The Dream

• It would be great if machines could
  – Process our emails
  – Translate languages accurately
  – Help us manage, summarize, and aggregate information
  – Understand phone conversation
  – Talk to us / listen to us

• But they cannot:
  – Language is complex, ambiguous, flexible, and subtle
  – Good solutions need linguistics and machine learning knowledge
What is NLP?

- Goal: intelligent processing of human language
  - Not just string and keyword matching

- End systems we want to build:
  - Less ambitious: spelling correction, name entity extractors
  - Ambitious: machine translation, information extraction, question answering, summarization...

Information Extraction

- Goal: build database entries from unstructured text
- Simple Task: Named Entity Extraction

---

Information Extraction

• Goal: build database entries from unstructured text
• Advanced: Multi-sentence template extraction

A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb-allegedly detonated by urban guerrilla commandos blew up a power tower in the north western part of San Salvador at 0650.

Incident type:
Date:
Location:
Perpetrator:
Physical target:
Effect on physical target:
Effect on human target:
Instrument:
Information Extraction

- Goal: build database entries from unstructured text
- Advanced: Multi-sentence template extraction

A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb-allegedly detonated by **urban guerrilla commandos** blew up a power tower in the north western part of San Salvador at 0650.

<table>
<thead>
<tr>
<th>Incident type:</th>
<th>bombing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date:</td>
<td>March 11, 2010</td>
</tr>
<tr>
<td>Location:</td>
<td>San Salvador (city)</td>
</tr>
<tr>
<td>Perpetrator:</td>
<td>urban guerrilla commandos</td>
</tr>
<tr>
<td>Physical target:</td>
<td>power tower</td>
</tr>
<tr>
<td>Effect on physical target:</td>
<td>destroyed</td>
</tr>
<tr>
<td>Effect on human target:</td>
<td>no injury or death</td>
</tr>
<tr>
<td>Instrument:</td>
<td>bomb</td>
</tr>
</tbody>
</table>

Information Retrieval

- Given a huge collection of text and a query
- Goal: find documents that are relevant to the query

![Google search results](https://via.placeholder.com/150)
Question Answering

- Find answers to general comprehension questions in a document collection

Text Summarization

http://emm-labs.jrc.it/EMMLabs/NewsGist.html
Machine Translation

Japan buys 20% of the eurozone bond

The country Japan is planning to acquire 20% of the debt issue that the eurozone finance minister says the operation is performed in order to boost confidence.

Japón comprará el 20% de los bonos que a fines de mes

El país nipón tiene previsto adquirir el 20% de la emisión de deuda que Finanzas del país asegura que la operación se realiza con el fin de imp॰

Speech Processing

• Automatic Speech Recognition

“will you move the clinic there?”

• Performance: 5% for dictation, 50%+TV
Sentiment Analysis

• Tracking sentiment towards politicians, movies, products.
• Assisting in writing e-mails, documents, and other text to convey desired emotion (and avoiding misinterpretation).
• Detecting how people use emotion-bearing-words to persuade and coerce others
• Deception detection

Text Categorization

• Build a system that groups text based on certain criteria
  – Spam Filtering
  – Share the same Topic (news clustering)
Linguistics Levels of Analysis

- Phonology: sounds / letters / pronunciation
- Morphology: construction of words
- Syntax: structural relationships between words
- Semantics: meaning of strings (words, phrases)
- Discourse: relationships across different sentences
- Pragmatics: how we use language to communicate
- World Knowledge: facts about the world, common sense

MORPHOLOGY
Morphological Analysis

- *Morphology* studies the internal structure of words
- A *morpheme* is the smallest linguistic unit that has semantic meaning (Wikipedia)
- *Morphological Analysis* is the task of segmenting a word into its morphemes
  - carried => carry + ed (past tense)
  - disconnect => dis (not) + connect
- Challenging for morphologically rich languages like Finnish and Turkish
Part-of-Speech Tagging (POS)

- Annotate each word in a sentence with a part-of-speech tag

  \[ \text{I ate the spaghetti with meatballs.} \]

  \[ \text{Pro V Det N Prep N} \]

- Useful for syntactic parsing and word sense disambiguation

- English POS tagging 95% accurate

Phrase Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.

  \[ \text{[NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs]} \]
Syntactic Parsing

• Produce syntactic parse tree of a sentence

I ate the spaghetti with meatballs.

• Help figuring out questions like: *Who did what and when?*

More issues in Syntax

• Prepositional Attachment
  “I saw the man with the telescope”

Syntax does not tell us much about meaning
SEMANTIC TASKS

Word Sense Disambiguation

• Understand language! How?

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

• Useful for machine translation, information retrieval
How to learn the meaning of words?

• From dictionaries, lexical repository like WordNet
  bank -- *sloping land, especially the slope beside a body of water*
  ex. "they pulled the canoe up on the bank"
  bank – *a financial institution that accepts deposits and channels the money into lending activities*
  ex. "he cashed a check at the bank"

• Automatically from the Web

Semantic Role Labeling

• For each clause, determine the semantic role played by each noun phrase that is an argument to the verb

  agent     patient     source     destination
  John      drove      Mary      from      LA      to      San Diego.
Textual Entailment

- Determine whether one natural language sentence entails another

  The glass is half empty.
  The glass is half full.

  Google bought Youtube.
  Google acquired Youtube.

DISCOURSE, PRAGMATICS AND WORLD KNOWLEDGE
Anaphora Resolution

• Determine which phrases in a document refer to the same entity

  “George woke up. He went to the kitchen.”
  “Peter put the carrot on the plate and ate it.”

Pragmatics

• Studies how language is used to accomplish goals

  What can we conclude from the following sentences?
  “Could you please pass me the salt?”
  “I am afraid I cannot do this”
World Knowledge

“George woke up. He went to the bathroom and started shaving. He took the car key and left.”

WHERE WE STAND TODAY
What can NLP do (robustly) today?

- Surface-level **preprocessing** (POS tagging, word segmentation, named entity extraction): 94%+
- Shallow syntactic **parsing**: 92%+ for English
- **IE**: ~40% for well-behaved topics (MUC, ACE)
- **Speech**: ~80% large vocab; 20%+ open vocab, noisy input
- **IR**: 40% (TREC)
- **MT**: ~70% depending on what you measure
- **Summarization**: ? (~60% for extracts; DUC)
- **QA**: ? (~60% for factoids; TREC)

What cannot NLP do today?

- Do general-purpose **text generation**
- Deliver **semantics**—either in theory or in practice
- Deliver **long/complex answers** by extracting, merging, and summarizing web info
- Handle extended **dialogues**
- **Read and learn** (extend own knowledge)
- Use **pragmatics** (style, emotion, user profile...)
- Provide significant contributions to a **theory of Language** (in Linguistics or Neurolinguistics) or of **Information** (in Signal Processing)
CLASS DETAILS

What is in this Class?

• Some linguistic basics
  – structure of English
  – sentence segmentation, tokenization
  – Morphology

• Syntactic parsing

• Semantics
  – Word sense disambiguation
  – Semantic relations
  – Textual Entailment
  – Paraphrases
  – Noun Compounds
What is in this Class?

• Applications:
  – Information Extraction
    • Named entity extraction
    • Relation extraction
  – Information Retrieval
  – Text Categorization
    • Clustering
    • Latent Semantic Analysis
  – Question Answering
    • Question classification
  – Sentiment Analysis
  – Text Summarization
  – Machine Translation

What is in this Class?

• Overview of Tools
  – Weka
  – Mallet

• Machine Learning Algorithms
  – Supervised learning
  – Semi-supervised learning
  – Unsupervised learning

• Graph Theory
What will we learn?

• We will learn issues and techniques in NLP
• We will learn various tools
• We will learn about applications that can benefit from NLP
• We will understand issues involved in processing natural language
• We will develop skills necessary to build NLP tools
• We will learn to solve real-world problems

Class Requirements

• Class requirements:
  – Basic linguistics background
  – Basic linear algebra
  – Decent coding skills
Course Work

• Recommended Readings:

• Assignments:
  – 2 coding assignments
    • brief 2-3 paged description
    • code
  – 1 final project
    • power point presentation / poster
    • 8 paged report which will follow ACL guidelines
    • code (demo visualizing input-output of your system with graphs etc)

NLP AT USC, ISI AND ICT
Ph.D. Researchers and Topics

At ISI:
- David Chiang — parsing, statistical processing
- Ulf Hermjakob — parsing, QA, language learning
- Jerry Hobbs — semantics, ontologies, discourse
- Eduard Hovy (on leave) — summarization, ontologies, NLG
- Kevin Knight — MT, NLG, decipherment
- Zornitsa Kozareva — IE, text mining, lexical semantics, sentiment analysis
- Daniel Marcu — MT, QA, summarization, discourse

At ICT:
- David DeVault — NL generation
- Andrew Gordon — cognitive science and language
- Anton Leuski — IR
- Kenji Sagae — parsing
- Bill Swartout — NLG
- David Traum — dialogue

At USC/EE:
- Shri Narayanan — speech recognition