Assignment: Named Entity Recognition
Data Intensive Linguistics 2005

Philipp Koehn and Sebastian Riedel

30. January 2006
Outline

1. Introduction
   - Information Extraction
   - Named Entity Recognition
   - CoNLL Shared Task

2. Choices

3. Assessment
Information Extraction

- Extract structured information (say, a Database) from unstructured or semi structured natural language data
  - Terrorist attack information from newspapers
  - Protein-protein interactions from biomedical papers
  - Character relations from novels
- Initially driven by the Message Understanding Conferences (MUC)
- Somewhere between Natural Language Understanding and Information Retrieval
The Information Extraction Subtasks

- 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

Example

...Frank Kuhn, CEO of Whatever You Say, said in San Diego...

- detected 3 segments
- classified into PERSON, ORGANISATION and LOCATION
CoNLL Shared Task 2003

- Given: Training, development and test sets for NER in German and English
- Identify entity phrases and classify into PERSON, LOCATION, ORGANISATION and MISC.
- Bring together researchers in the area of Computational Natural Language Learning
- Aims at evaluating the millions of Machine Learning approaches out there
### BIO Scheme in CoNLL 2003

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>American</td>
<td>Frank</td>
<td>Kuhn</td>
<td>CEO</td>
<td>of</td>
<td>Whatever</td>
<td>You</td>
<td>Say</td>
<td></td>
</tr>
<tr>
<td>I-ORG</td>
<td>B-ORG</td>
<td>I-PER</td>
<td>I-PER</td>
<td>O</td>
<td>O</td>
<td>I-ORG</td>
<td>I-ORG</td>
<td>I-ORG</td>
<td></td>
</tr>
</tbody>
</table>

- Very confusing since B tags are only used if a new entity of the same type starts after an entity
A Graphical Model for NER

The general NER framework covers
- Features,
- local classifiers and
- sequential constraints
Features

- The most important aspect of almost every ML system
- The easier it is to incorporate your own intuitions in the form of many “overlapping” features the more powerful a learner is!
- Features to consider: POS tags, Words, Gazetteer information, capitalization information, contextual versions of the above.
Local Classifier

Find

\[ p(tag|features) \]

- Maximum Entropy Classifier (Berger et al. 1996)
- Large Margin approach, such as SVM (Vapnik 1995)
- Naive Bayes (strong independence assumption)
- Whatever you like
Ensemble Methods

- Take a set of sufficiently diverse classifiers (different views)
- 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 (A “B” tag can not follow an “O” tag etc.) or

\[ p(t_1, t_2, t_3 \ldots \mid f_1, f_2, f_3 \ldots) \neq \prod_{i} p(t_i \mid f_i) \]

- Could for instance use Viterbi and a model such as

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

- Use any programming language you like!
- Try to find good toolkits!
  - Maxent Toolkit of Zhang Le! (very good and fast training)
  - CRF++ framework (supports sequential modelling)
  - Weka (easy to use but memory intensive, slow)
  - $SVM_{light}$, LibSVM (long training time, usually good performance)
Timetable

7.2.2006  Presentation of preliminary results for the baseline system you developed

13.3.2006  Hand in of paper and code!
Assessment Criteria

- Quality of the paper
  - Structure
  - Use of Literature
  - Error Analysis
- Performance of your system
- Creativity