Overview

1. Who? What? Why?
2. Formalities
3. Agenda
4. Named Entity Recognition
5. GATE
Overview

1. Who? What? Why?
2. Formalities
3. Agenda
4. Named Entity Recognition
5. GATE
Who

Johannes Knopp
- Magister in Computational Linguistics and Computer Science
- Interested in Named Entities, Word Sense Disambiguation
- Contact: johannes@informatik.uni-mannheim.de

Cäcilia Zirn
- Bachelor + Master in Computational Linguistics and Computer Science
- Interested in Opinion Mining, Stylometry, Information Retrieval
- caecilia@informatik.uni-mannheim.de

The “○” hopefully keeps spammers away while you just replace it with “@”
Definition: Computational Linguistics

“Computational linguistics (CL) is a discipline between linguistics and computer science which is concerned with the computational aspects of the human language faculty. It belongs to the cognitive sciences and overlaps with the field of artificial intelligence (AI), a branch of computer science aiming at computational models of human cognition. Computational linguistics has applied and theoretical components.”

\(^a\)Hans Uszkoreit,

http://www.colic.uni-saarland.de/~hansu/what_is_cl.html
We will try to help you automatizing some of your tasks.

Goals

- Don’t worry, we won’t to turn you into a geeky programmer
Goals

Participants will be able to . . .

- . . . automatize repetitive tasks with basic programming skills
- . . . extract information from semi-structured documents
- . . . use software for language processing and content analysis
Example: Estimate party positions with WordFish

Example: Estimate fine grained party positions
Example: Twitter sentiment

Sentiment analysis for berlusconi

<table>
<thead>
<tr>
<th>Sentiment by Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive (54%)</td>
</tr>
<tr>
<td>Negative (46%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sentiment by Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive (3939)</td>
</tr>
<tr>
<td>Negative (3315)</td>
</tr>
</tbody>
</table>

Evaluating content of twitter feeds automatically and finding trends in opinion
Do you have tasks that are beyond the scope of this seminar?

- There is a support center offering its service
- [http://service.informatik.uni-mannheim.de/de/das-center/](http://service.informatik.uni-mannheim.de/de/das-center/)
Overview

1. Who? What? Why?

2. Formalities

3. Agenda

4. Named Entity Recognition

5. GATE
### Properties
- 2 SWS + homework
- 6 ECTS
- Language: English

### Requirements
- Lecture-part: *voluntary* attendance
- Project presentation part: *mandatory* attendance (all dates)
- Implementation of individual project

### Further information & teaching materials
### Organization

- 1st part: Lecture (+ ungraded exercises)
- 2nd part: Individual projects

### Grading

Individual project:
- Presentation (1/3)
- Content (extent and quality of project) (1/3)
- Written report (ca. 15 pages) (1/3)
Overview

1. Who? What? Why?
2. Formalities
3. Agenda
4. Named Entity Recognition
5. GATE
Introduction to Named Entity Recognition & GATE

- Today! :) 
- Automatically recognizing persons, organizations & locations in text
Regular Expressions

- Regular expressions search for patterns and mark (or replace) parts of text

Here is an example for regular expressions.
We can just search for a specific letter.
Or if you have file names, you can choose all files with a specific ending.

- loveSong.mp3
- myDiarySecret.doc
- mySong1.mp3
- mySong2.mp3
- privateStuff.doc
Regular Expressions

- Regular expressions search for patterns and mark (or replace) parts of text

Here is an example for regular expressions. We can just search for a specific letter. Or if you have file names, you can choose all files with a specific ending.

loveSong.mp3
myDiarySecret.doc
mySong1.mp3
mySong2.mp3
privateStuff.doc
Regular Expressions

- Regular expressions search for patterns and mark (or replace) parts of text

```
./doc
```

- Checkboxes: `global`, `ignoreCase`, `extended`, `dotall`, `multiline`

Here is an example for regular expressions.
We can just search for a specific letter.
Or if you have file names, you can choose all files with a specific ending.

```
loveSong.mp3
myDiarySecret.doc
mySong1.mp3
mySong2.mp3
privateStuff.doc
```
Introduction to Python

- Object-oriented programming language
- Easy syntax

You will learn the general basic principles of programming

```python
>>> if student == 'diligent':
...     print 'passed course :)
...     else:
...         print 'failed'
... passed course :)
```
Text preprocessing with NLTK

- Toolkit for Python for Natural Language Processing
- Tokenizing, Stemming, Parsing, ...
Crawling Websites & Document Conversion

- Getting masses of data from the web without using copy & paste

### Annex I countries

<table>
<thead>
<tr>
<th>Annex I countries</th>
<th>First national communication</th>
<th>Second national communication</th>
<th>Third national communication (due 30 November 2001)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Australia</strong></td>
<td>☃️ NC 1 (1164kB)</td>
<td>☃️ NC 2 (1098kB)</td>
<td>☃️ NC 3 (788kB) Submission date: 16/08/02 GCOS report</td>
</tr>
<tr>
<td><strong>Austria</strong></td>
<td>☃️ NC 1 (309kB)</td>
<td>☃️ NC 2 (3516kB)</td>
<td>☃️ NC 3 (3360kB) Submission date: 29/11/01 GCOS report</td>
</tr>
</tbody>
</table>
Information Retrieval

- Finding relevant information
- Classic application: search engine
- Finding documents that are similar to a query
- Representations of texts as Word Vectors

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>change</th>
<th>climate</th>
<th>degree</th>
<th>high</th>
<th>rather</th>
<th>the</th>
</tr>
</thead>
<tbody>
<tr>
<td>text 1</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>text 2</td>
<td>10</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>text 3</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

Example:
Searching all documents talking about climate change in a collection of documents.
Text Classification & Machine Learning

- Distinguishing different classes of text (automatically)
- Train a classifier on labelled data
- Predict classes for unseen data

Classify topics, sentiment, party positions, authors,....
Project Assignments

- Bring your own problem ;)
- Have an idea
- Pick a project
Overview

1. Who? What? Why?
2. Formalities
3. Agenda
4. Named Entity Recognition
5. GATE
Pope Benedict XVI waded into a crowd of well-wishers in Rome on Sunday, just days after he was knocked down by a woman at a Christmas Eve Mass. It was the 82-year-old pontiff’s first appearance outside the Vatican since the attack...
Pope Benedict XVI waded into a crowd of well-wishers in Rome on Sunday, just days after he was knocked down by a woman at a Christmas Eve Mass. It was the 82-year-old pontiff’s first appearance outside the Vatican since the attack...
Pope Benedict XVI waded into a crowd of well-wishers in Rome on Sunday, just days after he was knocked down by a woman at a Christmas Eve Mass. It was the 82-year-old pontiff’s first appearance outside the Vatican since the attack...
Pope Benedict XVI waded into a crowd of well-wishers in Rome on Sunday, just days after he was knocked down by a woman at a Christmas Eve Mass. It was the 82-year-old pontiff’s first appearance outside the Vatican since the attack...
Pope Benedict XVI waded into a crowd of well-wishers in Rome on Sunday, just days after he was knocked down by a woman at a Christmas Eve Mass. It was the 82-year-old pontiff’s first appearance outside the Vatican since the attack...
Pope Benedict XVI waded into a crowd of well-wishers in Rome on Sunday, just days after he was knocked down by a woman at a Christmas Eve Mass. It was the 82-year-old pontiff’s first appearance outside the Vatican since the attack...
Why Named Entities?

Named Entities (NEs) are the actors in a text. Thus Named Entity Recognition (NER) is an important subtask of information extraction that allows for advanced search queries.

Search Query

*Find all speeches of Barack Obama where he is talking about Gaza.*
How are Named Entities recognized?

There are linguistic grammar-based approaches as well as statistical models that have to be trained on given data. The classifiers can be supervised, semi-supervised or unsupervised. For supervised and semi-supervised methods a set of manually annotated data is needed which can be costly and cumbersome to create. Unsupervised methods do not need annotated data which results in less precision but higher recall. Usually all classifiers rely on linguistic preprocessing like tokenization, part-of-speech tagging, phrase chunking or even parsing. All of those are already error prone so the used tools can have a big impact on the final results of a system.
How are Named Entities recognized?

There are linguistic grammar-based approaches as well as statistical models that have to be trained on given data. The classifiers can be supervised, semi-supervised or unsupervised. For supervised and semi-supervised methods, a set of manually annotated data is needed which can be costly and cumbersome to create. Unsupervised methods do not need annotated data which results in less precision but higher recall. Usually, all classifiers rely on linguistic preprocessing like tokenization, part-of-speech tagging, phrase chunking or even parsing. All of those are already error prone so the used tools can have a big impact on the final results of a system... 

There are algorithms who do that for you :-)

31/46
## Overview

1. **Who? What? Why?**
2. **Formalities**
3. **Agenda**
4. **Named Entity Recognition**
5. **GATE**
GATE
General Architecture for Text Engineering

- Automatized preprocessing and annotation of text corpora
- Custom annotations using gazetteers or self-defined rules

Software & Documentation: http://gate.ac.uk/
Using GATE

- Load Corpus
- Run preprocessing (*ANNIE*)
- Analyze Annotations
GATE 6.0 build 3764 started at Wed Sep 07 14:04:02 CEST 2011
and using java 1.6.0_23 Sun Microsystems Inc. on Linux amd64 3.0.0-1-amd64.
CREDE plugin loaded: file:/home/knopp/programs/GATE-6.0/plugins/ANNIE/
As we face the current economic crisis we will retain our strong international alliances to secure the action needed to create jobs, fight climate change and build a fairer world.

Today we live in an age of unprecedented change. As the economic hurricane sweeping the world has shown, events outside our borders have a direct effect on us here in Britain making our strong engagement in the European Union more important than ever.

As we face the current economic crisis we will retain our strong international alliances to secure the action needed to create jobs, fight climate change and build a fairer world.

In these difficult times of economic uncertainty our membership of the European Union is more important than ever with millions of jobs depending on it. A confident, outward-looking European Union can become a global player, taking advantage of the opportunities of globalisation, protecting us from its dangers and setting a new agenda that emphasises openness, fairness and the importance of co-operation between nations. Isolation and withdrawal have never been the answer and that applies now more than ever. By acting decisively and determinedly at all levels: local, national and European - we will win the fight for Britain's future.

Over the last 12 years Labour has been working hard to deliver on what matters to people across Britain. Labour's action, from record investment in education and the National Health Service to making our streets safer, has been driven by our core value of fairness. Now more than ever, in the changing world in which we live and facing unprecedented global economic challenges, that core value will not change. Our priority, as it has always been, is to create a fairer, safer, more prosperous Britain in which everyone can take advantage of their full potential.

Labour has a strong record at home and in Europe of fighting for measures to improve life for hardworking people. Our recent decision to sign up to the EU Social Chapter has given new rights and benefits to British workers and their families, including the right to take guaranteed paid holiday, maternity leave and parental leave. We have also worked hard in Europe to secure better consumer rights, higher standards and lower prices for people across Britain.

Today the threats that countries around the world face are different from those of the past. Thanks to our action in Europe to tackle shared problems, Britain is more secure. We will continue working with our European partners to tackle crime, illegal immigration and the threat of terrorism to keep our streets safer.

Together with our partners in Europe we have led the fight against climate change. The commitments we agreed, along with other European countries, have placed the European Union firmly at the forefront of the battle against climate change and the drive to build a low-carbon economy. Remaining engaged at the heart of Europe to tackle climate change remains vital to protect our environment for generations to come.

In these difficult times, isolation and cutting ourselves off from European allies is not the answer. The Conservatves’ policies would lead to years of instability in our relations with Europe.

Working to build a fairer world for all remains central to Labour’s core values. The European Union should and can be a model of how in the world, building on its leading position on international development, climate change and economic reform to help spread peace and democracy worldwide.

We know that the European Union needs further reform. Through the Labour Party’s full engagement in Europe we can ensure that the European Union becomes more effective and accountable to its citizens, insisting on transparency and reform to secure change in the interests of our citizens will remain one of our priorities in Europe.

In these difficult times, isolation and cutting ourselves off from European allies is not the answer. The Conservatves’ policies would lead to years of instability in our relations with Europe.

Working to build a fairer world for all remains central to Labour’s core values. The European Union should and can be a model of how in the world, building on its leading position on international development, climate change and economic reform to help spread peace and democracy worldwide.

We know that the European Union needs further reform. Through the Labour Party’s full engagement in Europe we can ensure that the European Union becomes more effective and accountable to its citizens, insisting on transparency and reform to secure change in the interests of our citizens will remain one of our priorities in Europe.

In these difficult times, isolation and cutting ourselves off from European allies is not the answer. The Conservatves’ policies would lead to years of instability in our relations with Europe.

Working to build a fairer world for all remains central to Labour’s core values. The European Union should and can be a model of how in the world, building on its leading position on international development, climate change and economic reform to help spread peace and democracy worldwide.

We know that the European Union needs further reform. Through the Labour Party’s full engagement in Europe we can ensure that the European Union becomes more effective and accountable to its citizens, insisting on transparency and reform to secure change in the interests of our citizens will remain one of our priorities in Europe.

In these difficult times, isolation and cutting ourselves off from European allies is not the answer. The Conservatves’ policies would lead to years of instability in our relations with Europe.

Working to build a fairer world for all remains central to Labour’s core values. The European Union should and can be a model of how in the world, building on its leading position on international development, climate change and economic reform to help spread peace and democracy worldwide.

We know that the European Union needs further reform. Through the Labour Party’s full engagement in Europe we can ensure that the European Union becomes more effective and accountable to its citizens, insisting on transparency and reform to secure change in the interests of our citizens will remain one of our priorities in Europe.
ANNIE Components

ANNIE English Tokenizer

- Splits text into tokens
- Distinguishes between *Punctuation*, *SpaceToken*, *Word*, …
- Adds additional attributes like
  - kind: word/number
  - orth: lowercase/UPPERCASE/CaMeIlCase/Upperinitial
  - length
  - …
ANNIE Components

ANNIE Gazetter

- Looks up words in predefined wordlists
- Adds *Lookup* annotations which often have a minorType and a majorType attribute
- Lookup lists are stored in text files which can easily be extended
ANNIE Components

ANNIE Sentence Splitter

- Identifies sentences and splits them
- Produces annotations *Sentence* and *Split*
ANNIE POS Tagger

- Assigns the *Token* annotations an attribute *category* that stores the part of speech (NN = singular or mass noun, . . .). There are 36 POS tags
ANNIE Components

ANNIE NE Transducer

- Set of JAPE grammars that identify semantic units like name, date, address...
ANNIE Components

- Detects synonyms and aliases
Next week: Regular Expressions

- Efficient string matching technique
- Search for character series or patterns of characters
- Written in a formal language (can differ between implementations)
- Can be used in GATE by writing JAPE rules