Development of Bengali Named Entity Tagged Corpus and its Use in NER Systems

Asif Ekbal
Department of Computer Science and
Engineering, Jadavpur University
Kolkata-700032, India
asif.ekbal@gmail.com

Sivaji Bandyopadhyay
Department of Computer Science and
Engineering, Jadavpur University
Kolkata-700032, India
sivaji_cse_ju@yahoo.com

Abstract

The rapid development of language tools using machine learning techniques for less computerized languages requires appropriately tagged corpus. A Bengali news corpus has been developed from the web archive of a widely read Bengali newspaper. A web crawler retrieves the web pages in Hyper Text Markup Language (HTML) format from the news archive. At present, the corpus contains approximately 34 million wordforms. The *date, location, reporter* and *agency tags* present in the web pages have been automatically named entity (NE) tagged. A portion of this partially NE tagged corpus has been manually annotated with the sixteen NE tags with the help of *Sanchay Editor*[1], a text editor for Indian languages. This NE tagged corpus contains 150K wordforms. Additionally, 30K wordforms have been manually annotated with the twelve NE tags as part of the IJCNLP-08 NER Shared Task for South and South East Asian Languages[2]. A table driven semi-automatic NE tag conversion routine has been developed in order to convert the sixteen-NE tagged corpus to the twelve-NE tagged corpus. The 150K NE tagged corpus has been used to develop Named Entity Recognition (NER) system in Bengali using pattern directed shallow parsing approach, Hidden Markov Model (HMM), Maximum Entropy (ME) Model, Conditional Random Field (CRF) and Support Vector Machine (SVM). Experimental results of the 10-fold cross validation test have demonstrated that the SVM based NER system performs the best with an overall F-Score of 91.8%.

1 Introduction

The mode of language technology work has been changed dramatically since the last few years with the web being used as a data source in a wide range of research activities. The web is anarchic, and its use is not in the familiar territory of computational linguistics. The web walked into the ACL meetings started in 1999. The use of the web as a corpus for teaching and research on language technology has been proposed a number of times (Rundel, 2000; Fletcher, 2001; Robb, 2003; Fletcher, 2003). There is a long history of creating a standard for western language resources. The human language technology (HLT) society in Europe has been particularly zealous for the standardization, making a series of attempts such as EAGLES[3], PROLE/SIMPLE (Lenci et al., 2000), ISLE/MILE (Calzolari et al., 2003; Bertagna et al., 2004) and more recently multilingual lexical database generation from parallel texts in 20 European languages (Giguet and Luquet, 2006). On the other hand, in spite of having great linguistic and cultural diversities, Asian language resources have received much less attention than their western counterparts. A new project (Takenobou et al., 2006) has been started to create a common standard for Asian language resources. They have extended an existing description framework, the

[1] sourceforge.net/project/nlp-sanchay

[2] http://ltrc.iiit.ac.in/ner-ssea-08

[3] http://www.ilc.cnr.it/Eagles96/home.html

MILE (Bertagna et al., 2004), to describe several lexical entries of Japanese, Chinese and Thai. India is a multilingual country with the enormous cultural diversities. (Bharati et al., 2001) reports on efforts to create lexical resources such as transfer lexicon and grammar from English to several Indian languages and dependency tree bank of annotated corpora for several Indian languages. Corpus development work from web can be found in (Ekbal and Bandyopadhyay, 2007d) for Bengali.

Named Entity Recognition (NER) is one of the core components in most Information Extraction (IE) and Text Mining systems. During the last few years, the probabilistic machine learning methods have become state of the art for NER (Bikel et al., 1999; Chieu and Ng, 2002) and for field extraction (McCallum et al., 2000). Most prominently, Hidden Markov Models (HMMs) have been used for the information extraction task (Bikel et al., 1999). Beside HMM, there are other systems based on Support Vector Machine (Sun et al., 2003) and Naïve Bayes (De Sitter and Daelemans, 2003). Maximum Entropy (ME) conditional models like ME Markov models (McCallum et al., 2000) and Conditional Random Fields (CRFs) (Lafferty et al., 2001) were reported to outperform the generative HMM models on several IE tasks.

The existing works in the area of NER are mostly in non-Indian languages. There has been a very little work in the area of NER in Indian languages (ILs). In ILs, particularly in Bengali, the work in NER can be found in (Ekbal and Bandyopadhyay, 2007a; Ekbal and Bandyopadhyay, 2007b; Ekbal et al., 2007c). Other than Bengali, the work on NER can be found in (Li and McCallum, 2003) for Hindi.

Newspaper is a huge source of readily available documents. In the present work, the corpus has been developed from the web archive of a very well known and widely read Bengali newspaper. Bengali is the seventh popular language in the world, second in India and the national language of Bangladesh. A code conversion routine has been written that converts the proprietary codes used in the newspaper into the standard Indian Script Code for Information Interchange (ISCII) form, which can be processed for various tasks. A separate code conversion routine has been developed for converting ISCII codes to UTF-8 codes. A portion of this corpus has been manually annotated with the sixteen NE tags as described in Table 3. Another portion of the corpus has been manually annotated with the twelve NE tags as part of the IJCNLP-08 NER Shared Task for South and South East Asian Languages. A table driven semi-automatic NE tag conversion routine has been developed in order to convert this corpus to a form tagged with the twelve NE tags. The NE tagged corpus has been used to develop Named Entity Recognition (NER) system in Bengali using pattern directed shallow parsing approach, HMM, ME, CRF and SVM frameworks.

A number of detailed experiments have been conducted to identify the best set of features for NER in Bengali. The ME, CRF and SVM based NER models make use of the language independent as well as language dependent features. The language independent features could be applicable for NER in other Indian languages also. The system has demonstrated the highest F-Score value of 91.8% with the SVM framework. One possible reason behind its best performance may be the flexibility of the SVM framework to handle the morphologically rich Indian languages.

2 Development of the Named Entity Tagged Bengali News Corpus

2.1 Language Resource Acquisition

A web crawler has been developed that retrieves the web pages in Hyper Text Markup Language (HTML) format from the news archive of a leading Bengali newspaper within a range of dates provided as input. The crawler generates the Universal Resource Locator (URL) address for the index (first) page of any particular date. The index page contains actual news page links and links to some other pages (e.g., Advertisement, TV schedule, Tender, Comics and Weather etc.) that do not contribute to the corpus generation. The HTML files that contain news documents are identified and the rest of the HTML files are not considered further.

2.2 Language Resource Creation

The HTML files that contain news documents are identified by the web crawler and require cleaning to extract the Bengali text to be stored in the corpus along with relevant details. The HTML file is scanned from the beginning to look for tags like <fontFACE=BENGALI_FONT_NAME>...<font>, where the BENGALI_FONT_NAME is the name of one of the Bengali font faces as defined in the news archive. The Bengali text enclosed within font tags are retrieved and stored in the database after appropriate tagging. Pictures, captions and tables may exist anywhere within the actual news. Tables are integral part of the news item. The pictures, its captions and other HTML tags that are not relevant to our text processing tasks are discarded during the file cleaning. The Bengali news corpus has been developed in both ISCII and UTF-8 codes. The tagged news corpus contains 108,305 number of news documents with about five (5) years (2001-2005) of news data collection. Some statistics about the tagged news corpus are presented in Table 1.

Total number of news documents in the corpus	108, 305
Total number of sentences in the corpus	2, 822, 737
Avgerage number of sentences in a document	27
Total number of wordforms in the corpus	33, 836, 736
Avgerage number of wordforms in a document	313
Total number of distinct wordforms in the corpus	467, 858

Table 1. Corpus Statistics

2.3 Language Resource Annotation

The Bengali news corpus collected from the web is annotated using a tagset that includes the type and subtype of the news, title, date, reporter or agency name, news location and the body of the news. A part of this corpus is then tagged with a tagset, consisting of sixteen NE tags and one non-NE tag. Also, a part of the corpus has been tagged with a tagest of twelve NE tags[4], defined for the IJCNLP-08 NER Shared Task for South and South East Asian Languages.

A news corpus, whether in Bengali or in any other language, has different parts like title, date, reporter, location, body etc. To identify these parts in a news corpus the tagset, described in Table 2, has been defined. The reporter, agency, location, date, bd, day and ed tags help to recognize the person name, organization name, location name and the various date expressions that appear in the fixed places of the newspaper. These tags are not able to recognize the various NEs that appear within the actual news body.

In order to identify NEs within the actual news body, we have defined a tagset consisting of seventeen tags. We have considered the major four NE classes, namely 'Person name', 'Location name', 'Organization name' and 'Miscellaneous name'. Miscellaneous names include the date, time, number, percentage and monetary expressions. The four major NE classes are further divided in order to properly denote each component of the multiword NEs. The NE tagset is shown in Table 3 with the appropriate examples.

We have also tagged a portion of the corpus as part of the IJCNLP-08 NER Shared Task for South and South East Asian Languages. This tagset has twelve different tags. The underlying reason for adopting these tags was the necessity of a slightly finer tagset for various natural language processing (NLP) applications and particularly for machine translation. The IJCNLP-08 NER shared task tagset is shown in Table 4.

One important aspect of IJCNLP-08 NER shared task was to annotate only the maximal NEs and not the structures of the entities. For example, *mahatma gandhi road* is annotated as location and assigned the tag 'NEL' even if *mahatma* and *gandhi* are NE title person and person name, respectively, according to the IJCNLP-08 shared task tagset. These internal structures of the entities need to be identified during testing. So, *mahatma gandhi road* will be tagged as *mahatma*/NETP *gandhi/*NEP *road/*NEL. The structure of the tagged element using the *Shakti Standard Format* (SSF)[5] will be as follows:

```
1        ((      NP      <ne=NEL>
1.1      ((      NP      <ne=NEP>
1.1.1    ((      NP      <ne=NETP>
1.1.1.1  mahatma
         ))
1.1.2    gandhi
         ))
1.2      road
         ))
```

[4] http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=3

[5] http://shiva.iiit.ac.in/SPSAL 2007/ssf.html

2.4 Partially Tagged News Corpus Development

A news document is stored in the corpus in XML format using the tagset, mentioned in Table 2. In the HTML news file, the date is stored at first and is divided into three parts. The first one is the date according to Bengali calendar, second one is the day in Bengali and the last one is the date according to English calendar. Both Bengali and English dates are stored in the form 'day month year'.

A sequence of four Bengali digits separates the Bengali date from the Bengali day. The English date starts with one/two digits in Bengali font. Bengali date, day and English date can be distinguished by checking the appearance of the numerals and these are tagged as <bd>, <day> and <ed>, respectively. For e.g., *25 sraban 1412 budhbar 10 august 2005* is tagged as shown in Table 5.

Tag	Definition	Tag	Definition	Tag	Definition
header	Header of the news documents	day	Day	body	Body of the news document
title	Headline of the news document	ed	English date	p	Paragraph
t1	1st headline of the title	reporter	Reporter name	table	Information in tabular form
t2	2nd headline of the title	agency	Agency providing news	tc	Table column
date	Date of the news document	location	News location	tr	Table row
bd	Bengali date				

Table 2. News Corpus Tagset

Tag	Meaning	Example
PER	Single word person name	*sachin* / PER, *manmohan*/PER
LOC	Single word location name	*jadavpur* / LOC, *delhi*/LOC
ORG	Single word organization name	*infosys* / ORG, *tifr*/ORG
MISC	Single word miscellaneous name	*100%* / MISC, *100*/MISC
B-PER I-PER E-PER	Beginning, Internal or the end of a multiword person name	*sachin*/ B-PER *ramesh* / I-PER *tendulkar* / E- PER
B-LOC I-LOC E-LOC	Beginning, Internal or the end of a multiword location name	*mahatma*/ B-LOC *gandhi* / I-LOC *road* / E-LOC
B-ORG I-ORG E-ORG	Beginning, Internal or the end of a multiword organization name	*bhaba* / B-ORG *atomic* / I-ORG *research* / I-ORG *centre* / E-ORG
B-MISC I-MISC E-MISC	Beginning, Internal or the end of a multiword miscellaneous name	*10 e* / B-MISC *magh* / I-MISC *1402* / E-MISC
NNE	Words that are not named entities	*neta*/NNE, *bidhansabha*/NNE

Table 3. Named Entity Tagset

NE tag	Meaning	Example
NEP	Person name	*sachin ramesh tendulkar* / NEP
NEL	Location name	*mahatma gandhi road* / NEL
NEO	Organization name	*bhaba atomic research centre* / NEO
NED	Designation	*chairman*/NED, *sangsad*/NED
NEA	Abbreviation	*b a*/NEA, *c m d a*/NEA, *b j p*/NEA
NEB	Brand	*fanta*/NEB, *windows/*NEB
NETP	Title-person	*sriman*/NED, *sree*/NED
NETO	Title-object	*american beauty/*NETO
NEN	Number	*10*/NEN, *dash/*NEN
NEM	Measure	*tin din/*NEM, *panch keji/*NEM
NETE	Terms	*hidden markov model*/NETE
NETI	Time	*10 e magh 1402*/ NETI

Table 4. IJCNLP-08 NER Shared Task Tagset

Original date pattern	Tagged date pattern
	<date>
25 sraban 1412	<bd>*25 sraban 1412* </bd>
budhbar	<day>*budhbar* </day>
10 august 2005	<ed>*10 august 2005* </ed>
	</date>

Table 5. Example of a Tagged Date Pattern

2.5 Named Entity Tagged Corpus Development

The partially NE tagged corpus contains 34 million wordforms and are in both ISCII and UTF-8 forms. A portion of this corpus, containing 150K wordforms, has been manually annotated with the sixteen NE tags that are listed in Table 3. The corpus has been annotated with the help of *Sanchay Editor*, a text editor for Indian languages. The detailed statistics of this NE-tagged corpus are given in Table 6. The corpus is in SSF form, which has the following structure:

```
<Story id="">
<Sentence id="">
1        biganni NNE
2        newton  PER
3        .
</Sentence id="">
              .
</Story id="">
```

Another portion of the partially NE tagged Bengali news corpus has been manually annotated as part of the IJCNLP-08 NER shared task with the twelve NE tags, as listed in Table 4. The annotation process has been very difficult due to the presence of a number of ambiguous NE tags. The potential ambiguous NE tags are: NED vs NETP, NEO vs NEB, NETE vs NETO, NETO vs NETP and NEN vs NEM. For example, it is difficult to decide whether ‘Agriculture’ is ‘NETE’, and if no then whether ‘Horticulture’ is ‘NETE’ or not. In fact, this the most difficult class to identify. This NE tagged corpus contains approximately 30K wordforms. Details statistics of this tagged corpus are shown in Table 7. This NE tagged corpus is in the following SSF form.

```
<Story id="">
<Sentence id="">
 1      ((        NP      <ne=NEP>
 1.1  ((          NP      <ne=NED>
 1.1.1    biganni
        ))
1.1.2  newton NEP
        ))
 2      .
</Sentence id="">
</Story id="">
```

NE Class	Number of wordforms	Number of distinct wordforms
Person name	20, 455	15, 663
Location name	11, 668	5, 579
Organization name	963	867
Miscellaneous name	11,554	3, 227

Table 6. Statistics of the 150K-tagged Corpus

2.6 Tag Conversion

A tag conversion routine has been developed in order to convert the sixteen-NE tagged corpus of 150K wordforms to the corpus, tagged with the IJCNLP-08 twelve-NE tags. This conversion is a semi-automatic process. Some of our sixteen NE tags can be automatically mapped to some of the IJCNLP-08 shared task tags. The tags that represent person, location and organization names can be directly mapped to the NEP, NEL and NEO tags, respectively. Other IJCNLP-08 shared task tags can be obtained with the help of gazetteer lists and simple heuristics. In our earlier NER experiments, we have already developed a number of gazetteer lists such as: lists of person, location and organization names; list of prefix words (e.g., *sree, sriman* etc.) that predict the left boundary of a person name; list of designation words (e.g., *mantri, sangsad* etc.) that helps to identify person names. The lists of prefix and designation words are helpful to find the NETP and NED tags. The sixteen-NE tagged corpus is searched for the person name tags and the previous word is matched against the lists of prefix and designation words. The previous word is tagged as NED or NETP if there is a match with the lists of designation words and prefix words, respectively. The NEN and NETI tags can be obtained by looking at our miscellaneous tags and using some simple heuristics. The NEN tags can be simply obtained by checking whether the MISC tagged element consists of digits only. The lists of cardinal and ordinal numbers have been also kept to recognize NETI tags. A list of words that denote the measurements (e.g., *kilogram, taka, dollar* etc.) has been kept in order to get the NEM tag. The lists of words, denoting the brand names, title-objects and terms, have been prepared to get the NEB, NETO and NETE tags. The NEA tags can be obtained with the help of a gazetteer list and using some simple heuristics (whether the word contains the dot and there is no space between the characters). The mapping from our NE tagset to the IJCNLP-08 NER shared task tagset is shown in Table 8.

3 Use of Language Resources

The NE tagged news corpus, developed in this work, has been used to develop the Named Entity Recognition (NER) systems in Bengali using pattern directed shallow parsing, HMM, ME, CRF and SVM frameworks.

NE Class	Number of wordforms	Number of distinct wordforms
Person name	5, 123	3, 201
Location name	1, 675	1, 119
Organization name	168	131
Designation	231	102
Abbreviation	32	21
Brand	15	12
Title-person	79	51
Title-object	63	42
Number	324	126
Measure	54	31
Time	337	212
Terms	46	29

Table 7. Statistics of the 30K-tagged Corpus

Sixteen-NE tagset	IJCNLP-08 tagset
PER, LOC, ORG	NEP, NEL, NEO
B-PER, I-PER, E-PER	NEP
B-LOC, I-LOC, E-LOC	NEL
B-ORG, I-ORG, E-ORG	NEO
MISC	NEN
B-MISC, I-MISC, E-MISC	NETI, NEM
Brand name gazetteer	NEB
Title-object gazetteer	NETO
Term gazetteer	NETE
Person prefix word + PER/B-PER, I-PER, E-PER	NETP
Designation word +PER/B-PER, I-PER, E-PER	NED
Abbreviation gazetteer + Heuristics	NEA

Table 8. Tagset Mapping Table

We have considered the sixteen NE tags to develop these systems. Named entity recognition in Indian Languages (ILs) in general and particularly in Bengali is difficult and challenging as there is no concept of capitalization in ILs.

The Bengali NER systems that use pattern directed shallow parsing approach can be found in

(Ekbal and Bandyopadhyay, 2007a) and (Ekbal and Bandyopadhyay, 2007b). An HMM-based Bengali NER system can be found in (Ekbal and Bandyopadhyay, 2007c). These systems have been trained and tested with the corpus tagged with the sixteen NE tags.

A number of experiments have been conducted in order to find out the best feature set for NER in Bengali using the ME, CRF and SVM frameworks. In all these experiments, we have used a number of gazetteer lists such as: first names (72, 206 entries), middle names (1,491 entries) and sur names (5,288 entries) of persons; prefix (245 entries) and designation words (947 entries) that help to detect person names; suffixes (45 and 23 entries) that help to identify person and location names; clue words (94 entries) that help to detect organization names; location name (7, 870 entries) and organization name (2,225 entries). Apart from these gazetteer lists, we have used the prefix and suffix (may not be linguistically meaningful suffix/prefix) features, digit features, first word feature and part of speech information of the words etc. We have used the C++ based Maximum Entropy package[6], C^{++} based OpenNLP CRF++ package[7] and open source YamCha[8] tool for ME based NER, CRF based NER and SVM based NER, respectively. For SVM based NER system, we have used TinySVM[9] classifier, pair wise multi-class decision method and the second-degree polynomial kernel function. The brief descriptions of all the models are given below:

- A: Pattern directed shallow parsing approach without linguistic knowledge.
- B: Pattern directed shallow parsing approach with linguistic knowledge.
- HMM based NER: Trigram model with additional context dependency, NE suffix lists for handling unknown words.
- ME based NER: Contextual window of size three (current, previous and the next word), prefix and suffix of length upto three of the current word, POS information of the current word, NE information of the previous word (dynamic feature), different digit features and the various gazetteer liststs.
- CRF based NER: Contextual window of size five (current, previous two words and the next two words), prefix and suffix of length upto three of the current word, POS information of window three (current word, previous word and the next word), NE information of the previous word (dynamic feature), different digit features and the various gazetteer lists.
- SVM based NER: Contextual window of size six (current, previous three words and the next two words), prefix and suffix of length upto three of the current word, POS information of window three (current word, previous word and the next word), NE information of the previous two words (dynamic feature), different digit features and the various gazetteer lists.

Evaluation results of the 10-fold cross validation test for all the models are presented in Table 9. Evaluation results clearly show that the SVM based NER model outperforms other models due to it's efficiency to handle the non-independent, diverse and overlapping features of Bengali language.

Model	F-Score (in %)
A	74.5
B	77.9
HMM	84.5
ME	87.4
CRF	90.7
SVM	91.8

Table 9.Results of 10-fold Cross Validation Test

4 Conclusion

In this work we have developed a Bengali news corpus of approximately 34 million wordforms from the web archive of a leading Bengali newspaper. The *date, location, reporter* and *agency tags* present in the web pages have been automatically NE tagged. Around 150K wordforms of this partially NE tagged corpus has been manually annotated with the sixteen NE tags. We have also tagged around 30K wordforms with the twelve NE tags, defined for the IJCNLP-08 NER shared task. A tag conversion routine has also been developed in order to convert any sixteen-NE tagged corpus to the twelve-NE tagged corpus. The sixteen-NE tagged corpus of 150K wordforms has been used to

[6] http://homepages.inf.ed.ac.uk/s0450736/software/maxent/maxent-20061005.tar.bz2

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

[8] http://chasen.org/~taku/software/yamcha/

[9] http://cl.aist-nara.ac.jp/taku-ku/software/TinySVM

develop the NER systems using various machine-learning approaches.

This NE tagged corpus can be used for other NLP research activities such as machine translation, information retrieval, cross-lingual event tracking, automatic summarization etc.

References

Bertagna, M. and A. Lenci, M. Monachini and N. Calzolari. 2004. CotentInteroperability of Lexical Resources, Open Issues and "MILE" Perspectives, In *Proceedings of the LREC*, 131-134.

Bharthi, A., D.M. Sharma, V. Chaitnya, A. P. Kulkarni and R. Sanghal. 2001. LERIL: Collaborative Effort for Creating Lexical Resources. In *Proceedings of the 6th NLP Pacific Rim Symposium Post-Conference Workshop*, Japan.

Bikel, D. M., Schwartz, R., Weischedel, R. M. 1999. An Algorithm that Learns What's in a Name. *Machine Learning*, 34, 211-231.

Calzolari, N., F. Bertagna, A. Lenci and M. Monachni. 2003. Standards and best Practice for Miltilingual Computational Lexicons, MILE (the multilingual ISLE lexical entry). *ISLE Deliverable D2.2 &3.2*.

Chieu, H. L., Tou Ng, H. 2002. Named Entity Recognition: A Maximum Entropy Approach Using Global Information, In *Proc. of the 6th Workshop on Very Large Corpora.*

De Sitter, A., Daelemans W. 2003. Information Extraction via Double Classification. In *Proeedings of International Workshop on Adaptive Text Extraction and Mining*, Dubronik.

Ekbal, Asif, and S. Bandyopadhyay. 2007a. Pattern Based Bootstrapping Method for Named Entity Recognition. In *Proceedings of the 6th International Conference on Advances in Pattern Recognition,* 2007, India, 349-355.

Ekbal, Asif, and S. Bandyopadhyay. 2007b. Lexical Pattern Learning from Corpus Data for Named Entity Recognition. In *Proceedings of the 5th International Conference on Natural Language Processing (ICON)*, 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 Bengali News Corpus for Named Entity Recognition. *Language Resources and Evaluation Journal* (Accepted and to appear by December 2007).

Fletcher, W. H. 2001. Making the Web More Useful as Source for Linguistics Corpora. In Ulla Conor and Thomas A. Upton (eds.), Applied corpus Linguistics: A Multidimensional Perspective, 191-205.

Fletcher, W. H. 2003. Concording the Web with KwiCFinder. In *Proceedings of the Third North American Symposium on Corpus Linguistics and Language Teaching, Boston.*

Giguet, E., and P. Luquet. 2006. Multilingual Lexical Database Generation from Parallel Texts in 20 European Languages with Endogeneous Resources. In *Proceedings of the COLING/ACL*, Sydney, 271-278.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, In *Proceedings of the 18th International Conference on Machine Learning*, 282-289.

Lenci, A., N. Bel, F. Busu, N. Calzolari, E. Gola, M. Monachini, A. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas and A. Zampolli. 2000. SIMPLE: A general Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography, Special Issue, Dictionaries, Thesauri and Lexical-Semantic Relations*, XIII(4): 249-263.

Li, Wei and Andrew McCallum. 2004. Rapid Development of Hindi Named Entity Recognition Using Conditional Random Fields and Feature Inductions. *ACM TALIP*, Vol. 2(3), 290-294.

McCallum, A., Freitag, D., Pereira, F. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proceedings of the 17th International Conference Machine Learning.*

Robb, T. 2003. Google as a Corpus Tool? *ETJ Journal,* 4(1), Spring 2003.

Rundell, M. 2000. The Biggest Corpus of All. *Humanising Language Teaching*, 2(3).

Sun, A., et al. 2003. Using Support Vector Machine for Terrorism Information Extraction. In *Proceedings of 1st NSF/NIJ Symposium on Intelligence and Security.*

Takenobou, T., V. Sornlertlamvanich, T. Charoenporn, N. Calzolari, M. Monachini, C. Soria, C. Huang, X. YingJu, Y. Hao, L. Prevot and S. Kiyoaki. 2006. Infrastructure for Standardization of Asian Languages Resources. In *Proceedings of the COLING/ACL* 2006, Sydney, 827-834.