ISM@FIRE -2013
NER Indian Languages Task

Dinesh Kumar Prabhakar
Gopashree Panda
Sukomal Pal

Department of Computer Science & Engineering
Indian School of Mines, Dhanbad, India

Contents

- Introdution
- FIRE Task
- Approaches
- Result
- Analysis
- Conclusion
- References

Introduction

- Process of identifying a proper noun in a given document
- Can be name of a person, place, organisation
- Important in designing system such as information extraction systems, machine translation systems
- Participated in NER for english

FIRE Task

- Corpus contains 80 text file (training data)
- Phrase (test data) was been given in column format along with **POS** tags, **Chunk** tags
  - (e.g.- Agra NNP B-NP)
- We have to design a system that can produce the output with different level **NE** tags
  - (e.g.- Agra NNP B-NP B-LOCATION B-PLACE B-CITY)

Approach-I

- Preprocessing
  - We observed 21 type of named-entities are there in training corpus (e.g.- location, person etc.)
  - Created 21 different text file
  - Stored the entities in respective file based on 1st level of NE tag in training data file

Tagging

- Step 1: Read a word along with its POS and chunk tags
- Step 2: Find match in entitie's files along with POS and chunk tag
  - If match found
    - Print word with POS tag, Chunk tag, NE1,NE2,NE3
      - Agra NNP B-NP B-LOCATION B-PLACEB-CITY
  - Else
    - Print word with POS tag, Chunk tag, 0,0,0
      - Agra NNP B-NP 0 0 0

Approach-II

- Preprocessing
  - We observed 21 type of named-entities are there in training corpus (e.g.- location, person, artifact etc.)
  - Stored the entities in respective file based on 1st level of NE tag in training data file

Tagging

Step 1: Read words before “B” chunk comes along with its POS and chunk tags

Step 2: Find match in entitie's files along with POS and chunk tags

- If match found
  - Print words with POS tag, Chunk tag, NE1,NE2,NE3
- Else
  - Print word with POS tag, Chunk tag, 0,0,0

Results

Institute	Precision	Recall	F-Score
TRDDC Sys-1	64.79	67.23	65.99
TRDDC Sys-2	64.92	68.63	66.73
ISM Sys-1	14.89	32.02	20.33
ISM2 Sys-2	39.33	34.46	36.47

Analysis

- 1st system may give inappropriate tags
  - For out-of-dictionary word assigning 0,0,0 to NE tags
  - We have not considerd the chunk order while matching(B-begning,I-intermediate)
- 2nd system may give inappropriate tags
  - For out-of-dictionary word it will give 0,0,0 to NE tags
  - If similar phrase are there gives good result

Conclusion

- System is lookup-based
- It is language independent
- Performance may improve if words are stored based on sub-categories
- Need to consider ML-based approach

References

1. David Nadeau, Satoshi Sekine: A survey of named entity recognition and classification National Research Council Canada / New York University, 2007
2. Lev Ratinov and Dan Roth: Design challenges and misconceptions in named entity recognition.

THANK YOU