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

Linguistics
- words
- morphology
- parts of speech
- syntax
- semantics
- discourse

Computational methods
- finite state machines (morphological analysis, POS tagging)
- grammars and parsing (CKY, statistical parsing)
- probabilistic models (HMMS, PCFGs, MaxEnt models)
- vector spaces (distributional semantics)
- lambda calculus (compositional semantics)

Words

This is a simple sentence
Morphology

This is a simple sentence

be
3sg
present

Parts of Speech

This is a simple sentence

DT VBZ DT JJ NN

Syntax

This is a simple sentence

NP
VP
S
NP

Semantics

This is a simple sentence

be
3sg
present

SIMPLE1 string of words satisfying the grammatical rules of a language

sharon goldwater ANLP Lecture 1 15 September 2014
Discourse

This is a simple sentence.

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:

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>
<th>nouns</th>
<th>Frequency in text</th>
<th>Content word</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1,929,379</td>
<td>the</td>
<td>European</td>
<td>129,851</td>
<td>Mr</td>
</tr>
<tr>
<td>,</td>
<td>1,297,736</td>
<td>,</td>
<td>Mr</td>
<td>110,072</td>
<td>commission</td>
</tr>
<tr>
<td>.</td>
<td>956,902</td>
<td>.</td>
<td>president</td>
<td>98,073</td>
<td>parliament</td>
</tr>
<tr>
<td>of</td>
<td>901,174</td>
<td>of</td>
<td>parliament</td>
<td>71,111</td>
<td>union</td>
</tr>
<tr>
<td>to</td>
<td>841,661</td>
<td>to</td>
<td>parliament</td>
<td>67,518</td>
<td>report</td>
</tr>
<tr>
<td>and</td>
<td>684,869</td>
<td>and</td>
<td>report</td>
<td>64,620</td>
<td>council</td>
</tr>
<tr>
<td>in</td>
<td>582,592</td>
<td>in</td>
<td>council</td>
<td>58,506</td>
<td>states</td>
</tr>
<tr>
<td>that</td>
<td>452,491</td>
<td>that</td>
<td>states</td>
<td>57,490</td>
<td>member</td>
</tr>
<tr>
<td>is</td>
<td>424,895</td>
<td>is</td>
<td>member</td>
<td>54,079</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>424,552</td>
<td>a</td>
<td></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

why a line in log-scales? \( fr = k \Rightarrow f = \frac{k}{r} \Rightarrow \log f = \log k - \log r \)

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/
- 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.