Assignment:
Named Entity Recognition

Empirical Methods in Natural Language Processing

Philipp Koehn and Annette Leonhard

29 January 2007

based on the 2006 slides by Sebastian Riedel

Outline

1. **Introduction**
   - Information Extraction
   - Named Entity Recognition
   - CoNLL Shared Task
2. **Choices**
3. **Assessment**

Information Extraction

- **Extract information salient to the needs of the users**
  - Information about house prices from real estate magazines
  - Character relations from novels
  - Location of terrorist attacks from newspapers
- **Extract structured data from unstructured or semi structured natural language data, e.g. from newspapers**
- **Task involving Natural Language Understanding and Information Retrieval**

Information Extraction Tasks

- **Named Entity Recognition**
  - Which phrases refer to **what kind of entities**
- **Coreference Resolution**
  - Which phrases refer to the **same entity**
- **Relation Extraction**
  - Which entities are related in **what kind of relationships**
- **Event Extraction**
  - Which events are mentioned with which attributes

Named Entity Recognition

- **Named entity** is an object of interest such as a person, organization, or location
- Identifying word sequences
- Labelling those sequences

Example:

Meg Whitman, CEO of **eBay**, said in **New York**…

- Label Meg Whitman as **PERSON**
- Label eBay as **ORGANISATION**
- Label New York as **LOCATION**

CoNLL Shared Task 2003

- Brings together researchers in Computational Natural Language Learning
- Aims at evaluating different Machine Learning approaches
- Gives training, development and test sets for NER in German and English
- Identify entities and classify as **PERSON**, **LOCATION**, **ORGANISATION** and **MISC**

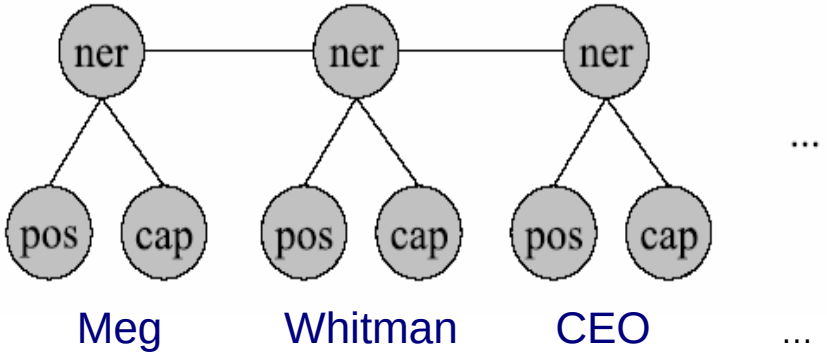
IOB Scheme in CoNLL

- **I**nside, **O**utside, **B**egin
- For each type of entity there is an I-XXX and a B-XXX tag
- Non-entities are tagged O
- B-XXX only used if two entities of same type next to each other
- Assumes that named entities are non-recursive and don‘t overlap

Example:

Meg	Whitman	CEO	of	eBay
I-PER	B-PER	O	O	I-ORG

A Graphical Model for NER



- The NER framework covers
  - Features
  - Local classifiers
  - Sequential constraints

Features

- **Features are the most important aspect of almost every Machine Learning system**
  - Is the word capitalised?
  - Is the word at the start of a sentence?
  - What is the POS tag?
  - Info from gazetteers
- **The more useful features you incorporate, the more powerful your learner gets**

Local Classifier

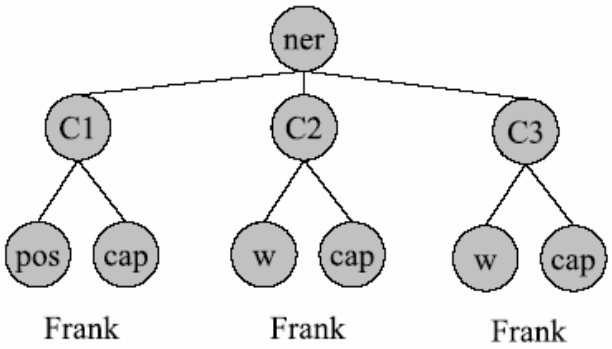
Find

p(*tag*|*features*)

- Maximum Entropy Classifier (Berger et al. 1996)
- Large Margin approach such as support vector machines (SVMs) (Vapnik 1995)
- Naive Bayes (strong independence assumption)
- Whatever you like

Ensemble Methods

- Take a set of diverse classifiers
- Let them vote on the tag of a single token (or average their probabilistic output)
- Diversity through different feature sets, different learners, different training data (Dietterich 2000)



Sequential Modelling

- Tags interdepend

$$p(t_1, t_2, t_3 \ldots | f_1, f_2, f_3 \ldots) \neq \prod_i^n p(t_i | f_i)$$

- Could use a model such as:

$$p(t_1, t_2, t_3 \ldots | f_1, f_2, f_3 \ldots) = p(t_1 | f_1) \prod_{i=2}^n p(t_i | f_i)\, p(t_i | t_{i-1})$$

Software

- Use any programming language you want
- Try to find good toolkits
  - Maxent Toolkit of Zhang Lee (very good and fast training)
  - CRF++ framework (supports sequential modelling)
  - Weka (easy to use but memory intensive and slow)
  - SVM light, LibSVM (long training time, usually good performance)

Timetable

20 & 21/02 Presentation of the results for your baseline system

16/03 Hand in your paper and code!

Assessment Criteria

- **Quality of paper**
  - Structure
  - Use of literature
  - Error Analysis
- **Performance of your system**
- **Creativity**