Advanced Natural Language Processing
Lecture 1
Introduction

Sharon Goldwater
(based on slides by Philipp Koehn)
Other lecturer: Henry Thompson

15 September 2014
What is Natural Language Processing?
Rajendra Prasad
The 1st President of India

List of Presidents of India - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/List_of_Presidents_of_India
The President of India is the head of state and first citizen of India. The President is also the Commander-in-Chief of the Indian Armed Forces. Although the ... Zakhir Hussain - Rajendra Prasad - VV Giri - R. Venkataraman

Rajendra Prasad - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Rajendra_Prasad
listen (help · info); 3 December 1884 – 28 February 1963) was the first President of the Republic of India. An Indian political leader, lawyer by training, Prasad ...
What is Natural Language Processing?

**Applications**
- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

**Core technologies**
- Morphological analysis
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Coreference resolution
- Word sense disambiguation
- Textual entailment
- ...

Sharon Goldwater  
ANLP Lecture 1  
15 September 2014
NLP or computational linguistics?

- Scientist vs. engineer
- Explaining language vs. building applications
- Insight vs. empirical results
**Quotes**

It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

Noam Chomsky, 1969

Whenever I fire a linguist our system performance improves.

Frederick Jelinek, 1988
## This Course

<table>
<thead>
<tr>
<th><strong>Linguistics</strong></th>
<th><strong>Computational methods</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>- words</td>
<td>- finite state machines (morphological analysis, POS tagging)</td>
</tr>
<tr>
<td>- morphology</td>
<td>- grammars and parsing (CKY, statistical parsing)</td>
</tr>
<tr>
<td>- parts of speech</td>
<td>- probabilistic models (HMMS, PCFGs, MaxEnt models)</td>
</tr>
<tr>
<td>- syntax</td>
<td>- vector spaces (distributional semantics)</td>
</tr>
<tr>
<td>- semantics</td>
<td>- lambda calculus (compositional semantics)</td>
</tr>
<tr>
<td>- discourse</td>
<td></td>
</tr>
</tbody>
</table>
Words

This is a simple sentence

Sharon Goldwater  ANLP Lecture 1  15 September 2014
Morphology

This is a simple sentence

be
3sg
present

WORDS
MORPHOLOGY

Sharon Goldwater
ANLP Lecture 1
15 September 2014
Parts of Speech

This is a simple sentence

be

3sg

present

PART OF SPEECH

WORDS

MORPHOLOGY
Syntax

This is a simple sentence

be
3sg
present

SYNTAX
PART OF SPEECH
WORDS
MORPHOLOGY
Semantics

This is a simple sentence
be 3sg present
having few parts
string of words satisfying the grammatical rules of a language

SYNTAX
PART OF SPEECH
WORDS
MORPHOLOGY
SEMANTICS
Discourse

This is a simple sentence be 3sg present

But it is an instructive one.
Why is Language Hard?

- Ambiguities on many levels, need context to disambiguate
- Rules, but many exceptions
- Language is infinite, cannot see examples of everything
Data: Words

- Definition: strings of letters separated by spaces

- But how about:
  - punctuation: commas, periods, etc. typically separated (tokenization)
  - hyphens: high-risk
  - clitics: Joe’s
  - compounds: website, Computerlinguistikvorlesung

- And what if there are no spaces:
  伦敦每日快报指出,两台记载黛安娜王妃一九九七年巴黎死亡车祸调查资料的手提电脑,被从前大都会警察总长的办公室里偷走.
## Word Counts

Most frequent words in the English Europarl corpus (54m wds)

<table>
<thead>
<tr>
<th>any word</th>
<th>Frequency in text</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1,929,379</td>
<td></td>
</tr>
<tr>
<td>,</td>
<td>1,297,736</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>956,902</td>
<td></td>
</tr>
<tr>
<td>of</td>
<td>901,174</td>
<td></td>
</tr>
<tr>
<td>to</td>
<td>841,661</td>
<td></td>
</tr>
<tr>
<td>and</td>
<td>684,869</td>
<td></td>
</tr>
<tr>
<td>in</td>
<td>582,592</td>
<td></td>
</tr>
<tr>
<td>that</td>
<td>452,491</td>
<td></td>
</tr>
<tr>
<td>is</td>
<td>424,895</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>424,552</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nouns</th>
<th>Frequency in text</th>
<th>Content word</th>
</tr>
</thead>
<tbody>
<tr>
<td>European</td>
<td>129,851</td>
<td></td>
</tr>
<tr>
<td>Mr</td>
<td>110,072</td>
<td></td>
</tr>
<tr>
<td>commission</td>
<td>98,073</td>
<td></td>
</tr>
<tr>
<td>president</td>
<td>71,111</td>
<td></td>
</tr>
<tr>
<td>parliament</td>
<td>67,518</td>
<td></td>
</tr>
<tr>
<td>union</td>
<td>64,620</td>
<td></td>
</tr>
<tr>
<td>report</td>
<td>58,506</td>
<td></td>
</tr>
<tr>
<td>council</td>
<td>57,490</td>
<td></td>
</tr>
<tr>
<td>states</td>
<td>54,079</td>
<td></td>
</tr>
<tr>
<td>member</td>
<td>49,965</td>
<td></td>
</tr>
</tbody>
</table>
Word Counts

But also:

There is a large tail of words that occur only once.

33,447 words occur once, for instance

• cornflakes
• mathematicians
• Tazhikhistan
Zipf’s law

\[ f \times r = k \]

- \( f \): frequency of a word
- \( r \): rank of a word (if sorted by frequency)
- \( k \): a constant
Zipf’s law as a graph

\[ fr = k \Rightarrow f = \frac{k}{r} \Rightarrow \log f = \log k - \log r \]
Linguistics and Data

- Data
  - looking at real use of language in text
  - can learn a lot from empirical evidence
  - but: Zipf’s law: there will be always instances that are rarely seen

- Linguistics
  - build a better understanding of language structure
  - linguistic analysis points to what is important
  - but: many ambiguities cannot be explained easily
Course organization

- Lecturers: Sharon Goldwater, Henry Thompson

- Lectures:
  - Mondays 1510-1600, David Hume Tower, Fac Room South
  - Tuesdays 1110-1200 14 Buccleuch Place, 1.01
  - Thursdays 1510-1600, David Hume Tower, Fac Room South

- Labs:
  - This week only: Thursday 1700-1800, Appleton Tower 5.04
  - Thereafter: Tuesdays 1600-1700, Appleton Tower 5.04

- Demonstrators: Luke Shrimpton, Bharat Ram Ambati
Course materials

• All info online at
  http://www.inf.ed.ac.uk/teaching/courses/anlp/


• Further reading: "Foundations of Statistical Natural Language Processing", Manning and Schuetze, 1999, MIT Press., and others listed on web page.

• Labs, assignments, code: all on web page.
Labs and assignments

- Three assignments, worth 10% each (30% total)
- Exam in December, worth 70% of the grade
- Weekly 1-hr labs for exploring data and developing practical skills
- Both labs and assignments will be done in pairs
Background needed for this course?

- Know or currently learning Python

- Background in Linguistics and prepared to learn maths (mainly probability) and algorithms

- Background in CS and prepared to learn linguistics
Preparing for the first lab

- **This Thursday**, 1700-1800, Appleton Tower 5.04

- **Before then**: go through Preliminaries section of Lab 1.
  - Get your DICE account and make sure you can log in to the lab machines in AT (or find a partner who can).
  - Read/work through the Introduction to DICE (linked from the lab) while at a DICE machine.