



Named Entity Recognition for South Asian Languages

by

Sathish Chandra Pammi
Veera Raghavendra
Kishore Prahallad

[Agenda]

- Introduction & Motivation
- Data Representation
- Techniques Used
- Experiments
- Results
- Issues

Introduction & Motivation

- Identifying proper names and classify them into predefined categories
 - First step for Information Extraction
 - Example: “APSRTC” is an Organization
- When NER meets the WEB
 - Find structured data from the unstructured data
 - Help user information search tasks
- Identifying the Named Entities for South Asian Languages
- Rule based or Statistical based approaches can be used for NER

Statistical Approaches for NER

- Hidden Markov Models (HMM).
- Decision Forest.
- Maximum Entropy.
- Support Vector Machines (SVM).
- Conditional Random Fields (CRF).

Data Representaion

| | |
|----------------|---------------------------------|
| prashna NULL | bhaaratiiya BNETE |
| gambhiira NULL | san:giita INETE |
| hai NULL | aura NULL |
| aura NULL | san:skrxti NULL |
| isakii NULL | kei NULL |
| mahimaa NULL | vishhaya NULL |
| koo NULL | mein: NULL |
| sviikaara NULL | doo-eika BNEN |
| karatei NULL | baatein: NULL |
| huei NULL | kahanaa NULL |
| apanei NULL | chaahuun~gaa NULL |
| savaala NULL | . NULL |
| para NULL | . O |
| aanei NULL | |
| sei NULL | |
| pahalei NULL | |
| , NULL | |

[Techniques Used]

- Hidden Markov Models
- Conditional Random Fields
- Decision Forest

NE Boundary Identification

- Named Entity Boundaries are predicted using HMM.
- Ex:
- Results: 68.674

| | |
|----------------|---------------------|
| prashna NULL | bhaaratiya B |
| gambhiira NULL | san:giita I |
| hai NULL | aura NULL |
| aura NULL | san:skrxti NULL |
| isakii NULL | kei NULL |
| mahimaa NULL | vishhaya NULL |
| koo NULL | mein: NULL |
| sviikaara NULL | doo-eika B |
| karatei NULL | baatein: NULL |
| huei NULL | kahanaa NULL |
| apanei NULL | chaahuun~gaa |
| savaala NULL | NULL |
| para NULL | . NULL |
| aanei NULL | . O |
| sei NULL | |
| pahalei NULL | |
| , NULL | |

[Experiment with HMM]

- Predicting the NE of the word using the input word

- Feature set:

<word> <ner - predictee>

⋮

. O

- Result: 47.6867

Experiments With CRF (Experiment1)

- Named Entities are predicted using the word and POS tag of the word.
- Feature: <word> <POS> <predictee>
:
:
.. O
- Result: 47.564

More about CRF : <http://crfpp.sourceforge.net/#source>

[Experiment2]

- Named Entities are predicted using the word, NE boundary and POS tag of the word.
- Feature: <word> <NE Boundary> <POS> <predictee>
:
:
... O
- Result: 48.213

[Experiment3]

- Named Entities are predicted using the word, POS tag and sub-word units of the word.
- Feature: <word> <POS> <1st syllable> <2nd syllable onset> <2nd syllable> <3rd syllable onset> <3rd syllable> <last 2 syllables> <predictee>
:
:
..... O
- Result: 46.767

[Experiment4]

- Named Entities are predicted using the word, POS tag and sub-word units of the word.
- Feature: <word> <POS> <1st syllable> <2nd syllable onset> <2nd syllable> <3rd syllable onset> <3rd syllable> <last 2 syllables> <predictee>
:
:
..... O
- Result: 47.547

[Decision Forest]

- A decision forest is a collection of decision trees
- These trees can be formed by various methods by different sub-samples of observations
- Using a voting method, a class attributed to observation x is a class which is preferred by majority of trees.
- We applied slightly modified decision forest algorithm
- Majority of non-NULL observation is considered instead of all the observations.

Results

Tag Level:

| HMM | CRF1 | CRF2 | CRF3 | CRF4 | DF | Mod.DF |
|---------|--------|--------|--------|--------|---------|---------------|
| 47.6867 | 47.564 | 48.213 | 46.767 | 47.547 | 49.6765 | 59.413 |

Chunk Level:

| Development Data | Test Data |
|--|---|
| Maximal Precision: 0.51411 Maximal Recall: 0.52462 Nested Precision: 0.51411 Nested Recall: 0.52462 Maximal F-Measure: 0.51931 Nested F-Measure: 0.519316 | Maximal Precision: 0.64762 Maximal Recall: 0.42503 Nested Precision: 0.67133 Nested Recall: 0.41312 Maximal F-Measure: 0.51323 Nested F-Measure: 0.51149 |

[Issues]

- Need to improve POS tagging, and NER performance
- Need to explore more features.
- Typographical mistakes.
- Improper Data.



Questions