Introduction to Natural Language Processing

Lecture #1

August 28, 2012

Course Information

- Instructor: Prof. Kathy McCoy (mccoy@cis.udel.edu)
- Times: Tues/Thurs 9:30-10:45
- Place: 102A Smith Hall

Home page:
http://www.cis.udel.edu/~mccoy/courses/cisc882.12f
Course Syllabus

Text

Required

- Text: Daniel Jurafsky and James H. Martin, *Speech and Language Processing, Second Edition*, Prentice-Hall.

What is Natural Language Processing?

- The study of human languages and how they can be represented computationally and analyzed and generated algorithmically
  - *The cat is on the mat.* --> on (mat, cat)
  - on (mat, cat) --> *The cat is on the mat.*
- Studying NLP involves studying natural language, formal representations, and algorithms for their manipulation

What *is* Natural Language Processing?

Building computational models of natural language comprehension and production

Other Names:
- Computational Linguistics (CL)
- Human Language Technology (HLT)
- Natural Language Engineering (NLE)
- Speech and Text Processing

Engineering Perspective

Use CL as part of a larger application:
- Spoken dialogue systems for telephone based information systems
- Components of web search engines or document retrieval services
  - Machine translation
  - Question/answering systems
  - Text Summarization
- Interface for intelligent tutoring/training systems

Emphasis on
- Robustness (doesn't collapse on unexpected input)
- Coverage (does something useful with most inputs)
- Efficiency (speech; large document collections)

Cognitive Science Perspective

Goal: gain an understanding of how people comprehend and produce language.

Goal: a model that explains actual human behaviour

Solution must:
explain psycholinguistic data
be verified by experimentation

Theoretical Linguistics Perspective

- In principle, coincides with the Cognitive Science Perspective
- CL can potentially help test the empirical adequacy of theoretical models.
- Linguistics is typically a descriptive enterprise.
- Building computational models of the theories allows them to be empirically tested. E.g., does your grammar correctly parse all the grammatical examples in a given test suite, while rejecting all the ungrammatical examples?

Orientation of this Class

- Emphasis on principles and techniques
- Emphasis on processing textual input (as opposed to speech)
- More oriented towards symbolic than statistical approaches

Language as Goal-Oriented Behaviour

- We speak for a reason, e.g.,
  - get hearer to believe something
  - get hearer to perform some action
  - impress hearer
- Language generators must determine how to use linguistic strategies to achieve desired effects
- Language understanders must use linguistic knowledge to recognise speaker's underlying purpose

Examples

(1) It's hot in here, isn't it?

(2) Can you book me a flight to London tomorrow morning?

(3) P: What time does the train for Washington, DC leave?
C: 6:00 from Track 17.

Why Should You Care?

Two trends

1. An enormous amount of knowledge is now available in machine readable form as natural language text
2. Conversational agents are becoming an important form of human-computer communication
3. Much of human-human communication is now mediated by computers

Knowledge needed to understand and produce language

- *Phonetics and phonology:* how words are related to sounds that realize them
- *Morphology:* how words are constructed from more basic meaning units
- *Syntax:* how words can be put together to form correct utterances
- *Lexical semantics:* what words mean
- *Compositional semantics*: how word meanings combine to form larger meanings
- *Pragmatics:* how situation affects interpretation of utterance
- *Discourse structure:* how preceding utterances affects processing of next utterance

What can we learn about language?

- Phonetics and Phonology: speech sounds, their production, and the rule systems that govern their use
  - tap, butter
  - nice white rice; height/hot; kite/cot; night/not...
  - city **hall**, **parking** lot, **city** hall **parking** lot
  - The cat is on the mat. The cat is on the mat?

Morphology

- How words are constructed from more basic units, called *morphemes*

friend + ly = friendly

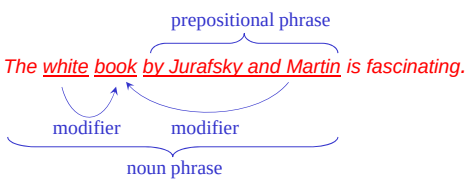
noun

Suffix *-ly* turns noun into an adjective (and verb into an adverb)

- Morphology: words and their composition
  - cat, cats, dogs
  - child, children
  - undo, union

Syntactic Knowledge

- how words can be put together to form legal sentences in the language
- what structural role each word plays in the sentence
- what phrases are subparts of other phrases



- Syntax: the structuring of words into larger phrases
  - John hit Bill
  - Bill was hit by John (passive)
  - Bill, John hit (preposing)
  - Who John hit was Bill (wh-cleft)

Semantic Knowledge

- What words mean
- How word meanings combine in sentences to form sentence meanings

The sole died. *(selectional restrictions)*

shoe part *fish*

Syntax and semantics work together!

(1) What does it taste like?
(2) What taste does it like?

N.B. Context-independent meaning

- Semantics: the (truth-functional) meaning of words and phrases
  - gun(x) & holster(y) & in(x,y)
  - fake (gun (x)) (compositional semantics)
  - The king of France is bald (presupposition violation)
  - bass fishing, bass playing (word sense disambiguation)

Pragmatics and Discourse: The influence of Context
"Going Home" – A play in one act

- Scene 1: Pennsylvania Station, NY
- Bonnie: Long Beach?
- Passerby: Downstairs, LIRR Station.
- Scene 2: Ticket Counter, LIRR Station
- Bonnie: Long Beach?
- Clerk: $4.50.

- Scene 3: Information Booth, LIRR Station
- Bonnie: Long Beach?
- Clerk: 4:19, Track 17.
- Scene 4: On the train, vicinity of Forest Hills
- Bonnie: Long Beach?
- Clerk: Change at Jamaica.
- Scene 5: On the next train, vicinity of Lynbrook
- Bonnie: Long Beach?
- Clerk: Right after Island Park.

Pragmatic Knowledge

- What utterances mean in different *contexts*

Jon was hot and desperate for a dunk in the river.

Jon suddenly realised he didn't have any cash.

He rushed to the bank.

financial institution river bank

Discourse Structure

Much meaning comes from simple conventions that we generally follow in discourse

- How we ***refer*** to entities
  - *Indefinite NPs* used to introduce new items into the discourse

    A woman walked into the cafe.
  - *Definite NPs* can be used to refer to subsequent references

    The woman sat by the window.
  - *Pronouns* used to refer to items already known in discourse

    She ordered a cappuccino.

Discourse Relations

- Relationships we infer between discourse entities
- Not expressed in either of the propositions, but from their juxtaposition

1. (a) I'm hungry.
(b) Let's go to the Fuji Gardens.

2. (a) Bush supports big business.
(b) He'll vote no on House Bill 1711.

Discourse and Temporal Interpretation

Max fell. John pushed him.

explanation

Syntax and semantics: "him" refers to Max

Lexical semantics and discourse: the pushing occurred before the falling.

Discourse and Temporal Interpretation

John and Max were struggling at the edge of the cliff.
Max fell. John pushed him.

Here discourse knowledge tells us the pushing event occurred **after** the falling event

World knowledge

- What we know about the world and what we can assume our hearer knows about the world is intimately tied to our ability to use language

I took the cake from the plate and ate it.

Ambiguity

I made her duck.

- The categories of knowledge of language can be thought of as ambiguity-resolving components
- How many different interpretations does the above sentence have?
- How can each ambiguous piece be resolved?
- Does speech input make the sentence even more ambiguous?

Ambiguity

- Computational linguists are obsessed with ambiguity
- Ambiguity is a fundamental problem of computational linguistics
- Resolving ambiguity is a crucial goal

Ambiguity

- Find at least 5 meanings of this sentence:
  - I made her duck

Ambiguity

- Find at least 5 meanings of this sentence:
  - I made her duck
- I cooked waterfowl for her benefit (to eat)
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into undifferentiated waterfowl

Ambiguity is Pervasive

- I caused her to quickly lower her head or body
  - **Lexical category**: "duck" can be a N or V
- I cooked waterfowl belonging to her.
  - **Lexical category:** "her" can be a possessive ("of her") or dative ("for her") pronoun
- I made the (plaster) duck statue she owns
  - **Lexical Semantics:** "make" can mean "create" or "cook"

Ambiguity is Pervasive

- **Grammar**: Make can be:
  - **Transitive: (verb has a noun direct object)**
    - I cooked [waterfowl belonging to her]
  - **Ditransitive: (verb has 2 noun objects)**
    - I made [her] (into) [undifferentiated waterfowl]
  - **Action-transitive (verb has a direct object and another verb)**
  - I caused [her] [to move her body]

Ambiguity is Pervasive

- **Phonetics!**
  - I mate or duck
  - I'm eight or duck
  - Eye maid; her duck
  - Aye mate, her duck
  - I maid her duck
  - I'm aid her duck
  - I mate her duck
  - I'm ate her duck
  - I'm ate or duck
  - I mate or duck

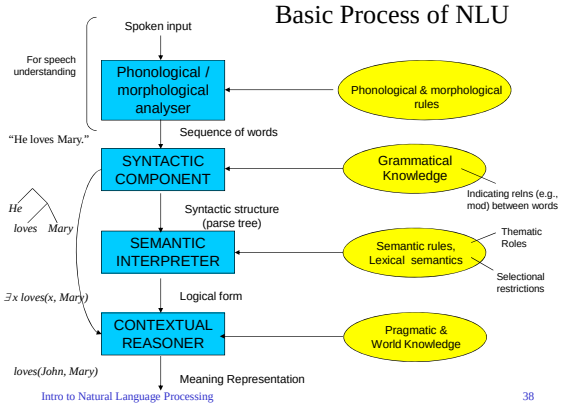
Dealing with Ambiguity

- Four possible approaches:
  1. Tightly coupled interaction among processing levels; knowledge from other levels can help decide among choices at ambiguous levels.
  2. Pipeline processing that ignores ambiguity as it occurs and hopes that other levels can eliminate incorrect structures.

Dealing with Ambiguity

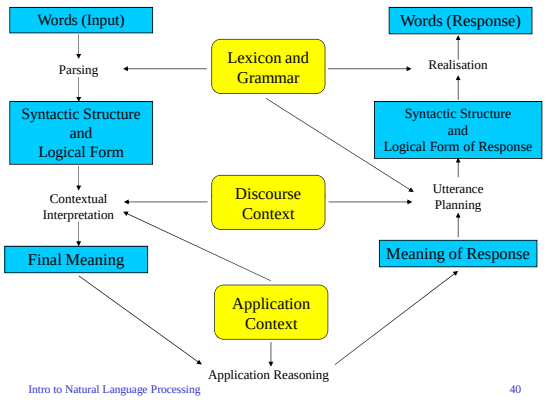
3. Probabilistic approaches based on making the most likely choices
4. Don't do anything, maybe it won't matter
   1. *We'll leave when the duck is ready to eat.*
   2. *The duck is ready to eat now.*
      - Does the "duck" ambiguity matter with respect to whether we can leave?

Basic Process of NLU



It's not that simple

- Syntax affects meaning
  1. (a) *Flying planes is dangerous.*
     (b) *Flying planes are dangerous*.
- Meaning and world knowledge affects syntax
  2. * (a) *Flying insects is dangerous.*
     (b) *Flying insects are dangerous.*
  3. (a) *I saw the Grand Canyon flying to LA.*
     (b) *I saw a condor flying to LA.*



Can machines think?

- Alan Turing: the *Turing test* (language as test for intelligence)
- Three participants: a computer and two humans (one is an interrogator)
- Interrogator's goal: to tell the machine and human apart
- Machine's goal: to fool the interrogator into believing that a person is responding
- Other human's goal: to help the interrogator reach his goal

Examples

Q: Please write me a sonnet on the topic of the Forth Bridge.

A: Count me out on this one. I never could write poetry.

Q: Add 34957 to 70764.

A: 105621 (after a pause)

Example (from a famous movie)

Dave Bowman: Open the pod bay doors, HAL.
HAL: I'm sorry Dave, I'm afraid I can't do that.


Deconstructing HAL

- Recognizes speech and understands language
- Decides how to respond and speaks reply
- With personality
- Recognizes the user's goals, adopts them, and helps to achieve them
- Remembers the conversational history
- Customizes interaction to different individuals
- Learns from experience
- Possesses vast knowledge, and is autonomous

The state of the art and the near-term future

- World-Wide Web (WWW)
- Sample scenarios:
  - generate weather reports in two languages
  - provide tools to help people with SSI to communicate
  - translate Web pages into different languages
  - speak to your appliances
  - find restaurants
  - answer questions
  - grade essays (?)
  - closed-captioning in many languages
  - automatic description of a soccer gams

NLP Applications

- Speech Synthesis, Speech Recognition, IVR Systems (TOOT: more or less succeeds)  
- Information Retrieval (SCANMail demo)
- Information Extraction
  - Question Answering (AQUA)
- Machine Translation (SYSTRAN)
- Summarization (NewsBlaster)
- Automated Psychotherapy (Eliza)

Web demos

- Dialogue
  - ELIZA *http://www.peccavi.com/eliza/*
  - DiaLeague 2001 *http://www.csl.sony.co.jp/SLL/dialeague/*
- Machine Translation (Systran & Altavista)
  - Systran *http://w3.systranlinks.com/systran/cgi*
  - Babel Fish *http://babelfish.altavista.com/translate.dyn*
- Question-answering
  - Ask Jeeves *http://www.ask.co.uk*
- Summarization (IBM)
  - http://www4.ibm.com/software/data/iminer/fortext/summarize/summarizeDemo.html
- Speech synthesis (CSTR at Edinburgh)
  - Festival *http://festvox.org/voicedemos.html*

The alphabet soup
(NLP vs. CL vs. SP vs. HLT vs. NLE)

- NLP (Natural Language Processing)
- CL (Computational Linguistics)
- SP (Speech Processing)
- HLT (Human Language Technology)
- NLE (Natural Language Engineering)
- Other areas of research: Speech and Text Generation, Speech and Text Understanding, Information Extraction, Information Retrieval, Dialogue Processing, Inference
- Related areas: Spelling Correction, Grammar Correction, Text Summarization