Named-Entity Recognition in Bengali @FIRE NER 2013

Apurbalal Senapati **Arjun Das** **Utpal Garain**

CVPR Unit, Indian Statistical Institute
203, B.T. Road, Kolkata – 700108
{apurbalal.senapati|arjundas.cs}@gmail.com,
utpal@isical.ac.in

Abstract

This paper describes performance of two systems for Named Entity Recognition (NER) task of FIRE 2013. The first system is a rule-based one whereas the second one is statistical (based on CRF) in nature. The systems vary in some other aspects too, for example, the first system works on untagged data (not even POS tag is done) to identify NER whereas the second system makes use of a POS tagger and a chunker. The rules used by the first system are mined from the training data. The CRF-based classification does not require any explicit linguistic rules but it uses a gazetteer built from Wiki and other sources.

1 Introduction

The Name Entity Recognizer (NER) plays a significant role in many NLP (natural language Processing) applications namely in machine translation, question-answering, information retrieval, anaphora resolution etc. Unlike in English development of an NER system is difficult for most of the Indic languages due to some inherent problems as explained in the next section. Hence, it is still an active area of research. Though there are a few research works on developing Bengali NER system but more research is needed to develop a better insight of the problem and thereby making a practical system.

To do NER a minimum level of preprocessing required (i.e. at least parts of speech, chunking) and these tools are not easily available for many Indic languages[1]. This has motivated us to develop one of our systems based on a set of very simple rules that does not require any pre-processing tools including POS tagger, chunker, etc. The second system however uses these tools but does not require any linguistic knowledge. Therefore, the requirements of these two systems are complementary in nature.

2 Problem in NER in Bengali

There are so many issues in NER in Bengali and has been identified earlier [1, 2]. Based on their important with respect to NER some of these are stated below.

Capitalization: Unlike English and most of the European languages, Bengali lacks capitalization information, which plays a very important role in identifying name entities.

Multiple meaning of person: Indian person names are more diverse compared to the other languages and a lot of these words can be found in the dictionary with some other specific meanings. Example: *kabitA* may be a name of person or is the meaning of poem in Bengali.

Agglutinative nature: Bengali is a highly inflectional language providing one of the richest and most challenging sets of linguistic and statistical features resulting in long and complex word forms. Example: কটক/*katak* is the proper name i.e. the name of location while কটকি/*katak**i*** appears as adjective in কটকি জুতো/*katak**i** juto (katki shoes)* and hence implies a special kind (made of Katak) of shows.

Sentence structure: Bengali is a relatively free order language within the sentence. Thus, name entities can appear in any position of the sentence.

Spelling variations: In Bengali proper name, especially in case of person variation of spelling is permitted.

Resource scare: Bengali is resource scare language. The basic NLP tools (annotated corpora, name dictionaries, good morphological analyzers, POS taggers, chunker etc) are not available in satisfactory accuracy.

Web source: Web sources for name lists are available in English, but such lists are not available in Bengali forcing the use of transliteration for creating such lists.

[1] A Bengali POS tagger is available at: http://www.isical.ac.in/~utpal/resources.php

3 Existing Approaches and Bengali NER

There are many approaches reported in NER. Broadly classify three main categories, namely rule-based NER, machine learning-based NER and hybrid NER. Ekbal et al [1, 3, 4] has focused the issues in Bengali and adopted the machine learning as well as hybrid system.

In Indian language, first workshop in NER (Workshop on NER for South and South East Asian Languages) was conducted in five Indic languages (Hindi, Bengali, Oriya, Telugu and Urdu) by IJCNLP 2008[5]. Where, Saha et al. [6], who were able to achieve the best results in the shared task, describe a hybrid system that applies maximum entropy models, language specific rules, and gazetteers. Gali et al [7]. also combined machine learning with language specific heuristics. Ekbal et al [8]. also used an approach based on CRFs. They also used some language specific features for Hindi and Bengali. Chaudhuri and Bhattacharya [2] also experimented on a news corpus for Bengali using a three stage NER system. The three stages were based on an NE dictionary, rules and contextual co-occurrence statistics. They only tried to identify the NEs, not classify them.

4 System-I: A Rule Based Approach

In Bengali, most of the named entities (e.g. person, date, time, year, quantity, etc.) having some context either before or after the word. For example, in case of person, most of the cases it appears with some honorific addressing terns (*Dr.*, *Mr.*, *Mrs.*, *Sri.*, ….) or followed by middle name (*Chandra, Boron, Lal, Nath,* ….) or surname (*Ghosh, Bosu, Roy, Mukherjee,* ….). Similarly in case of distance it must have some numeric value (in numeric or written form in Bengali script) followed by with some distance measuring term (*mile, inch, feet, km, kilometer,* ….) before (*mile tinek*) or after (*tin mile*) the numeric value. Such types of context available almost all the named entity type in Bengali. Based on our observation such context related word is limited (except person and location) for most of the named entity type. This is the main clue for the system. We have defined a context dictionary containing all such context words category wise for Bengali. The details of the dictionary and the resolution approach are given below.

The context dictionaries

As stated earlier, the dictionaries contain the context words of each name entity categories. The system use this dictionary to identify the name entity type for a given context word. The dictionaries have been building from the training data for NER provided by FIRE 2013 and ICON 2013. The detail descriptions of the (entries of the dictionary) contexts are given below.

Honorific context: In Bengali a set of honorific addressing terms used to define the honorificity of a person. These terms is either before (*Dr.*, *Mr.*, *Mrs.*, *Sri.*, ….) or after the name (*Babu*, *Mosai*, *Saheb*, ….). In our system we have used 30 such terms found from training data.

Surname context: Surname always comes after the name of a person and hence collection of common surnames (*Ghosh, Bosu, Roy, Mukherjee,* ….) helps to resolve the name entity.

Middle name context: Like surname a predefined common middle name (*Chandra, Boron, Lal, Nath,* ….) also helps to resolve the name entity.

Relational term context: Specially in Bengali community some relational terms (দা/*da,*দাদা /*dada,* দি/*di,* দিদি/*didi,* কাকা/*kaka,* কাকু/*kaku,* ….) used together with name or next to the name entity. Example: বাদল-দা/*badal-da* …, মিঠুদি/*mithudi* মিঠু দি/*mithu di* …., etc.

Abbreviation form of person: The dot (.) separated abbreviation forms (আর.এন. টেগোর/R.N. Tagore) with an honorific context earlier or followed by a surname also represent the person.

Suffix context for location: Some suffixes *(-gram, -pur, -nagar, -gung,* ….) are used to identify the location specially location of some place. Some suffixes (or next word) *(-dip, -dippunja, -saikat, -bich,* ….) are used to identify the location with sub category landscape. Some suffixes *(-mandir, -mall ,-market,* ….) (or next word) followed by a location are used to identify the location with sub category manmade.

Distance context: Numerical value followed by distance measuring unit (*mile, inch, feet, km, kilometer,* ….) or the distance measuring unit followed by numerical value is representing distance.

Quantity context: Numerical value followed by quantity measuring unit (*kilo, keji, liter, cc, degree, %,* ….) or the quantity measuring unit followed by numerical value is representing quantity.

Month context: All the name of month used in Bengali (জানুয়ারী/*January,* জানু /*Janu,* ফেব্রুয়ারী/*February,* ফেব্রু/*Febru,* …. বৈশাখ/*boisakh,* জৈষ্ঠ/*jasta,* ...)

and month related terms (*-mas, mase, ...*) used to identify the month name entity.

Year context: Numerical value followed by year related term (*sal, satabdi, khristabdo, krishta purba,* ….) or year related term followed by numerical value is representing year.

Period context: Numerical value followed by period related term (*bochhor, mas, din, jug aamal, saptaha,* ….) or period related term followed by numerical value is representing period.

Time context: Numerical value followed by time measuring unit (*second, ghanta, din, sokal, bikal,* ….) or the time measuring unit followed by numerical value is representing time related name entity.

Date context: Numerical value followed by month name (জানুয়ারী/*January,* জানু /*Janu,* ফেব্রুয়ারী/*February,* ফেব্রু/*Febru,* .... বৈশাখ/*boisakh,* জৈষ্ঠ/*jasta,* ...) or date related terms (সোমবার/*somber,* মঙ্গলবার/*mangalbar,* ....) representing the date name entity.

Money context: Numerical value followed by money measuring unit (*taka, hazar taka, dollar, lakh,* ….) or the money measuring unit followed by numerical value is representing quantity.

Note that the numeric value has been consider for Bengali digit (০,১,২,৩,৪,৫, ….), English digit (0,1,2,3,4,5, …) and numeric values written in Bengali scripts (*ek, dui, tin,* ....) and their variation (*tinek, charek, panchek,* ....২ *din,* ...).

Count context: Numerical value followed by count related term (*ti, ta, jon, jone, jora*….) or count related term followed by numerical value is representing period. Example: ৩১টি/*31ti*, তিনজোড়া/*tinjora* etc.

Entertainment context: The entertainment *(-festival, -mela, -paragliding, -safari, ...*) contest used to identify the entertainment related name entity.

Organization contest: The organization *(-corporation, -sanstha, -limited, -daptar, ...*) contest used to identify the organization relative name entity. The dot (.) separated abbreviation forms (C.M.C) also represent the organization.

Facilities contest: The facilities *(-university, -college, -resort ...)* contest used to identify the facilities relative name entity.

Skip word: Reduplicate words (repetition of same word, *ram-ram*, *hari-hari*), words with classifier (*-ti, -guli, -khani, -take,* ....), eco words (*gachh-tachh, khabar-dabar,* ....) etc. are not named entities. We maintain a list of redupl cated and eco words obtained from the training data. This feature has been used in the Chaudhuri, and Bhattacharya [2] system. In our system we have also included the all pronouns and conjunction as skip words.

Resolution approach

In the above discussion, it is noticed that for the entire name entities broadly can be categorized two types, i.e. with numeric value and without numeric value. The name entities containing numerical values should be one of the among distance, quantity, month, year, period, time, date, money, count types and the name entities without numerical values should be one of the among person, location, entertainment, organization, facilities etc.

The algorithmic structure of the resolution strategy is follows:

Step-1: For an input word find its context.

Step-2: If word is skip word then skip and go to Step-1 else go to Step-3.

Step-3: If words having numeric value then go to Step-4 else go to Step-5.

Step-4: With the context (±1-word & ±2-word context) search the context dictionaries (corresponding to distance, quantity, month, …..).

If the larger context (±2-word) matches (in that case ±1-word context may be match or unmatched) then resolve with respective name entities for larger context and go to Step-1.

If context ±1-word matches but ±2-word context does a not match then resolve with respective name entities for smaller context and go to Step-1.

If not match then skip and go to Step-1.

Step-5: If the word with context form the dot (.) separated abbreviation name take its context and search the context dictionaries corresponding to organization else search the context dictionaries (corresponding to person, location, entertainment, …..).

If the larger context (±2-word) matches (in that case ±1-word context may be match or unmatched) then resolve with respective name entities for larger context and go to Step-1.

If context ±1-word matches but ±2-word context does a not match then resolve with respective name entities for smaller context and go to Step-1.

If not match then skip and go to Step-1.

The elastration of the algorithm is with example given below.

Example-1: ... দীর্ঘ ৪৫০ বছরের সাতবাহন রাজতেব ...

In this example while identified word (৪৫০) as the numeric value then taking its context ±1 word

context i.e. (দীর্ঘ , বছরের) and ±2 word context i.e. (… দীর্ঘ , বছরের, সাতবাহন) search in dictionary and it match with period for ±1 word context (but does not match ±2 word context) and hence it is resolve as name entity period.

Example-2: ... সাড়ে পাঁচ হাজার বছর পূর্বে ....

In this example while identified word (পাঁচ) as the numeric value then taking its context ±1 word context i.e. (সাড়ে, হাজার) and search in dictionary and it match with count context. The ±2 word context i.e. (… , সাড়ে, হাজার,বছর) also match with period context. Finally choose the second case i.e. highest match and tag সাড়ে, পাঁচ হাজার বছর as period.

Example-3: ... এই আইল কে টেংগু সাহেবের আইল বলে উল্লেখ করতো ...

In this example while identified word টেংগু with its context ±1 word i.e. (কে,সাহেবের) and ±2 word context i.e. (আইল, কে,সাহেবের, আইল) and it only match ±1 word context in dictionary with the honorific context after the name and hence টেংগু tag with the name entity person.

5 System-II: A CRF-based Approach

Data Preprocessing

As preprocessing steps we have applied a tokenizer, part-of-speech tagger and a chunker for both training and testing dataset. We have trained Stanford POS tagger [9] and YamCha, a support vector machine based chunker [10] for that purpose.

Method

We have applied Conditional Random Field (CRF's) [11] for the NER task. In our experiment, we have used the CRF++[2], an open-source package.

Since the given NER task involves predictions at three different levels, we have applied three different CRF-based classifiers for predicting the three levels of a named-entity. For each level the output of the previous level is added as a feature for the next level of classification.

We have used different combination of language independent features and finally selected those features for which we got the highest accuracy on a development set which is created by randomly selecting 400 sentences from ICON 2013[3] training data. Following are the details of the features used in classification:

- **Context Words:** The previous and next word of a particular word
- **Word Prefix and Suffix:** We have considered word prefix and suffix from 3 character length to 5 character length for all NNP's.
- **POS and Chunk Information:** We have used POS and Chunk information of a particular word as a feature.
- **First and Last Words**: We have used the first word of a sentence and the last word i.e. (n-1)th token of the sentence as a feature.
- **Digit:** This is a binary valued feature which is been defined depending upon the presence or absence of a digit in a token.
- **Token ID:** This is a real valued feature which represents the current token id. This feature provides useful information about the position of the NE tag.
- **Associated Verb:** This feature provides the information about the nearest verb for all token. This feature is useful for differentiation between PERSON and LOACATION tag.
- **Gazetteer:** We have mined person names from Wikipedia [12] and the training data provided by ICON 2013[3].

6 Conclusion

FIRE 2013 evaluation will bring out the potential of the two systems described here. The context dictionaries (i.e. the rule base) play significant role for determining the performance of the first system. However, it is indeed an experience to work with a system which does not require any pre-processing modules not even POS tagging, chunking, etc. which often form the basic requirements for any NER system. The gazetteer used in the second system also play vital role. Given the present task, the second system will show inferior performance for decoding the second and third level tags of an NE as the gazetteer used was primarily developed to decode the first level. On the other hand, decoding the finer levels of an NE is slightly better with rule-based approach as creating rules like presence of the words উপকুল, সাগর, লেক, etc. refer to water body or presence of the words উদয়্যান, শিখর, উপতয়্কা, etc. refer to landform, could easily identify the second and third level tags for many NEs. However, more in-depth analysis of these systems is yet to be done.

[2]CRF++ is available at : http://crfpp.googlecode.com/svn/trunk/doc/index.html?source=navbar

[3] ICON 2013 dataset is available at: http://ltrc.iiit.ac.in/icon/2013/nlptools/

References

[1] A. Ekbal and S. Bandyopadhyay (2010). Named Entity Recognition using Appropriate Unlabeled Data, Post-processing and Voting. In Informatica, Volume (34), Number (1), PP.55-76.

[2] B.B. Chaudhuri, and S. Bhattacharya (2008). An Experiment on Automatic Detection of Named Entities in Bangla. In Proceedings of the Workshop on NER for SASEAL in IJCNLP-08, India, PP: 85-91.

[3] A. Ekbal and S. Bandyopadhyay (2010). A Multi-engine NER System with Context Pattern Learning and Post-processing Improves System Performance. International Journal of Computer Processing of Languages (IJCPOL), Vol. 22(2 and 3), World Scientific Press, Singapore, Volume (22:2-3), PP.171-204.

[4] A. Ekbal and S. Bandyopadhyay (2009). A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi. Linguistic Issues in Language Technology (LiLT), Volume (2:1), November 2009, PP.1-44, CSLI Publication, Stanford University

[5] Proceedings of the Workshop on Named Entity Recognition for South and In Proceedings of the Workshop on NER for SASEAL in IJCNLP-08, India, India

[6] Saha et al (2008). A Hybrid Named Entity Recognition System for South and South East Asian Languages. In Proceedings of the Workshop on Named Entity Recognition for In Proceedings of the Workshop on NER for SASEAL in IJCNLP-08, India, India, PP: 27-34.

[7] Gali et al (2008). Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition. In Proceedings of the Workshop on NER for SASEAL in IJCNLP-08, India, PP: 35-41.

[8] Ekbal et al (2008). Language Independent Named Entity Recognition in Indian Languages. In Proceedings of the Workshop on NER for SASEAL in IJCNLP-08, India, PP: 43-50.

[9] Kristina Toutanova and Christopher D. Manning (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

[10] Taku Kudu and Yuji Matsumoto (2000). Use of Support Vector Learning for Chunk Identification, CoNLL-2000.

[11] John D. Lafferty, Andrew McCallum and Fernando C. N. Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the International Conference on Machine Learning (ICML'01). 282–289.

[12] Utpal Garain, Arjun Das, David S. Doermann and Douglas W. Oard (2012). Leveraging Statistical Transliteration for Dictionary-Based English-Bengali CLIR of OCR'd Text. In Proceedings of the Proceedings of COLING (Poster). PP:339-348.