Natural Language Processing

What is Natural Language Processing?

- **Natural Language Processing (NLP)** : The process of computer analysis of input provided in a human language (natural language), and conversion of this input into a useful form of representation.
- The field of NLP is primarily concerned with getting computers to perform useful and **interesting tasks** with human languages.
- The field of NLP is secondarily concerned with helping us come to a better understanding of human language.

- The **goal of NLP field** is to get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech.

Forms of Natural Language

- The input/output of a NLP system can be:
  - **written text**
  - **speech**
- We will mostly concerned with written text in this course (not speech).
- To process written text, we need:
  - **lexical, syntactic, semantic knowledge about the language**
  - **discourse information, real world knowledge**
- To process spoken language, we need everything required to process written text, plus the challenges of speech recognition and speech synthesis.

NLP Tasks

- An application that requires the use of knowledge about human languages can be seen as **a NLP task**.
  - *Word count* is a NLP application since we need to know what a word is. That's knowledge of language.
  - Line or byte count is not a NLP application.

- Some big NLP Tasks require a tremendous amount of knowledge of language.
  - Conversational agents
  - Machine translation
  - Question answering
  - Information extraction

- … and many more NLP tasks

NLP Tasks: Conversational agents

- HAL computer in the movie ``*2001: A Space Odyssey*`` is an artificial agent capable of such advanced language-processing behavior as speaking and understanding English.
- We call programs like HAL that converse with humans via natural language **conversational agents** or **dialogue systems**.

- These kinds of applications require a tremendous amount of knowledge of language.
  - Speech recognition and synthesis
  - Knowledge of the English words involved
  - How groups of words clump and what the clumps mean?
  - Discourse

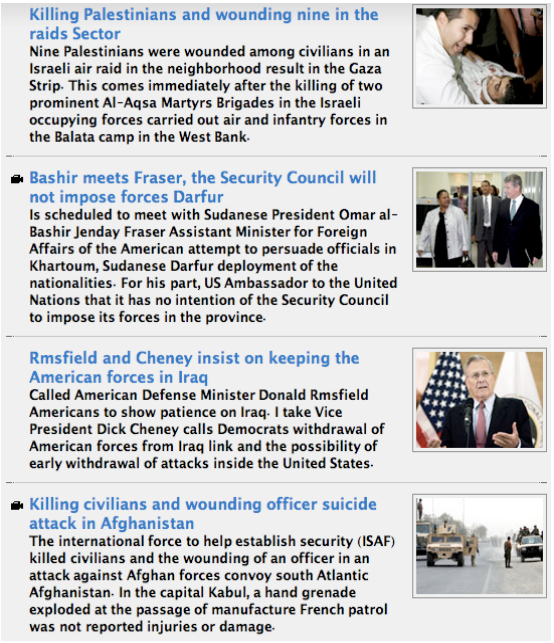
NLP Tasks: Machine translation

- The goal of **machine translation** is to automatically translate a document from one language to another.
- Translation from *Stanford's Phrasal*:

这 不过 是 一 个 时间 的 问题 . ➔ This is only a matter of time.

- 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.

NLP Tasks: Question answering

- Question answering task is to find answers for the complete questions ranging from easy to hard questions.
  - What does “divergent” mean?
  - What year was Abraham Lincoln born?
  - How many states were in the United States that year?
  - How much Chinese silk was exported to England by the end of the 18th century?
  - What do scientists think about the ethics of human cloning?
- Some of these question, such as **definition** questions, or simple **factoid** questions like dates and locations can be easily answered.
- Answering more complicated questions might require **extracting information** that is embedded in the text, or doing **inference** (drawing conclusions based on known facts), or synthesizing and **summarizing information** from multiple sources.

NLP Tasks: Information extraction

- **Information extraction** is the extraction of events and its attributes from natural language texts.

Subject: **curriculum meeting**
Date: January 15, 2012
To: Dan Jurafsky

Hi Dan, we've now scheduled the curriculum meeting.
It will be in Gates 159 tomorrow from 10:00-11:30.

➔

Event: Curriculum mtg
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

Language Technology

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA ... ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

⇨ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Knowledge in Language Processing

- What distinguishes *language processing applications* from other data processing systems is their use of **knowledge of language**.
- Some simple NLP tasks require limited knowledge of language.
- Big NLP tasks such as conversational agents, machine translation systems, robust question-answering systems, require much broader and deeper knowledge of language.

- **Phonology** – concerns how words are related to the sounds that realize them.
- **Morphology** – concerns how words are constructed from more basic meaning units called morphemes. A ***morpheme*** is the primitive unit of meaning in a language.
- **Syntax** – concerns how can be put together to form correct sentences and determines what structural role each word plays in the sentence and what phrases are subparts of other phrases.

Knowledge in Language Processing

- **Semantics** – concerns what words mean and how these meaning combine in sentences to form sentence meaning. The study of context-independent meaning.

- **Pragmatics** – concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

- **Discourse** – concerns how the immediately preceding sentences affect the interpretation of the next sentence.
  - For example, interpreting pronouns and interpreting the temporal aspects of the information.

- **World Knowledge** – includes general knowledge about the world.
  - What each language user must know about the other's beliefs and goals.

Why NLP is hard?

- Natural language is extremely rich in form and structure, and **very ambiguous**.
  - How to represent meaning,
  - Which structures map to which meaning structures.
- One input can mean many different things and **Ambiguity** can be at different levels.
  - Lexical (word level) ambiguity -- different meanings of words
  - Syntactic ambiguity -- different ways to parse the sentence
  - Interpreting partial information -- how to interpret pronouns
  - Contextual information -- context of the sentence may affect the meaning of that sentence.
- Many input can mean the same thing.
- Interaction among components of the input is not clear.

Ambiguity

I made her duck.

- How many different interpretations does this sentence have?
- What are the reasons for the ambiguity?
- The categories of knowledge of language can be thought of as ambiguity resolving components.
- How can each ambiguous piece be resolved?
- Does speech input make the sentence even more ambiguous?
  - Yes – deciding word boundaries

Ambiguity (cont.)

- Some interpretations of : **`I made her duck.`**
  1. I cooked *duck* for her.
  2. I cooked *duck* belonging to her.
  3. I created a toy duck which she owns.
  4. I caused her to quickly lower her head or body.
  5. I used magic and turned her into a *duck*.

- `duck` – morphologically and syntactically ambiguous: noun or verb.
- `her` – syntactically ambiguous: dative or possessive.
- `make` – semantically ambiguous: cook or create.
- `make` – syntactically ambiguous:
  - Transitive – takes a direct object. => 2
  - Di-transitive – takes two objects. => 5
  - Takes a direct object and a verb. => 4

Ambiguity in a Turkish Sentence

- Some interpretations of: **Adamı gördüm.**
  1. I saw the man.
  2. I saw my island.
  3. I visited my island.
  4. I bribed the man.
- Morphological Ambiguity:
  - ada-m-ı ada+P1SG+ACC
  - adam-ı adam+ACC
- Semantic Ambiguity:
  - gör to see
  - gör to visit
  - gör to bribe

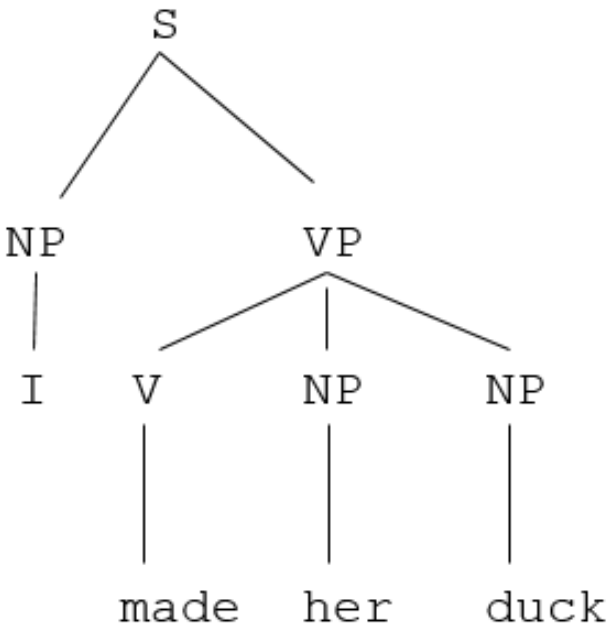
Resolve Ambiguities

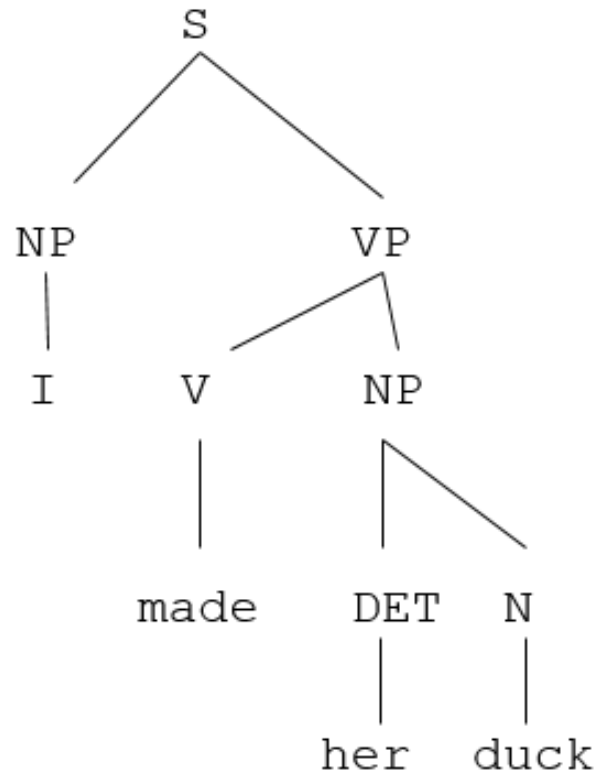
- We will introduce *models* and *algorithms* to resolve ambiguities at different levels.

- **part-of-speech tagging** -- Deciding whether `duck` is verb or noun.
- **word-sense disambiguation** -- Deciding whether `make` is `create` or `cook`.
- **lexical disambiguation** -- Resolution of part-of-speech and word-sense ambiguities are two important kinds of lexical disambiguation.
- **syntactic ambiguity** -- `her duck` is an example of syntactic ambiguity, and can be addressed by probabilistic parsing.

Resolve Ambiguities (cont.)

I made her duck





Models to Represent Linguistic Knowledge

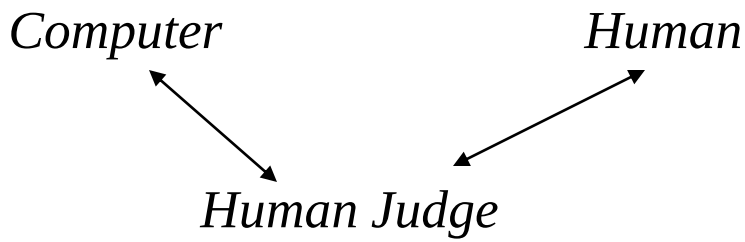
- We will use certain formalisms (*models*) to represent the required linguistic knowledge.
- **State Machines** -- FSAs, FSTs, HMMs, ATNs, RTNs
- **Formal Rule Systems** -- Context Free Grammars, Unification Grammars, Probabilistic CFGs.
- **Logic-based Formalisms** -- first order predicate logic, some higher order logic.
- **Models of Uncertainty** -- Bayesian probability theory.
- **Vector-space models –** to represent meanings of words

Algorithms to Manipulate Linguistic Knowledge

- We will use *algorithms* to manipulate the models of linguistic knowledge to produce the desired behavior.
- Most of the algorithms we will study are **transducers** and **parsers**.
  - These algorithms construct some structure based on their input.
- Since the language is ambiguous at all levels, these algorithms are never simple processes.
- Categories of most algorithms that will be used can fall into following categories.
  - state space search
  - dynamic programming

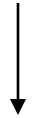
Language and Intelligence

Turing Test



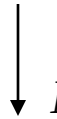
- *Human Judge* asks tele-typed questions to *Computer* and *Human.*
- *Computer's* job is to act like a human.
- *Human's* job is to convince Judge that he is not machine.
- *Computer* is judged "intelligent" if it can fool the judge
- Judgment of intelligence is linked to appropriate answers to questions from the system.

Natural Language Understanding



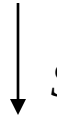
Words

Morphological Analysis



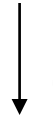
Morphologically analyzed words *(another step: POS tagging)*

Syntactic Analysis



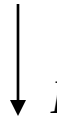
Syntactic Structure

Semantic Analysis



Context-independent meaning representation

Discourse Processing



Final meaning representation

Morphological Analysis

- Analyzing words into their linguistic components (morphemes).
- Morphemes are the smallest meaningful units of language.

cars	car+PLU	
giving	give+PROG	
geliyordum	gel+PROG+PAST+1SG	- I was coming

- Ambiguity: More than one alternatives

flies	fly_{VERB}+PROG	
	fly_{NOUN}+PLU	
adamı	adam+ACC	- the man (accusative)
	adam+P1SG	- my man
	ada+P1SG+ACC	- my island (accusative)

Morphological Analysis (cont.)

- Relatively simple for English. But for some languages such as Turkish, it is more difficult.

  uygarlaştıramadıklarımızdanmışsınızcasına

  uygar-laş-tır-ama-dık-lar-ımız-dan-mış-sınız-casına

  uygar +BEC +CAUS +NEGABLE +PPART +PL +P1PL +ABL +PAST +2PL +AsIf

  "(behaving) as if you are among those whom we could not civilize/cause to become civilized"

+BEC	is "become" in English
+CAUS	is the causative voice marker on a verb
+PPART	marks a past participle form
+P1PL	is 1st person plural possessive marker
+2PL	is 2nd person plural
+ABL	is the ablative (from/among) case marker
+AsIf	is a derivational marker that forms an adverb from a finite verb form
+NEGABLE	is "not able" in English

- Inflectional and Derivational Morphology.
- Common tools: Finite-state transducers

Part-of-Speech (POS) Tagging

- Each word has a part-of-speech tag to describe its category.
- Part-of-speech tag of a word is one of major word groups (or its subgroups).
  - **open classes** -- noun, verb, adjective, adverb
  - **closed classes** -- prepositions, determiners, conjuctions, pronouns, particples
- POS Taggers try to find POS tags for the words.
- `duck` is a verb or noun? (morphological analyzer cannot make decision).
- A POS tagger may make that decision by looking the surrounding words.
  - `Duck!` (verb)
  - `Duck is delicious for dinner.` (noun)

Lexical Processing

- The purpose of lexical processing is to determine meanings of individual words.
- Basic methods is to lookup in a database of meanings -- **lexicon**
- We should also identify non-words such as punctuation marks.
- Word-level ambiguity -- words may have several meanings, and the correct one cannot be chosen based solely on the word itself.
  - bank in English
  - yüz in Turkish
- Solution -- resolve the ambiguity on the spot by POS tagging (if possible) or pass-on the ambiguity to the other levels.

Syntactic Processing

- **Parsing** -- converting a flat input sentence into a hierarchical structure that corresponds to the units of meaning in the sentence.
- There are different parsing formalisms and algorithms.
- Most formalisms have two main components:
  - **grammar** -- a declarative representation describing the syntactic structure of sentences in the language.
  - **parser** -- an algorithm that analyzes the input and outputs its structural representation (its parse) consistent with the grammar specification.
- CFGs are in the center of many of the parsing mechanisms. But they are complemented by some additional features that make the formalism more suitable to handle natural languages.

Semantic Analysis

- Assigning meanings to the structures created by syntactic analysis.
- Mapping words and structures to particular domain objects in way consistent with our knowledge of the world.
- Semantic can play an import role in selecting among competing syntactic analyses and discarding illogical analyses.
  - `I robbed the bank` -- bank is a river bank or a financial institution
- We have to decide the formalisms which will be used in the meaning representation.

Knowledge Representation for NLP

- Which knowledge representation will be used depends on the application .
  - Requires the choice of representational framework, as well as the specific meaning vocabulary (what are concepts and relationship between these concepts -- ontology)
  - Must be computationally effective.
- Common representational formalisms:
  - first order predicate logic
  - conceptual dependency graphs
  - semantic networks
  - Frame-based representations
  - Vector-space models

Discourse

- Discourses are collection of coherent sentences (not arbitrary set of sentences)
- Discourses have also hierarchical structures (similar to sentences)
- **anaphora resolution** -- to resolve referring expression
  - `Mary bought a book for Kelly. She didn't like it.`
    - **She** refers to Mary or Kelly. -- possibly Kelly
    - **It** refers to what -- book.
  - `Mary had to lie for Kelly. She didn't like it.`
- Discourse structure may depend on application.
  - Monologue
  - Dialogue
  - Human-Computer Interaction

Natural Language Generation (NLG)

- Natural Language Generation (NLG) is the process of constructing natural language outputs from non-linguistic inputs.
- NLG can be viewed as the reverse process of NL understanding.
- A NLG system may have two main parts:
  - **Discourse Planner** -- what will be generated: *which sentences*.
  - **Surface Realizer** -- realizes a sentence from its internal representation.
    - **Lexical Selection** -- selecting the correct words describing the concepts.