Sequence Labeling: Shallow Parsing

Natural Language Understanding Lab

Evgeny A. Stepanov,
Mahed Mousavi, Gabriel Roccabruna

SISL, DISI, UniTN & VUI, Inc.
evgeny.stepanov@unitn.it

Objectives

- Understanding:
  - relation between Sequence Labeling and Shallow Parsing
  - IOB Notation
  - Joint Segmentation and Classification
  - Feature Engineering
- Learning how to:
  - use Named Entity Recognition with `spacy` and `NLTK`
  - train, test, and evaluate Conditional Random Fields models
  - perform feature engineering with CRF

Outline

1. Sequence Labeling and Shallow Parsing
2. Shallow Parsing
3. Encoding Segmentation Information: CoNLL Corpus Format
   - IOB Scheme
4. Named Entity Recognition with `NLTK`
   - *Exercise*: 15 min
5. Named Entity Recognition with `spacy`

Signals & Interactive Systems

Outline

6. Shallow Parsing with Conditional Random Fields
   - Python CRF Tutorial
   - Baseline Shallow Parsing with CRFs
7. Feature Engineering
   - SpaCy Token Features
   - Adding Features to CRF
8. ***Lab Exercise***: Feature Engineering with CRF: 40 min

Recommended Reading

- Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed. draft)
  - Chapter 8: Sequence Labeling for Parts of Speech and Named Entities
- Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python
  - Chapter 7: Extracting Information from Text
- Conditional Random Fields
  - Lafferty et al. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data (original paper)
  - Sutton & McCallum's An Introduction to Conditional Random Fields
  - Edwin Chen's Introduction to Conditional Random Fields
  - Michael Collin's Log-Linear Models, MEMMs, and CRFs