A brief introduction to.
Natural Language Processing

Claire Cardie
Professor
Computer Science and Information Science

Plan for the lecture

- What is NLP?
  - Levels of linguistic analysis
  - Issues that arise
- NLP applications
  - Focus: information extraction
  - Current trend
Natural Language Processing (NLP)

Linguistic analysis

- Phonetics and phonology
  - Consider the phone [ni]
  - Corresponds to which word(s) in a corpus of phone conversations (AT&T)?
    - knee
    - But this was the LOWEST probability option!!!!
    - need
    - neat
    - new
    - Yeah, I live in New York.
Linguistic analysis

- **Syntax**
  - I ate spaghetti with meatballs.
  - I ate spaghetti with friends.
  - I ate spaghetti with a fork.
  - I ate spaghetti with glee.

- **Semantics**
  - Lexical (word-level) semantics
    » Word-sense disambiguation
      ◆ with ➔ using
      ◆ with ➔ co-participant
      ◆ ...
      ◆ Mirabelle drank from the **banks** of lake Cayuga.
        – Ground alongside a body of water?
        – Financial institution?
Linguistic Problems

- **Semantics**
  - Concerns what words mean and how these meanings combine to form sentence meanings.
    » Compositionality

- **Discourse (and dialogue)**
  - Concerns how the immediately preceding sentences affect the interpretation of the next sentence
    » Jack saw Sam arrive at the party. Then he went back inside for some chips.
    » Jack saw Sam arrive at the party. He was driving a Subaru Outback.

---

Linguistic Problems

- **Pragmatics**
  - Concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

  “I just came from Collegetown Bagels.”

  » Do you want to go find some lunch?
  » Boy, you look frazzled.
Key issue

- Ambiguity!!!!
  - At all levels of linguistic analysis
- Difficult (impossible) to design algorithms based on our intuitions
- Solution:
  - Rely on machine learning methods

Plan for the lecture

- What is NLP?
  - Levels of linguistic analysis
  - Issues that arise
- NLP applications
  - Focus: information extraction
  - Current trend
Text categorization

Is the document about politics? sports? economics? fashion?

Sentiment categorization

Is the document positive? negative?

Pang & Lee [2008]
Summarization

- headlines (from around the world)
- outlines (notes for students)
- minutes (of a meeting)
- previews (of movies)
- synopses (soap opera listings)
- reviews (of a book, CD, movie, etc.)
- digests (TV guide)
- biography (resumes, obituaries)
- abridgments (Shakespeare for children)
- bulletins (weather forecasts/stock market reports)
- sound bites (politicians on a current issue)
- histories (chronologies of salient events)

Involves natural language generation!!!

Question answering

- Task
  - How many calories are there in a Big Mac?
  - Who is the voice of Miss Piggy?
  - Who was the first American in space?
  - Retrieve not just relevant documents, but return the answer

? → text collection → answer + supporting text
Machine translation

- one of the first applications envisioned for NLP techniques
  - vodka good meat rotten
  - "The spirit is willing, but the flesh is weak."
  - "open"

Focus on one application

- Information extraction
Information extraction

- Unstructured text → structured representation
- Usually domain-specific, usually fact- or event-oriented

Opinion extraction

- Unstructured text → structured representation

The White House press corps launched a bitter attack on Trump…
IE subproblems

- Named Entity identification
  - White House → location? Organization?
  - Trump → person
- Relation extraction
  - Press corps located-In? / employed-by? White House

Feature-based ML methods

- E.g., naïve Bayes, CRFs
- Features based on the output of many NLP components
Trend: RNNs + Word Embeddings

- Many powerful variations

word embeddings The proposal is criticized …

Trend: RNNs + Word Embeddings

- Better or comparable performance than feature-based approaches
  - **Without** NLP components, **without** feature engineering
    - Sentiment/opinion analysis: **without** manually procured sentiment lexicons

Constituent-level parser Dependency parser N-grams
POS tagger Semantic class tagger Sentiment word lexicons Opinion word lexicons
Opinion analysis

<table>
<thead>
<tr>
<th>opinion expressions</th>
<th>proportional overlap</th>
<th>exact match</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF</td>
<td>64.4 F</td>
<td>57.7 F</td>
</tr>
<tr>
<td>CRF + word embeddings</td>
<td>66.4 F</td>
<td>59.6 F</td>
</tr>
<tr>
<td>Deep bidirectional RNNs</td>
<td>71.7 F</td>
<td>66.0 F</td>
</tr>
</tbody>
</table>

[Irsoy & Cardie, EMNLP2014]

Opinion expressions

1. The situation obviously remains fluid from hour to hour but it seems to be [issuing a] report in the right direction. [Irsoy & Cardie, EMNLP2014]

2. The situation obviously remains fluid from hour to hour but it seems to be [going into] the right direction. [Irsoy & Cardie, EMNLP2014]
...There ARE complications

- NLP
  - AI-complete
    » To “solve” NLP, you’d need to solve all of the problems in AI

If you are interested in learning more, see the list of courses at Cornell on NLP:

https://nlp.cornell.edu/courses/

THANKS FOR YOUR ATTENTION!!!