1; otherwise, it is set to 0. This 'nominalPOS' feature has been used additionally with the 7-tag POS feature. Sometimes, postpositions play an important role in NER as postpositions occur very frequently after a NE. A binary valued feature 'nominalPREP' is defined as: If the current word is nominal and the next word is PREP then the feature 'nomianlPREP' of the current word is set to 1, otherwise set to 0.

• Named Entity Information: The NE tag of the previous word is also considered as the feature. This is the only dynamic feature in the experiment.

• First word: If the current token is the first word of a sentence, then the feature 'FirstWord' is set to 1. Otherwise, it is set to 0.

• Digit features: Several binary digit features have been considered depending upon the presence and/or the number of digits in a token (e.g., ContainsDigit [token contains digits], FourDigit [token consists of four digits], TwoDigit [token consists of two digits]), combination of digits and punctuation symbols (e.g., ContainsDigitAndComma [token consists of digits and comma], ConatainsDigitAndPeriod [token consists of digits and periods]), combination of digits and symbols (e.g., ContainsDigitAndSlash [token consists of digit and slash],

[2] http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

ContainsDigitAndHyphen [token consists of digits and hyphen], ContainsDigitAndPercentage [token consists of digits and percentages]). These binary valued features are helpful in recognizing miscellaneous NEs such as time expressions, monetary expressions, date expressions, percentages, numerical numbers etc.

• Gazetteer Lists: Various gazetteer lists have been developed from the partially NE tagged Bengali news corpus (Ekbal and Bandyopadhyay, 2007d). These lists have been used as the binary valued features of the CRF. If the current token is in a particular list then the corresponding feature is set to 1 for the current and/or the surrounding word(s); otherwise, set to 0. The following is the list of gazetteers:

(i) Organization suffix word (94 entries): This list contains the words that are helpful in identifying organization names (e.g., *kong*, *limited* etc). The feature 'OrganizationSuffix' is set to 1 for the current and the previous words.

(ii) Person prefix word (245 entries): This is useful for detecting person names (e.g., *sriman*, *sree*, *srimati* etc.). The feature 'PersonPrefix' is set to 1 for the current and the next two words.

(iii) Middle name (1,491 entries): These words generally appear inside the person names (e.g., *chandra*, *nath* etc.). The feature 'MiddleName' is set to 1 for the current, previous and the next words.

(iv) Surname (5,288 entries): These words usually appear at the end of person names as their parts. The feature 'SurName' is set to 1 for the current word.

(v) Common location word (547 entries): This list contains the words that are part of location names and appear at the end (e.g., *sarani*, *road*, *lane* etc.). The feature 'CommonLocation' is set to 1 for the current word.

(vi) Action verb (221 entries): A set of action verbs like *balen*, *ballen*, *ballo*, *shunllo*, *haslo* etc. often determines the presence of person names. The feature 'ActionVerb' is set to 1 for the previous word.

(vii) Frequent word (31,000 entries): A list of most frequently occurring words in the Bengali news corpus has been prepared using a part of the corpus. The feature 'RareWord' is set to 1 for those words that are not in this list.

(viii) Function words (743 entries): A list of function words has been prepared manually. The feature 'NonFunctionWord' is set to 1 for those words that are not in this list.

(ix) Designation words (947 entries): A list of common designation words has been prepared. This helps to identify the position of the NEs, particularly person names (e.g., *neta*, *sangsad*, *kheloar* etc.). The feature 'DesignationWord' is set to 1 for the next word.

(x) Person name (72, 206 entries): This list contains the first name of person names. The feature 'PersonName' is set to 1 for the current word.

(xi) Location name (7,870 entries): This list contains the location names and the feature 'LocationName' is set to 1 for the current word.

(xii) Organization name (2,225 entries): This list contains the organization names and the feature 'OrganizationName' is set to 1 for the current word.

(xiii) Month name (24 entries): This contains the name of all the twelve different months of both English and Bengali calendars. The feature 'MonthName' is set to 1 for the current word.

(xiv) Weekdays (14 entries): It contains the name of seven weekdays in Bengali and English both. The feature 'WeekDay' is set to 1 for the current word.

4 Experimental Results

A partially NE tagged Bengali news corpus (Ekbal and Bandyopadhyay, 2007d) has been used to create the training set for the NER experiment. Out of 34 million wordforms, a set of 150K wordforms has been manually annotated with the 17 tags as shown in Table 1 with the help of *Sanchay Editor* [3], a text editor for Indian languages. Around 20K NE tagged corpus has been selected as the development set and the rest 130K wordforms has been used as the training set of the CRF based NER system.

We define the *baseline* model as the one where the NE tag probabilities depend only on the current word: $P(t_1, t_2, \ldots, t_n|w_1, w_2, \ldots, w_n) = \prod_{i=1,\ldots,n} P(t_i, w_i)$.

In this model, each word in the test data will be assigned the NE tag which occurred most frequently for that word in the training data. The unknown word is assigned the NE tag with the help of various gazetteers and NE suffix lists.

Ninety-five different experiments were conducted taking the different combinations from the set 'F' to

[3] Sourceforge.net/project/nlp-sanchay

Feature (word, tag)	FS (in %)
pw, cw, nw, FirstWord	71.31
pw2, pw, cw, nw, nw2, FirstWord	72.23
pw3, pw2, pw, cw, nw, nw2, nw3, FirstWord	71.12
pw2, pw, cw, nw, nw2, FirstWord, pt	74.91
pw2, pw, cw, nw, nw2, FirstWord, pt, $\|pre\| \leq 4$, $\|suf\| \leq 4$	77.61
pw2, pw, cw, nw, nw2, FirstWord, pt, $\|suf\| \leq 3$, $\|pre\| \leq 3$	79.70
pw2, pw, cw, nw, nw2, FirstWord, pt, $\|suf\| \leq 3$, $\|pre\| \leq 3$, Digit features	81.50
pw2, pw, cw, nw, nw2, FirstWord, pt, $\|suf\| \leq 3$, $\|pre\| \leq 3$, Digit features, pp, cp, np	83.60
pw2, pw, cw, nw, nw2, FirstWord, pt, $\|suf\| \leq 3$, $\|pre\| \leq 3$, Digit features, pp2, pp, cp, np, np2	82.20
pw2, pw, cw, nw, nw2, FirstWord, pt, $\|suf\| \leq 3$, $\|pre\| \leq 3$, Digit features, pp, cp	83.10
pw2, pw, cw, nw, nw2, FirstWord, pt, $\|suf\| \leq 3$, $\|pre\| \leq 3$, Digit features, cp, np	83.70
pw2, pw, cw, nw, nw2, FirstWord, pt, $\|suf\| \leq 3$,$\|pre\| \leq 3$, Digit features, pp, cp, np, nominalPOS, nominalPREP, Gazetteer lists	**89.30**

Table 2: Results on Development Set

identify the best suited set of features for the NER task. From our empirical analysis, we found that the following combination gives the best result with 744 iterations:

F=[$w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}$, $|\text{prefix}| \leq 3$, $|\text{sufix}| \leq 3$, NE information of the previous word, POS information of the window three, nominalPOS of the current word, nominalPREP, FirstWord, Digit features, Gazetteer lists].

The meanings of the notations, used in experimental results, are defined as below:
cw, pw, nw: Current, previous and next word; pwi, nwi: Previous and the next ith word from the current word; pre, suf: Prefix and suffix of the current word; pt: NE tag of the previous word; cp, pp, np: POS tag of the current, previous and the next word; ppi, npi: POS tag of the previous and the next ith word.

Evaluation results of the system for the development set in terms of overall F-Score (FS) are presented in Table 2. It is observed from Table 2 that word window $[-2, +2]$ gives the best result with 'FirstWord' feature only and the further increase of the window size reduces the overall F-Score value.

Results of Table 2 (3rd and 5th rows) show that the inclusion of NE information of the previous word increases the overall F-Score by 2.68%. It is also indicative from the evaluation results that the performance of the system can be improved by including the prefix and suffix features. Results (6th and 7th rows) also show the fact that prefix and suffix of length upto three of the current word is more effective. In another experiment, it has been also observed that the surrounding word suffixes and/or prefixes do not increase the F-Score value. The overall F-Score value is further improved by 1.8% (7th and 8th rows) with the inclusion of various digit features.

Results (8th and 9th rows) show that POS information of the words improves the overall F-score by 2.1%. In the above experiment, the POS tagger was developed with 26 POS tags. Experimental results (9th, 10th, 11th and 12th rows) suggest that the POS tags of the previous, current and the next words, i.e., POS information of the window $[-1, +1]$ is more effective than POS information of the window $[-2, +2]$, $[-1, 0]$ or $[0, +1]$. In another experiment, we also observed that the POS information of the current word alone is less effective than the window $[-1, +1]$. The modified POS tagger that is developed with 7 POS tags increases the overall F-Score to 85.2%, while other set of features are kept unchanged. So, it can be decided that smaller POS tagset is more effective than the larger POS tagset in NER. We have observed from two separate experiments that the overall F-Score values can further be improved by 0.4% and 0.2%, respectively, with the 'nominalPOS' and 'nominalPREP' features. Finally, an overall F-Score value of 89.3% is obtained by including the gazetteer lists.

The best set of features is identified by training the system with 130K wordforms and testing with the development set of 20K wordforms. Now, the development set is included as part of the training set and resultant training set is thus consists of 150K wordforms. The training set has 20,455 person names, 11,668 location names, 963 organization

names and 11,554 miscellaneous names. We have performed 10-fold cross validation test on this training set. The Recall, Precision and F-Score values for the 10 different experiments in the 10-fold cross validation test are presented in Table 3. The overall average Recall, Precision and F-Score values are 93.8%, 87.8% and 90.7%, respectively.

The other existing Bengali NER systems along with the *baseline* model are also trained and tested under the same experimental setup. The *baseline* model has demonstrated the overall F-Score value of 56.3%. The overall F-Score value of the CRF based NER system is 90.7%, which is an improvement of more than 6% over the HMM based system, best reported Bengali NER system (Ekbal et al., 2007c). The reason behind the rise in overall F-Score value might be its better capability than HMM to capture the morphologically rich and overlapping features of Bengali language. The system has been evaluated also for the four individual NE classes and it has shown the average F-Score values of 91.2%, 89.7%, 87.1% and 99.2%, respectively, for person, location, organization and miscellaneous names.

Test set no.	Recall	Precision	FS (%)
1	92.4	87.3	89.78
2	92.3	87.4	89.78
3	91.4	86.6	88.94
4	95.2	87.7	91.29
5	91.6	86.7	89.08
6	92.2	87.1	89.58
7	94.5	87.9	91.08
8	93.8	89.3	91.49
9	96.9	88.4	92.45
10	97.7	89.6	93.47
Average	93.8	87.8	90.7

Table 3: Results for the 10-fold Cross Validation Test

5 Conclusion

In this paper, we have developed a NER system using CRF with the help of a partially NE tagged Bengali news corpus, developed from the archive of a leading Bengali newspaper available in the web. Experimental results with the 10-fold cross validation test have shown reasonably good Recall, Precision and F-Score values. It has been shown that the contextual window [-2, +2], prefix and suffix of length upto three, first word of the sentence, POS information of the window [-1, +1], current word, NE information of the previous word, different digit features and the various gazetteer lists are the best-suited features for the Bengali NER.

Analyzing the performance using other methods like MaxEnt and Support Vector Machines (SVMs) will be other interesting experiments.

References

Daniel M. Bikel, Richard L. Schwartz, and Ralph M. Weischedel. 1999. An Algorithm that Learns What's in a Name. *Machine Learning*, 34(1-3):211–231.

A. Borthwick. 1999. *Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, New York University.

A. Ekbal and S. Bandyopadhyay. 2007a. Lexical Pattern Learning from Corpus Data for Named Entity Recognition. In *Proceedings of ICON*, pages 123–128, India.

A. Ekbal and S. Bandyopadhyay. 2007b. Pattern Based Bootstrapping Method for Named Entity Recognition. In *Proceedings of ICAPR*, pages 349–355, India.

A. Ekbal and S. Bandyopadhyay. 2007d. A Web-based Bengali News Corpus for Named Entity Recognition. *Language Resources and Evaluation Journal (accepted)*.

A. Ekbal, S.K. Naskar, and S. Bandyopadhyay. 2007c. Named Entity Recognition and Transliteration in Bengali. *Named Entities: Recognition, Classification and Use, Special Issue of Lingvisticae Investigationes Journal*, 30(1):95–114.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, pages 282–289.

Wei Li and Andrew McCallum. 2004. Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction. *ACM TALIP*, 2(3):290–294.

A. McCallum and W. Li. 2003. Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons. In *Proceedings of CoNLL*, pages 188–191, Canada.

Fei Sha and Fernando Pereira. 2003. Shallow Parsing with Conditional Random Fields. In *Proceedings of NAACL '03*, pages 134–141, Canada.