LING 120, Fall 2017: Language and Computers

Topic: Overview of Natural Language Processing

Instructor: Sowmya Vajjala

Iowa State University, USA

2 October 2017 (Week 7)

Class outline

- Reminders, Announcements etc.
- Natural Language Processing: What is it, and why now?
- Examples of bottlenecks in automatically understanding human language
- Tasks in NLP: an overview

Reminders and Announcements

- Assignment 3 is due this week! (7th October)

Reminders and Announcements

- Assignment 3 is due this week! (7th October)
- Midterm related stuff:
  - Presentations are due next week (3 teams on Monday, 3 on Wednesday)
  - You can get your laptops or use mine (If you use mine, send a PDF - I don't have powerpoint)
  - Suggestion: Do the presentations using Google slides or some such collaborative online tool.
  - I will accomodate some time in the class on Wednesday (and may be Friday) for this.

Reminders and Announcements

- Assignment 3 is due this week! (7th October)
- Midterm related stuff:
  - Presentations are due next week (3 teams on Monday, 3 on Wednesday)
  - You can get your laptops or use mine (If you use mine, send a PDF - I don't have powerpoint)
  - Suggestion: Do the presentations using Google slides or some such collaborative online tool.
  - I will accomodate some time in the class on Wednesday (and may be Friday) for this.
- Assignment 4: description in Friday's class

What is NLP? and Why now?

- What: It is a field concerned with the interaction between computers and human languages.
- Address issues like - how can computers process and understand large amounts of data in various human languages.

What is NLP? and Why now?

- What: It is a field concerned with the interaction between computers and human languages.
- Address issues like - how can computers process and understand large amounts of data in various human languages.
- In Week 1, I gave an overview of where is Language understanding is relevant in real life computer based applications.
- Now that you gained some knowledge of different applications of language, it is time to think about those initial questions again, in more detail.

Example from Week 1

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune (`http://goo.gl/zvx9Uw`)

1. When she says "I" in the first sentence, does she mean herself literally?
2. What is she referring to? When will we know what is she referring to?
3. Who is "She"?
4. What is "Chinese Homestyle Cooking" referring to?

Let us take a small text snippet -2

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune `http://goo.gl/zvx9Uw`

1. What is the main event of this text?
2. What is the relationship between "Chinese Homestyle cooking" and Tina?
3. Is Lincoln Way something related to President Lincoln?

So, what is it about language that makes it difficult for a computer to process the text and answer such questions?

1. Language is ambiguous

Some ambiguous sentences

- Newspaper headlines
  - "Children make delicious snacks"
  - "Dead expected to rise"
  - "Republicans grill IRS chief over lost emails"
- Normal, grammatical sentences can be ambiguous too:
  - "I saw a man on a hill with a telescope."
  - "Look at the man with one eye"

We are not even talking about ambiguities involving speech or alternative interpretations due to stress/emphasis on some word.

Some types of ambiguity

1. Lexical ambiguity: due to multiple meanings or senses of word usage
   e.g., He stood near the **bank**

Some types of ambiguity

1. Lexical ambiguity: due to multiple meanings or senses of word usage
   e.g., He stood near the **bank**
2. Structural ambiguity: due to syntactic structure
   e.g., I saw the man on the hill with telescope.

Some types of ambiguity

1. Lexical ambiguity: due to multiple meanings or senses of word usage
   e.g., He stood near the **bank**
2. Structural ambiguity: due to syntactic structure
   e.g., I saw the man on the hill with telescope.
3. Semantic ambiguity: more interpretations possible
   e.g., John and Mary are married (to each other? or to different people?)

Some types of ambiguity

1. Lexical ambiguity: due to multiple meanings or senses of word usage
   e.g., He stood near the **bank**
2. Structural ambiguity: due to syntactic structure
   e.g., I saw the man on the hill with telescope.
3. Semantic ambiguity: more interpretations possible
   e.g., John and Mary are married (to each other? or to different people?)
4. Referential ambiguity
   e.g., She dropped the *plate* on the *table* and broke **it**

Some types of ambiguity

1. Lexical ambiguity: due to multiple meanings or senses of word usage
   e.g., He stood near the **bank**
2. Structural ambiguity: due to syntactic structure
   e.g., I saw the man on the hill with telescope.
3. Semantic ambiguity: more interpretations possible
   e.g., John and Mary are married (to each other? or to different people?)
4. Referential ambiguity
   e.g., She dropped the *plate* on the *table* and broke **it**
5. Ambiguity due to the use of non-literal language
   e.g., Time flies like an arrow

Good source to read more:
`http://cs.nyu.edu/faculty/davise/ai/ambiguity.html`

Ambiguity for humans

personal experience

- "ISU sends out alert as devastating weed spreads in Iowa" - read a local newspaper headline last year. What did you understand?

Ambiguity for humans

personal experience

- "ISU sends out alert as devastating weed spreads in Iowa" - read a local newspaper headline last year. What did you understand?
- My immediate reaction: Some new drug is circulating and ISU wants its students to be aware and not consume it.

Ambiguity for humans

personal experience

- "ISU sends out alert as devastating weed spreads in Iowa" - read a local newspaper headline last year. What did you understand?
- My immediate reaction: Some new drug is circulating and ISU wants its students to be aware and not consume it.
- However, the actual news is about some poisonous wild plant, and the alert was issued by ISU weed and crop specialists to farmers.

... so some humans also cannot disambiguate certain things, due to... may be.. cultural differences?

2. "common" knowledge for humans

Look at these two sentences:
Dog bit man.
Man bit dog.
- For a computer, both of them are linguistically the same. We know only the first one is "normal" English sentence (I hope!) because we have "world knowledge".

3. Language is creative

Literary texts have their own language style: long sentences, neologisms, creative usage of words etc.

4. Language can be complex to understand

Legal documents, Writing style of some authors, propaganda materials, etc.

5. Language is Diverse

Some Examples

- There are different types of text online: news, tweets, SMS, email, forum posts, speech transcripts etc.
- Each genre has some specific characteristics of its own
- NLP methods should generalize to all genres, and at the same time capture such specific characteristics
- Example: a machine translation model created by training examples from European parliament speeches should also be able to translate casual day to day conversations.

6. Other Issues

- different text formats (pdf, doc, txt etc)
- spelling variations
- sarcasm
- slang
- using synonyms, paraphrases etc

.. and so on.

7. Languages are many

.. and each language has its own special characteristics apart from similarities with other languages. NLP should handle both these aspects.

Tasks in NLP: An overview

NLP tasks: Word Collocations and Concordances

- Task: compiling lists of words, or word sequences occuring in documents.
- Simplest and least ambiguous form of language processing.
- Can get to more advanced collocations beyond surface forms.
- Established methods exist for collecting different kinds of collocations

NLP tasks: Pattern Extraction

- Task: Extract the language patterns that exist in textual data.
- Regular expressions are very useful for this
- More advanced methods (which rely on machine learning) exist to extract unknown patterns from unstructured text documents.

NLP tasks: POS Tagging

What is the big deal about automatic tagging?

- Task: Given a sequence of words, return the POS tags for each word.
- An example problem: What is the best tag for a word in a context?
  - I wish to cite this work.
    PRP/I VBP/wish TO/to VB/cite DT/this NN/work ./.
  - He has a wish.
    PRP/He VBZ/has DT/a NN/wish ./.
- Largely considered solved for English, but there are still issues if we go beyond typical newspaper language (e.g., tagging speech or tweets). Still an unsolved problem for several languages.

(to be continued)

Attendance Exercise

- Form into your mid-term teams, do the exercise in the given worksheet, write the names of your teammates and return to me.
- This counts as your attendance for today.
- Source: `http://www.nacloweb.org/resources/problems/2011/E.pdf`