Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Brief Survey about Work Done in Named Entity Recognition for Indian Languages

Varinder Kaur, Amandeep Kaur Randhawa
Department of Computer Science, BGIET,
Sangrur, Punjab, India

Abstract - This paper reports about the development of Named Entity Recognition for some Indian languages like Punjabi, Hindi, Bengali, Kannada, Manipuri, Assamese, Oriya, Malayalam etc. Firstly, we give the brief overview of various approaches used in NER followed by the performance metrics which helpful in evaluate the NER systems. Next, we present the work done in each language with the help of its tagset, features, experiments and results.

Keywords: Named Entity Recognition, Maximum Entropy model , Hidden Markov Models, Conditional Random Fields, Named entities.

I. INTRODUCTION

Named Entity Recognition (NER) has important applications in almost all Natural Language Processing (NLP) application areas that include Information Retrieval, Information Extraction, Machine Translation, Question Answering and Automatic Summarization etc. The objective of NER is to identify and classify every word/term in a document intosome predefined categories like person name, location name, organization name, miscellaneous name (date, time, number, percentage, monetary expressions etc.) and "none-of-the above".

APPROACHES TO NER

There are several approaches to NER. They can be categorized into two broad categories:

A. Rule based (Linguistic) approaches

Rule based approaches rely on hand-crafted rules, which focuses on extracting names. Rule-based approaches may contain gazetteer lists, list of triggered words etc.

B. Machine learning (Statistical) approaches

Machine learning approaches rely on statistical models to make predictions about NEs in given text. Large amounts of annotated training data are required .There are three main machine learning approaches: Unsupervised, Supervised and Semi-supervised. In unsupervised learning approaches, model work without any feedback. Supervised learning approaches build predictive models based on the labelled data In semi-supervised approach, a model is trained on an initial set of labeled data, then, predictions are made on a separate set of unlabeled data. Commonly used supervised statistical approaches are: Hidden Markov Models (HMM), Conditional Random Fields (CRF), Maximum Entropy model (ME) and Support Vector Machine (SVM).

1) *Hidden Markov Models*: Hidden Markov model is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved states. Here, the state is not directly visible, but output, which depends on the state, is visible. HMMs have difficulty modeling overlapping, non- independent features. Conditional Random Fields (CRF) solve this problem.
2) *Conditional Random Fields* :These are undirected graphical models used to calculate the conditional probability of values on designated output nodes, given values assigned to other designated input nodes. They are conditionally trained probabilistic finite state automata. Since they are conditionally trained, these CRFs can easily incorporate a large number of arbitrary, non-independent features, while still having efficient procedures for non-greedy finite-state inference and training.
3) *Maximum Entropy model* : Maximum Entropy models are conditional probabilistic sequence models. The maximum entropy framework estimates probabilities based on the principle of making as few assumptions as possible, other than the constraints imposed. Such constraints are derived from training data, expressing some relationship between the features and outcomes. The probability distribution that satisfies this property, is the one with the highest entropy.
4) *Support Vector Machines* : SVM solves two-class pattern recognition problem. It takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. It gives best results when the data set is small, and with extended algorithms it can be used in multi-class problems.

PERFORMANCE METRICS

The following metrics are used to evaluate an NER system:

A. Precision: Precision is the fraction of the documents retrieved that are relevant to the user's information need. It is given by

Precision (P) = Correct answers / Answers produced

B. Recall: Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

Recall (R) = Correct answers / Total possible correct answers

C. F-measure: F-measure is the weighted harmonic mean value of precision and recall.

II. RELATED WORK IN VARIOUS INDIAN LANGUAGES

A. Punjabi

1) Named entity recognition for Punjabi language using HMM and MEMM:

This paper introduces Named Entity Recognition for Punjabi language using two Machine learning models Hidden Markov Model (HMM) and Maximum Entropy Markov Model (MEMM). Here the developed system identifies four Named Entity classes Person name, Location name, Organization name, and Date\Time. A number of language dependent and independent features are used. Results are measured using F-measure metric for each different entity class.

Features

Word suffix feature: The 4 character long suffix feature is used to recognize the Named Entities. Punjabi is an inflectional language so the suffix feature is more important to use for recognition of entities.

Word prefix feature: Punjabi language has two types of affixes: Prefix and Suffix. So the 4 character long word prefix feature is also used for identification of Named Entities based on their prefixes.

Previous map feature: Generates features indicating the outcome associated with a pervious occurrence of the word in the document.

Surrounding word feature: The previous and next words of a particular word w are used as features.

Bigram feature: This will generate the features consists of bigrams of words that precede the Named Entity class.

Gazetteers list: Various gazetteer lists has been developed in Punjabi language from various sources like Wikipedia, either manually or using transliteration tools, to improve the performance of system.

Experiments and results

They have experimented on two machine learning techniques HMM and MEMM. The training corpus of HMM models contains 42k words and MEMM training corpus contains 61k words from various news articles. Test data is also taken from Punjabi newspapers. Both systems are tested on test data which contains up to 20k words. The performance of both NER systems for different entity classes is measured using F-measure.

Table 1 Results computed using f-measure metric for four different entities

NE Class	Hidden markov model	MaxEnt model
Person	.8346	.8793
Location	.8220	.8332
Organization	.8613	.8992
DTime	.9154	.9374

Both HMM and MEMM system make use of different type of contextual and orthographic word level features. Both models shows good results for date/time, person name and organization names.

2) Named Entity Recognition in Punjabi Using hidden Markov Model

They have developed a NER based system for Punjabi. They took Punjabi text from the web which are stories written in Punjabi and then performed the task of Corpus Development

Table 2 Various named entity tags NE TAG: Named Entity Tags , PER: Name of Person, LOC-Location, CO-Country and QTY-Quantity

NE Tag	Example
PER	ਚਰਨਕੋਰੇ, ਲਾਜੋ
LOC	ਛਾਰਪੁਰ, ਗੁਰਦੁਆਰੇ
CO	ਮਰੀਕੇ, ਕਨੇਡੇ
TIME	ਚਾਰਿਦਨ, ਸਵੇਰੇ, ਤਕਾਲ□
DRUG	ਫੀਮ, ਭੱ□ਕੀ, ਜਰਦਾ
DRYFRUIT	ਿਮਸਰੀ, ਕਾਜੂ, ਬਦਾਮ, ਖਰੇਟ
GARAM_MASALA	ਲੋਗਾਂ, ਲੈਚੀਆਂ
QTY	ਦੂਣੀ
FOOD	ਦੱ□ਧਾਂ, ਘੋਏ, ਦੱ□ਧ

The tags chosen for the annotation depends on the content of the document under consideration and the choice of tags may vary from one person to another.

RESULTS:

Table 3 Results of ner in punjabi using hmm

Total training sentences	631		
Total testing sentences	10		
No. of testing sentences Giving wrong result	1		
	No. of tags in training sentences	No. of tags in testing sentences	No. of correct tags identified
PER	13	6	6
LOC	4	0	0
OTHER	3846	57	50
CO	2	1	0
TIME	7	4	4
DRUG	4	0	0
DRYFRUIT	5	1	1
GARAM_MASALA	2	0	0
QTY	1	0	0
FOOD	3	0	0
TOTAL	3887	69	61
Accuracy=(61/69)*100=88.4%	Recall=88.4%	Precision=88.4%	f-score =88.4%

3) Named Entity Recognition for Punjabi Language Text Summarization

This paper explains the Named Entity Recognition System for Punjabi language text summarization. A Condition based approach has been used for developing NER system for Punjabi language. Various rules have been developed like prefix rule, suffix rule, proper name rule, middle name rule and last name rule. For implementing NER, various resources in Punjabi, have been developed like a list of prefix names, a list of suffix names, a list of proper names, middle names and last names. The Precision, Recall and F-Score for condition based NER approach are 89.32%, 83.4% and 86.25% respectively.

- *Punjabi Names Prefix List :* The Prefix list contains various prefixes of names for checking whether next word is a proper name or not. In the Punjabi corpus of 11.29 million words, the frequency count of these prefix words is 17127, which covers 0.15% of the corpus.
- *Punjabi Names Suffix List* :The Suffix list contains various suffixes of names for checking whether the current word is a proper name or not. In the Punjabi corpus of 11.29 million words, the frequency count of suffix words is 225306, which covers 1.99% of the corpus.
- *Punjabi Middle Names List :*The Punjabi middle name list contains various middle names of persons for checking whether that word is proper name or not. After manually analyzing unique words of Punjabi corpus, we have identified mainly 8 middle names and a list is developed by creating a frequency list from corpus
- *Punjabi Last Names List :*The Punjabi last name list contains various last names of persons for checking whether that word is proper name or not, like ਖੁਰਾਨਾ khurānā. After manually analyzing unique words of Punjabi corpus, we have identified 310 last names and a list is developed by creating a frequency list from corpus.
- *Punjabi Proper Names List :*Proper names are very much important in deciding a sentence's importance. Those sentences containing proper names are important.

Implementation and results

It is producing Precision=89.32%, Recall=83.4% and F-score=86.25%. An in depth error analysis of condition based system has been done over 50 news documents and it is giving 13.75% errors. Prefix rule is producing no errors, Suffix rule is producing 1% errors. Middle name rule is producing 0.25% errors. Last name rule is producing 10% errors. Proper names rule is producing 0.25% errors, Rest 2.25% errors are due to those proper names who do not lie under any of rules.

B. Hindi

1) Hindi Named entity Recognition by aggregating rule based heuristics and hidden markov model

Tagset used

Some NE Tags and their meanings(PER: Name of Person, CO-Country, ORG-Organization, VEH-vehicle and QTY-Quantity)

Table 6 Various named entity tags

NE TAG	EXAMPLE
PER	Deepti, Sudha, Rohit
CITY	Jaipur, Mumbai, Kolkata
CO	India, China, Pakistan
STATE	Rajasthan, Maharashtra
SPORT	Hockey, Badminton
ORG	TCS, Infosys, Accenture
RIVER	Ganga, Krishna, kaveri
DATE	27-04-2012, 31/01/1989
TIME	10:10

Performance Metrics

The output of a NER system may be termed as "response" and the interpretation of human as the "answer key". We consider the following terms:

1. Correct-If the response is same as the answer key.
2. Incorrect-If the response is not same as the answer key.
3. Missing-If answer key is found to be tagged but response is not tagged.
4. Spurious-If response is found to be tagged but answer key is not tagged.

Hence, we define Precision, Recall and F-Measure as follows:

Precision (P): Correct / (Correct + Incorrect + Missing)

Recall (R): Correct/ (Correct + Incorrect + Spurious)

F-Measure: (2*P*R)/(P+R)

Resuts and discussions

Table 7 Results of rule based heriustics and shallow parsing techniques

Named entities	Total Named Entities	Named Entities identified	Accuracy (%)
LOC	247	125	50.60
PER	56	29	51.79
QTY	79	40	50.63
TIME	67	34	50.75
ORG	135	68	50.37
SPORT	45	23	51.11
RIVER	11	6	54.54
VEH	25	0	0
MONTH	22	0	0
TOTAL=	687	325	47.5

Table 8 Results of hidden makov model

Named entities	Total Named Entities undetected	Named Entities identified	Accuracy (%)
LOC	122	107	87.70
PER	27	24	88.89
QTY	39	34	87.18
TIME	33	29	87.88
ORG	67	59	88.06
SPORT	22	20	90.90
RIVER	5	5	100
VEH	25	25	100
MONTH	22	22	100
TOTAL=	362	325	89.78

Table 9 Results of combination of approaches or hybrid approach

Total Named Entities	Named Entities identified	Accuracy (%)
687	650	94.61

Here it obtained accuracy of about 94.61% by aggregating rule based heuristics and HMM, as shown in Table 9. Table 7 depicts that if we applied only Rule Based Heuristics, then it performed very poorly, and the accuracy obtained by this approach was 47.5%. Similarly, Table 8 depicts that if we applied only HMM, then its performance was average, and the accuracy obtained by this approach was 89.78% .This shows that if we apply combined approach, then it gives very good results in Named Entity Recognition.

2) Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction (Short Paper)

This paper describes application of Conditional Random Fields (CRFs) with feature induction to a Hindi named entity recognition task. With only five days development time and little knowledge of this language, they automatically discover relevant features by providing a large array of lexical tests and using feature induction to automatically construct the features that most increase conditional likelihood.

Their training set for the Hindi task is composed of 601 BBC and 27 EMI documents after we remove the ones with no tag files or containing the No-Annotation tags. It contains about 340k words, 4540 Person, 7342 Location and 3181 Organization entities. In the 25 documents in the NIST test set, there are 10k words and the entity counts are 152, 232 and 92 respectively. To train the CRF, they experimented with various options, such as first-order versus second-order models, using feature induction or not and using lexicons or not. In an effort to reduce overfitting, they also tried different Gaussian priors and early-stopping. Finally,a first-order CRF is trained with the whole training set, inducing 500 or fewer features (down to a gain threshold of 5.0) every 10 iterations.

Results and discussions

Table 10 Experiments and results

% training data	10	50	100	100	100	100	100	100	100	100
Markov order	1	1	1	2	1	1	1	1	1	1
feature induction	Y	Y	Y	Y	N	Y	Y	Y	Y	Y
using lexicons	Y	Y	Y	Y	Y	N	Y	Y	Y	Y
early-stopping	N	N	N	N	N	N	Y	N	N	N
Gaussian prior	100.0	100.0	100.0	100.0	100.0	100.0	100.0	10.0	1.0	0.1
validation set	65.82	77.13	81.16	79.51	-	81.31	82.55	80.73	80.66	78.80
test set	56.68	66.46	71.50	-	62.94	70.77	68.80	70.62	69.27	63.16

The experiment results for validation and test sets are summarized in table 10. The first-order model performs slightly better than the second-order model on the validation set, and the testing performance is significantly better when using feature induction. Using lexicons or not does not make much difference, and tight Gaussian priors do not improve the performance. While an early stopping point of 240 iterations of L-BFGS obtains the highest average F1 score for the 10-fold cross validation experiments, early-stopping actually hurts the performance on the test set. Although performance is similar to an HMM on a validation set, their model does not perform as well on the test set.

3) A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition

This paper describes the effort in developing a Named Entity Recognition (NER) system for Hindi using Maximum Entropy (Max-Ent) approach. They developed a NER annotated corpora for the purpose. We have tried to identify the most relevant features for Hindi NER task to enable us to develop an efficient NER from the limited corpora developed. Apart from the orthographic and collocation features, we have experimented on the efficiency of using gazetteer lists as features. We also worked on semi-automatic induction of context patterns and experimented with using these as features of the MaxEnt method. We have evaluated the performance of the system against a blind test set having 4 classes - Person,Organization, Location and Date and system achieved a f-value of 81.52%.

Training data is composed of about 243K words which is collected from the popular daily Hindi newspaper "Dainik Jagaran". This corpus has been manually annotated and has about 16,482 NEs. In this development we have considered 4 types of NEs, these are Person(P), Location(L), Organization(O) and Date(D). To recognize entity boundaries each name class N is subdivided into 4 sub-classes, i.e., N Begin, N Continue, N End, and N Unique. Hence, there are a total of 17 classes including 1 class for not-name. The corpus contains 6, 298 Person, 4, 696 Location, 3, 652 Organization and 1, 845 Date entities.

Features

The features which we have identified for Hindi Named Entity Recognition are:

Static Word Feature: The previous and next words of a particular word are used as features. During their experiment, different combinations of previous 4 to next 4 words are used.

Context Lists: Context words are defined as the frequent words present in a word window for a particular class

Dynamic NE tag: Named Entity tags of the previous words (ti−m...ti−1) are used as features.

First Word: If the token is the first word of a sentence, then this feature is set to 1. Otherwise, it is set to 0.

Contains Digit: If a token 'w' contains digit(s) then the feature ContainsDigit is set to 1. This feature is helpful for identifying company product names.

Numerical Word: For a token 'w' if the word is a numerical word i.e. a word denoting a number then the feature NumWord is set to 1.

Word Suffix: Word suffix information is helpful to identify the named NEs. Two types of suffix features have been used. Firstly a fixed length word suffix of the current and surrounding words are used as features. Secondly we compiled lists of common suffixes of person and place names in Hindi.

Word Prefix: Prefix information of a word may be also helpful in identifying whether it is a NE. A fixed length word prefix of current and surrounding words are treated as a features.

Parts-of-Speech (POS) Information: The POS of the current word and the surrounding words may be useful feature for NER.

Enhancement using Gazetteer Feature

Lists of names of various types are helpful in name identification. We have compiled some specialized name lists from different web sources. But the names in these lists are in English, not in Hindi. So we have transliterated these English name lists to make them useful for our Hindi NER task.

Context Pattern based Features

Context patterns are helpful for identifying NEs. As manual identification of context patterns takes much manual labour and linguistic knowledge, we have developed a module for semi-automatically learning of context pattern

Results and discussions

Table 11 F-values for different features

Feature	Class	F-value
f1 = Word, NE Tag	PER LOC	63.33
	ORG	69.56
	DAT	58.58
	TOTAL	91.76
		69.64
f2 = Word, NE Tag Suffix (≤ 2)	PER LOC	69.75
	ORG	75.8
	DAT	59.31
	TOTAL	89.09
		73.42
f3 = Word, NE Tag Suffix (≤ 2), Prefix	PER LOC	70.61
	ORG	71
	DAT	59.31
	TOTAL	89.09
		72.5
F4 = Word, NE Tag, Digit, Suffix (≤ 2)	PER LOC	70.61
	ORG	75.8
	DAT	60.54
	TOTAL	93.8
		74.26
f5 = Word, NE Tag, POS	PER LOC	64.25
	ORG	71
	DAT	60.54
	TOTAL	89.09
		70.39
F4 = Word, NE Tag, Digit, Suffix (≤ 2), *NomPSP*	PER LOC	72.26
	ORG	78.6
	DAT	51.36
	TOTAL	92.82
		75.6

Table 12 F-values for different features with gazetteers and context patterns

Feature	F-value				
	Class	No gaz or pat	With gaz	With patt	With gaz and patt
F2	PER	69.75	74.2	75.61	76.03
	LOC	75.8	82.02	79.94	82.02
	ORG	59.3	72.61	73.4	74.63
	DAT	89.09	94.29	95.32	95.32
	TOTAL	73.4	79.8	80.06	80.69

F6	PER	72.26	76.03	75.61	78.41
	LOC	78.6	82.02	80.49	83.26
	ORG	51.36	72.61	74.1	75.43
	DAT	92.8	94.28	95.87	96.5
	TOTAL	75.6	80.24	80.37	81.52

It is shown that MaxEnt based NER system is able to achieve a f-value of 81.52%, using a hybrid set of features including traditional NER features augmented with gazetteer lists and extracted context patterns. The system outperforms the existing NER systems in Hindi.

4) Named Entity Recognition in Hindi using Maximum entropy and Transliteration

In this paper, they described a Maximum Entropy based NER system for Hindi. They have explored different features applicable for the Hindi NER task and incorporated some gazetteer lists in the system to increase the performance of the system. These lists are collected from the web and are in English. To make these English lists useful in the Hindi NER task, they have proposed a two-phase transliteration methodology. A considerable amount of performance improvement is observed after using the transliteration based gazetteer lists in the system. The proposed transliteration based gazetteer preparation methodology is also applicable for other languages. Apart from Hindi, they also applied this approach in Bengali NER task and achieved performance improvement.

Feature Description

The features that we have identified for the Hindi NER task are:

Surrounding Words

As the surrounding words are very important to recognize a NE, previous and next words of a particular word are used as features.

Binary Word Feature

The multi-valued feature can be modified as a set of binary feature to reduce the feature space. Class specific lists are compiled taking the frequent words present in a particular position.

Context Lists

The idea of binary word feature is used to define the class context features. Context words are defined as the frequent words present in a word window for a particular class. In their experiment they have listed all the frequent words present anywhere in wi-3...wi+3 window for a particular class. Then this list is manually edited to prepare the context word list for a class.

Named Entity Tags of Previous Words

Named entity (NE) tags of the previous words (ti-m...ti-1) are used as feature. This feature is dynamic. The value of the feature for wi is available after obtaining the NE tag of wi-1.

First Word

If the word is the first word of a sentence, then this feature is set to 1. Otherwise, it is set to 0.

Containing Digit

If a word contains digit(s) then the binary feature
ContainsDigit is set to 1.

Made up of 4 Digits

For a word *w* if all the characters are digits and having only 4 digits in *w*, then the feature fourDigit is set to 1. This feature is helpful for identifying year.

Numerical Word

If a word is a numerical word, i.e. it is a word denoting a then the feature NumWord is set to 1.

Word Suffix

Suffix information is useful to identify the named entities. *Word Prefix*

Prefix information of a word is also useful. A fixed length word prefix of current and surrounding words can be treated as feature.

Parts-of-Speech (POS) Information

The POS of the current word and the surrounding words are important to recognize names. For this task ,they used the POS tagger developed at IIT Kharagpur, India. The tagset of the tagger contains 28 tags. Firstly we have used the POS values of current and surrounding tokens as feature.

Experiments and results

About 80 different experiments are conducted taking several combinations from the mentioned features to identify the best feature set for the NER task. We have evaluated the system using a blind test file of size 25 K words, which is totally different from the training file.

First of all, they used only the current and surrounding words as feature of MaxEnt. We have experimented with several combinations of previous 4 to next 4 words (wi-4...wi+4) to identify the best word-window.

Table 13 Results based on features for each category

Feature	**Per**	**Loc**	**Org**	**Date**	**Total**

wi, wi-1, wi+1	61.36	68.29	52.12	88.9	67.26
wi, wi-1, wi-2, wi+1, wi+2	64.10	67.81	58	92.30	69.09
wi, wi-1, wi-2, wi-3, wi+1,wi+2,wi+3	60.42	67.81	51.48	90.18	66.84
wi, wi-1, wi-2, wi-3, wi-4,wi+1,wi+2, wi+3, wi+4	58.42	64.12	47.97	84.69	61.27
wi, wi-1inList, wi-2inList, wi+1inList, wi+2inList	65.37	70.33	47.37	83.72	66.17

From Table 13 we can observe that word window (wi-2...wi+2) gives the best result. When the window size is increased, the performance degrades. List based binary word features are not effective. In the table, the notation *wi-ninList* is used to indicate binary word features for all classes for wi-n. Manual editing of the lists might help the binary word feature to perform better. Similar experiments are conducted to find the best feature set for the Hindi NER task. The features described earlier are applied separately or in combination to build the MaxEnt based model. In Table 14 we have summarized the results. Only the best values of each feature category are given in the table. This result is considered as the *baseline* in this study.

Table 14 F-values of ne tags for each feature category

Feature	Per	Loc	Org	Date	Total
words, previous NE tags	63.33	69.56	58.58	91.76	69.64
words,tags,prefix(≤4)	66.67	71	58.58	87.8	70.02
words,tags,suffix(≤4)	70	76.92	59.18	88.9	73.5
words, tags, suffix (≤4), prefix(≤4)	70.44	70.33	59.18	90.18	72.64
words, tags, digit information	62.94	69.56	50	91.76	67.63
words, tags, suffix (≤4), digit	70.44	76.92	60.44	93.02	74.51
words, tags, POS (28 tags)	66.67	72.84	60	88.9	71.22
words, tags, POS(coarsegrained)	69.62	80.74	58.7	91.76	75.22
words, tags, POS(coarsegrained), suffix (≤4), digit	72.23	78.1	62.37	93.02	75.67
words, tags, 'nominalPSP', suffix (≤4), digit	72.5	80.74	58.7	93.02	75.89

From the table they observed that some of the features are able to improve the system accuracy separately, but when applied in combination with other features, they cause decreasing of the accuracy.

C. Assamese

1) A survey of NER in Assamese and other Indian languages

Assamese like other Indian languages is agglutinative and suffers from lack of appropriate resources as Named Entity Recognition requires large data sets, gazetteer list, dictionary etc and some useful feature like capitalization as found in English cannot be found in Assamese. Apart from this we also describe some of the issues faced in Assamese while doing Named Entity Recognition.

Key issues in assamese NER

- Ambiguity in Assamese : Ambiguity occurs between common noun and proper noun as most of the names are taken from dictionary. For example, জোন(Jon) indicates moon which is a common noun but may also indicate the name of a person which is a proper noun.

- Agglutinative nature : Assamese language suffers from agglutination and complex words are created by adding additional features to change the meaning of the word. For example, অসম (Assam) is the name of a place which is a location named entity but অসমীয়া (AssamIYA) is produced by adding suffix ীীয়া(IYA) to অসম (Assam) which signifies people residing in Assam which is not a location named entity.
- *Lack of capitalization :*In Assamese there is no capitalization that can help to recognize the proper nouns as found in English.
- *Nested Entities :*When two or more proper nouns are present then it becomes difficult to assign the proper named entity class. For example, In গুৱাহাটী বিশ্ববিদ্যালয় (Gauhati bishabidyaly) গুৱাহাটী (Gauhati) is a location named entity and বিশ্ববিদ্যালয় (bishabidyalay) refers to organization thus creating problem in assigning the proper class.
- *Spelling Variation* :Changes in the spelling of proper names are another problem in Assamese Named Entity Recognition. For example, In শ্রী শ্রীচোন্থ (Shree Shreesanth) there is a confusion whether শ্রী (Shree) in শ্রীচোন্থ (Shreesanth) is a Pre-nominal word or person named entity.

Features in NER

Features commonly used for Named Entity Recognition are:

*Surrounding words :*Various combinations of previous to next words of a word which are the surrounding words can be treated as a feature.

*Context word feature :*Specific list for class can be created for the words occurring quite often at previous and next positions of a word belonging to a particular class. This feature is set to 1 if the surrounding words are found in the class context list.

Digit features :Binary valued digit features can be helpful in detecting the miscellaneous named entities such as:

- ContDigitPeriod: Word contains digits and periods.
- ContDigitComma: Word contains of digits and commas.
- ContDigitSlash: Word contains digits and slash.
- ContDigitHyphen: Word contains digits and hyphen.
- ContDigitPercentage: Word contains digits and percentage.
- ContDigitSpecial: Word contains digits and special symbols.

*Infrequent word :*Infrequent or rare word can be found by calculating the frequencies of the words in the corpus of the training phase and by selecting a cut off frequency to allow only those words as rare that occurs less than the cut off frequency.

Word suffix : The word suffix feature can be defined in two ways. The first way is to use fixed length suffix information of the surrounding words and the other way is to use suffixes of variable length and matching with the predefined lists of valid suffixes of named entities.

Word prefix : The word prefix feature can be defined in two ways. The first way is to use fixed length prefix information of the surrounding words and the other way is to use prefixes of variable length and matching with the predefined lists of valid prefixes of named entities.

*Part of speech information :*The part of speech information of the surrounding words can be used for recognizing named entities.

*Numerical word :*This feature can be set to 1 if a token signifies a number.

*Word length :*This feature can be used to check if the length of a word is less than three or not because it is found that very short words are not named entities.

Observations and discussions

In this section we provide a survey of the research done in Indian languages

Table 15. F-measure achieved in Hindi for different statistical approach.

Approach Used	F-measure (%)
MEMM	65.13
Character based n-gram technique	45.18
Language dependent features	33.12

Table 16 F-measure achieved in Bengali for different statistical approach.

Approach Used	F-measure (%)
MEMM	65.96
Language dependent features	59.39
SVM	91.8

Table 17 F-measure achieved in telugu for different statistical approach

Approach Used	F-measure (%)
MEMM	18.75
CRF	92
Language dependent features	47.49
Character based n-gram technique	48.93

Table 18 F-measure achieved oriya for different statistical approach

Approach Used	**F-measure (%)**
MEMM	44.65
Language independent features	28.71

Table 19 F-measure achieved in assamese for suffix stripping based approach

Approach Used	**F-measure (%)**
Suffix stripping based approach [6]	88

Table 20 F-measure achieved in urdu for different statistical approach

Approach Used	**F-measure (%)**
MEMM [4]	35.47
Language independent features [1]	35.52

They found that rule based approaches with gazetteer list along with some language independent rules combined together with statistical approach may give satisfactory results for Named Entity Recognition in Indian languages because of insufficient data available for training. So if sufficient training data is not available, a hybrid model where combination of rule based, statistical and language independent rules are used will be a better approach to perform NER in Indian languages.

2) Named Entity Recognition In Assamese using CRFs and Rules

This paper discusses work on NER in Assamese using both Conditional Random Fields and a Rule-Based approach .The data used for the training of the model was taken from the reputed Assamese Newspaper 'Asomiya Pratidin' and 'Emille corpus' consisting of 0.15 million word forms. Their work in Assamese NER using CRFs involves the following steps.

– The first phase involves using a standard tool developed at Stanford University, which is a Java based tool of CRFs for NER.

– The second layer involves post-processing work the output of CRFs using heuristics or rules.

In the first phase, the collected data is manually tagged with named entities manually with three different annotators namely person, location and organization. The tagged data is in tab-separated columns e.g., word is in column 0 and its label is in column 1. We consider three classes of named entities, namely, person, location and organization. A word other than a named entity is labeled as NNE. The system makes use of various features that helps predict the named entity classes. Gazetteer lists are also often been used to identify the NEs. The problem that we have encountered in our experiments is that CRFs can find only single word named entities but no multiword named entities. Some post processing work is performed on the output of the CRFs in order to assign labels to multiword named entities.Performing a rule-based analysis on the output of the statistical analysis, shows improvement in the result. The following pre-processing steps are performed for

the rule-based analysis-"

– Find the words that belong to a closed class part of speech in Assamese.

– Prepare look-up table for 5 parts of speech namely verb, pronoun, conjunction, adjective and common noun in Assamese.

– Based on this look-up table the output file of the CRFs is tagged.

Some of the rules that derived are:

- If the previous and the succeeding words are verbs then the current word is most likely to be a person name.
- If any two/three untagged words in sequence are preceded and succeeded by a verb then those untagged words are most likely a person if the second (and third) word falls under title Category
- If there exist a word like (TF:nogor),(TF:jiLa), (TF:chAhar), etc., then the previous word represents a location name Identity.
- If the current word is a number and the next word represents a unit of measurement such as (TF:kiLo) , (TF:gRam) etc then it represent NE measurement.
- If the current word is a digit and the following word is a month name then it represents NE date.

- If the current word is a number and the next word is a month name followed by digit, then it represent NE date.
- If two words in sequence are both verbs, then the previous word is most likely to be a person name.
- If there exists a suffix like (TF:bAri),(TF:pur), (TF:Ali), etc, then the current word is a location named entity. And if the next word can be found in the organization gazetteer list, it is considered to be an organization named entity.
- If there exists a dot between each consecutive letters then it most likely to be an organization named entity.
- If there exist a prenominal word, it is always at the beginning of a name. Now jumping over the next word is a title, we need to search forward until we obtain a value in the title gazetteer list. Now if the next word is not is not in a title list, we need to check if it occurs in the organization gazetteer list. Ifyes, we will put an end marked here and whole word will represent an organization. Otherwise,we put an end marked in the last title word and the whole word will represents a person name.

Results and discussions

We have used 0.15 million word forms for training purpose and the testing was done on 0.1 million words and our system obtained an accuracy of 83% using only Statistical approach, i.e. CRFs and using both the combined approach our system gives an improvement to 93.22% F-measure shown in Table 21

Table 21 Comparison with other languages

Language	**f-measure**
Hindi	94%
Bengali	85%
Telugu	90%
Kannada	87%
Assamese	93.22%

Taking into account different features of the Assamese language, and adopting a combination of strategies gives improved results compared to the individual approaches. There is scope for considering more features, beyond what have been considered in our experiments, to get better results. They have encountered some challenges in our language such as ambiguity exists between a proper noun and the other POS which makes it difficult to identify the Named Entities which can be future work for other researchers.

D. Bengali

1) Named Entity Recognition in Bengali: A Conditional Random Field Approach

This paper reports about the development of a Named Entity Recognition (NER) system for Bengali using the statistical Conditional Random Fields (CRFs). The system makes use of the different contextual information of the words along with the variety of features that are helpful in predicting the various named entity (NE) classes. A portion of the partially NE tagged Bengali news corpus, developed from the archive of a leading Bengali newspaper available in the web, has been used to develop the system. The training set consists of 150K words and has been manually annotated with a NE tagset of seventeen tags.

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 newspaper 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.

Table 22 Ne TAGSET

NE tag	Meaning	Example
PER	Single-word person name	*sachin*/ PER
LOC	Single-word Location name	*jadavpur*/LOC
ORG	Single-word organization name	*infosys*/ ORG
MISC	Single-word	100%/ MISC
B-PER I-PER	Beginning, Internal or	*sachin*/B-PER *ramesh*/I-PER

E-PER	End of a multiword person name	*tendulkar*/E-PER
B-LOC I-LOC E-LOC	Beginning, Internal or End of a multiword location name	*mahatma*/B-LOC *gandhi*/I-LOC *road*/E-LOC
B-ORG I-ORG E-ORG	Beginning, Internal or End of a multiword organization name	*bhaba*/B-ORG *atomic*/I-ORG *research*/I-ORG *center*/E-ORG
B-MISC I-MISC E-MISC	Beginning, Internal or End of a multiword miscellaneous name	*10e*/B-MISC *magh*/ I-MISC *1402*/E-MISC
NNE	Words that are not NEs	*neta*/NNE

Features

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.

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

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

.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).

Named Entity Information: The NE tag of the previous

word is also considered as the feature. This is the only dynamic feature in the experiment.

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

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

symbols (e.g., ContainsDigitAndComma [token consists of digits and comma], ConatainsDigitAndPeriod[token consists of digits and periods]),combination of digits and symbols (e.g., Contains- DigitAndSlash [token consists of digit and slash],ContainsDigitAndHyphen [token consists of digits

and hyphen], ContainsDigitAndPercentage [token consists of digits and percentages]).

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

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

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

(iii) Middle name (1,491 entries): These words generally

appear inside the person names (e.g., *chandra*, *nath* etc.). The feature `MiddleName' is set to 1 for the current, previous and the next words.

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

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

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

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

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

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

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

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

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

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

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

Experimental Results

Table 23 Feature set with f-score

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

NER system is developed using CRF with the help of a partially NE tagged Bengali news corpus. Experimental results with the 10-fold cross validation test have shown reasonably good Recall, Precision and F-Score values. It has been shown that the contextual window [-2, +2], pre_x and suf_x of length upto three, _rst word of the sentence, POS information of the window [-1, +1], current word, NE information of the previous word, different digit features and the various gazetteer lists are the best-suited features for the Bengali NER. Analyzing the performance using other methods like MaxEnt and Support Vector Machines (SVMs) will be other interesting experiments.

2) Named Entity Recognition in Bengali: A Multi-engine Approach Named Entity Recognition in Bengali: A Multi-Engine Approach

This paper reports about a multi-engine approach for the development of a Named Entity Recognition (NER) system in Bengali by combining the classifiers such as Maximum Entropy (ME), Conditional Random Field (CRF) and Support Vector Machine (SVM) with the help of weighted voting techniques. The training set consists of approximately 272K wordforms, out of which 150K wordforms have been manually annotated with the four major named entity (NE) tags, namely Person name, Location name, Organization name and Miscellaneous name.. The individual classifiers make use of the different contextual information of the words along with the variety of features that are helpful to predict the various NE classes. Lexical context patterns, generated from an unlabeled corpus of 3 million wordforms in a semi-automatic way, have been used as the features of the classifiers in order to improve their performance. In addition, we propose a number of techniques to post-process the output of each classifier in order to reduce the errors and to improve the performance further. Finally, we use three weighted voting techniques to combine the individual models

Table 24 Named entity tagset for indian languages(ijcnlp-08 nersseal shared task tagset)

NE Tag	Meaning	Example
NEP	Person name	sachIna/NEP, sachIna rameshatenDUlkara / NEP
NEL	Location name	kolkAtA/NEL, mahatmA gAndhi roDa / NEL
NEO	Organization name	adabpUra bishVbidyAlYa/NEO, bhAbA eytOmika risArcha sentAra / NEO
NED	Designation	cheYArmAn/NED, sA.msada/NED
NEA	Abbreviation	bi e/NEA, ci em di a/NEA, bi je pi/NEA, Ai bi em/ NEA
NEB	Brand	fYAntA/NEB
NETP	Title-person	shrImAna/NED, shrI/NED, shrImati/NED
NETO	Title-object	AmericAn biUti/NETO
NEN	Number	10/NEN, dasha/NEN
NEM	Measure	tina dina/NEM, p.NAch keji/NEM
NETE	Terms	hidena markbha madela/NETE, kemikYAla riYYAkchYAna/NETE
NETI	Time	10 i mAgha 1402 / NETI, 10 ema/NETI shared task

Language Independent Named Entity Features

Following are the descriptions of the set of language independent features that have been applied to the NER task:

- *Context words*: Preceding and following words of a particular word can be used as the features. This is based on the observation that the surrounding words are very effective in the identification of NEs.
- *Word suffix*: Word suffix information is helpful to identify NEs. This is based on the observation that the NEs share some common suffix strings.
- *Word prefix*: Word prefixes are also helpful and based on the observation that NEs share some common prefix strings.Fixed length word prefixes are basically the character strings of fixed lengths that are stripped from the beginning positions of the various wordforms.
- *Named Entity Information*: The NE tag(s) of the previous word(s) carry very effective information in determining the NE tag of the current word. This is the only dynamic feature in the experiment.
- *First word:* This feature is used to check whether the current token is the first word of the sentence or not. The first word of the sentence is most likely a NE.
- *Position of the word*: Position of the word in a sentence is a good indicator of NEs. Generally, verbs occur at the last position of the sentence.
- *Length of the word*: The training corpus has been analyzed to prepare a list containing each wordform along with its 'type' (whether NE or not) and length.
- *Infrequent word:* The frequencies of the words in the training corpus have been calculated.
- *Digit features*: Several digit features have been considered depending upon the presence and/or the number of digit(s) in a token (e.g., ContainsDigit, FourDigit, TwoDigit), combination of digits and punctuation symbols (e.g., ContainsDigitAndComma, Contains-DigitAndPeriod), combination of digits and symbols (e.g.,ContainsDigitAndSlash, ContainsDigitAndHyphen, ContainsDigitAndPercentage).
- *Part of Speech (POS) Information*: We have used a CRF-based POS tagger (Ekbal et al. 2007a) that was originally developed with a POS tagset of 26 POS tags, defined for the Indian languages. The SVM based NER systems make use of the POS information extracted from this fine-grained POS tagger.

Language Dependent Named Entity Features

Language dependent features have been identified based on the earlier experiments (Ekbal and Bandyopadhyay 2007a), (Ekbal and Bandyopadhyay 2007b) on NER. Additional NE
features have been identified from the Bengali news corpus (Ekbal and Bandyopadhyay 2008b). Various gazetteers used in the experiment are presented. These gazetteer lists have been used as the features in each of the classifiers. A number of these gazetteers have been also used to post-process the outputs of the classifiers to improve their performance further. Some of the gazetteers are briefly described as below:

- *NE Suffix list* (variable length suffixes): Variable length suffixes of a word are matched with the predefined lists of useful suffixes that are helpful to detect person (e.g., -bAbU[- babu], -dA[-da], -di[-di] etc.) and location (e.g., lYAnDa[land], -pUra[pur], -liYA[-liya] etc.) names.
- *Organization suffix word list:* This list contains the words that are helpful to identify organization names (e.g., kO.m[Co.], limiteDa[Limited] etc.). These are also the part of organization names.
- *Person prefix word list*: This is useful for detecting person names (e.g., shrImAna[Mr.], shrI[Mr.], shrImati[Mrs.] etc.). Person name generally appears after these clue words.

- *Common location word list*: This list contains the words (e.g., saranI[Sarani], rOda[Road], lena[Lane] etc.) that are part of the multiword location names and usually appear at their end.
- *Action verb list*: A set of action verbs like balena[says], balalena[told], ballO[says],sUnllO[heared], h.AsalO[smiled] etc. often determines the presence of person names. Person names generally appear before the action verbs.
- *Designation words*: A list of common designation words (e.g., netA[leader], sA.msada[MP], khelOYAra[player] etc.) has been prepared manually. This helps to identify the position of person names.

Table 25 Overall evaulation resultson the development set for the baseline models

Model	Recall (in %)	Precision (in %)	F-Score (in %)
ME	73.57	73.07	73.32
CRF	75.97	75.45	75.71
SVM-F	77.14	75.48	76.30
SVM-B	77.09	75.14	76.10

Use of context patterns as features

High ranked patterns of the Accept Pattern set can be used as thefeatures in each of the classifiers. Words in the left and/or the right contexts of person, location and organization names carry effective information that could be helpful for their

identification. These words are used as the trigger words. A particular trigger word may appear in more than one pattern type.

A feature '*ContextInformation'* is defined as below

by observing the three preceding and following words of the current word:

- If the window W[−3,+3] (three words spanning to left and right) of the current word contains any trigger word of Person name then the feature value is set to 1.
- If the window W[−3,+3] contains any trigger word of Location name then the feature value is set to 2.
- If the window W[−3,+3] contains any trigger word of Organization name then the feature value is set to 3.
- If the window W[−3,+3] contains any trigger word that appears in more than one NE type pattern then feature value is set to 4.
- Otherwise, the value of the feature is set to 0.

Table 26 Results on the development set including context patterns

Model	**Recall (in %)**	**Precision (in %)**	**F-Score (in %)**
ME	78.59	77.58	78.08
CRF	82.07	83.75	82.90
SVM-F	84.56	82.60	83.57
SVM-B	84.42	82.58	83.49

Comparison with other Systems

Some of the existing Bengali NER systems, i.e., Ekbal et al. (2007b), Ekbal et al.(2007a), Ekbal et al. (2008) and Ekbal and Bandyopadhyay (2008a) have been trained

Table 27 Results of the voted system for the test set

Voting Scheme	**Recall**	**Precision**	**F-Score**
Majority	93.21	89.75	91.45
Total F-Score	93.92	90.11	91.98
Tag F-Score	93.98	90.63	92.28

Table 28 Results of the individual ne tag in the voted

NE tag	**Recall**	**Precision**	**F-Score**
Person name	96.12	93.26	94.67
Location name	89.03	87.62	88.32
Organization name	88.12	85.97	87.03
Miscellaneous name	99.15	98.89	99.02

and tested with the same datasets. Evaluation results are presented in Table 24. Results show the effectiveness of the proposed multi-engine NER system that outperforms the other existing Bengali NER systems based on HMM, CRF and SVM by the impressive margins of 19%, 12.13% and 11.99% F-Scores, respectively. Thus, it can be decided that purely statistical approaches cannot yield very good performance always. Comparative evaluation results suggest that the contextual words along with their information and several post-processing methods can yield reasonably good performance in each of the individual models. Results also suggest that combination of several classifiers is more effective than the single classifier.

Table 29 Comparisons with other bengali ner system

Model	Recall	Precision	F-Score
HMM (Ekbal et al. 2007b)	74.02	72.55	73.28
CRF (Ekbal et al. 2008)	80.02	80.21	80.15
SVM (Ekbal and Bandyopadhyay 2008a)	81.57	79.05	80.29
Voted System (proposed)	93.98	90.63	92.28

In this paper, they reported a multi-engine NER system for Bengali by combining the outputs of the classifiers such as ME, CRF and SVM. Two different systems have been developed with the SVM approach based on the forward and backward parsing directions. Performance of the individual classifier has been improved significantly with the use of context patterns learned from an unlabeled corpus of 3 million wordforms and the various post-processing methodologies developed by observing the different kinds of errors involved in each classifier. All the four systems are then combined together into a final system by the three different weighted voting techniques. The voted system has demonstrated the overall Recall, Precision and F-Score values of 93.98%, 90.63% and 92.28%, respectively. This is actually an improvement of 18.63% in F-Score over the least performing baseline ME system and 14.92% in F-Score over the best performing baseline SVM based system.

E. Kannada

Description of kannada language

Dravidian languages have a history of more than 2,000 years. Kannada is a Dravidian language spoken mainly in southern part of India and ranks third among Indian languages in terms of number of speakers as notified in census information. Kannada is a highly inflectional and agglutinating language providing one of the richest and challenging set of linguistic and statistical features. There are few languages in the world that match Kannada in this regard.

1) Rule based Methodology for Recognition of Kannada Named Entities

In this paper, they proposed a rule based methodology to recognize Kannada named entities like person name, location name, organization name, number, measurement, time. They have manually developed suffix, prefix list and proper noun dictionary of 5000 words.

In Kannada proper nouns are indistinguishable from forms that are common nouns and adjectives. This ambiguity makes Kannada NER a challenging. Famous Kannada news paper Prajavani corpus is used to carry out experiments. The tool is language independent as there is no hard coding can be used for other Dravidian language group.

Proposed method

The main aim here is to design a system which takes a Kannada sentence as input and identifies and categorize named entities in the input. The design process for the system involves the following tasks:

1. Read transliterated file
2. Tokenization module.
3. Dictionary lookup.
4. NER Module.

Transliteration–

First the raw corpus is converted in to transliterated corpus by using converter program and is given as input. The Unicode text file in Kannada font is converted to romanized or transliterated file intermediate map file is used for conversion between and English text, we have both Iscii to romanized conversion file vice-versa. The operation of channel separation is applied on the watermarked color image to generate its sub images, and then 2-level discrete wavelet transform is applied on the sub images to generate the approximate coefficients and detail coefficients.

Tokenization –

Tokenization process divides the text file in to sequence of tokens. And removes the delimiters like [.! ?:;] which are not part of words.They also formulated some rules to identify abbreviations in the text and listed some abbreviations and if the sentence boundary delimiter comes with this abbreviation; it may or may not be the exact sentence boundary. etc. are tokens

NER Recognizer –

The named entity recognizer takes an input like "Ram" and output Ram as person name. The system uses set of pre - defined rules, suffix and prefix list, gazetteer list and dictionary.

Dictionary –

Each token is searched in the dictionary, if is found in the dictionary then tag corresponding to that is assigned to that token. If word is not found then it is passed to the NER module for further processing.

Features for NER

The suffix and prefix information works well for highly inflected language like Kannada. Sliding window size feature, F= {wi-1, wi, wi +1, wi+ 2 etc.} they have manually prepared various gazetteer lists for use in NER like 11 location suffixes, 49 designation prefixes, 74 organization prefixes, 50 person prefixes, 32 measurement prefixes, 32 next word clues, and 5000 words propernoun dictionary.

Following features are most often used for the recognition and classification of named entities.

I. Context word feature: Previous and next words of a particular word have been used as a feature.

II. Dictionaries: Dictionaries are used, Prefix and suffix lists were also important.

III. Named Entity Information: It is the features in which NE tag, tag of the previous word is considered. It is the dynamic feature.

IV. Gazetteer Lists: Due to the scarcity of resources in electronic format for Kannada language, so the gazetteer lists are prepared manually. Seven different lists are prepared.

V. Word suffix and prefix: Word suffix information is useful to identify NE's

VI. Contains digit: This features is useful in date kind of expressions, numbers, time expressions all formats of time like 8.00 AM/am, named entity measurement like 10 kg. date expressions like 12/09/2009,12-09- 12,13-04-2012, 14 June 2009, Monday 14 January representing floating point values like 12,345. The features like digits and percentage, digits and hyphen, digits and period, digits and slash are handled.

VII. Word clue: This feature helps in identifying next or previous token as nes.

Table 30 Tagsets used for ner

Tag	Description	Example
N-PRP-PRSN	Person Name	Ram
N-PRP-ORG	Organization Name	Basaveshwar sugar ltd.
N-PRP-LOC	Place Name	Bombay
N-PRP-NUM	Numeric Value	12,345
N-PRP-TIM	Time	14 june Monday
N-PRP-MEA	Measurement	10 kg.

Table 31 Evaluation of ner output

Number of Total NEs in Test Data	713
Number of NEs identified	655
Number of Correct identified NEs	567
Precision	86%
Recall	90%
F-measure	87.95%

We carried out the experiments by varying the testing data , no of words changed as 2423, 4203, 6537, and observe that, our precision and recall are good, and having nearer values.

Table 6 Confusion matrix

Corpus Size (Words)	TP	FP	FN	TN	Precision	Recall	F-measure
2423	136	29	18	2269	82.42%	88%	85%
4203	230	50	29	3944	83.42%	89%	86%
6537	403	70	41	6093	85.14%	90.7%	87%

Kannada poses challenge in NER due to its inherent ambiguity nature and lack of capitalization feature. The Kannada named entity recognition is a difficult task, especially to achieve human like performance. No work is cited in the literature on NER using rule based approach for Kannada. Our's is the first attempt, the proposed rule based methodology for recognition of Kannada named entities has good recognition rate and precision around 86%.. It is observed that use of suffix, prefix lists is important in identification of named entities. The performance can further be improved by improving gazetteer lists like proper noun dictionary, prefix and suffix.

2) Named Entity Recognition and Classification in Kannada Language

Kannada is a highly inflectional and agglutinating language providing one of the richest and most challenging sets of linguistic and statistical features resulting in long and complex word forms, which is large in number. It is primarily a suffixing Language and inflected word starts with a root and may have several suffixes added to the right. It is also a Freeword order Language. The work related to NERC in Kannada is not yet reported. In recent years, automatic named entity recognition and extraction systems have become one of the popular research areas. Building NERC for Kannada is challenging. It seeks to classify words which represent names in text into predefined categories like person name, location, organization, date, time etc. This paper deals with some attempts in this direction. This work starts with experiments in building Semi-Automated Statistical Machine learning NLP Models based on Noun Taggers. In this paper we have developed an algorithm based on supervised learning techniques that include Hidden Markov Model (HMM). Some sample results are reported.

Challenges and Issues specific to Kannada language

- No capitalization
- Brahmi script- It has high phonetic characteristic which could be utilized by NER system.
- Non-availability of large gazetteer
- Lack of standardization and spelling
- Number of frequently used words (common nouns), which can also be used as names are very large. Also the frequency with which they can be used as common noun as against person name is more or less unpredictable.
- Lack of labeled data
- Scarcity of resources and tools
- Freeword order languagE

General Procedural steps for NERC

Step1. Read raw text file and divide into sentences.

Step2. Read sentences and divide into tokens (words) (segmentation)

Step3. Read each token and perform parts of speech tagging (POS tagging)

Step4. Identify named entities from POS tag

Step5. Recognize relation of named entity (person, place, org etc

F. Malayalam

1) Named Entity Identifier for Malayalam Using Linguistic Principles Employing Statistical Methods

The system presented here is a Named Entity (NE) Identifier created using Statistical methods based on linguistic grammar principles. Malayalam NER is a difficult task as each word of named entity has no specific feature such as Capitalization feature in English. NERs in other languages are not suitable for Malayalam language since its morphology, syntax and lexical semantics is different from them. For testing this system, documents from well known Malayalam news papers and magazines containing passages from five different fields are selected. Experimental results show that the average precision recall and F-measure values are 85.52%, 86.32% and 85.61% respectively.

NE's are identified by using phonological, morphological, semantic, and syntactic properties of linguistic forms and that act as the targets of linguistic rules and operations. Two kinds of features that have been commonly used are internal and external, internal features are provided from within the sequence of words that constitute the entity-, in contrast, external features are those that can be obtained by the context in which entities appear .Based on the above investigation we have categorized an entity as either sole-entity, Constituent-entity, Dependant-entity or Not-an-entity.

Tokenizer :Input to the Tokenizer block in Fig 1 is a document in Malayalam. During the tokenization process each sentence of the document is taken and split into words or co-occurrence patterns.

NE Marker :This block checks each token to see whether it is present in the lexicon or not. Lexicon has all the root words along with its POS information. If it is present in the lexicon then it is a simple word, then the word details are retrieved from the lexicon. Based on this information token is marked with the possible NE tags.

*NE Identifier :*If the token M is a compound word then it is to be decomposed into its constituents M to Mi. To find each constituent, the longest match method is adopted. When one component Mi is separated the remaining portion is sent to modification algorithm. The component M is searched in the lexicon, if it is not found transformation algorithm is called to obtain various forms of M and again searching is carried out. If not found process is repeated with next smaller string M. Based on the constituents, NE Identifier assigns suitable tags to each token

NE Tag Disambiguator

Previous blocks assigns each input token a single/multiple NE tags. Tokens with multiple tags are sent to the Disambiguator to solve the tag ambiguity which removes all tags except one. Output of tag Disambiguator is a string of all tokens along with their NE tags.

Tests and Discussions

NER is designed and implemented using J2SDK1.4.2 and MySQL. Its performance is evaluated using standardized techniques precision, recall and F-score.

Table 32 Performance of ne identifier

Token type	NE/NAN		
	Precsion	Recall	F-measure
Proper Noun	60.0	73.0	65.86
Pronoun	85.2	87.4	86.29
Common Noun	81.4	83.5	82.44
Locative	86.3	85.0	85.64
Accusative	87.6	88.1	87.85
Dative	84.1	87.2	85.62
Instrumental	89.3	88.3	88.79
Reason	81.0	82.4	81.69
Sociative	91.7	89.0	90.33
Car-Num	90.6	91.2	90.89
Ord-Num	93.1	92.3	92.69
Adj-Num	89.0	87.5	88.24
Adj-Quantity	86.0	87.3	86.64
Other Tokens	92.0	90.5	91.24

Agglutinative Nature Malayalam is a highly inflectional and agglutinative language. 85% of words in Malayalam text are compound words and hence role of these words can be decided only by knowing its components and their types. Role of an entity depends on the importance of the word which is decided by local and global information. To derive local information, each word is analyzed and collected its component details.

Word Order

Malayalam sentence is a sequence of words where words may appear in any order and each word can be a combination of any number of stems and affixes. Even though there is no specific order for the words in the sentence, within a chunk word categories are related. In Malayalam language there is no distinction between uppercase and lowercase. Hence proper techniques are to be adopted to overcome such challenges.

Most of the systems have concentrated on three kinds of NEs ie on the roles of proper nouns, Time and percentage expressions. But these entity types are not sufficient for many question answering systems where entities like reason, cause, instrument etc are to be identified. The NER system described here is designed incorporating these types. Also this paper addresses the problem of NER in a query which involves the detection and classification of the named entity in a given query into predefined classes. They have selected formal text since this is developed as a part of QA system based on health IR. For this application text from various textbooks, journals and magazines and web sites are selected which are mostly formal texts.

2) A Hybrid Statistical Approach for Named Entity Recognition for Malayalam Language

Several Statistical methods with supervised and unsupervised learning have applied English and some other Indian languages successfully. Malayalam has a distinct feature in nouns having no subject-verb agreement, which is of free order, makes the NER identification a complex process. In this paper, a hybrid approach combining rule based machine learning with statistical approach is proposed and implemented, which shows 73.42% accuracy.

Statistical Apporach

The statistical methods are mainly based on the probability measures including the unigram, bigram, trigram and n-grams. TnT Trigrams n Tags is a very efficient statistical part of speech tagger that can be trained on any language with any tagset. The parameter generation component trains on tagged corpora. The system uses several techniques for smoothing and handling of unknown words. TnT can be used for any language, adapting the tagger to a new language, new domain or new tagset very easy.

Experiments and Results

Under the same domain, a comparison on two supervised taggers namely TnT and SVM was conducted. In our experiment, for known words, SVM shows better performance but for unknown words TnT outperformed. However, for embedded tags, it is required to generate rules that combining with TnT shows better result. So our proposed hybrid supervised machine learning approach with the combination of TnT and Rule based is a good strategy for NER especially for embedded tags.The corpus was tagged using the NER tagset for Malayalam. The TnT was learned using the tagged corpus. When learned, the dictionary file was created for the corpus. Once learning

process is done, then the input text file was given to the tool and tagging was performed. The system gives an accuracy of 73.42% .The accuracy can be increased by increasing the amount of training data.

Table 33 Result of ne tagging using TnT

Size of training corpus(in tokens)	Size of test corpus(in tokens)	Automated accuracy obtained	precision	recall	f-measure

100	150	57.59	37.5	26.09	30.77
200	150	56.96	56.25	39.13	46.15
500	150	60.76	58.33	30.43	39.10
2000	150	68.99	87.5	30.43	45.16
5000	150	73.42	100	30.43	46.66
10000	150	73.42	100	43.48	60.61

Considering the various issues like classifying ambiguous strings correctly, detecting the boundaries of an NE correctly, categorizing NERs, and availability of Unicode data, the proposed hybrid model achieves 73.42% accuracy. The domains considered for tagging were health and tourism. Accuracy can be further increased by increasing the number of words in the training corpus. The work shows that a hybrid statistical approach, combining TnT and rule based suit better for highly morphologically and inflectionally rich languages like Malayalam.

G. Manipuri

1) CRF Based Name Entity Recognition (NER) in Manipuri: A Highly Agglutinative Indian Language

This paper deals about the Name Entity Recognition(NER) of Manipuri, a highly agglutinative Indian Language. This language is an Eight Scheduled Language of Indian Constitution. Feature selection is an important factor in recognition of Manipuri Name Entity using Conditional Random Field (CRF).

Features:

Current word: The current word is the focal point of Name Entity Recognition so selecting the current word as a feature is important.

Surrounding Stem words as feature: Stemming is done as mentioned in Section III and the preceding and following stem words of a particular word with the stem of the current word are used as features since the preceding and following words influence the present word in case of Name Entities.

Acceptable suffixes: 61 suffixes have been manually identified in Manipuri and the list of suffixes is used as one feature.

Acceptable prefixes as feature: 11 prefixes have been manually identified in Manipuri and the list of prefixes is used as one feature. For every word the prefix is identified and a column is created mentioning the prefix if the prefix is presents, otherwise the "0" notation is used.

Binary notation if a suffix(es) is present: The suffixes play an important role in Manipuri since it is a highly agglutinative language. For every word if a suffix(es) is/are present during stemming a binary notation '1' is use otherwise '0'.

Number of acceptable suffixes as feature: For every word the number of suffixes is identified during stemming, if any and the number of suffixes is used as a feature.

Binary notation if a prefix(es) is present: The prefixes play an important role in Manipuri since it is a highly agglutinative language. For every word if a prefix(es) is/are present during stemming a binary notation '1' is use otherwise '0'.

Digit features: Date, currency, weight, time etc are generally digits. Thus the digit feature is an important feature. A binary notation of '1' is used if the word consist of a digit else '0'.

Binary Notation of general salutations/preceding word of Name Entity: Salutations like Mr., Miss, Mrs, Shri, Lt., Captain, Rs., St., Date etc precedes the Name Entity so the preceding word can also be considered as a feature for the Name Entity Recognition. A binary notation of '1' if found else '0' is used.

Binary notation of general follow up words of Name Entity: The following word of the current word can also be consider as a feature since a name may have end up with clan name, surname also word like 'organization', 'Lup' etc for organization, word like 'Leikai', 'City' etc for places and so on. A binary notation of '1' if found else '0' is used. Length of the word: Length of the word is set to 1 if it is greater than 3 otherwise, it is set to 0. Very short words are rarely Name Entity.

Word frequency: A range of frequency for words in the training corpus is set: those words with frequency <100 occurrences are set the value 0, those words which occurs >=100 but less than 400 are set to 1 and so on. The word frequency is considered as one feature since Name Entities are rare in occurrence compared to those of determiners, conjunctions and pronouns.

Surrounding POS tag: Name Entity is generally a combination of nouns, so the POS of the surrounding words are considered as an important feature.

Table 34. Example ner tagging

NE Tag	Meaning
B-DAT	Beginning, Internal or the End of a multiword date
I_DAT	
E-DAT	
B-DES	Beginning, Internal or the End of a multiword designation
I-DES	
E-DES	
B-LOC	Beginning, Internal or the End of a multiword location name
I-LOC	

E-LOC	
B-MISC	Beginning, Internal or the End of a multiword miscellaneous name
I-MISC	
E-MISC	
B-ORG	Beginning, Internal or the End of a multiword organization name
I-ORG	
E-ORG	
B-PER	Beginning, Internal or the End of a multiword person name
I-PER	
E-PER	
B-TIM	Beginning, Internal or the End of a multiword time format
I-TIM	
E-TIM	
CUR	Currency
DES	Designation
LOC	Single word location name
MISC	Single word miscellaneous Name
ORG	Single word organization Name
PER	Single word person name

Notation Meaning
Wi -The current word,
SW[-I,+J]- Stem Words spanning from the i-th left position to the j-th right position
POS[-I, +J]- POS tags of the words spanning from the ith left to the jth right positions
MS- Binary value '1' if a suffix(es) is/are present else '0'
MP- Binary value '1' if prefix(es) is/are present else '0'
NoSuf- Number of suffix present in the word
Pre- Prefixes present in the word
PW- General salutations/preceding word of Name Entity
FW- General follow up words of Name Entity
Sufj -Suffixes present in the word, where j= 1 to 10
DF- Digit feature
Len- Length of the word
Frq -Frequency of the word

Table 34 Feature set with recall, precision, f-score

Feature	**R(%)**	**P(%)**	**FS(%)**
W,SW[-2,+2], Sufj=(1-10), NoSuf, MS, Pre, MP, PW, FW, DF, Len, Frq, POS [-2,+2]	81.12	85.67	83.33
W,SW[-3,+3], Sufj=(1-10), NoSuf, MS, Pre, MP, PW, FW, DF, Len, Frq, POS[-3,+3]	80.32	84.24	82.23
W,SW[-3,+2], Sufj=(1-10), NoSuf, MS, Pre, MP, PW, FW, DF, Len, Frq, POS[-3,+2]	78.87	76.87	77.86
W,SW[-4,+3], Sufj=(1-10), NoSuf, MS, Pre, MP, PW, FW, DF, Len, Frq, POS[-4,+3]	70.33	74.65	72.43
W,SW[-4,+3], Sufj=(1-10), NoSuf, MS, Pre, MP, PW, FW, DF, Len, Frq, POS[-4,+3]	63.78	67.86	65.76

W,SW[-2,+2], Sufi=(1-10), NoSuf, MS, Pre, MP, PW, FW, DF, Len, Frq, POS[-2,+2]	57.75	60.03	58.87
W,SW[-3,+3], Sufi=(1-10), NoSuf, MS, Pre, MP, PW, FW, DF, Len, Frq, POS[-3,+3],	43.77	45.84	44.78
W,SW[-4,+3], Sufi=(1-10), NoSuf, MS, Pre, MP, PW, FW, DF, Len, Frq, POS[-4,+2],	32.03	31.87	31.95
W,SW[-4,+4], Sufi=(1-10), NoSuf, MS, Pre, MP, PW, FW, DF, Len, Frq, POS[-4,+4],	20.00	21.30	20.63

In their model feature selection is done through manual assumption but implementation of a Genetic Algorithm (GA) or other technique in feature selection could be the future road
map. A Manipuri gazetteer list can be formed using the NER and this model can be useful.

2) Named Entity Recognition for Manipuri Using Support Vector Machine

This paper reports about the development of a Manipuri NER system, a less computerized Indian language. Two different models, one using an active learning technique based on the context patterns generated from an unlabeled news corpus and the other based on the well known Support Vector Machine (SVM), have been developed. The active learning technique has been considered as the *baseline* system. The Manipuri
news corpus has been manually annotated with the major NE tags, namely *Person name*, *Location name*, *Organization name* and *Miscellaneous name* to apply SVM. The SVM
based system makes use of the different contextual information of the words along with the variety of orthographic word-level features which are helpful in predicting the NE classes. In addition, lexical context patterns generated using the active learning technique have been used as the features of SVM in order to improve performance. The system has been trained and tested with 28,629 and 4,763 wordforms, respectively

Named Entity Tagset

In the present work, the NE tagset used have been further subdivided into the detailed categories in order to denote the boundaries of NEs properly.

Table 35 Named entity tagset

NE Tag	Meaning
B-LOC	Beginning, Internal or the End of a multiword location name
I-LOC	
E-LOC	
PER	Single word person name
LOC	Single word location name
ORG	Single word organization name

Features for Manipuri Named Entity Recognition

Context word feature: Preceding and following words of a particular word since the surrounding words carry effective information for the identification of NEs.

Word suffix: Word suffix information is helpful to identify NEs. This is based on the observation that the NEs share some common suffixes. The fixed length (say, n) word suffix of the current and/or the surrounding word(s) can be treated as the feature. If the length of the corresponding word is less than or equal to n − 1 then the feature values are not defined and are denoted by 'ND'.

Word prefix: Word prefixes are also helpful to identify NEs. It is based on the observation that NEs share some common prefix strings. This feature has been defined in a similar way as that of the fixed length suffixes.

Named Entity Information: The NE tag(s) of the previous word(s) have been used as the only dynamic feature in the experiment. The output tag of the previous word is very
informative in deciding the NE tag of the current word.

Digit features: Several binary valued digit features have been defined depending upon the
(i). Presence and/or the exact number of digits in a token.
(a). CntDgtCma: Token consists of digits and comma
(b). CntDgtPrd: Token consists of digits and periods
(ii). Combination of digits and symbols. For example,
(a). CntDgtSlsh: Token consists of digit and slash

(b). CntDgtHph: Token consists of digits and hyphen
(c). CntDgtPrctg: Token consists of digits and percentages
(iii). Combination of digit and special symbols. For example,
(a). CntDgtSpl: Token consists of digit and special symbol such as $, # etc. These binary valued digit features are helpful in recognizing miscellaneous NEs such as measurement expression and percentage expression.

Infrequent word: The frequencies of the words in the training corpus have been calculated. A cut off frequency has been chosen in order to consider the words that occur with less
than the cut off frequency in the training corpus. A binary valued feature 'Infrequent' is defined to check whether the current word appears in this infrequent word list or not. This
is based on the observation that the infrequent words are most probably NEs.

Length of a word: This binary valued feature is used to check whether the length of the current word is less than three or not. We have observed that very short words are most
probably not the NEs.

Part of Speech (POS) information: We have used an SVM-based POS tagger (Doren et al.,2008) that was originally developed with 26 POS tags, defined for the Indian languages.

The POS information of the current and/or the surrounding words can be effective for NE identification.

Table 36 Feature set with recall, precision, f-score

Feature	R(in %)	P(in %)	FS(in %)
Static: W[-2,+2], POS[-2,+2], |Pre|<=3, |Suf|<=3, Length, Infrequent, FirstWord, Digit Dynamic: NE[-2,-1]	94.2	98.47	96.29
Static: W[-3,+3], POS[-3,+3], |Pre|<=3, |Suf|<=3, Length, Infrequent, FirstWord, Digit Dynamic: NE[-3,-1]	88.91	97.82	93.15
Static: W[-3,+2], POS[-3,+2], |Pre|<=3, |Suf|<=3, Length, Infrequent, FirstWord, Digit Dynamic: NE[-3,-1]	91.3	96.99	94.06
Static: W[-4,+3], POS[-4,+3], |Pre|<=3, |Suf|<=3, Length, Infrequent, FirstWord, Digit Dynamic: NE[-2,-1]	87.05	97.66	92.05
Static: W[-4,+3], POS[-4,+3], |Pre|<=3, |Suf|<=3, Length, Infrequent, FirstWord, Digit Dynamic: NE[-3,-1]	85.28	98.03	91.21
Static: W[-2,+2], POS[-2,+2], |Pre|<=4, |Suf|<=4, Length, Infrequent, FirstWord, Digit Dynamic: NE[-2,-1]	88.70	98.49	93.34
Static: W[-3,+3], POS[-3,+3], |Pre|<=4, |Suf|<=4, Length, Infrequent, FirstWord, Digit Dynamic: NE[-3,-1]	87.05	97.09	91.79
Static: W[-4,+3], POS[-4,+2], |Pre|<=4, |Suf|<=4, Length, Infrequent, FirstWord, Digit Dynamic: NE[-2,-1]	78.55	97.54	87.02
Static: W[-4,+4], POS[-4,+4], |Pre|<=4, |Suf|<=4, Length, Infrequent, FirstWord, Digit Dynamic: NE[-3,-1]	71.71	97.44	82.62

A number of experiments have been carried out to find out the best set of features for NER in Manipuri. The system has been trained and tested with the 28,629 and 4,763 wordforms, respectively. We get effective result from news domain. The system has demonstrated the *Recall*, *Precision* and *F-Score* values of 93.91%, 95.32 and % 94.59%, respectively

H. Odia

1) Case Study of Named Entity Recognition in Odia Using Crf++ Tool

NER have been regarded as an efficient strategy to extract relevant entities for various purposes. The aim of this paper is to exploit conventional method for NER in Odia by parameterizing CRF++ tool in different ways. As a case study, we have used gazetteer and POS tag to generate different feature set in order to compare the performance of NER task. Comparison study demonstrates how proposed NER system works on different feature set.

EXPERIMENTAL SET UP

Part Of Speech Tag :In corpus linguistics, part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph.

Gazetteer : We have prepared 4 different gazetteers. The words belongs to the person, location, organization are stored in 3 different gazetteers respectively. Another gazetteer contains only NE without any classification and it contains around 730 NEs.

	Gazetteer			All features (gazetteer+ POS tag)			F measure comparison
	P	R	F	P	R	F	
person	0.87	0.81	0.84	0.97	0.44	0.63	25% decrease
location	0.88	0.82	0.85	0.75	0.50	0.60	18% Decrease
organization	0.50	0.82	0.62	0.66	0.25	0.35	43% Decrease

The performance of the system is quite good when we experiment with individual case (f-measure for NEs only is 71% and f-measure for NEs with classification is 84% for PER, 85% for LOC and 62% for ORG). The performance of system decreases when combine both POS tag and Gazetteer to generate feature. The reason for decrease in performance may be the average accuracy of POS Tagger tool.

III. CONCLUSION

This paper discusses about NER task done in some Indian languages and its various factors like different machine learning models ,independent or dependent features and performance metrics like precision, recall, f-score by which system performance is evaluated.

REFERENCES

[1] Singh J., Lehal J.S.,(March 2015) *"Named entity recognition for Punjabi language using HMM and MEMM"* In Proceedings of 21st IRF International Conference, pp 4-8.

[2] Chopra D., Morwal S.,(2012) "*Named Entity Recognition in Punjabi Using Hidden Markov Mode*l" International Journal of Computer Science & Engineering Technology (IJCSET) Vol. 3, pp 616-620.

[3] Gupta V.,LEHAL G. S. ,*Named Entity Recognition for Punjabi Language Text Summarization*, International Journal of Computer Applications (0975 – 8887) Volume 33– No.3, November 2011,pp 28-32

[4] Chopra D., Jahan N., Morwal S. ,Hindi *Named entity Recognition by aggregating rule based heuristics and hidden markov model*, International Journal of Information Sciences and Techniques (IJIST) Vol.2, No.6, November 2012, pp 43-52

[5] Li W., McCallum A., *Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction (Short Paper),* In Proc. Of ACM Transactions on Computational Logic, Vol. V, No. N, February 2004, Pages 1–4

[6] Sarkar S. ,Mitra P. ,*A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition* , In Proceedings of 3rd International joint Conference in Natural Language Processing,

[7] Talukdar G., Borah P. P., Baruah A.,*A survey of NER in Assamese and other Indian languages*, International Conference On Natural Language Processing and Cognitive Computing (ICONACC)-2014

[8] Sharma P , Kalita J., Sharma U., *Named Entity Recognition In Assamese using CRFs and Rules*, In Proc. International Conference on Asian Language Processing (IALP), 2014, pp 15-18

[9] Ekbal A., Bandyopadhyay S., Haque R., *Named Entity Recognition in Bengali: A Conditional Random Field Approach*,In the proceedings of IJCNLP-2008, pages 589-594

[10] Ekbal A., Bandyopadhyay S. (2002), *Named Entity Recognition in Bengali: A Multi-Engine Approach* , Northern European Journal of Language Technology, Vol. 1, Article 2, pp 26–58.

[11] Melinamath B. C., *Rule based Methodology for Recognition of Kannada Named Entities*, International Journal of Latest Trends in Engineering and Technology (IJLTET) , Vol. 3 Issue 4 March 2014 ,pp 50-59

[12] Amarappa S., and Sathyanarayana S. V., *Named Entity Recognition and Classification in Kannada Language*, International journal of electronics and computer science engineering (IJECSE),Volume2,Number 1,PP-281-289

[13] S.B.M. and Idicula S.M. *Named Entity Identifier for Malayalam Using Linguistic Principles Employing Statistical Methods*, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, September 2011,pages 185-191

[14] Jayan J.P., R R R., Sherly E., *A Hybrid Statistical Approach for Named Entity Recognition for Malayalam Language*, International Joint Conference on Natural Language Processing, Nagoya, Japan, 14-18 October 2013, pages 58-63

[15] Nongmeikapam K., Singh L.N. , Shangkhunem T. , Salam B. ,*CRF Based Name Entity Recognition (NER) in Manipuri: A Highly Agglutinative Indian Language*, in the proceedings of NCETACS 2011 Volume 1,

[16] Singh T. D. , Nongmeikapam K., Ekbal A. and Bandyopadhyay S. "*Named Entity Recognition for Manipuri Using Support Vector Machine"* In Proc. of 23rd Pacific Asia Conference on Language, Information and Computation, pages 811–818

[17] Balabantaray R.C.,Das S. , Mishra K.T., *Case Study of Named Entity Recognition in Odia Using Crf++ Tool*, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.6, 2013.

