Named Entity Recognition for South and South East Asian Languages: Taking Stock

Anil Kumar Singh
Language Technologies Research Centre
IIIT, Hyderabad, India
anil@research.iiit.ac.in

Abstract

In this paper we first present a brief discussion of the problem of Named Entity Recognition (NER) in the context of the IJCNLP workshop on NER for South and South East Asian (SSEA) languages[1]. We also presents a short report on the development of a named entity annotated corpus in five South Asian language, namely Hindi, Bengali, Telugu, Oriya and Urdu. We present some details about a new named entity tagset used for this corpus and describe the annotation guidelines. Since the corpus was used for a shared task, we also explain the evaluation measures used for the task. We then present the results of our experiments on a baseline which uses a maximum entropy based approach. Finally, we give an overview of the papers to be presented at the workshop, including those from the shared task track. We discuss the results obtained by teams participating in the task and compare their results with the baseline results.

1 Introduction

One of the motivations for organizing a workshop (NERSSEAL-08) focused on named entities (NEs) was that they have a special status in Natural Language Processing (NLP) because they have some properties which other elements of human languages do not have, e.g. they refer to specific things or concepts in the world and are not listed in the grammars or the lexicons. Identifying and classifying them automatically can help us in processing text because they form a significant portion of the types and tokens occurring in a corpus. Also, because of their very nature, machine learning techniques have been found to be very useful in identifying them. In order to use these machine learning techniques, we need corpus annotated with named entities. In this paper we describe such a corpus developed for five South Asian languages. These languages are Hindi, Bengali, Oriya, Telugu and Urdu.

This paper also presents an overview of the work done for the IJCNLP workshop on NER for SSEA languages. The workshop included two tracks. The first track was for regular research papers, while the second was organized on the lines of a shared task.

Fairly mature named entity recognition systems are now available for European languages (Sang, 2002; Sang and De Meulder, 2003), especially English, and even for East Asian languages (Sassano and Utsuro, 2000). However, for South and South East Asian languages, the problem of NER is still far from being solved. Even though we can gain much insight from the methods used for English, there are many issues which make the nature of the problem different for SSEA languages. For example, these languages do not have capitalization, which is a major feature used by NER systems for European languages.

Another characteristic of these languages is that most of them use scripts of Brahmi origin, which have highly phonetic characteristics that could be utilized for multilingual NER. For some languages, there are additional issues like word segmentation

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

Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 5–16,
Hyderabad, India, January 2008.

(e.g. for Thai). Large gazetteers are not available for most of these languages. There is also the problem of lack of standardization and spelling variation. The number of frequently used words (common nouns) which can also be used as names (proper nouns) is very large for, unlike for European languages where a larger proportion of the first names are not used as common words. For example, 'Smith', 'John', 'Thomas' and 'George' etc. are almost always used as person names, but 'Anand', 'Vijay', 'Kiran' and even 'Manmohan' can be (more than often) used as common nouns. And the frequency with which they can be used as common nouns as against person names is more or less unpredictable. The context might help in disambiguating, but this issue does make the problem much harder than for English.

Among other problems, one example is that of the various ways of representing abbreviations. Because of the alpha-syllabic nature of the SSEA scripts, abbreviation can be expressed through a sequence of letters or syllables. In the latter case, the syllables are often combined together to form a pseudo-word, e.g. BAjapA (bhaajapaa) for Bharatiya Janata Party or BJP.

But most importantly, there is a serious lack of labeled data for machine learning. As part of this workshop, we have tried to prepare some data but we will need much more data for really accurate NER systems.

Since most of the South and South East Asian languages are scarce in resources as well as tools, it is very important that good systems for NER be available, because many problems in information extraction and machine translation (among others) are dependent on accurate NER.

The need for a workshop specifically for SSEA languages was felt because the South and South East Asian region has many major and numerous minor languages. In terms of the number of speakers there are at least four in any list of top ten languages of the world. For practical reasons, we focus only on the major languages in the workshop (and in this paper). Most of the major languages belong to two families: Indo-European and Dravidian. There are a lot of differences among these languages, but there are a lot of similarities too, even across families (**?**; **?**). For the reasons mentioned above, NER is perhaps more difficult for SSEA languages than for European languages. For better or for worse, there too many languages and too few resources. Moreover, these languages are also comparatively less studied by researchers. However, we can benefit from the similarities across these languages to build multilingual systems so as to reduce the overall cost and effort required.

All the issues mentioned above show that we might need different methods for solving the NER problem for SSEA languages. However, for comparing the results of these different methods, we will need a reasonably good baseline. A mature system tuned for English but trained on SSEA language data can become such a baseline. We will describe such a baseline in a later section. This baseline system has been tested on the data provided for the shared task. We present the results for all five languages under the settings required for the shared task.

2 Related Work

Various techniques have been used for solving the NER problem (Mikheev et al., 1999; Borthwick, 1999; Cucerzan and Yarowsky, 1999; Chieu and Ng, 2003; Klein et al., 2003; Kim and Woodland, 2000) ranging from naively using gazetteers to rules based techniques to purely statistical techniques, even hybrid approaches. Several workshops consisting of shared tasks (Sang, 2002; Sang and De Meulder, 2003) have been held with specific focus on this problem. In this section we will mention some of techniques used previously.

Most of the approaches can be classified based on the features they use, whether they are rule based or machine learning based or hybrid approaches. Some of the commonly used features are:

- Word form and part of speech (POS) tags
- Orthographic features like capitalization, decimal, digits
- Word type patterns
- Conjunction of types like capitalization, quotes, functional words etc.
- Bag of words
- Trigger words like New York *City*

Tag	Name	Description
NEP	Person	Bob Dylan, Mohandas Gandhi
NED	Designation	General Manager, Commissioner
NEO	Organization	Municipal Corporation
NEA	Abbreviation	NLP, B.J.P.
NEB	Brand	Pepsi, Nike (ambiguous)
NETP	Title-Person	Mahatma, Dr., Mr.
NETO	Title-Object	Pride and Prejudice, Othello
NEL	Location	New Delhi, Paris
NETI	Time	3rd September, 1991 (ambiguous)
NEN	Number	3.14, 4,500
NEM	Measure	Rs. 4,500, 5 kg
NETE	Terms	Maximum Entropy, Archeology

Table 1: The named entity tagset used for the shared task

- Affixes like Hydera*bad*, Ram*pur*, Mehdi*patnam*, Lingam*pally*
- Gazetteer features: class in the gazetteer
- Left and right context
- Token length, e.g. the number of letters in a word
- Previous history in the document or the corpus
- Classes of preceding NEs

The machine learning techniques tried for NER include the following:

- Hidden Markov Models or HMM (Zhou and Su, 2001)
- Decision Trees (Isozaki, 2001)
- Maximum Entropy (Borthwick et al., 1998)
- Support Vector Machines or SVM (Takeuchi and Collier, 2002)
- Conditional Random Fields or CRF (Settles, 2004)

Different ways of classifying named entities have been used, i.e., there are more than one tagsets for NER. For example, the CoNLL 2003 shared task[2] had only four tags: persons, locations, organizations and miscellaneous. On the other hand, MUC-6[3] has a near ontology for information extraction purposes. In this (MUC-6) tagset, there are three[4] main kinds of NEs: ENAMEX (persons, locations and organizations), TIMES (time expressions) and NUMEX (number expresssions).

There has been some previous work on NER for SSEA languages (McCallum and Li, 2003; Cucerzan and Yarowsky, 1999), but most of the time such work was an offshoot of the work done for European languages. Even including the current workshop, the work on NER for SSEA languages is still in the initial stages as the results reported by papers in this workshop clearly show.

3 A New Named Entity Tagset

The tagset being used for the NERSSEAL-08 shared task consists of more tags than the four tags used for the CoNLL 2003 shared task. The reason we opted for these tags was that we needed a slightly finer tagset for machine translation (MT). The initial aim was to improve the performance of the MT system.

As annotation progressed, we realized that there were some problems that we had not anticipated. Some classes were hard to distinguish in some contexts, making the task hard for annotators and bringing in inconsistencies. For example, it was not always clear whether something should be marked as

[2]http://www.cnts.ua.ac.be/conll2003/ner/

[3]http://cs.nyu.edu/cs/faculty/grishman/muc6.html

[4]http://cs.nyu.edu/cs/faculty/grishman/NEtask20.book_6.html

Number or as Measure. Similarly for Time and Measure. Another difficult class was that of (technical) terms. Is 'agriculture' a term or not? If no (as most people would say), is 'horticulture' a term or not? In fact, Term was the most difficult class to mark.

An option that we explored was to merge the above mentioned confusable classes and ignore the Term class. But we already had a relatively large corpus marked up with these classes. If we merged some classes and ignored the Term class (which had a very large coverage and is definitely going to be useful for MT), we would be throwing away a lot of information. And we also had some corpus annotated by others which was based on a different tagset. So some problems were inevitable. Finally, we decided to keep the original tagset, with one modification. The initial tagset had only eleven tags. The problem was that there was one Title tag but it had two different meanings: 'Mr.' is a Title, but 'The Seven Year Itch' is also a Title. This tag clearly needed to be split into two: Title-Person and Title-Object

We should mention here that we considered using another tagset developed at AUKBC, Chennai. This was based on ENAMEX, TIMEX and NUMEX. The total number of tags in this tagset is more than a hundred and it is meant specifically for MT and only for certain domains (health, tourism). Moreover, this is a tagset for entities in general, not just named entities.

The twelve tags in our tagset are briefly explained in Table-1. In the next section we mention the constraints under which the annotated corpus was created, using this tagset.

4 Annotation Constraints

The annotated corpus was created under severe constraints. The annotation was to be for five languages by different teams, sometimes with very little communication during the process of annotation. As a result, there were many logistical problems.

There were other practical constraints like the fact that this was not a funded project and all the work was mainly voluntary. Another major constraint for all the languages except Hindi was time. There was not enough time for cross validation as the corpus was required by a deadline. To keep annotation reasonably consistent, annotation guidelines were created and a common format was specified.

5 Annotation Guidelines

The annotation guidelines were of two kinds. One was meant for preparing training data through manual annotation. The other one was meant for preparing reference data as well as for automatic annotation. The main guidelines for preparing the training data are as follows:

- *Specificity*: The most important criterion while deciding whether some expression is a named entity or not is to see whether that expression specifies something definite and identifiable as if by a name or not. This decision will have to be based on the context. For example, 'aanand' (in South Asian languages, where there is no capitalization) is not a named entity in 'saba aanand hii aanand hai' ('There is bliss everywhere'). But it is a named entity in 'aanand kaa yaha aakhiri saala hai' ('Anand is in the last year (of his studies)'). Number, Measure and Term may be seen as exceptions (see below).

- *Maximal Entity*: Only the maximal entities have to be annotated for training data. Structure of entities will not be annotated by the annotators, even though it has to be learnt by the NER systems. For example, 'One Hundred Years of Solitude' has to be annotated as one entity. 'One Hundred' is not to be marked as a Number here, nor is 'One Hundred Years' to be made marked as a Measure in this case. The purpose of this guideline is to make the task of annotation for several languages feasible, given the constraints.

- *Ambiguity*: In cases where an entity can have two valid tags, the more appropriate one is to be used. The annotator has to make the decision in such cases. It is recommended that the annotation be validated by another person, or even more preferably, two different annotators have to work on the same data independently and inconsistencies have to be resolved by an adjudicator. Abbreviation is an exception to the Ambiguity guideline (see below).

Some other guidelines for specific tags are listed below:

- *Abbreviations*: All abbreviations have to be marked as Abbreviations, Even though every abbreviation is also some other kind of named entity. For example, APJ is an Abbreviation, but also a Person. IBM is also an Organization. Such ambiguity cannot be resolved from the context because it is due to the (wrong?) assumption that a named entity can have only one tag. Multiple annotations were not allowed. This is an exception to the third guideline above.

- *Designation and Title-Person*: An entity is a Designation if it represents something formal and official status with certain responsibilities. If it is just something honorary, then it is a Title-Object. For example, 'Event Coordinator' or 'Research Assistant' is a Designation, but 'Chakravarti' or 'Mahatma' are Titles.

- *Organization and Brand*: The distinction between these two has to be made based on the context. For example, 'Pepsi' could mean an Organization, but it is more likely to mean a Brand.

- *Time and Location*: Whether something is to be marked as Time or Location or not is to be decided based on the Specificity guideline and the context.

- *Number, Measure and Term*: These three may not be strictly named entities in the way a person name is. However, we have included them because they are different from other words of the language. For problems like machine translation, they can be treated like named entities. For example, a Term is a word which can be directly translated into some language if we have a dictionary of technical terms. Once we know a word is a Term, there is likely to be less ambiguity about the intended sense of the word, unlike for other normal words.

The second set of guidelines are different from the first set mainly in one respect: the corpus has to be annotated with not just the maximal NEs, but with all levels of NEs, i.e., nested NEs also have to be marked.

Nested entities were introduced because one of the requirements was that the corpus be useful for building systems which can become parts of a machine translation (MT) system. Nested entities can be useful for MT systems because, quite often, parts of the entities can need to be translated, while the others can just be transliterated. An example of a nested named entity is 'Mahatma Gandhi International Hindi University'. This would be translated in Hindi as *mahaatmaa gaandhii antarraashtriya hindii vishvavidyaalaya*. Only 'International' and 'University' are to be translated, while the other words are to be transliterated. The nested named entities in this case are: 'Mahatma' (NETO), 'Gandhi' (NEP), 'Mahatma Gandhi' (NEP), and 'Mahatma Gandhi International Hindi University' (NEO).

6 Named Entity Annotated Corpus

For Hindi, Oriya and Telugu, all the annotation was performed at IIIT, Hyderabad. For Bengali, the corpus was developed at IIIT, Hyderabad and Jadavpur University (Ekbal and Bandyopadhyay, 2008b), Calcutta. For Urdu, annotation was performed at CRULP, Lahore (Hussain, 2008) and IIIT, Allahabd. Even though all the annotation was done by native speakers of respective languages, named entity annotation was a new task for everyone involved. This was because of practical constraints as explained in an earlier section.

The corpus was divided into two parts, one for training and one for testing. The testing corpus was annotated with nested named entities, while the training corpus was only annotated with 'maximal' named entities.

Since different teams were working on different languages, in some cases even the same language, and also because most of the corpus was created on short notice, each team made its own decisions regarding the kind of corpus to be annotated. As a result, the characteristics of the corpus differ widely among the five languages. The Hindi and Bengali (partly) text that was annotated was from the multilingual comparable corpus known as the CIIL (Central Institute of Indian Languages) corpus. The Oriya corpus was part of the Gyan Nidhi corpus.

NE	Hindi		Bengali		Oriya		Telugu		Urdu	
	Trn	**Tst**	**Trn**	**Tst**	**Trn**	**Tst**	**Trn**	**Tst**	**Trn**	**Tst**
NEP	4025	199	1299	728	2079	698	1757	330	365	145
NED	935	61	185	11	67	216	87	77	98	41
NEO	1225	44	264	20	87	200	86	12	155	40
NEA	345	7	111	9	8	20	97	112	39	3
NEB	5	0	22	0	11	1	1	6	9	18
NETP	1	5	68	57	54	201	103	2	36	15
NETO	964	88	204	46	37	28	276	118	4	147
NEL	4089	211	634	202	525	564	258	751	1118	468
NETI	1760	50	285	46	102	122	244	982	279	59
NEN	6116	497	407	144	124	232	1444	391	310	47
NEM	1287	17	352	146	280	139	315	53	140	40
NETE	5658	843	1165	314	5	0	3498	138	30	4
NEs	26432	2022	5000	1723	3381	2421	8178	3153	2584	1027
Words	503179	32796	112845	38708	93173	27007	64026	8006	35447	12805
Sentences	19998	2242	6030	1835	1801	452	5285	337	1508	498
Trn: Training Data, **Tst**: Testing Data										

Table 2: Statistics about the corpus: counts of various named entity classes and the size of the corpus as the number of words and the number of sentences. Note that the values for the testing part are of nested NEs. Also, the number of sentences, especially in the case or Oriya is not accurate because the sentences were not correctly segmented as there was no automatic sentence splitter available for these languages and manual splitting would have been too costly: without much benefit for the NER task.

Both of these (CIIL and Gyan Nidhi) corpora consist of text from educational books written on various topics for common readers. The Urdu text was partly news corpus. The same was the case with Telugu, but the text for both these languages included text from other domains too.

Admittedly, the texts selected for annotation were not the ideal ones. For example, many documents had very few named entities. Also, the distribution of domains as well as the classes of NEs was not representative. The size of the annotated corpora for different languages is also widely varying, with Hindi having the largest corpus and Urdu the smallest. However, this corpus is hopefully just a starting point for much more work in the near future.

Some statistics about the annotated corpus are given in Table-2.

7 Shared Task

In the shared task, the contestants having their own NER systems were given some annotated test data. The contestants had the freedom to use any technique for NER, e.g. a purely rule based technique or a purely statistical technique.

The contestants could build NER systems targeted for a specific language, but they were required to report results for their systems on all the languages for which training data had been provided. This condition was meant to provide a somewhat fair ground for comparison of systems, since the amount of training data is different for different languages.

The data released for the shared task has been made accessible to all for non-profit research word, not just for the shared task participants, with the hope others will contribute in improving this data and adding to it.

The task in this contest was different in one important way. The NER systems also had to identify nested named entities. For example, in the sentence The Lal Bahadur Shastri National Academy of Administration is located in Mussoorie, 'Lal Bahadur Shastri' is a Person, but 'Lal Bahadur Shastri National Academy of Administration' is an Organization. In this case, the NER systems had to identify both 'Person' and 'Organization' in the given sentence.

An evaluation script was also provided to evaluate the performance of different systems in a uniform way.

8 Evaluation Measures

As part of the evaluation process for the shared task, precision, recall and F-measure had to be calculated for three cases: maximal named entities, nested named entities and lexical matches. Thus, there were nine measures of performance:

- Maximal Precision: $P_m = \frac{c_m}{r_m}$
- Maximal Recall: $R_m = \frac{c_m}{t_m}$
- Maximal F-Measure: $F_m = \frac{2*P_m*R_m}{P_m+R_m}$
- Nested Precision: $P_n = \frac{c_n}{r_n}$
- Nested Recall: $R_n = \frac{c_n}{t_n}$
- Nested F-Measure: $F_n = \frac{2P_nR_n}{P_n+R_n}$
- Lexical Precision: $P_l = \frac{c_l}{r_l}$
- Lexical Recall: $R_l = \frac{c_l}{t_l}$
- Lexical F-Measure: $F_l = \frac{2P_lR_l}{P_l+R_l}$

where c is the number of correctly retrieved (identified) named entities, r is the total number of named entities retrieved by the system being evaluated (correct plus incorrect) and t is the total number of named entities in the reference data.

The participants were encouraged to report results for specific classes of NEs. Evaluation was automatic and was against the manually prepared reference data given to the participants. An evaluation script for this purpose was also provided. This script assumes that there are single test and reference file and the number and order of sentences is the same in both. The format accepted by the evaluation script (which was also the format used for annotated data) was explained in an online tutorial[5].

9 Experiments on a Baseline

For our baseline experiments, we used an open source implementation of maximum entropy based Natural Languages Processing tools which are part of the OpenNLP[6] package. This package includes a name finder tool.

This name finder was trained for all the twelve classes of NEs and for all the five languages. The test data, which was the same as that given to the shared task participants, was run through this name finder. Note that this NER tool is tuned for English in terms of the features used, even though it was trained on different SSEA languages in our case. Since the goal of the shared task was to encourage investigation of techniques (especially features) specific to the SSEA languages, this fairly mature NER system (for English) could be used as a baseline against which to evaluate systems tuned (or specially designed) for the five South Asian languages.

The overall results of the baseline experiments are shown in Table-3. The performance on specific NE classes is given in Table-4. It can be seen from the tables that the results are drastically low in comparison to the state of the art results reported for English. These results clearly show that even a machine learning based system cannot be directly used for SSEA languages even when it has been trained with annotated data for these languages.

In the next section we present a brief overview of the papers selected for the workshop including the shared task papers.

10 An Overview of the Papers

In all, twelve papers were selected for the workshop, out of which four were in the shared task track. Saha et al., 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. For Hindi, the features they utilized include orthographic features, information about suffixes and prefixes, morphological features, part of speech information, and information about the surrounding words. They used rules for numbers, measures and time classes. For designation, title-person and some terms (NETE), they built lists or gazetteers. They also used gazetteers for person and location. They did not use rules or gazetteers for Oriya, Urdu and Telugu. To identify some kinds of nested entities, they applied a set of rules.

Gali et al. also combined machine learning with language specific heuristics. In a separate section, they discussed at some length the issues relevant to NER for SSEA languages. Some of these have al-

[5] http://ltrc.iiit.ac.in/ner-ssea-08/NER-SAL-TUT.pdf
[6] http://opennlp.sourceforge.net/

Measure →	**Precision**			**Recall**			**F-Measure**		
Language ↓	P_m	P_n	P_l	R_m	R_n	R_l	F_m	F_n	F_l
Bengali	50.00	44.90	52.20	07.14	06.90	06.97	12.50	11.97	12.30
Hindi	75.05	73.61	73.99	18.16	17.66	15.53	29.24	28.48	25.68
Oriya	29.63	27.46	48.25	09.11	07.60	12.18	13.94	11.91	19.44
Telugu	00.89	02.83	22.85	00.20	00.67	5.41	00.32	01.08	08.75
Urdu	47.14	43.50	51.72	18.35	16.94	18.94	26.41	24.39	27.73
m**: Maximal,** n**: Nested,** l**: Lexical**									

Table 3: Results for the experiments on a baseline for the five South Asian languages

	Bengali	**Hindi**	**Oriya**	**Telugu**	**Urdu**
NEP	06.62	26.23	28.48	00.00	04.39
NED	00.00	12.20	00.00	00.00	00.00
NEO	00.00	15.50	03.30	00.00	11.98
NEA	00.00	00.00	00.00	00.00	00.00
NEB	NP	NP	00.00	00.00	00.00
NETP	00.00	NP	11.62	00.00	00.00
NETO	00.00	05.92	04.08	00.00	00.00
NEL	03.03	44.79	25.49	00.00	40.21
NETI	34.00	47.41	22.38	01.51	38.38
NEN	62.63	62.22	10.65	03.51	09.52
NEM	13.61	24.39	08.03	00.71	07.15
NETE	00.00	00.18	00.00	00.00	00.00
NP: Not present in the reference data					

Table 4: Baseline results for specific named entity classes (F-Measures for nested lexical match)

ready been mentioned, but two others are the agglutinative property of these (especially Dravidian) languages and the low accuracy of available part of speech taggers, particularly for nouns. They used a Conditional Random Fields (CRF) based method for machine learning and applied heuristics to take care of the language specific issues. They also point out that a very high percentage of NEs in the Hindi corpus were marked as NETE and machine learning failed to take care of this class of NEs. This has been validated by our results on the baseline too (Table-4) and is understandable because terms are hard to identify even for humans.

Ekbal et al. also used an approach based on CRFs. They also used some language specific features for Hindi and Bengali. Srikanth and Murthy describe the results of their experiments on NER using CRFs for Telugu. They concentrated only on person, place and organization names and used newspaper text as the corpus. In this focused setting, they were able to achieve overall F-measures between 80% and 97% in various experiments. Chaudhuri and Bhattacharya 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. For this task, they were able to achieve an overall F-measure of 89.51%.

Praveen and Ravi Kumar present the results of experiments (as part of the shared task) using two approaches: Hidden Markov Models (HMM) and CRF. Interestingly, they obtained better results with HMM for all the five languages. Goyal described experiments using a CRF based model. He also used part of speech information. He experimented only on Hindi and was able to achieve results above 60%. One notable fact about this paper is that it also de-

Language ↓	BL	IK	IH1	IH2	JU
Bengali	12.30	65.96	40.63	39.77	59.39
Hindi	25.68	65.13	50.06	46.84	33.12
Oriya	19.44	44.65	39.04	45.84	28.71
Telugu	08.75	18.74	40.94	46.58	04.75
Urdu	27.73	35.47	43.46	44.73	35.52
Average	18.78	45.99	42.83	44.75	32.30

BL: Baseline, **IK**: IIT Kharagpur
JU: Jadavpur University, Calcutta
IH1: Karthik et al., IIIT Hyderabad
IH2: Praveen and Ravi Kiran, IIIT Hyderabad

Table 5: Comparison of NER systems which participated in the NERSSEAL-08 shared task against a baseline that uses maximum entropy based name finder tuned for English but trained on data from five South Asian languages

scribes experiments on the CoNLL 2003 shared task data for English, which shows that the significantly higher results for English are mainly due to the fact that the CoNLL 2003 data is already POS tagged and chunked with high accuracy. Goyal was also able to show that capitalization is a major clue for English, either directly or indirectly (e.g., for accurate POS tagging and chunking). He also indicated that the characteristics of the Hindi annotated corpus were partly responsible for the low results on Hindi.

Nayan et al. mainly describe how an NER system can benefit from approximate string matching based on phonetic edit distance, both for a single language (to account for spelling variations) and for cross-lingual NER. Shishtla et al. ('Experiments in Telugu NER') experimented only on Telugu and used the CoNLL shared task tagset. Using a CRF based approach, they were able to achieve an F-measure of 44.91%. Ekbal and Bandyopadhyay describe a method based on Support Vector Machines (SVMs) for Bengali NER. On a news corpus and with sixteen NE classes, they were able to achieve an F-measure of 91.8%. Vijayakrishna and Sobha describe a CRF based system for Tamil using 106 NE classes. Their system is a multi-level system which gave an overall F-measure of 80.44%. They also mention that their system achieved this level of performance on a domain focused corpus. Shishtla et al. ('Character n-gram Based Approach') used a character n-gram based method to identify NEs. They experimented on Hindi as well as English and achieved F-measure values up to 45.48% for Hindi and 68.46% for English.

Apart from the paper presentations, the workshop will also have two invited talks. The first one is titled "Named Entity Recognition: Different Approaches" by Sobha L. and the second one is "Multilingual Named Entity Recognition" by Sivaji Bandyopadhyay.

11 Shared Task Results

Five teams participated in the shared task. However, only four submitted papers for the workshop. All the teams tried to combine machine learning with some language specific heuristics, at least for one of the languages. The results obtained by the four teams are summarized in Table-5, which shows only the F-measure for lexical match. It can be seen from the table that all the teams were able to get significantly better results than the baseline. Overall, the performance of the IIT Kharagpur team was the best, followed by the two teams from IIIT Hyderabad.

Even though all the teams obtained results much better than the baseline, it is still quite evident that the state of the art for NER for SSEA languages leaves much to be desired. At around 46% maximum F-measure on lexical matching, the results mean that the NER systems built so far for SSEA languages are not quite practically useful. But, after this workshop, we at least know where we stand and how far we still have to go.

However, it may be noted that the conditions for

the shared task were very stringent compared to the previous shared tasks on NER, e.g. neither the corpus was tagged with parts of speech or chunks, nor were good POS taggers or chunkers available for the languages involved. This indicates that with progress in building better resources and basic tools for these languages, the accuracy of NER systems should also increase. Already, some very high accuracies are being reported under less stringent conditions, e.g. for domain focused NER.

12 Conclusions

We started by discussing the problem of NER for South and South East Asian languages and the motivations for organizing a workshop on this topic. We also described a named entity annotated corpus for five South Asian languages used for this workshop. We presented some statistics about the corpus and also the problems we encountered in getting the corpus annotated by teams located in distant places. We also presented a new named entity tagset that was developed for annotation of this corpus. Then we presented the results for our experiments on a reasonable baseline. Finally we gave an overview of the papers selected for the NERSSEAL-08 workshop and discussed the systems described in these papers and the results obtained, including those for the shared task which was one of the two tracks in the workshop.

References

Sivaji Bandyopadhyay. 2008. Invited talk: Multilingual named entity recognition. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 15–17, Hyderabad, India, January.

A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 152–160.

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

Bidyut Baran Chaudhuri and Suvankar Bhattacharya. 2008. An experiment on automatic detection of named entities in bangla. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 51–58, Hyderabad, India, January.

H.L. Chieu and H.T. Ng. 2003. Named entity recognition with a maximum entropy approach. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 160–163.

S. Cucerzan and D. Yarowsky. 1999. Language independent named entity recognition combining morphological and contextual evidence. *Proceedings of the Joint SIGDAT Conference on EMNLP and VLC 1999*, pages 90–99.

Asif Ekbal and Sivaji Bandyopadhyay. 2008a. Bengali named entity recognition using support vector machine. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 85–92, Hyderabad, India, January. Association for Computational Linguistics.

Asif Ekbal and Sivaji Bandyopadhyay. 2008b. Development of bengali named entity tagged corpus and its use in ner systems. In *Proceedings of the Sixth Workshop on Asian Language Resources*, Hyderabad, India, January.

Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka, and Sivaji Bandyopadhyay. 2008. Language independent named entity recognition in indian languages. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 33–40, Hyderabad, India, January.

Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth Shishtla, and Dipti Misra Sharma. 2008. Aggregating machine learning and rule based heuristics for named entity recognition. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 25–32, Hyderabad, India, January.

Amit Goyal. 2008. Named entity recognition for south asian languages. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 63–70, Hyderabad, India, January.

Sarmad Hussain. 2008. Resources for urdu language processing. In *Proceedings of the Sixth Workshop on Asian Language Resources*, Hyderabad, India, January.

Hideki Isozaki. 2001. Japanese named entity recognition based on a simple rule generator and decision tree learning. In *Meeting of the Association for Computational Linguistics*, pages 306–313.

J.H. Kim and PC Woodland. 2000. A rule-based named entity recognition system for speech input. *Proc. of ICSLP*, pages 521–524.

D. Klein, J. Smarr, H. Nguyen, and C.D. Manning. 2003. Named entity recognition with character-level models. *Proceedings of CoNLL*, 3.

Sobha L. 2008. Invited talk: Named entity recognition: Different approaches. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 13–14, Hyderabad, India, January.

A. McCallum and W. Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *Seventh Conference on Natural Language Learning (CoNLL)*.

A. Mikheev, M. Moens, and C. Grover. 1999. Named Entity recognition without gazetteers. *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 1–8.

Animesh Nayan, B. Ravi Kiran Rao, Pawandeep Singh, Sudip Sanyal, and Ratna Sanyal. 2008. Named entity recognition for indian languages. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 71–78, Hyderabad, India, January. Association for Computational Linguistics.

Praveen P and Ravi Kiran V. 2008. Hybrid named entity recognition system for south and south east asian languages. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 59–62, Hyderabad, India, January.

Vijayakrishna R and Sobha L. 2008. Domain focused named entity recognizer for tamil using conditional random fields. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 93–100, Hyderabad, India, January. Association for Computational Linguistics.

Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dandapat, Sudeshna Sarkar, and Pabitra Mitra. 2008. A hybrid named entity recognition system for south and south east asian languages. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 17–24, Hyderabad, India, January.

E.F.T.K. Sang and F. De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Development*, 922:1341.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.

Manabu Sassano and Takehito Utsuro. 2000. Named entity chunking techniques in supervised learning for japanese named entity recognition. In *Proceedings of the 18th conference on Computational linguistics*, pages 705–711, Morristown, NJ, USA. Association for Computational Linguistics.

B. Settles. 2004. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. *log*, 1:1.

Praneeth M Shishtla, Karthik Gali, Prasad Pingali, and Vasudeva Varma. 2008a. Experiments in telugu ner: A conditional random field approach. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 79–84, Hyderabad, India, January. Association for Computational Linguistics.

Praneeth M Shishtla, Prasad Pingali, and Vasudeva Varma. 2008b. A character n-gram based approach for improved recall in indian language ner. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 101–108, Hyderabad, India, January. Association for Computational Linguistics.

P Srikanth and Kavi Narayana Murthy. 2008. Named entity recognition for telugu. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 41–50, Hyderabad, India, January.

K. Takeuchi and N. Collier. 2002. Use of support vector machines in extended named entity recognition. In *Proceedings of the sixth Conference on Natural Language Learning (CoNLL-2002)*.

G.D. Zhou and J. Su. 2001. Named entity recognition using an HMM-based chunk tagger. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480.