TRDDC @ FIRE 2013

System for NER in Indian Languages

Forum for Information Retrieval Evaluation

December, 2013

Nitin Ramrakhiyani Sachin Pawar

Tata Research Development and Design Centre

Tata Consultancy Services Ltd.

Introduction

Introduction

- Submitted as a participation in the NER for Indian Languages Track

Introduction

- Submitted as a participation in the NER for Indian Languages Track
- Submitted two runs for English and one for Hindi

Introduction

- Submitted as a participation in the NER for Indian Languages Track
- Submitted two runs for English and one for Hindi
- Data

Introduction

- Submitted as a participation in the NER for Indian Languages Track
- Submitted two runs for English and one for Hindi
- Data
  - Training
    - POS tag and chunk tag available as features
    - Three level classification
    - We focus only on classification at level 1

Introduction

- Submitted as a participation in the NER for Indian Languages Track
- Submitted two runs for English and one for Hindi
- Data
  - Training
    - POS tag and chunk tag available as features
    - Three level classification
    - We focus only on classification at level 1
  - Testing
    - POS tag and chunk information available as features

Methodology

[1] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001

Methodology

- Named Entity Recognition is a sequential labeling task.

[1] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001

Methodology

- Named Entity Recognition is a sequential labeling task.
- Extending the given set of features to represent each token.

[1] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001

Methodology

- Named Entity Recognition is a sequential labeling task.
- Extending the given set of features to represent each token.
- Trained a Conditional Random Fields(CRF)[1] classifier using training data

[1] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001

Methodology

- Named Entity Recognition is a sequential labeling task.
- Extending the given set of features to represent each token.
- Trained a Conditional Random Fields(CRF)[1] classifier using training data
- Testing the classifier model on the test set

[1] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001

Methodology

- Named Entity Recognition is a sequential labeling task.
- Extending the given set of features to represent each token.
- Trained a Conditional Random Fields(CRF)[1] classifier using training data
- Testing the classifier model on the test set
- Evaluation

[1] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001

Feature Set

Feature Set

- Word Structure
  - Cc for token with first letter in capitals
  - bb for token with all letters in small
  - AA for token with all letters in capitals
  - 11 for token with all digits
  - a1 for token with digits and letters

Feature Set

- Word Structure
  - Cc for token with first letter in capitals
  - bb for token with all letters in small
  - AA for token with all letters in capitals
  - 11 for token with all digits
  - a1 for token with digits and letters
- Known Lower: To check if a Cc or AA word ever present in small

Feature Set

- Word Structure
  - Cc for token with first letter in capitals
  - bb for token with all letters in small
  - AA for token with all letters in capitals
  - 11 for token with all digits
  - a1 for token with digits and letters
- Known Lower: To check if a Cc or AA word ever present in small
- Prefix: Token prefixes of length 2, 3 and 4

Feature Set

- Word Structure
  - Cc for token with first letter in capitals
  - bb for token with all letters in small
  - AA for token with all letters in capitals
  - 11 for token with all digits
  - a1 for token with digits and letters
- Known Lower: To check if a Cc or AA word ever present in small
- Prefix: Token prefixes of length 2, 3 and 4
- Suffix: Token suffixes of length 2, 3 and 4

Feature Set

- Word Structure
  - Cc for token with first letter in capitals
  - bb for token with all letters in small
  - AA for token with all letters in capitals
  - 11 for token with all digits
  - a1 for token with digits and letters
- Known Lower: To check if a Cc or AA word ever present in small
- Prefix: Token prefixes of length 2, 3 and 4
- Suffix: Token suffixes of length 2, 3 and 4
- Next Verb and Previous Verb

Feature Set

- Word Structure
  - Cc for token with first letter in capitals
  - bb for token with all letters in small
  - AA for token with all letters in capitals
  - 11 for token with all digits
  - a1 for token with digits and letters
- Known Lower: To check if a Cc or AA word ever present in small
- Prefix: Token prefixes of length 2, 3 and 4
- Suffix: Token suffixes of length 2, 3 and 4
- Next Verb and Previous Verb
- Context Words: Bag-of-Words in a window of ± 2

Feature Set

[2] P. Bhattacharyya, "IndoWordNet. Lexical Resources Engineering Conference 2010," *Malta, May*, 2010

[3] IIT Bombay, "Hindi wordnet 1.3," `http://www.cfilt.iitb.ac.in/wordnet/webhwn/`, 2010

[4] H. L. Chieu and H. T. Ng, "A maximum entropy approach to information extraction from semi-structured and free text," *AAAI/IAAI*, vol. 2002, pp. 786–791, 2002

Feature Set

- WordNet
  - To check if the token has an ancestor hypernym as ORGANIZATION or PERSON or LOCATION
  - English WordNet (Java WordNet Library)
  - Hindi WordNet (Developed by IIT Mumbai [2] [3])

[2] P. Bhattacharyya, "IndoWordNet. Lexical Resources Engineering Conference 2010," *Malta, May*, 2010

[3] IIT Bombay, "Hindi wordnet 1.3," `http://www.cfilt.iitb.ac.in/wordnet/webhwn/`, 2010

[4] H. L. Chieu and H. T. Ng, "A maximum entropy approach to information extraction from semi-structured and free text," *AAAI/IAAI*, vol. 2002, pp. 786–791, 2002

Feature Set

- WordNet
  - To check if the token has an ancestor hypernym as ORGANIZATION or PERSON or LOCATION
  - English WordNet (Java WordNet Library)
  - Hindi WordNet (Developed by IIT Mumbai [2] [3])
- Point Wise Mutual Information
  - To capture affinity towards an output class
  - $PMI_k(Token, Class) = \log(\frac{Pr(Token \text{ and } Class \text{ occur within window of } k \text{ tokens})}{Pr(Class)Pr(Token)})$
  - Inspired from correlation in Chieu and Ng[4]

[2] P. Bhattacharyya, "IndoWordNet. Lexical Resources Engineering Conference 2010," *Malta, May*, 2010

[3] IIT Bombay, "Hindi wordnet 1.3," `http://www.cfilt.iitb.ac.in/wordnet/webhwn/`, 2010

[4] H. L. Chieu and H. T. Ng, "A maximum entropy approach to information extraction from semi-structured and free text," *AAAI/IAAI*, vol. 2002, pp. 786–791, 2002

Evaluation

Evaluation

- Divided the training set into 80:20 percent ratio. Used the 80% for Training and the rest as Development

Evaluation

- Divided the training set into 80:20 percent ratio. Used the 80% for Training and the rest as Development
- Trained a Conditional Random Fields(CRF) classifier with the above feature sets and data

Evaluation

- Divided the training set into 80:20 percent ratio. Used the 80% for Training and the rest as Development
- Trained a Conditional Random Fields(CRF) classifier with the above feature sets and data
- Different feature combinations for both Hindi and English were tried

Evaluation

- Divided the training set into 80:20 percent ratio. Used the 80% for Training and the rest as Development
- Trained a Conditional Random Fields(CRF) classifier with the above feature sets and data
- Different feature combinations for both Hindi and English were tried
- Performance Measurement: F1 measure calculated using the CoNLL evaluation script

Evaluation

Evaluation

- Performance on English Development Data
  - With all but PMI and WN: 0.58
  - With all but PMI: 0.59
  - All Features: 0.62

Evaluation

- Performance on English Development Data
  - With all but PMI and WN: 0.58
  - With all but PMI: 0.59
  - All Features: 0.62
- Performance on Hindi Development Data
  - With all but PMI and WN: 0.48
  - With all but PMI: 0.48
  - All Features: 0.51

Questions ?