Natural Language Processing

Lecture 1—1/13/2015
CSCI 5832
Susan W. Brown

We’re going to study what goes into getting computers to perform useful and interesting tasks involving human language.
More specifically, it’s about the structure of human languages, the algorithms that exploit that structure to process language, and the formal basis for those algorithms.

Why Should You Care?

Three trends

1. An enormous amount of information is now available in machine readable form as natural language text (newspapers, web pages, medical records, financial filings, etc.)
2. Conversational agents are becoming an important form of human-computer communication
3. Much of human-human interaction is now mediated by computers via social media
Applications

• Let’s take a quick look at three important application areas
  ◦ Text analytics
  ◦ Question answering
  ◦ Machine translation

Text Analytics

• Data-mining of weblogs, microblogs, discussion forums, message boards, user groups, and other forms of user generated media
  ◦ Product marketing information
  ◦ Political opinion tracking
  ◦ Social network analysis
  ◦ Buzz analysis (what’s hot, what topics are people talking about right now)
Question Answering

- Traditional *information retrieval* provides documents/resources that provide users with what they need to satisfy their information needs.
- *Question answering* on the other hand directly provides an answer to information needs posed as questions.

Web Q/A

![Google search results for the population of Boulder](image-url)

- Population, Boulder County, CO:
  - 2011 estimate: 233,161
  - Source: U.S. Census Bureau, Population Division

- Boulder, Colorado - Wikipedia, the free encyclopedia
  - The United States Census Bureau estimates that in 2010 the population of the city of Boulder was 94,284, while the population of the Boulder Metropolitan Area...

  [Link to Wikipedia page](wiki-link)
The automatic translation of texts between languages is one of the oldest non-numerical applications in Computer Science. In the past 10 years or so, MT has gone from a niche academic curiosity to a robust commercial industry.
Google Translate

Killing Palestinians and wounding nine in the raids Sector
Nine Palestinians were wounded among civilians in an Israeli air raid in the neighborhood east of the Gaza Strip. This comes immediately after the killing of two prominent Al-Aqsa Martyrs' Brigades in the Israeli-occupied forces carried out an air and infantry forces in the Gush Katif camp in the West Bank.

• Bashir meets Fraser, the Security Council will not impose forces Darfur
In a scheduled meeting with Sudanese President Omar al-Bashir, Jenday Fraser, Assistant Minister for Foreign Affairs of the United Kingdom, attempts to persuade officials in Khartoum, Sudanese Darfur deployment of the international forces. For his part, US ambassador to the United Nations that it has no intention of the Security Council to impose its forces in the present.

Rumsfield and Cheney insist on keeping the American forces in Iraq
Called American Defense Minister Donald Rumsfield Americans to show patience on Iraq. Vice President Dick Cheney calls Democrats withdrawal of American forces from Iraq link and the possibility of early withdrawal of attacks inside the United States.

• Killing civilians and wounding officer suicide attack in Afghanistan
The international force to help establish security (ISAF) killed civilians and the wounding of an officer in an attack against Afghan forces convoy south of the Afghan capital. In the capital Kabul, a hand grenade exploded at the entrance of the building, the Ministry of Defense. A tire was not reported injuries or damage.
How?

All of these applications operate by exploiting underlying regularities inherent in human languages. Sometimes in complex ways, sometimes in pretty trivial ways.

Language structure → Formal models → Practical applications

Major Class Topics

1. Words
2. Syntax
3. Meaning
4. Texts
5. Applications exploiting each
Applications

First, what makes an application a language processing application (as opposed to any other piece of software)?

- An application that requires the use of knowledge about the structure of human language
  - Example: Is Unix wc (word count) an example of a language processing application?

Applications

- Word count?
  - When it counts words: Yes
    - To count words you need to know what a word is. That’s knowledge of language.
    - Note that the definition of “word” embodied in wc doesn’t work for Chinese or other languages that don’t delimit words with spaces
  - When it counts lines and bytes: No
    - Lines and bytes are computer artifacts, not linguistic entities
Caveat

NLP has an distinct **AI** aspect to it
- We’re often dealing with ill-defined problems
- We don’t often come up with exact solutions/algorithms
  - That is, we’re dealing with algorithms that don’t work.
- To make progress we need to have concrete metrics that tell us how well we’re doing, or at least whether our systems are improving or not

Administrative Stuff

- Waitlist
- Web page
  - [verbs.colorado.edu/~mpalmer/csci5832/](verbs.colorado.edu/~mpalmer/csci5832/)
- Reasonable preparation
- Requirements
Web Page

The course web page can be found at:
verbs.colorado.edu/~mpalmer/csci5832/

It will have the syllabus, lecture notes, assignments, announcements, etc.

You should check the News tab periodically for new stuff. I’ll be using this in preference to email.

Mailing List

• There is a automatically generated mailing list.
• Mail goes to your colorado.edu email address.
  ✦ I can’t alter it so don’t ask me to send your mail to gmail/yahoo/work or whatever
  ✦ You can set up a forward yourself
Preparation

• Some exposure to logic
• Exposure to basic concepts in probability
• Familiarity with linguistics
• Ability to write well in English
• Ability to program
• Basic algorithm and data structure analysis

Requirements

• Readings:
  ➡ Speech and Language Processing by Jurafsky and Martin, 2ed. Prentice-Hall 2009
  ➡ A few conference or journal papers
• 3 programming assignments
• Problem sets (about 10)
• 2 midterms
• Final report and presentation
Programming

• Most of the programming will be done in Python.
  ◆ It’s free and works on Windows, Macs, and Linux
  ◆ It’s easy to install
  ◆ Easy to learn

• Go to www.python.org to get started.
• The default installation comes with an editor called IDLE. It’s a serviceable development environment.
• Python mode in Emacs is pretty good. It’s what I use, but I’m a dinosaur.
• If you like Eclipse use that.
Grading

- Programming assignments – 30%
- Problem sets – 18%
- Midterms – 28%
- Final report 14%
- Participation – 10%

Questions?
Course Material

• We’ll be intermingling discussions of:
  ♦ Linguistic topics
    ▪ Morphology, syntax, semantics, discourse
  ♦ Formal systems
    ▪ Regular languages, context-free grammars, probabilistic models
  ♦ Applications
    ▪ Question answering, machine translation, information extraction

Course Material

• We won’t be doing speech recognition or synthesis.
Topics: Linguistics

- Word-level processing
- Syntactic processing
- Lexical and compositional semantics

Topics: Techniques

- Finite-state methods
- Context-free methods
- Probabilistic models

Supervised machine learning methods
Categories of Knowledge

- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse

Each kind of knowledge has associated with it an encapsulated set of processes that make use of it. Interfaces are defined that allow the various levels to communicate. This often leads to a pipeline architecture.

Morphological Processing → Syntactic Analysis → Semantic Interpretation → Context

Ambiguity

- Ambiguity is a fundamental problem in computational linguistics
- Hence, resolving, or managing, ambiguity is a recurrent theme
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 (ceramic?) duck she owns
- I caused her to quickly lower her upper 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 noun or verb
• I cooked waterfowl belonging to her.
  ✷ Lexical category: “her” can be a possessive (“of her”) or dative (“for her”) pronoun
• I made the (ceramic) duck statue she owns
  ✷ Lexical Semantics: “make” can mean “create” or “cook”, and about 100 other things as well

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

Problem

- Remember our pipeline...
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
   1. Or passing along n-best 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?

Models and Algorithms

• By models we mean the formalisms that are used to capture the various kinds of linguistic knowledge we need.
• Algorithms are then used to manipulate the knowledge representations needed to tackle the task at hand.
Models

- State machines
- Rule-based approaches
- Logical formalisms
- Probabilistic models

Algorithms

- Many of the algorithms that we’ll study will turn out to be transducers; algorithms that take one kind of structure as input and output another.
- Unfortunately, ambiguity makes this process difficult. This leads us to employ algorithms that are designed to handle ambiguity of various kinds.
Paradigms

• In particular..
  • State-space search
    ▪ To manage the problem of making choices during processing when we lack the information needed to make the right choice
  • Dynamic programming
    ▪ To avoid having to redo work during the course of a state-space search
      ▪ CKY, Earley, Minimum Edit Distance, Viterbi, Baum-Welch
  • Classifiers
    ▪ Machine learning based classifiers that are trained to make decisions based on features extracted from the local context

Next Time

• Read Chapters 1 and 2 of the textbook