A Project Report on

Named Entity Recognition (NER)

Under the guidance of

Professor Dipti Misra Sharma

Submitted by:

Ayush Agarwal (Indian Institute of Technology (BHU), Varanasi)

Anurag Dubey (National Institute of Technology, Durgapur)

At

International Institute of Information Technology, Hyderabad

Language Technologies Research Centre

2013

At the Language Technologies Research Centre (LTRC), International Institute of Information Technology, Hyderabad (IIIT-H) there has been substantial work in the area of Named Entity Recognition. The goal of Named Entity Recognition is to identify and classify the proper names appearing in the text and the number of meaningful phrases. This project report is a part of the on-going research in the field of Named Entity Recognition in different domains.

Problem Description:

The following is a quote of the problem description:

Named entity recognition (NER) is a technology for recognizing named entities in text and associating them with the appropriate types which is an important step towards resolving many issues for successful processing of subsequent modules in a Natural Language Processing application. Common types in NER systems are location, person name, date, address, etc. The purpose of this project is to develop NER systems for Hindi in a specific domain.

Introduction

Named Entity is an expression which refers to very specific things for which the referent is a rigid designator. Rigid Designator includes proper names as well as certain natural kind terms like biological species and substances.

Person names, Organizations (companies, government organisations, committees, etc.), Locations (cities, countries, rivers, etc.), Date and time expressions etc. are considered to be the name entities. Proper Nouns are mostly taken as Named Entity.

Named Entity Recognition (NER) involves identification of Named Entity in texts and their classification into a set of predefined categories of interest. Different categories are usually person names, location names, organization names, date & time expressions etc.

Named Entity Recognition is two-step process i.e. identification of named entities and its classification. Identification is concerned with marking of a word/phrase as named entity in the given sentence and classification is for denoting role of the identified Named Entity. The named entity is identification phase also involves the detection of their boundaries, i.e., the start and end of all possible spans of tokens that are likely to belong to a Named Entity. NER is a key part of information extraction system.

Background Study:

Named Entity Recognition is a precursor for many Natural Languages Processing tasks. An accurate Named Entity Recognition system is needed for Machine Translation, more accurate internet search engines, automatic indexing of documents, automatic question-answering, Information Retrieval etc. It is sub-task of Information Extraction where structured text is extracted from un-structured texts, such as Newspapers articles. It is also used in data classification and automatic summarization.

In recent years automatic NER Systems have become a popular research area in which considerable number of studies has been addressed on developing these systems. These studies can be classified into three main classes:

1: Rule Based NER
2: Machine Learning based NER
3: Hybrid NER
All the approaches may make use of gazetteer information to build system because it improves accuracy of the System.

Rule Based Approach:

It focuses on extracting names using a number of hand crafted rules and patterns. Generally, these system consists of a set of patterns using grammatical (Part of Speech), Syntactic (e.g. Word Precedence) and Orthographic features e.g. Capitalization in combination with dictionaries.

Handicraft Rule based systems usually give good results, How-ever they need months of development by experienced linguists.

Main disadvantages of Rule based technique are that these require huge experience and grammatical knowledge of particular languages and domains. These systems are not robust and are not transferable to other languages.

Machine Learning:

Machine learning (ML) based NER learns symbolic rules and statistical model with the help of data either annotated or un-annotated. It makes use of large amount of training data to acquire high level language knowledge. Machine learning approaches are popularly used in NER because these are easily trainable, adaptable to different domains and languages and their maintenances less expensive. Different methods are:

1. Supervised ML Technique:

It requires preparing labelled training data to construct statistical models. Some of the ML techniques are:

a.) Hidden Markov Model
b.) Maximum Entropy
c.) Conditional Random Field
d.) Support Vector Machine

2. Semi supervised ML Technique

The term semi-supervised is relatively recent. The main technique for SSL is called bootstrapping and involves a small degree of supervision, such as a set of seeds, for starting the learning process.

3. Unsupervised ML Technique:

These approaches do not require labeled training data i.e. training requires few seed lists and large un-annotated corpora

Un-supervised learning is not popular as Supervised Learning approaches and systems that do use unsupervised are usually not completely un-supervised. Supervised approaches are more expensive than unsupervised one in terms of time spent to pre-process the training data. Supervised approaches can achieve good performance only when large amount of high quality training data is available.

Hybrid NER System:

In Hybrid NER system, approach uses the combination of both rule based and ML technique and makes new methods using strongest points from each method. It is making use of essential feature from ML approach and uses the rules to make it more efficient.

Named Entity Tag-set:

According to the specifications defined by MUC, the NER tasks generally work on seven types of named entities as listed below:

Person Name
Location Name
Organization Name
Abbreviation
Time
Term Name
Measure

Features for Recognition and Classification of Named Entities

Features are characteristic attributes of words designed for algorithmic purpose. Following features are most often used for the recognition and classification of named entities. These are defined into three categories i.e. named entities i.e.

Word-level features
List lookup features
Document and Corpus features

Word-level features describe the character makeup of words i.e. the word case, punctuation, numerical value, part-of-speech (POS) and special characters. List look up features can be called also as the term “gazetteer”, “lexicon” and “dictionary”.

Related Works:

1. In the work done in [3] on Language Hindi and Bengali the results obtained were :

Language	Model	Recall	Precision
Bengali	Baseline	64.73	52.21
Hindi	Baseline	64.67	52.01
Bengali	SVM	82.21	76.07
Hindi	SVM	80.23	79.02

2. A work done on Hybrid approach on different Indian Languages gave the results as follows:

Language	Precision	Recall
Hindi	75.19	58.94
Bengali	52.92	68.07
Oriya	21.17	26.92
Urdu	26.12	29.69
Telugu	10.47	9.64

3. A work done on Hybrid Approach on named entity recognition for Hindi Language [4].Rule Based, CRF Machine Learning and Maximum Entropy ML approach have been combined to form the Hybrid Approach. Their results were as follows:

Model	Precision	Recall
Hybrid	81.11	84.08

4. A work done on hybrid approach on Aggregating rule based heuristics and hidden Markov model in Hindi Named Entity Recognition By Aggregating Rule Based Heuristics And Hidden Markov Model [5]. The results obtained were as follows:

Total Named Entities(NEs)	Named Entities Identified	ACCURACY
687	650	94.61

System Development

Named Entity Recognition System is developed for specific domain of “Electronic Gadgets” in Hindi Corpuses of 5300 sentences under domain of Electronic Gadgets were collected from different Newspapers websites. The data was tokenized and POS tagging was done with the help of “Shallow Parser”. Following Modules of shallow parser were used.

Sentence splitter

The sentence splitter, which is domain and application-independent, is a cascade of finite state transducers which segments the text into sentences. This module is required for the tagger. The splitter uses a gazetteer list of abbreviations to help distinguish phrase marked points and apart from other types.

Tokenizer

The tokenizer splits the text into very simple tokens such as numbers, punctuation and words of different types.

Morph Analyzer

A tool that identifies the root and the grammatical features of a given word. It looks at words independent of its context. The grammatical features of a word given by Morph analyzer are root, lexical category, gender, number, case, Vibhakti and the suffix.

Part-of Speech tagger

POS tagger was introduced before that used to recognize parts of speech to each word. It produces a part-of-speech tag as an annotation on each word or symbol. The tagger uses a default lexicon and rule set.

After running all these modules on the raw data collected. Manual Annotation was done on the final output files generated by running all the above mentioned modules. The Named Entity Tag-set used for manual annotation is as shown below.

Named Entity Tag set used:

Tagset	Named entity	Example
NEP	Person	Mark Twain
NED	Designation	Chairman
NEO	Organization	Microsoft
NEA	Abbreviation	3G, WWDC
NEB	Brand, Products	Windows, Linux
NETP	Title-Person	Mahatma, Shree
NETO	Title-Object	American Beauty
NEL	Location	Delhi
NETI	Time	5 pm
NEN	Number	1 Lakh
NEM	Measure	Seven Years

Problems for NER in Hindi:

- No Capitalization
- More diverse Indian Person names for e.g. KavitA (Person name vs. Common Noun with meaning 'poem').
- Lack of Standardization and spelling.
- Non-availability of large gazetteer..
- Lack of labeled data
- Scarcity of resource and tools
- Free word order Language.

Data Annotation:

- A Workshop on NER for South and South East Asian Languages(IJCNLP 2008) which defines above twelve different classes of NEs(available at http://ltrc.iiit.ac.in/ner-ssea-08) is used for annotation of data.
- BIO Format is used for the annotation of data
- No nested entities in manual annotation. Although they have to be marked by the system. e.g.: Samsung Galaxy Tab is annotated as product. Samsung is not separately marked as Organization in this above phrase.
  Maximal Entity: Only the longest sequence forming an NE is to be manually annotated.
  e.g. : (Samsung Galaxy Tab)
  Correct one: ((Samsung) (Galaxy Tab))
  Specificity: Whether the expression identifies something specific as if by a name
- **Ambiguity was** resolved by context

- Each sentence within <Sentence></Sentence> tag.

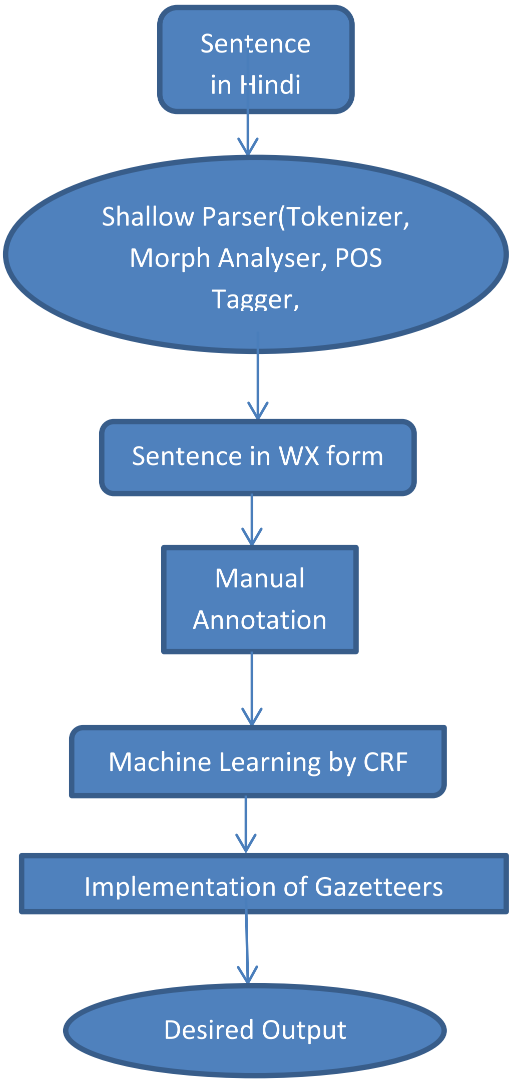
Example of manual annotation:

```
<Sentence id="4">
1   PujuwsI       JJ     B-NEO
2   teknolOYjI    NN     I-NEO
3   sOYlyUSana    NN     I-NEO
4   ne            PSP    O
5   bAjAra NN     O
6   meM           PSP    O
7   nayA          JJ     O
8   notabuka      NN     O
9   kama          QF     O
10  tEbaleta      NN     O
11  lAMca         VM     O
12  kiyA          VAUX   O
13  hE            VAUX   O
14  .             SYM    O
</Sentence>
```

Our Approach

- Rule based and Machine learning Hybrid Approach will be used.
- Machine learning based on Conditional Random Fields (CRFs).
- Gazetteer list prepared from the corpus will also be used.
- CRF++ Tool kit (An open source based on CRFs) was used for Machine learning.

Overview of our work:



CRF++ Tool Kit

CRF++ is a simple, customizable, and open source designed for generic purposes and is applied for various NLP tasks like Named Entity recognition, Information Extraction and test chunking. It is implementation of Conditional Random Fields (CRFs) for segmenting/labeling of sequential data.It requires:
a.) training data in a particular format.
Training Data format is as follows:
Word<tab>POS<tab>length<tab>Prefix<tab>Suffix<tab>NE Tag.

b.) Testing data in a particular format.
Test Data format is as follows:
Word<tab>POS<tab>length<tab>Prefix<tab>Suffix

c.) Feature templates describing features used for testing and training. We have use unigram feature templates (describes unigram feature only) in training and testing.

Features used for Machine Learning

- Context Word feature: Previous and next words of a particular word have been used as a feature. Generally, word window of size 2 or 3 is used.
- Word Suffix and Prefix: In this feature a length of 2 or 3 characters of the current and /or the surrounding words is taken.
- Part of Speech(POS) Information
- Named Entity Information: It is the feature in which NE Tag of surrounding and the current word is considered.

Analysis of Machine learning output

We have generated testing and training data on varied length of Affixes.
Calculation of Precision and Recall on each training and testing sets using 7 different feature templates.

Precision = No. of correct NE Tags/No. of Total NE Tags output by the System.

Recall = No. of correct NE Tags/Total No. of NE Tags on the data manually annotated.

Templates	2 Suff-2Pre	2Suff_3Pre	3Suff_2Pre	3Suff-3Pre
Template 1	Prec77.24 Recall 71.03	Prec77.92 Recall71.47	Prec 77.42 Recall70.79	Prec 77.59 Recall70.09
Template 2	Precision80.7 Recall 72.67	Prec 80.43 Recall72.44	Prec 79.39 Recall71.28	Prec 79.37 Recall71.18
Template 3	Prec78.09 Recall 71.04	Prec 79.01 Recall71.47	Prec 77.28 Recall70.20	Prec 78.12 Recall70.59
Template 4	Prec79.09 Recall 72.47	Prec 79.74 Recall72.05	Prec 78.57 Recall71.76	Prec 78.73 Recall70.98
Template 5	Prec78.57 Recall 71.69	Prec 77.77 Recall71.18	Prec 78.41 Recall71.08	Prec 77.93 Recall70.50
Template 6	Prec77.78 Recall 71.50	Prec 77.26 Recall70.79	Prec 76.30 Recall69.91	Prec 76.61 Recall69.52
Template 7	Prec77.67 Recall 70.79	Prec 77.09 Recall70.79	Prec 77.71 Recall70.59	Prec 78.42 Recall70.40

After data analysis, Reason for high Precision and Recall was found to be for test data from same source as of training data.
Statistics on test data from different source is as shown in the table below:

Templates	2 Suff-2Pre	2Suff-3Pre	3Suff-2Pre	3Suff-3Pre
Template 1	Prec 64.88 Recall 61.05	Prec 65.36 Recall 61.92	Prec 63.59 Recall 60.84	Prec 65.10 Recall 58.51
Template 2	Prec 65.66 Recall 62.43	Prec 65.94 Recall 62.34	Prec 63.96 Recall 61.51	Prec 65.62 Recall 60.06
Template 3	Prec 62.97 Recall 58.49	Prec 64.78 Recall 58.36	Prec 62.41 Recall 56.81	Prec 63.78 Recall 56.97
Template 4	Prec 64.79 Recall 5.97	Prec 65.22 Recall 60.37	Prec 65.01 Recall 58.67	Prec 64.88 Recall 58.05
Template 5	Prec 63.01 Recall 55.11	Prec 63.32 Recall 57.81	Prec 62.69 Recall 57.74	Prec 63.95 Recall 58.20
Template 6	Prec 62.19 Recall 54.49	Prec 64.52 Reacll 57.82	Prec 62.63 Recall 57.59	Prec 64.44 Recall 58.36
Template 7	Prec 62.32 Recall 54.80	Prec 63.42 Recall 57.43	Prec 62.88 Recall 57.43	Prec 63.72 Recall 56.81

- Template2 that is window of 2 words (2 next words and 2 previous words) gives best precision and recall.
- Length of affixes that gives best result is 2
- After removing NEM,NEN,NETI tags precision and recall are as follows:

  Test data from same source

  Precision=80.33
  Recall=67.67

  Test data from different source

  Precision=60.85
  Recall=53.68

Machine Learning Output Analysis

Output of Machine Learning was thoroughly analyzed and was compared with result of manually annotated test data. After the analysis, we found that the main reason for low precision and recall was that the machine is not able to identify most of the new names of products and organization that appeared in the test data. We also found that machine was not able to identify person names correctly. Some manual annotation errors were also found while comparing the output from the manually annotated data.

Net Result of Machine Learning after correction of Manual Annotation

Data from different sources:

Precision= 67.70
Recall=57.04

Data from same sources:

Precision=79.52
Recall=67.94

Hence we use our gazetteer approach for identifying new names of Product and Organizations. Statistics were derived from the data to develop a Rule based approach to identify person names.

Gazetteers

Gazetteers are list of names specific to named entities.

We developed gazetteers list from different websites. We crawled different websites in order to collect names of different products and companies related to our domain of "Electronic Gadgets". We also manually collected names from some of the websites. All the data available was in English. We collected about 3000 company names and 5000 product names of different companies (all in English). Part of both list were transliterated in Hindi and then were corrected manually. These lists were then converted into WX format. Finally we have around 500 names of organization names and around 2000 names of gadgets in WX format.

Problems faced in developing Gazetteers

- Non-Availability of names of products and Organizations in Hindi on various Websites.
- The tools which were available for transliteration do not have good accuracy.
  Almost each name needs to be corrected manually.
- WX form produced also does not match from that of the data.

After applying gazetteers, Output of gazetteers and Machine learning were combined giving more priority to the output of the gazetteers. The result were obtained as follows:

For data from the same sources:

Precision = 86.35
Recall = 83.71

For data from the same sources:

Precision = 80.80
Recall = 80.09

Accuracy of identifying organization and product names was highly increased.

Rule Based Approach

In the course of data analysis, it was found that some tools may be developed for person named entities. It was observed that at many occurrences of person named entity were followed by the post position word 'NE' followed by word

Having Part of Speech (POS) tag 'VM' i.e. main verb. For analyzing this information accurate statistics were obtained from the data regarding this and it was found that if a word is followed by ‘NE’ and ‘VM’ there are around 70% chances that the word belongs to person named entity (NEP).

But this observation fails in tabbing whether the word is beginning of named entity or intermediate of named entity person. Also the number of occurrences was around 100 in data of 5300 sentences, which is very less. Hence it was not applied as a rule.

In the data analysis, it was also found that most of the Named Entities are followed by post position (PSP).Accurate statistics were framed for calculating the number of occurrences for each Named Entity before each different Post-Position (PSP) word in the data. It also does not help in making out any general rule.

Experiments

After this, different experiments were done on Machine Learning and results were obtained by combining both the gazetteer output and Machine Learning output.
Experiments done are as follows:

1. Part of Speech (POS) tag ‘NNP’ were replaced by Part of Speech (POS) tag ‘NN’.
Machine was again trained on the basis of changes done on the training data.
Results obtained from this experiment were as follows:

Data from the different sources:

Precision = 79.45
Recall = 75.77

Data from the same sources:

Precision = 84.40
Recall =81.90

2. Machine was trained without POS tag information. Results obtained from this experiment were as follows:

Data from the different sources:

Precision = 81.75
Recall = 75.99

Data from the same sources:

Precision = 85.81
Recall =81.90

Result and Conclusion

Results for data from the same source:

	Machine Learning (CRF)		Gazetteer & ML combined	
	Precision	Recall	Precision	Recall
With all POS tags	80.97	66.91	86.35	83.71
POS NNP Replaced with POS NN	78.36	65.25	84.40	81.90
Without POS Tag information	79.03	67.66	85.85	81.90

Results for data from the different source:

	Machine Learning (CRF)		Gazetteer & ML combined	
	Precision	Recall	Precision	Recall
With all POS tags	69.17	58.81	80.80	80.09
POS NNP Replaced with POS NN	66.76	54.41	79.45	75.77
Without POS Tag information	71.91	56.39	81.75	75.99

After analyzing all of the above cases, it is concluded that the value of Precision and Recall was found to be around 81 , 85 and 80, 77 percent for the data from the same sources and different sources respectively.

Accuracy for NEP and NEA is low for our system and can be improved by implementing Rule based approach and improving gazetteers.

References

[1]. Darvinder Kaur and Vishal Gupta, 2010. "A Survey of Named Entity Recognition in English and other Indian Languages, Department of Computer Science and Engineering, Punjab University, Chandigarh, India.

[2]. Asif Iqbal and Sivaji Bandhopadhyay. Named Entity Recognition Using Appropriate Unlabeled Data, Post-processing and Voting. Department of Computer Science and Engineering, Jadavpur University, Kolkata.

[3]. Asif Iqbal and Sivaji Bandhopadhyay. Named Entity Recognition using Support Vector Machine: A Language Independent Approach. Department of Computer Science and Engineering, Jadavpur University, Kolkata.

[4]. Shilpi Srivastava, Mukund Sanglikar and D.C Kothari. University of Mumbai
Named Entity Recognition System for Hindi Language: A Hybrid Approach

[5]. Deepti Chopra, Nusrat Jahan, Sudha Morwal. Hindi Named Entity Recognition By Aggregating Rule Based Heuristics And Hidden Markov Model, 2012 . Banasthali Vidyapith Jaipur(Raj.), INDIA.