Introduction to Natural Language Processing

CMSC 470
Marine Carpuat
“Language is the ability to acquire and use complex systems of communication, particularly the human ability to do so, and a language is any specific example of such a system. The scientific study of language is called linguistics.”

From Wikipedia
Computational Linguistics (CL)
  • The science of doing what linguists do with language, but using computers

Natural Language Processing (NLP)
  • The engineering discipline of doing what people do with language, but using computers

Speech/Language/Text processing
Human Language Technology
NLP today
What does an NLP system need to “know”?

• Language consists of many levels of structure

• Humans fluently integrate all of these in producing and understanding language

• Ideally, so would a computer!
This is a simple sentence
This is a simple sentence.

But it is an instructive one.
Why is NLP hard?
Ambiguity

At the word level

• Part of speech
  • [V Duck]!
  • [N Duck] is delicious for dinner.

• Word sense
  • I went to the bank to deposit my check.
  • I went to the bank to look out at the river
Ambiguity

At the syntactic level

• PP Attachment ambiguity
  • I saw the man on the hill with the telescope

• Structural ambiguity
  • I cooked her duck
  • Visiting relatives can be annoying
  • Time flies like an arrow
Ambiguity

• Quantifier scope
  • Everyone on the island speaks two languages.

• Hard cases require world knowledge, understanding of speaker goals
  • The city council denied the demonstrators the permit because they advocated violence
  • The city council denied the demonstrators the permit because they feared violence
Ambiguity

• NLP challenge: how can we model ambiguity, and choose the correct analysis in context?

• Approach: learn from data
Word counts

- Most frequent words in the English Europarl corpus
- (out of 24M word tokens)

<table>
<thead>
<tr>
<th>any word</th>
<th>Frequency</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1,698,599</td>
<td>the</td>
</tr>
<tr>
<td></td>
<td>849,256</td>
<td>of</td>
</tr>
<tr>
<td></td>
<td>793,731</td>
<td>to</td>
</tr>
<tr>
<td></td>
<td>640,257</td>
<td>and</td>
</tr>
<tr>
<td></td>
<td>508,560</td>
<td>in</td>
</tr>
<tr>
<td></td>
<td>407,638</td>
<td>that</td>
</tr>
<tr>
<td></td>
<td>400,467</td>
<td>is</td>
</tr>
<tr>
<td></td>
<td>394,778</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>263,040</td>
<td>I</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nouns</th>
<th>Frequency</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>124,598</td>
<td>European</td>
</tr>
<tr>
<td></td>
<td>104,325</td>
<td>Mr</td>
</tr>
<tr>
<td></td>
<td>92,195</td>
<td>Commission</td>
</tr>
<tr>
<td></td>
<td>66,781</td>
<td>President</td>
</tr>
<tr>
<td></td>
<td>62,867</td>
<td>Parliament</td>
</tr>
<tr>
<td></td>
<td>57,804</td>
<td>Union</td>
</tr>
<tr>
<td></td>
<td>53,683</td>
<td>report</td>
</tr>
<tr>
<td></td>
<td>53,547</td>
<td>Council</td>
</tr>
<tr>
<td></td>
<td>45,842</td>
<td>States</td>
</tr>
</tbody>
</table>
Word counts

• But also, out of the 93,638 distinct words (word types), 36,231 occur only once

  • cornflakes, mathematicians, fuzziness, jumbling
  • pseudo-rapporteur, lobby-ridden, perfunctorily,
  • Lycketoft, UNCITRAL, H-0695
  • policyfor, Commissioneris, 145.95, 27a
Plotting word frequencies
Plotting word frequencies (with log-log axes)
Zipf’s law

\[ f \times r \approx k \]

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

• Even in a very large corpus, there will be a lot of infrequent words

• The same holds for many other levels of linguistic structure

• Core NLP challenge: we need to estimate probabilities or to be able to make predictions for things we have rarely or never seen
Variation and Expressivity

• The same meaning can be expressed with different forms
  
  • I saw the man
  • The man was seen by me

  • She needed to make a quick decision in that situation
  • The scenario required her to make a split-second judgment
6,800 living languages
600 with written tradition
100 spoken by 95% of population
Social Impact

• NLP experiments and applications can have a direct effect on individual users’ lives

• Some issues
  • Privacy
  • Exclusion
  • Overgeneralization
  • Dual-use problems

[Hovy & Spruit ACL 2016]
Today

• Levels of linguistic analysis in NLP
  • Morphology, syntax, semantics, discourse

• Why is NLP hard?
  • Ambiguity
  • Sparse data
    • Zipf’s law, corpus, word types and tokens
    • Variation and expressivity
  • Social Impact
This semester

• Words and their meanings
  • Distributional semantics and word sense disambiguation
  • Fundamentals of supervised classification

• Sequences
  • N-gram and neural language models
  • Sequence labeling tasks
  • Structured prediction and search algorithms

• Application: Machine Translation

• Trees
  • Syntax and grammars
  • Parsing
Course Logistics

http://www.cs.umd.edu/class/fall2018/cmsc470/
Exam dates

• 9/27 2pm-3:15pm Midterm #1
• 11/1 2pm-3:15pm Midterm #2
• 12/15 10:30am-12:30pm Final Exam
Before next class

• Read the syllabus

• Make sure you have access to piazza

• Get started on homework 1 – due Wednesday Sep 5 by 11:59pm.