Natural Language Processing: CIS 410/510

Introduction

Instructor: Thien Huu Nguyen

Based on slides from: Ralph Grishman
Logistics

• **Instructor:** Thien Huu Nguyen
  – Email: thien@cs.uoregon.edu
  – Office: Deschutes 330
  – Office hours: Tuesday, 3:30-4:30pm

• **Teaching Assistant:** NONE

• **Web:** [http://www.cs.uoregon.edu/Classes/19S/cis472nlp/](http://www.cs.uoregon.edu/Classes/19S/cis472nlp/)

• **Discussion Board:** Piazza (link on web page)
  Please use this for discussion!
Evaluation

• 4 homeworks (40%)
  – Three programming, one exercise

• One Midterm (30%)
  – 2/3rds of the way through

• One Project (30%)
  – Proposal due on Nov. 5
  – Code, written report due on Dec. 11
  – Presentations during the final week
So what are natural languages?

- A primary and natural mode of human communication and reasoning
- representation for most recorded human language
- a very rich and flexible representation (comparing to most formal representations)
Natural Language Processing (NLP)

- Processing natural languages with computers, enabling computers to understand, generate and analyze natural languages.

- Technically, designing methods and algorithms that take as input and/or produce as output unstructured, natural language data (“Hello World”).

  - More abstract: understanding the underlying principles of human languages to allow computers to mimic human capacity on languages.

  - More abstract: understanding the cognitive processes within human minds based on the human-generated languages

  - More abstract: the computational science for linguistics
The Turing Test

Distinguishing human vs. computer only through written language
Michael Jordan, leading machine learning researcher, UC Berkeley

I'd use the billion dollars to build a NASA-size program focusing on natural language processing (NLP), in all of its glory (semantics, pragmatics, etc).

Intellectually I think that NLP is fascinating, allowing us to focus on highly-structured inference problems, on issues that go to the core of "what is thought" but remain eminently practical, and on a technology that surely would make the world a better place.
### Baseline mutual information model (Li et al. 2015)

<table>
<thead>
<tr>
<th>A: Where are you going? (1)</th>
<th>B: I’m going to the restroom. (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: See you later. (3)</td>
<td>B: See you later. (4)</td>
</tr>
<tr>
<td>A: See you later. (5)</td>
<td>B: See you later. (6)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Proposed reinforcement learning model**

<table>
<thead>
<tr>
<th>A: Where are you going? (1)</th>
<th>B: I’m going to the police station. (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: I’ll come with you. (3)</td>
<td>B: No, no, no, no, you’re not going anywhere. (4)</td>
</tr>
<tr>
<td>A: Why? (5)</td>
<td>B: I need you to stay here. (6)</td>
</tr>
<tr>
<td>A: I don’t know what you are talking about. (7)</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A: how old are you? (1)</th>
<th>B: I’m 16. (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: 16? (3)</td>
<td>B: I don’t know what you are talking about. (4)</td>
</tr>
<tr>
<td>A: You don’t know what you are saying. (5)</td>
<td>B: What made you think so? (4)</td>
</tr>
<tr>
<td>B: I don’t know what you are talking about . (6)</td>
<td>A: I don’t know what you are talking about. (5)</td>
</tr>
<tr>
<td>A: You don’t know what you are saying. (7)</td>
<td>B: You don’t know what you are saying. (6)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

---

Li et al., 2016: Deep Reinforcement Learning for Dialogue Generation (EMNLP)
One thing I wanted to ask you about is this. A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. Huxley has recently told me that they are appalled by the magnitude and the importance of the translation job.

Warren Weaver (in his letter to Prof. Norbert Wiener at MIT in 1947)
NLP Applications: Machine Translation

- Long history of development (since early 1950’s)

- Still not good enough for polished translations, but effective for producing rough drafts for post-editing, or for getting the gist of a text

- Quality depends on similarity of language structures (Chinese to English much harder than French to English)

NLP Applications: Question Answering

• Originally passage retrieval systems, gradually enriched with NLP

• Example: MIT Start system: [http://start.csail.mit.edu/](http://start.csail.mit.edu/)

• Major web search tools originally operated as web page retrieval (document retrieval) using word-based strategies but now seek to detect questions they can answer directly (e.g., google.com)

• IBM Watson system to play Jeopardy demonstrated potential of QAS when coupled to a rich variety of knowledge sources; this technology now being applied to medical domain: [https://www.youtube.com/watch?v=WFR3lOm_xhE](https://www.youtube.com/watch?v=WFR3lOm_xhE)
NLP Applications: Interactive Systems

• One of the first interactive applications was data base query, but that had limited appeal with written input: people don't like to type a lot; GUIs have been more effective.

• Chatbots provide the impression of intelligent conversation (try Eliza: [https://www.cyberpsych.org/eliza/#.WIQIu2PDX8s](https://www.cyberpsych.org/eliza/#.WIQIu2PDX8s))
or: [https://www.tolearnenglish.com/free/celebs/alice.php](https://www.tolearnenglish.com/free/celebs/alice.php)

• Conversational agents support simple conversations (using text or speech) for order taking, information.

• Smartphones with speech recognition (e.g., SIRI) have greatly increased opportunities for speech input.
NLP Applications: Information Extraction

- Conversion of unstructured (or semi-structured) data to structured form
- Resumes: [http://monster.com](http://monster.com)
- Name Tagger: GATE: [http://services.gate.ac.uk/annie/](http://services.gate.ac.uk/annie/)
- More general entity/relation/event taggers
- For search and general news monitoring
  - [Google News](http://news.google.com)
  - [European Media Monitor](http://www.limes.uni-freiburg.de)
- For search and general news monitoring
  - Infectious disease reports (e.g., the NYU Proteus-Bio system)
  - Electronic health records
  - Scientific literature (e.g., [Paleo Deep-Dive](http://www.paleodeepdive.org))
NLP Applications: Text Mining

- Discovering patterns from large text collections
- Direct (“first generation”) approach uses “bag of words” representation of documents
- Second generation: information extraction + data mining = text mining
- Take advantage of social media (e.g., Dataminr)

Applications:
- Product monitoring (using sentiment analysis to collect detailed feedback from customers)
- Situational awareness (rapid detection of emergency situations)
- Finance (rapid response to financial news)
- Scientific research (mine large collections of research papers to identify trends and correlations like treatments and adverse reactions)
Challenges in NLP

• Language is discrete, making it more challenging to solve the optimization-related problems.

• Language is compositional (the meaning of a sentence involves understanding the individual words and how to combine/compose the individual meanings to form the overall meaning).

• Language is flexible (there are multiple ways to talk about the same things).
  – I hate spicy food
  – I don’t like spicy food
  – If I can choose, I won’t take spicy food

• Language is ambiguous (context is often necessary to accurately understand the meanings of a word/sentence).
The ambiguity of language

“One morning I shot an elephant in my pajamas”

So, I or the elephant is in my pajamas?

Animal Crackers
The ambiguity of language

“One morning I shot an elephant in my pajamas. How he got into my pajamas I’ll never know.”
Relation to other fields

• Linguistics
  – The goal is to describe language
    • Provide simple models which can *predict* language behavior
    • Understand what is *universal* about language
    • Through these formal models, understand how language can be *acquired*

  – Formal linguistic models have been helpful for NLP, but their goals are different from NLP’s:
    • A simple counterexample can invalidate a model as a linguistic theory, but would not significantly lessen the value of NLP models
    • NLP must address all phenomena that arise in an application, while linguistics may focus on selected phenomena that are of interest to the language faculty
Relation to other fields

• (symbolic) Artificial Intelligence
  – Classical “symbolic” AI is mostly concerned with generic problem solving strategies and suitable knowledge representation

  – The inherent link between AI and NLP: some NLP problems need the sort of deep reasoning provided by these AI models.

  – But NLP has found increasing success through avoiding such deep reasoning and turning instead to ...
Relation to other fields

• **Statistics and Machine Learning**
  
  – Early NLP systems (before 1990) were purely symbolic and hand-crafted
  
  – Statistical methods and models have become more widely used in NLP since the mid 1990’s
  
  – Easily trainable and easily computable models have proven much more effective than the more complex hand-crafted models
  
  – This approach has become more attractive now as lots of training data is available (the “big data” era)
What functionality we need to address NLP applications?

- Consider machine translation

  at first, people imagined that machine translation is mostly a “data processing” task where a system looks up the words one at a time in a bilingual dictionary, and then maybe has to fix up the translation a bit. However, there is a lot more to do for machine translation:
Machine Translation

- **Word segmentation**
  - For some languages (e.g., Chinese, Japanese) there are no spaces between words, so it’s not clear what the words are

- **Morphology**
  - Words appear in different forms, indicating singular vs. plural (for nouns), present tenses vs. past tense (for verbs), nominative vs. accusative case (different words forms for subjects and objects)

  - English has only a few morphological forms, so it’s possible to put them all in a dictionary. This doesn’t hold for most Western languages; for example, a Spanish verbs could have over 50 forms.
Machine Translation

- **Syntax**
  - Word-for-word translation only works if the word order in the two languages is about the same.
  
  - Otherwise, we need to understand the structure of the two languages (i.e., their syntax) to change from one word order to another.
  
  - English has a rather fixed subject-verb-object order (“SVO”), while many more inflected languages have more variable word order and some languages have basically different word orders (e.g., “SOV” for Japanese).
Machine Translation

- **Lexical Semantics**
  - Many words are polysemous (i.e., they have multiple meanings). A word will have to be translated differently depending on its meaning in a specific context.
  - For instance, “bill” can mean both a statement of charges (an “invoice”) and a part of a duck (i.e., its “beak”). It’s not likely that any foreign language has a word with both these senses.

- If we choose the wrong sense in translating “bill” into a foreign language, we will have things like:

  *At the end of the meal, the waiter presented the beak*
Machine Translation

• **Discourse**
  – Sometime we need to look beyond the individual sentences to obtain proper translation.
  
  – For instance, cross-sentence evidence might be needed to choose the correct word senses.
  
  – Another example arises in *translating into English from languages where subject pronouns can be omitted*, then we need to determine what the subject actually is, so we can pick the right pronoun (i.e., “he”, “she”, or “it”) in English.
**Similarly for Information Extraction**

- **Name recognition**: knowing “Fred Smith” is a person while “IBM” is a company (these are called named entities).
- **Syntax**: the information may be presented in the passive form (“Fred Smith was hired by IBM”) or in a relative clause (“Fred Smith, who was hired by IBM”); also, there maybe extra modifiers (“IBM yesterday hired Fred Smith as president”).
- **Lexical Semantics**: there might be lots of synonyms for “hired” (i.e., “appointed”, “named”, ...)
- **Discourse – pronoun**: the system might need to determine what the pronoun refers to (“Fred Smith left Compaq last week. IBM hired him yesterday as president.”).
So, for NLP we need:

- What is the meaning of a word (lexical semantics) or a sentence (compositional semantics) in isolation?
- What is the structure of a sentence?
- What is the appropriate internal structure of a word?
- What are words/tokens?
- How can a sentence be interpreted in context?
- What is the structure of a document?