Natural Language Processing

Huong Le Thanh
huong@wv.inf.tu-dresden.de
Why study natural language processing (NLP)?

Applications

• Line breakers, hyphenators, spell checkers, grammar & style checkers
• information retrieval
• question answering
• automatic speech recognition
• intelligent Web searching
• automatic text summarization and classification
• pseudo-understanding and generation of natural language; multi-lingual systems including machine translation
Some information

Time and Place
• Lectures and Tutorials are held on Wednesday, room GRU 350, 09.20am - 10.50am

References
• Dan Jurafsky and James Martin. 2000. *Speech and Language Processing*. PrenticeHall.

Recommended Reading
Course Goals

• Learn the basic principles and theoretical approaches underlying NLP

• Learn techniques and tools which can be used to develop practical, robust systems that can (partly) understand text or communicate with users in one or more languages

• Gain insight into many of the open research problems in natural language
Topics in NLP

- **Levels of Analysis:** syntax, semantics, discourse, pragmatics, world knowledge...

- **Subproblems:** part-of-speech tagging, syntactic parsing, word sense disambiguation, discourse processing...

- **Algorithms and Methodologies:** corpus-based methods, knowledge-based techniques,...

- **Applications:** information extraction, information retrieval, machine translation, question answering, natural language understanding....
Levels of Analysis and Knowledge Used in NLP

- **Morphology**: how words are constructed; prefixes & suffixes
- **Syntax**: structural relationships between words
- **Semantics**: meanings of words, phrases, and expressions
- **Discourse**: relationships across different sentences or thoughts; contextual effects
- **Pragmatic**: the purpose of a statement; how we use language to communicate
- **World Knowledge**: facts about the world at large; common sense
Morphology

- kick, kicks, kicked, kicking
- sit, sits, sat, sitting
- murder, murders

But it’s not just as simple as adding and deleting endings...

- gorge, gorgeous
- glass, glasses
- arm, army
Syntax: part-of-speech tagging

- The boy threw a ball to the brown dog.

- The/DT boy/NN threw/VBD a/DT ball/NN to/IN the/DT brown/JJ dog/NN./.

<table>
<thead>
<tr>
<th>DT</th>
<th>NN</th>
<th>VBD</th>
<th>IN</th>
<th>JJ</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>determiner</td>
<td>noun, single or mass</td>
<td>verb, past tense</td>
<td>preposition, sub-conj</td>
<td>adjective</td>
<td>sentence final punc</td>
</tr>
</tbody>
</table>
Syntax: structural ambiguity (part of speech)

Time flies like an arrow.

Time // flies ___________ like_____________ an arrow.

VBZ comparative proposition (IN)

Time flies // like an arrow.

NNS VBP
I saw the man on the hill with a telescope.
Syntax: structural ambiguity (attachment)

I saw the man on the hill with a telescope.
I saw the man on the hill with a telescope.
But syntax doesn’t tell us much about meaning

- Colorless green ideas sleep furiously. [Chomsky]
- fire match arson hotel
- plastic cat food can cover
Semantics: lexical ambiguity

• I walked to the bank ... of the river.
  to get money.

• The bug in the room ... was planted by spies.
  flew out the window.

• I work for John Hancock ... and he is a good boss.
  which is a good company.
Discourse: coreference

President John F. Kennedy was assassinated. The president was shot yesterday. Relatives said that John was a good father. JFK was the youngest president in history. His family will bury him tomorrow. Friends of the Massachusetts native will hold a candlelight service in Mr. Kennedy’s home town.
Pragmatics

What should you conclude from the fact that I said something? How should you react?

Rules of Conversation

• Can you tell me what time it is?
• Could I please have the salt?

Speech Acts

• I bet you $50 that the Jazz will win.
John went to the diner. He ordered a steak. He left a tip and went home.

- What did John eat for dinner?
- Who brought John his food?
- Who cooked the steak?
- Did John pay his bill?
Why NLP is difficult

- Complex phenomenon arising out of the interaction of many distinct kinds of knowledge
- What is this knowledge? (data structures - linguistics)
- How is it put to use? (algorithms)
- Example: “the dogs ate ice-cream”
Knowledge of language: What do we know about this sequence?

- Words must appear in a certain order:
  *Dogs icecream ate*

- Parts and divisions:
  dogs = Subject; ate icecream = Predicate

- Who did what to whom:
  agent(dogs), action(ate), object(ice-cream)
Anything else?

- The two sentences “John claimed the dogs ate icecream” and “John denied the dogs ate ice-cream” are logically incompatible.

- Sentence & the world: know whether the sentence is true or not - perhaps whether in some particular situation (possible world) the dogs did indeed eat icecream.

- “I had espresso this morning, but John is intelligent” looks odd.
What is the character of this knowledge?

• Some of it must be memorized:
  • Singing → Sing+ing; Bringing → bring+ing

• *Duckling* → ?? Duckl +ing

• So, must know *duckl* is not a word

• But it can’t all be memorized because there is too much to know
Besides memory, what else do we need?

English plural:

- Toy+s -> toyz ; add z
- Book+s -> books ; add s
- Church+s -> churchiz ; add iz
- Box+s-> boxiz ; add iz

➢ must be a rule system to generate/process infinite # of examples
“Parsing” = mapping from surface to underlying representation

- What makes NLP hard: there is not a 1-1 mapping between any of these representations!

- We have to know the data structures and the algorithms to make this efficient, despite exponential complexity at every point
Six sculptures – C, D, E, F, G, H – are to be exhibited in rooms 1, 2, and 3 of an art gallery.

- Sculptures C and E may not be exhibited in the same room.
- Sculptures D and G must be exhibited in the same room.
- If sculptures E and F are exhibited in the same room, no other sculpture may be exhibited in that room.
- At least one sculpture must be exhibited in each room, and no more than three sculptures may be exhibited in any room.

If sculpture D is exhibited in room 3 and sculptures E and F are exhibited in room 1, which of the following may be true?

A. Sculpture C is exhibited in room 1
B. Sculpture H is exhibited in room 1
C. Sculpture G is exhibited in room 2
D. Sculptures C and H are exhibited in the same room
E. Sculptures G and F are exhibited in the same room
U: Where is A Bug’s Life playing in Mountain View?
S: A Bug’s Life is playing at the Summit theater.
U: When is it playing there?
S: It’s playing at 2pm, 5pm, and 8pm.
U: I’d like 1 adult and 2 children for the first show. How much would that cost?

Knowledge sources:

• Domain knowledge
• Discourse knowledge
• World knowledge
Why is natural language computing hard?

Natural language is:

• highly ambiguous at all levels
• complex and fuzzy
• involves reasoning about the world
Making progress on this problem...

• The task is difficult! What tools do we need?
  • Knowledge about language
  • Knowledge about the world
  • A way to combine knowledge sources

• A potential solution:
  • probabilistic models built from language data
    • $P(\text{“maison” } \rightarrow \text{“house”})$ high
    • $P(\text{“L’avocat general” } \rightarrow \text{“the general avocado”})$ low