DRONACHARYA College of Engineering

INTELLIGENT SYSTEMS (CSE-303-F)

Section D

Natural Language Processing

Any Light at The End of The Tunnel?

- Yahoo, Google, Microsoft → Information Retrieval
- Monster.com, HotJobs.com (Job finders) → Information Extraction + Information Retrieval
- Systran powers Babelfish → Machine Translation
- Ask Jeeves → Question Answering
- Myspace, Facebook, Blogspot → Processing of User-Generated Content
- Tools for "business intelligence"
- All "Big Guys" have (several) strong NLP research labs:
  - IBM, Microsoft, AT&T, Xerox, Sun, etc.
- Academia: research in an university environment

Why Natural Language Processing ?

- Huge amounts of data
  - Internet = at least 20 billions pages
  - Intranet
- Applications for processing large amounts of texts

  require NLP expertise

- Classify text into categories
- Index and search large texts
- Automatic translation
- Speech understanding
  - Understand phone conversations
- Information extraction
  - Extract useful information from resumes
- Automatic summarization
  - Condense 1 book into 1 page
- Question answering
- Knowledge acquisition
- Text generations / dialogues

Natural?

- Natural Language?
  - Refers to the language spoken by people, e.g. English, Japanese, Swahili, as opposed to artificial languages, like C++, Java, etc.
- Natural Language Processing
  - Applications that deal with natural language in a way or another
- [Computational Linguistics
  - Doing linguistics on computers
  - More on the linguistic side than NLP, but closely related]

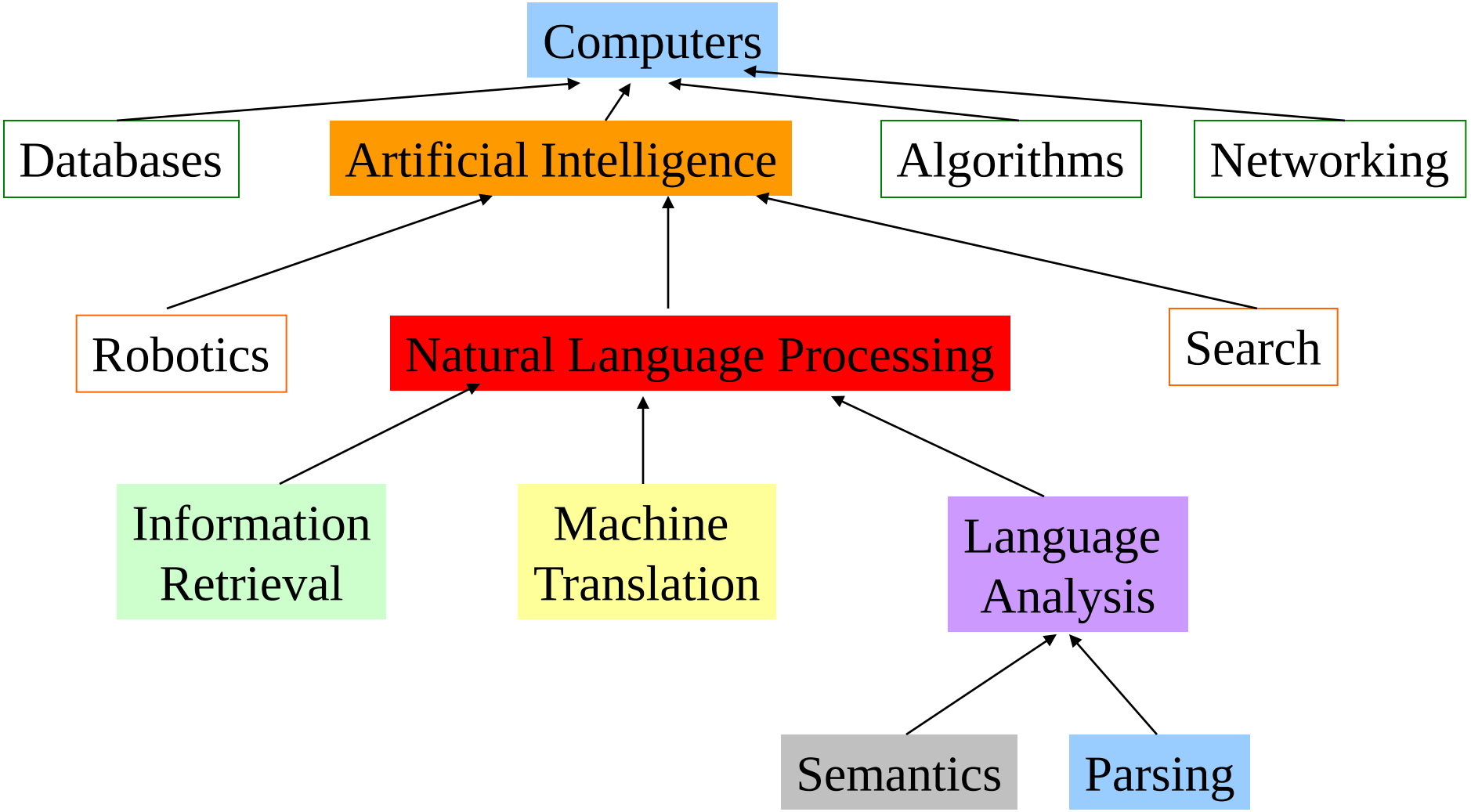
Why Natural Language Processing?

- kJfmmfj mmmvvv nnnffn333
- Uj iheale eleee mnster vensi credur
- Baboi oi cestnitze
- Coovoel2^ ekk; ldsllk lkdf vnnjfj?
- Fgmflmllk mlfm kfre xnnn!

Computers Lack Knowledge!

- Computers “see” text in English the same you have seen the previous text!
- People have no trouble understanding language
  - Common sense knowledge
  - Reasoning capacity
  - Experience
- Computers have
  - No common sense knowledge
  - No reasoning capacity

Where does it fit in the CS taxonomy?



Linguistics Levels of Analysis

- Speech
- Written language
  - Phonology: sounds / letters / pronunciation
  - Morphology: the structure of words
  - Syntax: how these sequences are structured
  - Semantics: meaning of the strings
- Interaction between levels

Issues in Syntax

"the dog ate my homework" - Who did what?

1. Identify the part of speech (POS)

   Dog = noun ; ate = verb ; homework = noun

   English POS tagging: 95%

2. Identify collocations

   mother in law, hot dog

   Compositional versus non-compositional collocates

Issues in Syntax

- Shallow parsing:

  “the dog chased the bear”

  “the dog” “chased the bear”

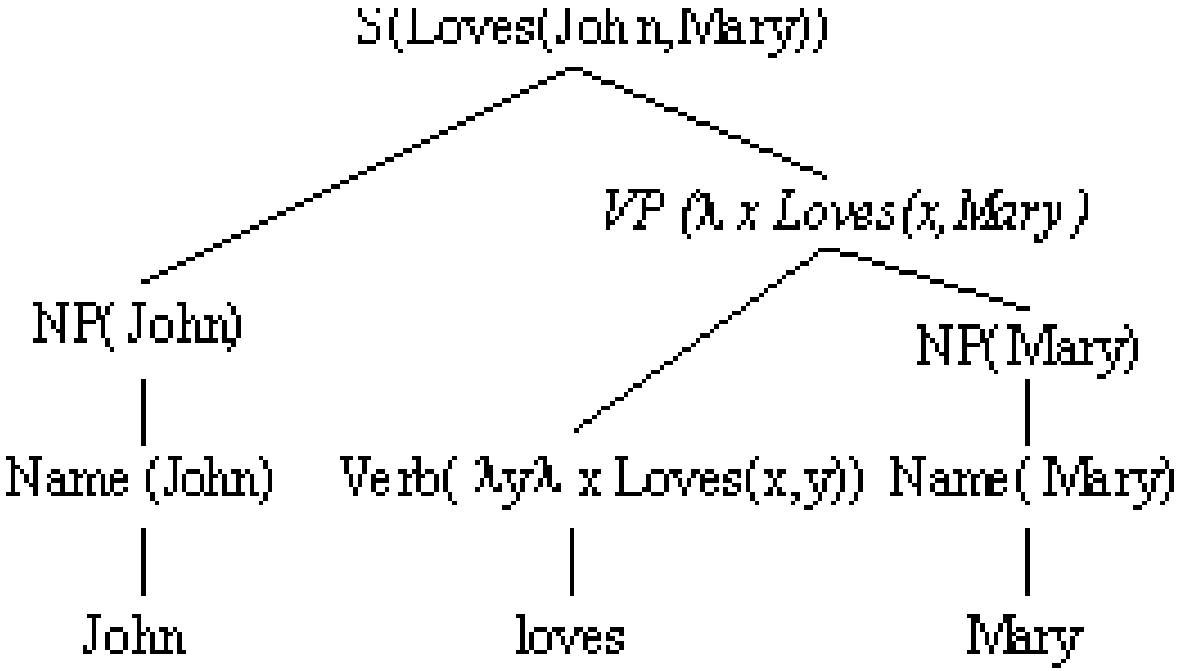
  subject - predicate

  Identify basic structures

  NP-[the dog] VP-[chased the bear]

Issues in Syntax

- Full parsing: John loves Mary



Help figuring out (automatically) questions like: Who did what and when?

More Issues in Syntax

- Anaphora Resolution:

"The dog entered my room. It scared me"

- Preposition Attachment

"I saw the man in the park with a telescope"

Issues in Semantics

- Understand language! How?
- *"plant" = industrial plant*
- *"plant" = living organism*
- Words are ambiguous
- Importance of semantics?
  - Machine Translation: wrong translations
  - Information Retrieval: wrong information
  - Anaphora Resolution: wrong referents

Why Semantics?

- The sea is at the home for billions factories and animals

- The sea is home to million of plants and animals
- English → French [commercial MT system]
- Le mer est a la maison de billion des usines et des animaux
- French → English

Issues in Semantics

- How to learn the meaning of words?
- From dictionaries:

  plant, works, industrial plant -- (buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles")

  plant, flora, plant life -- (a living organism lacking the power of locomotion)

They are producing about 1,000 automobiles in the new plant

The sea flora consists in 1,000 different plant species

The plant was close to the farm of animals.

Issues in Semantics

- Learn from annotated examples:
  - Assume 100 examples containing “plant” previously tagged by a human
  - Train a learning algorithm
  - How to choose the learning algorithm?
  - How to obtain the 100 tagged examples?

Issues in Information Extraction

- "There was a group of about 8-9 people close to the entrance on Highway 75"
- Who? "8-9 people"
- Where? "highway 75"

- Extract information
- Detect new patterns:
  - Detect hacking / hidden information / etc.
- Gov./mil. puts lots of money put into IE research

Issues in Information Retrieval

- General model:
  - A huge collection of texts
  - A query
- Task: find documents that are relevant to the given query
- How? Create an index, like the index in a book
- More …
  - Vector-space models
  - Boolean models
- Examples: Google, Yahoo, Altavista, etc.

Issues in Information Retrieval

- Retrieve specific information
- Question Answering
- “What is the height of mount Everest?”
- 11,000 feet

Issues in Information Retrieval

- Find information across languages!
- Cross Language Information Retrieval
- “What is the minimum age requirement for car rental in Italy?”
- Search also Italian texts for “eta minima per noleggio macchine”
- Integrate large number of languages
- Integrate into performant IR engines

Issues in Machine Translations

- Text to Text Machine Translations
- Speech to Speech Machine Translations

- Most of the work has addressed pairs of widely spread languages like English-French, English-Chinese

Issues in Machine Translations

- How to translate text?
  - Learn from previously translated data
- → Need parallel corpora
- French-English, Chinese-English have the Hansards
- Reasonable translations
- Chinese-Hindi – no such tools available today!

Even More

- Discourse
- Summarization
- Subjectivity and sentiment analysis
- Text generation, dialog [pass the Turing test for some million dollars] – Loebner prize
- Knowledge acquisition [how to get that common sense knowledge]
- Speech processing

What will we study this semester?

- Intro to Perl
  - Great great for text processing
  - Fast: one person can do the work of ten others
  - Easy to pick up
- Some linguistic basics
  - Structure of English
  - Parts of speech, phrases, parsing
- Morphology
- N-grams
  - Also multi-word expressions
- Part of speech tagging
- Syntactic parsing
- Semantics
  - Word sense disambiguation
  - Semantic relations

What will we study this semester?

- Information Retrieval
- Question answering
- Text classification
- Text summarization
- Sentiment analysis

- Depending on time, we may touch on
  - Speech recognition
  - Dialogue
  - Text generation
  - Other topics of your interest

Administrivia

- Instructor: Rada Mihalcea, F228, rada@cs.unt.edu
- Class meetings: TTh 11-12:20pm
- Office hours: TTh 4:00-5:00pm
- TA: TBA
- Textbook: Speech and Language Processing, by Jurafsky and Martin (2nd edition)
  - Recommended: Statistical Methods in NLP, by Manning and Schutze
- Other readings (papers) may be assigned throughout the semester
- Grading: Assignments, 2 exams, term project
  - Late submission policy for assignments: can submit up to three days late, with 10% penalty / day