Bengali Named Entity Recognition using Support Vector Machine

Abstract

Named Entity Recognition aims to classify each word of a document into predefined target named entity classes and is nowadays considered to be fundamental for many Natural Language Processing (NLP) tasks such as information retrieval, machine translation, information extraction, question answering systems and others. This paper reports about the development of a Named Entity Recognition (NER) system for Bengali using Support Vector Machine (SVM). Though this state of the art machine learning method has been widely applied to NER in several well-studied languages, this is our first attempt to use this method to Indian languages (ILs) and particularly for Bengali. 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 a 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 SVM-based NER system. The training set consists of approximately 150k words and has been manually annotated with sixteen NE tags. Experimental results with the 10-fold cross validation test have demonstrated the overall average Recall, Precision and F-Score of 94.3%, 89.4% and 91.8%, respectively. It has been shown that this system outperforms other existing Bengali NER systems.

1 Introduction

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

NER has drawn more and more attention from the named entity (NE) tasks (Chinchor 95; Chinchor 98) in Message Understanding Conferences (MUCs) [MUC6; MUC7]. The problem of correct identification of NEs is specifically addressed and benchmarked by the developers of Information Extraction System, such as the GATE system (Cunningham, 2001). NER also finds application in question-answering systems (Maldovan et al., 2002) and machine translation (Babych and Hartley, 2003).

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 HMM (BBN's IdentiFinder in (Bikel, 1999)), Maximum Entropy (New York University's MENE in (Borthwick, 1999)), Decision Tree (New York University's system in (Sekine 1998) and Conditional Random Fields (CRFs) (Lafferty et al., 2001). Support Vector Machines (SVMs) based NER system was proposed by Yamada et al. (2001) for Japanese. His system is an extension of Kudo's chunking system (Kudo and Matsumoto, 2001) that gave the best performance at CoNLL-2000 shared tasks. The other SVM-based NER systems can be found in (Takeuchi and Collier, 2002) and (Asahara and Matsumoto, 2003).

Named Entity (NE) identification in Indian languages in general and in Bengali in particular is difficult and challenging. In English, the NE always appears with capitalized letter but there is no concept of capitalization in Bengali. There has been very little work in the area of NER in Indian languages. In Indian languages particularly in Bengali, the work in NER can be found in (Ekbal and Bandyopadhyay, 2007a; Ekbal and Bandyopadhyay, 2007b). These two systems are based on the pattern directed shallow parsing approach. An HMM-based NER in Bengali can be found in (Ek-

bal et al., 2007c). Other than Bengali, a CRF-based Hindi NER system can be found in (Li and McCallum, 2004).

2 Support Vector Machines

Suppose we have a set of training data for a two-class problem: $\{(x_1, y_1), ...(x_N, y_N)\}$, where $x_i \varepsilon R^D$ is a feature vector of the i-th sample in the training data and $y_i \varepsilon \{+1, -1\}$ is the class to which x_i belongs. The goal is to find a decision function that accurately predicts class y for an input vector x. A non-linear SVM classifier gives a decision function f(x)=sign(g(x)) for an input vector where,

$$g(x) = \sum_{i=1}^{m} w_i K(x, z_i) + b$$

Here, f(x)=+1 means x is a member of a certain class and f(x)=-1 means x is not a member. z_i s are called support vectors and are representatives of training examples, m is the number of support vectors. Therefore, the computational complexity of g(x) is proportional to m. Support vectors and other constants are determined by solving a certain quadratic programming problem. $K(x, z_i)$ is a *kernel* that implicitly maps vectors into a higher dimensional space. Typical kernels use dot products: $K(x, z_i) = k(x.z)$. A polynomial kernel of degree d is given by $K(x, z_i) = (1+x)^d$. We can use various kernels, and the design of an appropriate kernel for a particular application is an important research issue.

We have developed our system using SVM (Jochims, 1999) and (Valdimir, 1995), which performs classification by constructing an N-dimensional hyperplane that optimally separates data into two categories. Our general NER system includes two main phases: training and classification. Both the training and classification processes were carried out by YamCha[1] toolkit, an SVM based tool for detecting classes in documents and formulating the NER task as a sequential labeling problem. Here, the pairwise multi-class decision method and *second degree polynomial kernel function* were used. We have used TinySVM-0.07[2] classifier that seems to be the best optimized among publicly available SVM toolkits.

[1]http://chasen-org/~taku/software/yamcha/
[2]http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM

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 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 sixteen NE tags as described in Table 1.

3.1 Named Entity Tagset

An SVM 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 properly denote the boundaries of the NEs and to apply SVM 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 Feature Descriptions

Feature selection plays a crucial role in Support Vector Machine (SVM) framework. Experiments were carried out to find out most suitable features for NE tagging task. 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 lan-

guages. In addition, various gazetteer lists have been developed for use in the NER task.

NE tag	Meaning	Example
PER	Single-word person name	শচীন [*sachin*] / PER
LOC	Single-word location name	যাদবপুর [*jadavpur*]/LOC
ORG	Single-word organization name	ইনফোসিস [*infosys*] / ORG
MISC	Single-word miscellaneous name	১০০%[100%]/ MISC
B-PER I-PER E-PER	Beginning, Internal or the End of a multi-word person name	শচীন [*sachin*]/ B-PER রমেশ [*ramesh*] / I-PER তেন্ডুলকর [*tendulkar*] / E- PER
B-LOC I-LOC E-LOC	Beginning, Internal or the End of a multi-word location name	মহাত্মা [*mahatma*] / B-LOC গান্ধী [*gandhi*] / I-LOC রোড [*road*] / E-LOC
B-ORG I-ORG E-ORG	Beginning, Internal or the End of a multi-word organization name	ভাবা [*bhaba*] / B-ORG এটোমিক [*atomic*] / I-ORG রির্সাচ [*research*] / I-ORG সেন্টার [*center*] / E-ORG
B-MISC I-MISC E-MISC	Beginning, Internal or the End of a multi-word miscellaneous name	১০ই [*10 e*] / B-MISC মাঘ [*magh*] / I-MISC ১৪০২ [*1402*] / E-MISC
NNE	Words that are not named entities (“none-of–the-above” category)	নেতা [*neta*]/NNE, বিধানসভা [*bidhansabha*]/NNE

Table 1: Named Entity Tagset

We have considered different combination from the following set for inspecting the best feature set for NER task:

F={ $w_{i-m},...,w_{i-1},w_i,w_{i+1},...,w_{i+n}$, |prefix|≤n, |suffix|≤n, previous NE tags, 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 surrounding words 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., -বাব [*-babu*], -দা [*-da*] , -দি [*-di*] etc.) and location names (e.g., - ল্যান্ড [*-land*], -পুর [*-pur*], -লিয়া [*-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 the surrounding words might be treated as features.

•Part of Speech (POS) Information: The POS of the current and the surrounding words might 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 coarse-grained POS tagger, which can identify if a word is nominal or not. Then the nominal feature can be defined as a two-valued feature. The value is ‘+1’, if the current/previous/next word is nominal. Otherwise, feature value is ‘-1’. Sometimes, the postpositions are also important in NER as postpositions occur very frequently after a NE.

Here, we have used a CRF-based POS tagger, which was originally developed with the help of 26 different POS tags[3], 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 tags) for incorporating the nominal POS information. This binary ‘nominalPOS’ feature is separate from the 7-tag POS feature. The POS features of the surrounding words can be considered.

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

•Named Entity Information: The NE tags of the previous words are also considered as the features. This is the only dynamic feature in the experiment.

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

•Contains digit: This feature is helpful for identifying the company names, house numbers and the numerical numbers.

•Four digit token: This is helpful in identifying the year (e.g., ২০০৭ সাল [*2007*]) and the numerical numbers (e.g., ২০০৭ [*2007*]).

•Two digit token: This is helpful for identifying the time expressions (e.g., ১২ টা [*12 ta*], ১০ এম [*8 am*], ১০ পি এম [*9 pm*]) in general.

•Contains digits and comma: This feature is helpful in identifying monetary expressions (e.g., ১২০,৪৫,৩৩০ টাকা [*120,45,330 taka*]), date information (e.g., ১৫ই আগষ্ট, ২০০৭ (*15 e august, 2007*) and numerical numbers (e.g., ১২০,৪৫,৩৩০ [*120,45,330]*).

•Contains digits and slash: This helps in identifying date expressions (e.g., ১৫/৮/২০০৭ [15/8/2007]).

•Contains digits and hyphen: This is helpful for the identification of date expressions (e.g., ১৫-৮-২০০৭ [15-8-2007]).

•Contains digits and period: This helps to recognize numerical quantities (e.g., ১২০৪৫৩.৩৫ [*120453.35*]) and monetary amounts (e.g., ১২০৪৫৩.৩৫ টাকা [*120453.35 taka*]).

•Contains digits and percentage: This helps to recognize numerical quantities (e.g., ১২০ % [*120%*]).

•Gazetteer Lists: Various gazetteer lists have been created from the partially NE tagged Bengali news corpus (Ekbal and Bandyopadhyay, 2007d). These lists have been used as the two-valued features of the SVM. If the current token is in a particular list, then the corresponding feature is set to '+1'for the current/previous/next word; otherwise, '-1'. The following is the list of gazetteers:

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

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

(iii). Middle name list (37 entries): These words generally appear inside the person names (e.g., চন্দ্র [*chandra*], নাথ [*nath*] etc). The feature 'MiddleName' is set to 1 for the current, previous and the next words.

(iv). Surname list (1823 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 list (547 entries): This list contains the words that are part of location names and appear at the end (e.g., সরনী (*sarani*), রোড (*road*), লেন (*lane*) etc.). The feature 'CommonLocation' is set to 1 for the current word.

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

(vii). Frequent word list (31,000 entries): A list of most frequently occurring words in the tagged Bengali news corpus has been prepared automatically 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 'Non-FunctionWord' 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., নেতা [*neta*], সাংসদ [*sangsad*], খেলোয়াড় [*kheloar*] etc.). The feature 'DesignationWord' is set to 1 for the next word.

(x). Person name list (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 list (7,234 entries): This list contains the location names and the feature 'LocationName' is set to 1 for the current word.

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

(xiii). Month name list: 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 list: It contains the name of seven weekdays in Bengali and English both. The feature 'WeekDay' is set to 1 for the current word.

3.3 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, 150k worforms has been manually annotated with the 16 different NE tags with the help of *Sanchay Editor*[4], a text editor for Indian languages. The non-NEs are marked with the NNE tags. 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 SVM based NER system.

We define the *baseline* model as the one where the NE tag probabilities depend only on the current word:

$$P(t_1, t_2, t_3 ... t_n \mid w_1, w_2, w_3 ... w_n) = \prod_{i=1...n} P(t_i, w_i)$$

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.

Seventy-four different experiments were conducted taking the different combinations from the set 'F' to identify the best-suited set of features for the SVM based NER system. From our empirical analysis, we found that the following combination gives the best result for the development set.

F={ $w_{i-3}w_{i-2}w_{i-1}w_i w_{i+1}w_{i+2}$, |prefix|<=3, |suffix|<=3 NE information of the previous two words, 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 in Table 2. Evaluation results of the system for the development test in terms of F-Score (FS) are presented in Tables 3-7.

It is observed from Table 3 that word window (-3…+2) gives the best result with 'FirstWord' feature and the further increase of the window size reduces the accuracy. Results of Table 4 shows that the word window (-3…+2) gives the best result with the named entity information of the previous two words. It is experimentally observed from Table 5 that suffix and prefix of length 3 of the current word gives the best results with the 'FirstWord' feature within the window (-3…+2). It is also evident that the surrounding word suffixes and/or prefixes do not increase accuracy. The F-Score value is increased by 1.6% with the inclusion of the various digit features. Results of Table 6 shows that POS information with the word is helpful but only the POS information without the word decreases the accuracy significantly. In the above experiment, the POS tagger was developed with 26 POS tags. Experimental results 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 the window (-2…+2), (-1…0), (0…+1) or the current word alone. It can be observed from this result (5th and 6th row) that the POS information of the window (-1…0) increases the F-Score by 0.6% compared to the window (0…+1).

Notation	Meaning
cw, pw, nw	Current, previous and next word repectively
pwi, nwi	Previous and next ith word from the current word
pt	NE tag of the previous word
pti	NE tag of the previous ith word
pre, suf	Prefix and suffix of the current word
ppre, psuf	Prefix and suffix of the previous word
npre, nsuf	Prefix and suffix of the next word
cp, pp, np	POS tags of the current, previous and next word
ppi, npi	POS tag of the previous and the next ith word
cwnl	Current word is nominal

Table 2: Notations used in the experimental results

Feature (word, tag)	FS (in %)
pw, cw, nw, First word	71.23
pw2, pw, cw, nw, nw2, First Word	73.23
pw3, pw2, pw, cw, nw, nw2, First Word	**74.87**
pw3, pw2, pw, cw, nw, nw2, nw3, First Word	74.12
pw4, pw3, pw2, pw, cw, nw, nw2, First Word	74.01

Table3: Experimental results on the development set

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

Feature (word, tag)	FS (in %)
pw3, pw2, pw, cw, nw, nw2, First Word, pt	75.3
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2	**76.23**
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, pt3	75.48

Table 4: Experimental results on the development set

Feature (word, tag)	FS (in %)
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |pre| <=6, |suf|<=6	77.51
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |pre|<=5, |suf|<=5	78.17
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |pre|<=4, |suf|<=4	78.72
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3	**81.2**
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3 |psuf|<=3	80.4
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, |psuf|<=3, |nsuf|<=3, |ppre|<=3, |npre|<=3	78.14
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, |nsuf|<=3, |npre|<=3	79.9
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, |psuf|<=3, |ppre|<=3,	80.1
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, Digit	**82.8**

Table 5: Experimental results on the development set

So, it can be argued that the POS information of the previous word is more helpful than the POS of the next word. In another experiment, the POS tagger was modified with 7 POS tags (NN, NNC, NNP, NNPC, QFNUM, PREP, Other). This modified POS tagger increases the F-Score to **87.3%** with the feature (word, tag)=[pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, Digit, pp, cp, np]. So, it can be decided that the smaller POS tagset is more effective than the larger tagset in NER. The POS tagger is further modified to a coarse-grained tagger with 2 POS tags, Nominal and Not-nominal. Now, a two-valued feature 'nominalPOS' is defined as: If the current/previous/next word is 'Nominal' then the feature 'nominalPOS' is set to '+1'; otherwise, it is set to '–1'. It has been observed from the experimental results that inclusion of nominal feature along with POS information (with 7 tags) of the window (-1…+1) gives better results. It was also observed that 'nominalPOS' feature of the current word is only helpful and not of the neighboring words. The F-Score value of the NER system increases to **87.8%** with the inclusion of this **'**nominalPOS' feature.

Feature (word, tag)	FS (in %)
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, Digit, pp, cp, np	**85.6**
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, Digit, pp2, pp, cp, np, np2	83.4
pp2, pp, cp, np, np2, pt, pt2, |pre|<=3, |suf|<=3, FirstWord, Digit	40.2
pp, cp, np, pt, pt2, |pre|<=3, |suf|<=3, FirstWord, Digit	35.7
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, Digit pp, cp	84.7
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, Digit, cp, np	84.1
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, Digit cp	84.4

Table 6: Experimental results on the development set

Postpositions could be helpful in NER and in order to use it, a two-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 '–1'. The accuracy of the NER system increases to **88.1%** with the feature: feature (word, tag)=[pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, Digit pp, cp, np, cwnl, nominalPREP].

Experimental results with the various gazetteer lists are presented in Table 7 for the development set. Results show that NER system with the inclusion of gazetteer lists produces the highest F-Score of **90.7%**, which is an improvement of 2.6%.

The best set of features is identified by training the system with 130k wordforms and tested with the help of development set of 20k wordforms. Now, the development set is included as part of the training set and resultant training set is thus consisting of 150k wordforms. The training set has 20,455 person names, 11,668 location names, 963 organization names and 11,554 miscellaneous names. This training set is distributed into 10 subsets of equal size. In the cross validation test, one subset is withheld for testing while the remaining 9

subsets are used as the training sets. This process is repeated 10 times to yield an average result, which is called as the 10-fold cross validation test. The Recall, Precision and F-Score values for the 10 different experiments in the 10-fold cross validation test are presented in Table 8. The overall average Recall, Precision and F-Score are **94.3%, 89.4%** and **91.8%**, respectively.

Feature (word, tag)	FS (%)
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, Digit pp, cp, np, cwnl, nominal-PREP, DesignationWord, Non-FunctionWord	89.2
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, Digit pp, cp, np, cwnl, nominal-PREP, DesignationWord, Non-FunctionWord, RareWord	89.5
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, Digit pp, cp, np, cwnl, nominal-PREP, DesignationWord, Non-FunctionWord, RareWord, OrganizationSuffix, PersonPrefix	90.2
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, Digit pp, cp, np, cwnl, nominal-PREP, DesignationWord, Non-FunctionWord, RareWord, OrganizationSuffix, PersonPrefix MiddleName, CommonLocation	90.5
pw3, pw2, pw, cw, nw, nw2, First Word, pt, pt2, |suf|<=3, |pre|<=3, Digit pp, cp, np, cwnl, nominal-PREP, DesignationWord, Non-FunctionWord, RareWord, OrganizationSuffix, PersonPrefix MiddleName, CommonLocation, Other gazetteers	**90.7**

Table 7: Experimental results on the development set

The other existing Bengali NER systems along with the *baseline* model were also trained and tested with the same data set. Comparative evaluation results have been presented in Table 9. It presents the F-Score values for the four major NE classes: 'Person', 'Location', 'Organization' and 'Miscellaneous'. Table 9 shows that the SVM based NER model has reasonably high F-Score value. The overall F-Score of this model is **91.8%**, which is an improvement of more than 6% over the HMM, best reported Bengali NER system (Ekbal et al., 2007c). The reason behind the rise in accuracy might be its better capability to capture the morphologically rich and overlapping features of Bengali language.

Test set no.	Recall	Precision	FS (%)
1	92.5	87.5	89.93
2	92.3	87.6	89.89
3	94.3	88.7	91.41
4	95.4	87.8	91.40
5	92.8	87.4	90.02
6	92.4	88.3	90.30
7	94.8	91.9	93.33
8	93.8	90.6	92.17
9	96.9	91.8	94.28
10	97.8	92.4	95.02
Average	**94.3**	**89.4**	**91.8**

Table 8: Results for the 10-fold cross validation test

Model	F_P	F_L	F_O	F_M	F_T
Baseline	61.3	58.7	58.2	52.2	56.3
A	75.3	74.7	73.9	76.1	74.5
B	79.3	78.6	78.6	76.1	77.9
HMM	85.5	82.8	82.2	92.7	84.5
SVM	91.4	89.3	87.4	99.2	**91.8**

Table 9: Experimental results for the 10-fold cross validation tests (F_P: Avg. f-score of 'Person', F_L: Avg. f-score of 'Location', F_O: Avg. f-score of 'Organization', F_M: Avg. f-score of 'Miscellaneous' and F_T: Overall avg. f-score of all classes)

Accuracy increases with the increment of training data. This fact is represented in Figure 1. Also, it is evident from Figure 1 that the F-Score/ F-measure value of 'Miscellaneous' name is nearly close to 100% followed by 'Person', 'Location' and 'Organization' NE classes with the training data of 150k words.

4 Conclusion

We have developed a named entity recognition system using Support Vector Machine with the help of a Bengali news corpus, developed from the archive of a leading Bengali newspaper available in the web. It has been shown that the contextual window of size six, prefix and suffix of length up to three of the current word, POS information of the window of size three, first word, NE information of the previous two words, different digit fea-

tures and the various gazetteer lists are the best-suited features for the Bengali NER. Experimental results with the 10-fold cross validation test have shown reasonably good F-Score. The performance of this system has been compared with the previous existing three Bengali NER systems and it has been shown that the SVM-based system outperforms other systems.

Analyzing the performance using other methods like MaxEnt and Conditional Random Fields (CRFs) will be other interesting experiments.

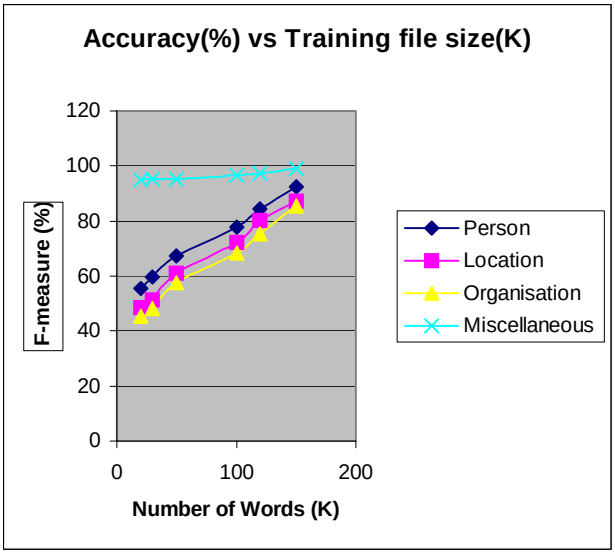

Fig. 1: Training file size VS Accuracy measure

References

Babych, Bogdan, A. Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of EAMT/EACL 2003 Workshop on MT and other language technology tools*, 1-8, Hungary.

Bikel, Daniel M., R. Schwartz, Ralph M. Weischedel. 1999. An Algorithm that Learns What's in Name, *Machine Learning (Special Issue on NLP)*, 1-20.

Bothwick, Andrew. 1999. A Maximum Entropy Approach to Named Entity Recognition, *Ph.D. Thesis*, New York University.

Chinchor, Nancy. 1995. MUC-6 Named Entity Task Definition (Version 2.1), *MUC-6*, Columbia, Maryland.

Chinchor, Nancy. 1998. MUC-7 Named Entity Task Definition (Version 3.5), *MUC-7*, Fairfax, Virginia.

Cunningham, H. 2001. GATE: A general architecture for text engineering, *Comput. Humanit.* (36), 223-254.

Ekbal, Asif, and S. Bandyopadhyay. 2007a. Pattern Based Bootstrapping Method for Named Entity Recognition, In *Proceedings of ICAPR 2007*, Kolkata, India, 349-355.

Ekbal, Asif, and S. Bandyopadhyay. 2007b. Lexical Pattern Learning from Corpus Data for Named Entity Recognition, In *Proceedings of ICON 2007*, Hyderabad, India, 123-128.

Ekbal, Asif, Naskar, Sudip 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 (2007), 95-114.

Ekbal, Asif, and S. Bandyopadhyay. 2007d. A Web-based Tagged Bengali News Corpus, *Language Resources and Evaluation Journal* (To appear December).

Hiroyasu Yamada, Taku Kudo and Yuji Matsumoto. 2002. Japanese Named Entity Extraction using Support Vector Machine, In *Transactions of IPSJ*, Vol. 43 No. 1, 44-53.

Koichi Takeuchi and Nigel Collier. 2002. Use of Support Vector Machines in Extended Named Entity Recognition, In *Proceedings of the 6th Conference on Natural Language Learning*, 119-125.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In *Proc. of 18th International Conference on Machine learning.*

Li, Wei and Andrew McCallum. 2003. Rapid Development of Hindi Named Entity Recognition Using Conditional Random Fields and Feature Inductions, *ACM TALIP*, 2(3), (2003), 290-294.

Masayuki, Asahara and Yuji Matsumoto. 2003. Japanese Named Entity Extraction with Redundant Morphological Analysis, In *Proceedings of HLT-NAACL.*

Moldovan, Dan I., Sanda M. Harabagiu, Roxana Girju, P. Morarescu, V. F. Lacatusu, A. Novischi, A. Badulescu, O. Bolohan. 2002. LCC Tools for Question Answering, In *Proceedings of the TREC*, 1-10, Maryland.

Sekine, Satoshi. 1998. Description of the Japanese NE System Used for MET-2, *MUC-7*, Fairfax, Virginia.

T. Joachims. 1999. Making Large Scale SVM Learning Practical, In *B. Scholkopf, C. Burges and A. Smola editions, Advances in Kernel Methods-Support Vector Learning, MIT Press.*

Taku Kudo and Yuji Matsumoto. 2001. Chunking with Support Vector Machines, In *Proceedings of NAACL, 192*-199.

Taku Kudo and Yuji Matsumoto. 2000. Use of Support Vector Learning for Chunk Identification, In *Proceedings of CoNLL-2000.*

Valdimir N. Vapnik. 1995. The Nature of Statistical Learning Theory, *Springer*.