Natural Language Processing

Introduction to NLP

Natural Language Processing

We're going to study what goes into getting computers to perform useful and interesting tasks involving human language.

Slides by James Martin, adapted by Diana Inkpen
for CSI 5386 @ uOttawa

Natural Language Processing

More specifically, it's about the algorithms that we use process language, the formal basis for those algorithms, and the facts about human language that allow those algorithms to work.

Speech and Language Processing - Jurafsky and Martin

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

Speech and Language Processing - Jurafsky and Martin

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)

Speech and Language Processing - Jurafsky and Martin

Text Analytics

Speech and Language Processing - Jurafsky and Martin

Text Analytics

Speech and Language Processing - Jurafsky and Martin

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.

Speech and Language Processing - Jurafsky and Martin

Web Q/A

Google   what's the population of Boulder   Search   Advanced Search

Web   ➕ Show options...   Results **1 - 10** of about **6,290,000** for **what's the population of Boulder**. (0.24 seconds)

Population, Boulder County, CO

293,161 - Jul 2008

Source: U.S. Census Bureau, **Population** Division

Disclaimer

300k / 150k / 0 / 1980 / 2008

www.google.com/publicdata

Boulder, Colorado - Wikipedia, the free encyclopedia

The United States Census Bureau estimates that in 2008 the **population** of the city of **Boulder** was 94268, while the **population** of the **Boulder** Metropolitan ...

Demographics - Geography and climate - History

en.wikipedia.org/wiki/**Boulder**,_Colorado - Cached - Similar - 💬 ⬆ ✖

Watson

Speech and Language Processing - Jurafsky and Martin

Machine Translation

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.

Speech and Language Processing - Jurafsky and Martin

Google Translate

Speech and Language Processing - Jurafsky and Martin

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 result in the Gaza Strip. This comes immediately after the killing of two prominent Al-Aqsa Martyrs Brigades in the Israeli occupying forces carried out air and infantry forces in the Balata camp in the West Bank.

Bashir meets Fraser, the Security Council will not impose forces Darfur

Is scheduled to meet with Sudanese President Omar al-Bashir Jenday Fraser Assistant Minister for Foreign Affairs of the American attempt to persuade officials in Khartoum, Sudanese Darfur deployment of the nationalities. For his part, US Ambassador to the United Nations that it has no intention of the Security Council to impose its forces in the province.

Rmsfield and Cheney insist on keeping the American forces in Iraq

Called American Defense Minister Donald Rmsfield Americans to show patience on Iraq. I take 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 Atlantic Afghanistan. In the capital Kabul, a hand grenade exploded at the passage of manufacture French patrol was not reported injuries or damage.

Speech and Language Processing - Jurafsky and Martin

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 |

Speech and Language Processing - Jurafsky and Martin
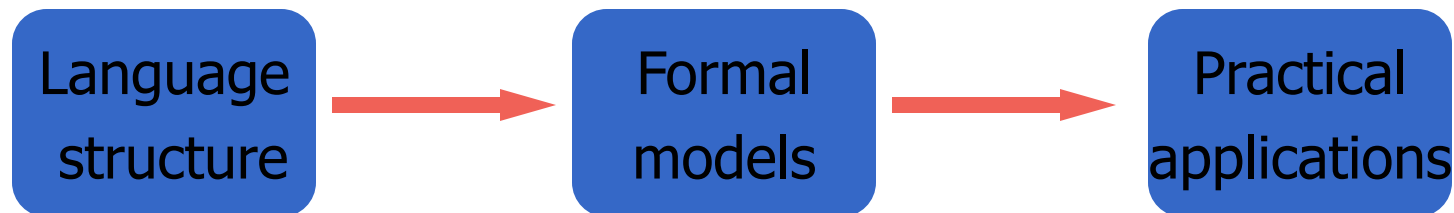
Major Class Topics

1. Words
2. Syntax
3. Meaning (sematics)
4. Texts (discourse)

5. Applications exploiting each

Speech and Language Processing - Jurafsky and Martin

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?

Speech and Language Processing - Jurafsky and Martin

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

Speech and Language Processing - Jurafsky and Martin

Questions?

Speech and Language Processing - Jurafsky and Martin

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

Speech and Language Processing - Jurafsky and Martin

Course Material

- We won't be doing speech recognition or synthesis.

Speech and Language Processing - Jurafsky and Martin

Topics: Linguistics

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

Speech and Language Processing - Jurafsky and Martin

Topics: Techniques

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

Supervised machine learning methods

Speech and Language Processing - Jurafsky and Martin

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

Speech and Language Processing - Jurafsky and Martin
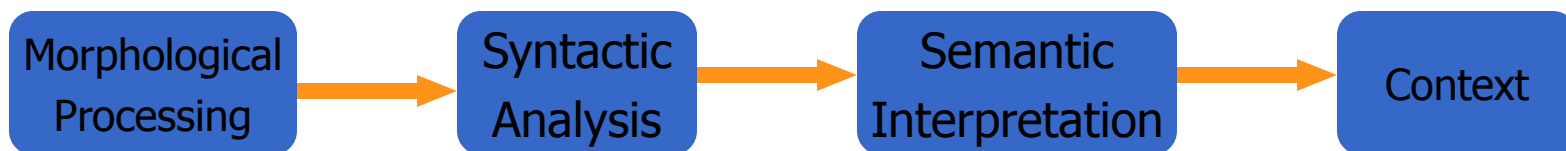
Ambiguity

- Ambiguity is a fundamental problem in computational linguistics

- Hence, resolving, or managing, ambiguity is a recurrent theme

Speech and Language Processing - Jurafsky and Martin

Ambiguity

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

Speech and Language Processing - Jurafsky and Martin

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

Speech and Language Processing - Jurafsky and Martin

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]

Speech and Language Processing - Jurafsky and Martin

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...

| Morphological Processing | → | Syntactic Analysis | → | Semantic Interpretation | → | Context |

Speech and Language Processing - Jurafsky and Martin

Really it's this

Speech and Language Processing - Jurafsky and Martin
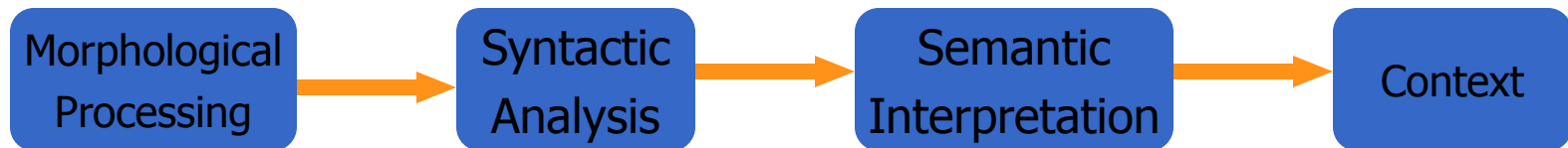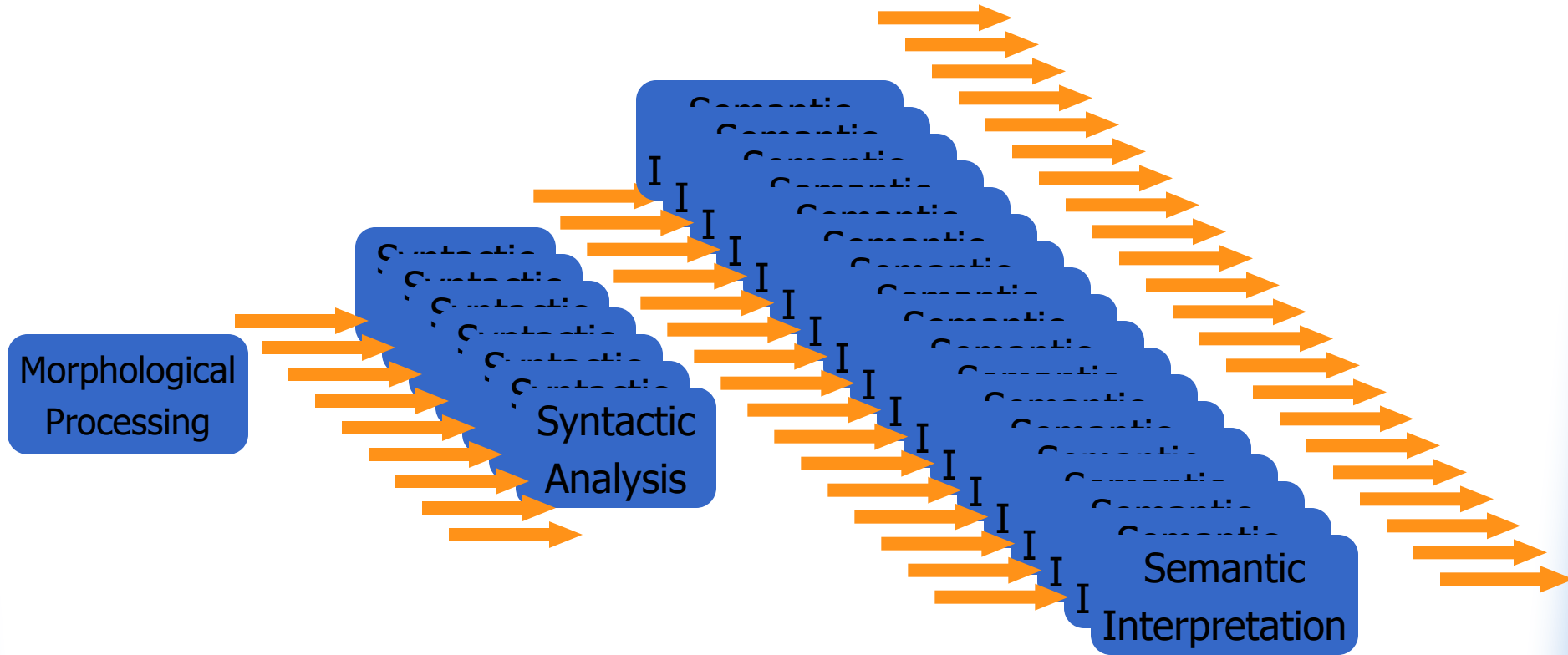
Dealing with Ambiguity

- 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.

Speech and Language Processing - Jurafsky and Martin

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?

Speech and Language Processing - Jurafsky and Martin

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

Speech and Language Processing - Jurafsky and Martin

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

Speech and Language Processing - Jurafsky and Martin

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

Speech and Language Processing - Jurafsky and Martin