NLP 201: Natural Language Processing 1
Introduction Part 2

Jeffrey Flanigan
October 6, 2020
University of California Santa Cruz
jmflanig@ucsc.edu
Last time: What is Natural Language Processing (NLP)?

- The set of methods for making human language accessible to computers (Eisenstein, 2018).
NLP has many end-user tasks (downstream tasks or applications)

- Machine translation
- Summarization
- Question answering*
- Conversational agents*
- Search (information retrieval)
- Recommender systems
- Document classification

*These two tasks are supertasks.
Downstream tasks sometimes benefit from intermediate tasks

- Knowing a word’s sense (i.e. duck – animal vs duck – action) could help translate it. This is sense disambiguation.
- Knowing if a word is a verb or noun (its part of speech) could help translate it (duck – noun vs duck – verb). This is part-of-speech (POS) tagging.
- Splitting text into sentences is often required before processing. This is sentence segmentation.
- Deciding what should count as a word ($100 vs $100 or it’s vs it’s) (tokenization) usually has a very large effect on performance.
Examples of intermediate tasks

- Tokenization
- Language modeling
- POS tagging
- Lemmatization
- Syntactic parsing
- Entity recognition
- Entity linking
- Relation extraction
- Semantic role labeling
- Semantic parsing
- Generation (from an intermediate representation)
Classes of tasks

- Document Classification
- Tagging
- Parsing
- Generation
- Sequence-to-sequence
The traditional NLP pipeline

1. Tokenization
2. Morphological analysis
3. Part-of-speech tagging
4. Syntactic Parsing
5. Semantic Parsing (optional)
6. Downstream task: classification, QA, summarization, etc
7. Generation (optional)

With deep learning, sometimes tasks are done end-to-end, without any intermediate steps.
Areas of Linguistics

- Phonology
- Orthography
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse
- Phonetics
- Text
- Speech
- Lexemes

"Shallower"

"Deeper"
Areas of Linguistics

- Phonetics and phonology
- Orthography
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse
Computers can understand programming languages. Could we do the same thing for natural language?
“At last, a computer that understands you like your mother.”
“At last, a computer that understands you like your mother.”

1. It understands you as well as your mother understands you.
2. It understands (that) you like your mother.
3. It understands you as well as it understands your mother.

1 and 3: Does this mean well, or poorly?
Ambiguity at Many Levels

At the *acoustic* level

1. “... a computer that understands you *like your* mother.”
2. “... a computer that understands you *lie cure* mother.”
At the **syntactic** level:

Different structures lead to different interpretations.
More Syntactic Ambiguity

```
VP
  \- V
    \- list
  \- NP
    \- DET
      \- all
    \- N
      \- flights
  \- PP
    \- on Tuesday

VP
  \- V
    \- list
  \- NP
    \- all
    \- flights
  \- PP
    \- on Tuesday
```
At the semantic (meaning) level:
Two definitions of “mother”

- a woman who has given birth to a child
- a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar

This is an instance of word sense ambiguity
More Semantic Ambiguity

At the semantic (meaning) level:

- They put money in the *bank* = buried in mud?
- I saw her duck with a telescope
At the **discourse** (multi-clause) level:

- Alice says they’ve built a computer that understands you like your mother
- But **she** . . .
  
  . . . doesn’t know any details
  . . . doesn’t understand me at all

This is an instance of **anaphora**, where she co-referees to some other discourse entity
Large growth in NLP in recent years
NLP applications are now commonplace

- Spam email filtering
- Google translate
- Built-in recommender systems (in Amazon, Ebay, Netflix, etc)
- Siri, Amazon Alexa
- Auto-completion suggestions
- Grammar checking
- Automatic essay grading (used by ETS)
- Inappropriate social media post filtering
- Fake news detection
- Lots you probably don’t even realize!
With widespread use, NLP has potential ethical issues such as:
- Bias
- Censorship
- Privacy
- Security

These issues are hot topics, very active area of research.
Relation of NLP to other fields

- Speech (both recognition and generation) are separate, not an NLP tasks
- Machine learning (computers learn from experience or examples)
- Linguistics (the study of language).
- Computational linguistics (CL)
  - Sometimes synonymous with NLP
  - CL often has larger emphasis on linguistics and linguistic theories. CL degree programs sometimes have a different curriculum than NLP degree programs
(5 min break)
References