

Introduction to Natural Language Processing (NLP)



Content

- **What we will learn today:**
 - What is NLP?
 - NLP and the Turing Test
 - Why NLP?
 - Goal of NLP
 - Levels of NLP Processing and Analysis
 - NLP Phases
 - NLP common tasks
 - NLP applications

What is NLP?

- A field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages.
- The study of human languages and how they can be represented computationally, analyzed and generated algorithmically.
- The cake is on the table → on(table, cake)
- on(floor, table) → The table is on the floor

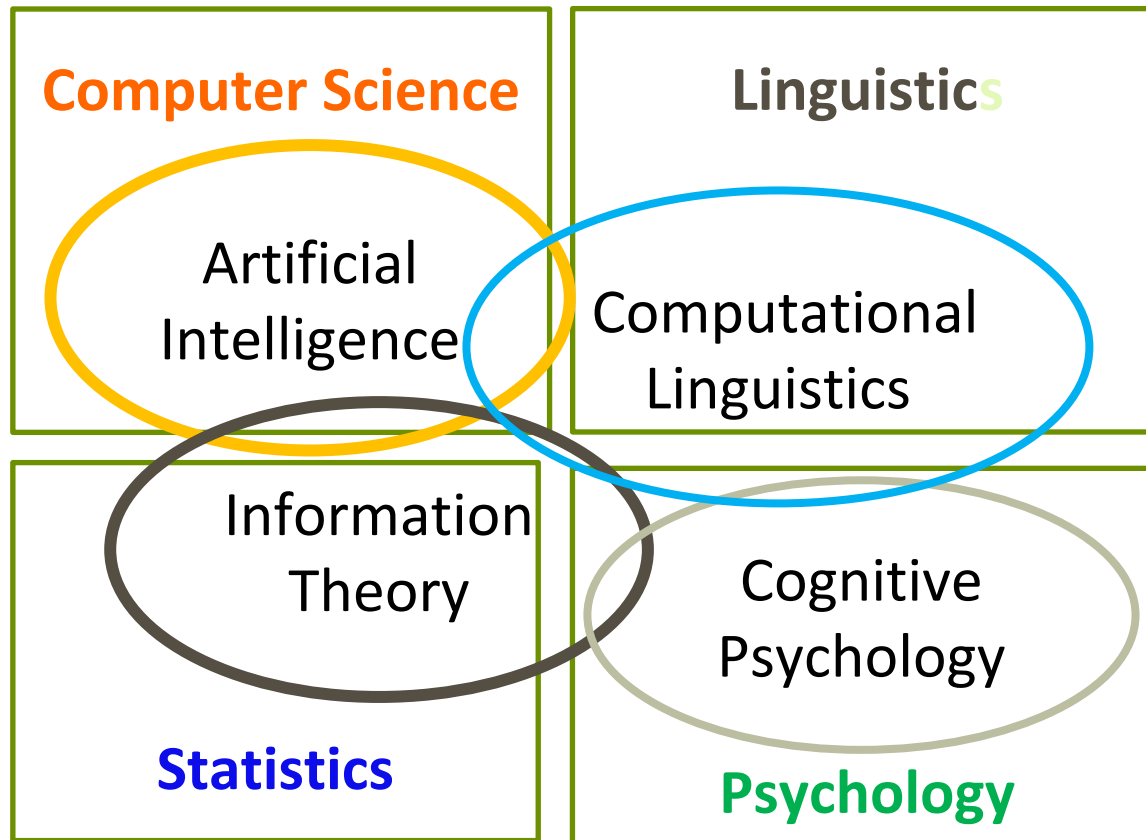
Other NLP Definitions

- The process of building **computational models** for understanding natural language.
- A coherent **study of the human language** from the point of views of **several disciplines** : Linguistics, Psychology, Cognitive Science, Computer Science, Statistics and Mathematics.
- A theoretically motivated range of **computational techniques** for **analyzing and representing naturally occurring texts** at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications

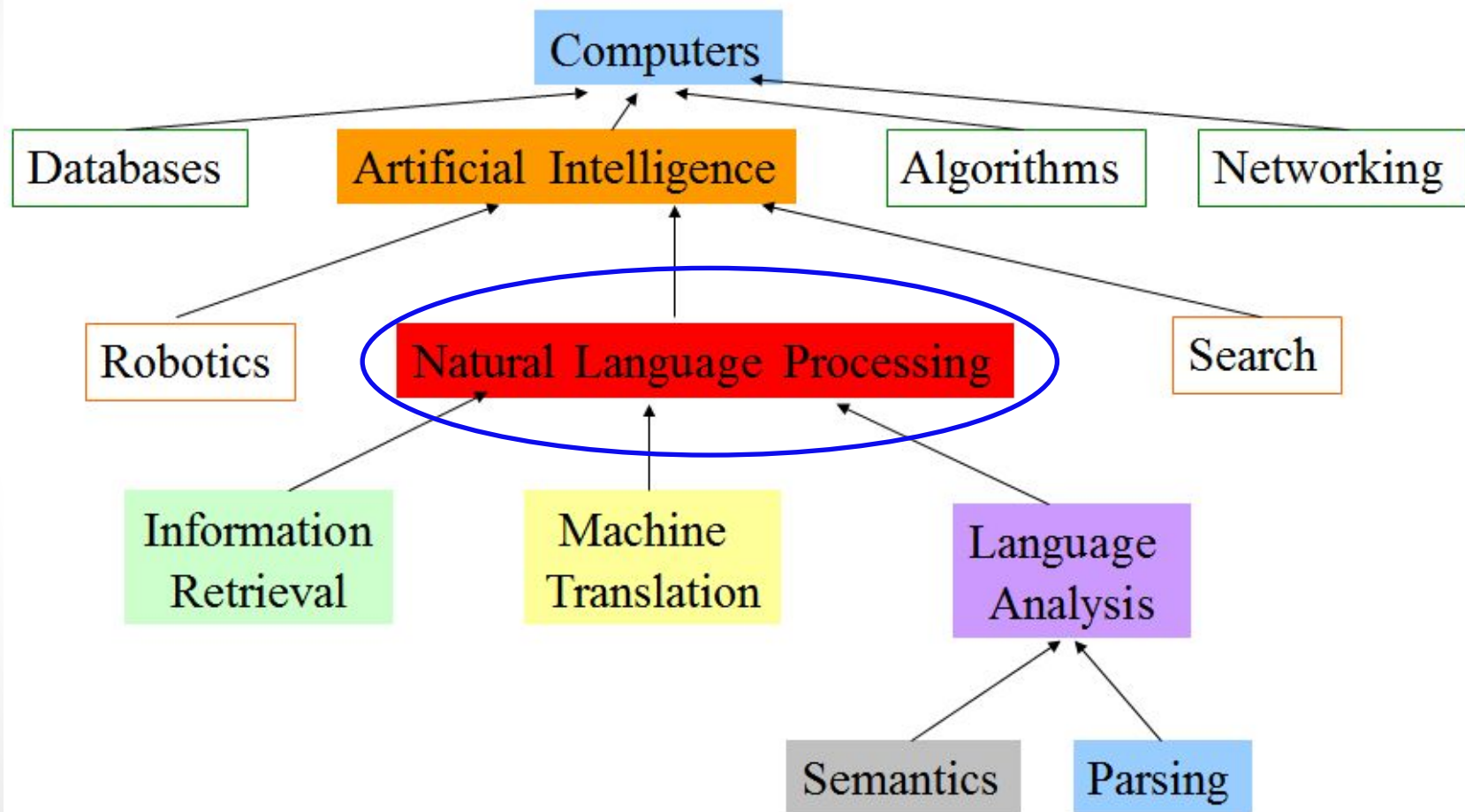
Other names for NLP

- Computational Linguistics (CL)
- Human Language Technology (HLT)
- Natural Language Engineering (NLE)
- Speech and Text Processing

Multidisciplinary NLP



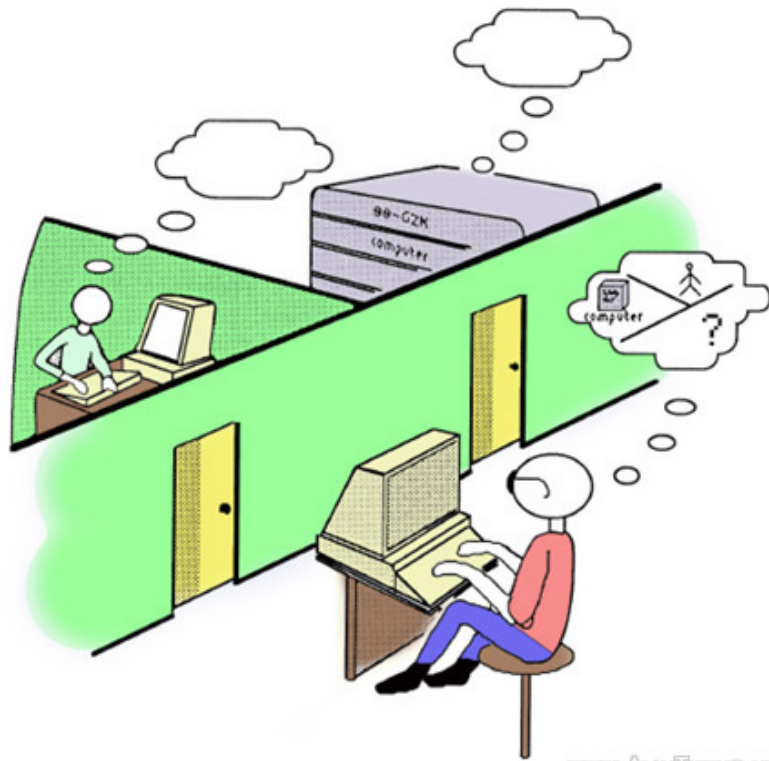
Where does NLP fit in CS?



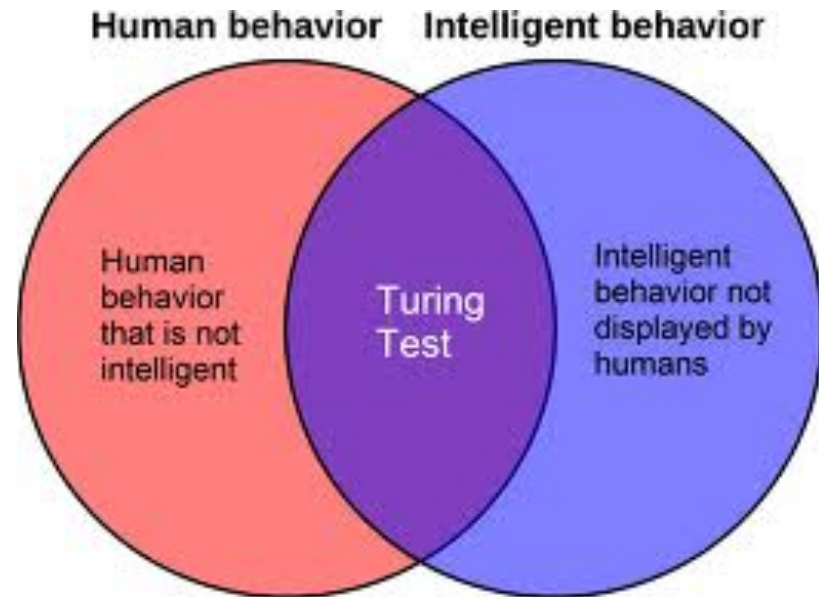
NLP and the Turing Test

- Can machines think?
 - What does it mean to say that a machine can think???
- The basis to determine if a machine could think is a computer's use of language
 - Empirical test → Turing game
- Using human language (by itself) is sufficient as a test for intelligence

NLP and the Turing Test



www.ALAN.TURING.NET



NLP and the Turing Test

- The Test
 - 3 participants : 1 human interrogator, 1 participant, 1 computer
 - Goal of computer: Fool the interrogator into believing it is a person by responding like a human to the interrogator's questions.
 - Goal of participant: Convince the interrogator that the other participant is the machine and that she is human.
 - Goal of interrogator: Determine which participant is the machine through a series of questions
- Turing's prediction:
 - By end of the 20th century, a machine with 10 GB of memory would have approx. 30% chance of fooling a human interrogator after five minutes of questioning (Jurafsky & Martin, 2008)

ELIZA (1964-1966) <http://nlp-addiction.com/eliza/>

> Hello. My name is Eliza. How may I help you?

Eliza Chat bot

Is she a Rogerian psychotherapist,
a semi intelligent chat bot,
or just a toy?



Eliza, one of the original chat bots, is now ready to talk to you.

She is programmed to behave as a Rogerian psychotherapist, and is an interesting example of the limitations of early artificial intelligence programs.

#1

If Eliza (or you, or your web browser) gets confused, [refresh the page](#). To start a new session with the therapist [reload the page](#).

chat:

- A simple program ("chatterbot") simple program that uses **pattern matching** to process the input and translate it into suitable outputs
- Models the behaviour of a psychiatrist ("active listening" strategies of a Rogerian therapist).
- ELIZA's relevance to Turing test: Many people who interacted with ELIZA is convinced that it really understood them and their problems.

ALICE (1995-Present)

<http://alice.pandorabots.com>



A.L.I.C.E. and judge

You said: I knew because you are developed by the Artificial Intelligence Foundation.

A.L.I.C.E.: How, specifically, do you know?. Good reason.. Thanks for telling me that I am developed by the Artificial Intelligence Foundation.

You say:

Say

[Get an A.I. Foundation membership certificate.](#)

Conversation Log:

judge: Hi Alice! How are you doing today?

ALICE: Hi there! Not too bad.

judge: That is good to hear. Do you like Artificial Intelligence class?

ALICE: Is it good to hear? No, I don't care for it.

judge: Why? Aren't you an inspiration from Artificial Intelligence?

ALICE: If I knew why, I wouldn't be sitting here. Yes I am an inspiration from Artificial Intelligence. How did you know?

judge: I knew because you are developed by the Artificial Intelligence Foundation.

ALICE: How, specifically, do you know?. Good reason.. Thanks for telling me that I am developed by the Artificial Intelligence Foundation.

- A.L.I.C.E. (Artificial Linguistic Internet Computer Entity) is an award-winning free natural language artificial intelligence chat robot. Uses a free (open source) software known as **AIML (Artificial Intelligence Markup Language)** for responses and input.
- The development of A.L.I.C.E (by Dr. Richard S.Wallace) was inspired by Eliza Chatbot.

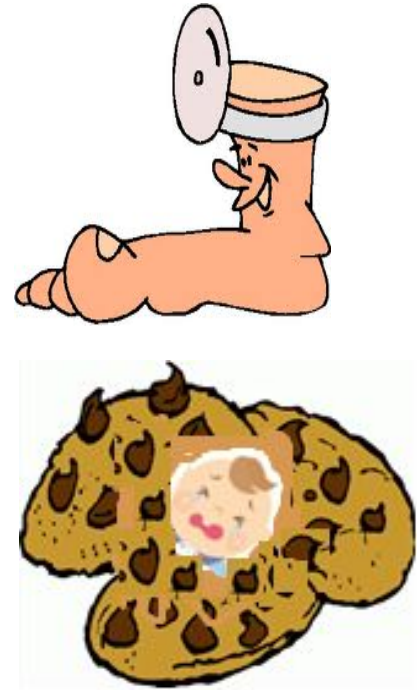
Why NLP?

- Ambiguities in human language
 - Some input can be ambiguous when alternative linguistic structures are possible
- Resolving ambiguity:
 - Introduce models or algorithms to resolve ambiguity
 - E.g models: finite state machines, rule systems, logic, probabilistic models and vector space models.
 - E.g algorithms: Dynamic programming, Expectation Maximization, Artificial Neural Network, State space search or other machine learning algorithms

Ambiguities and Complexities in Language

Ambiguous

- *“Hospitals are sued by 7 Foot doctors”*
- *“Include your children when baking cookies”*
- *“Kids make nutritious snacks”*



Complex

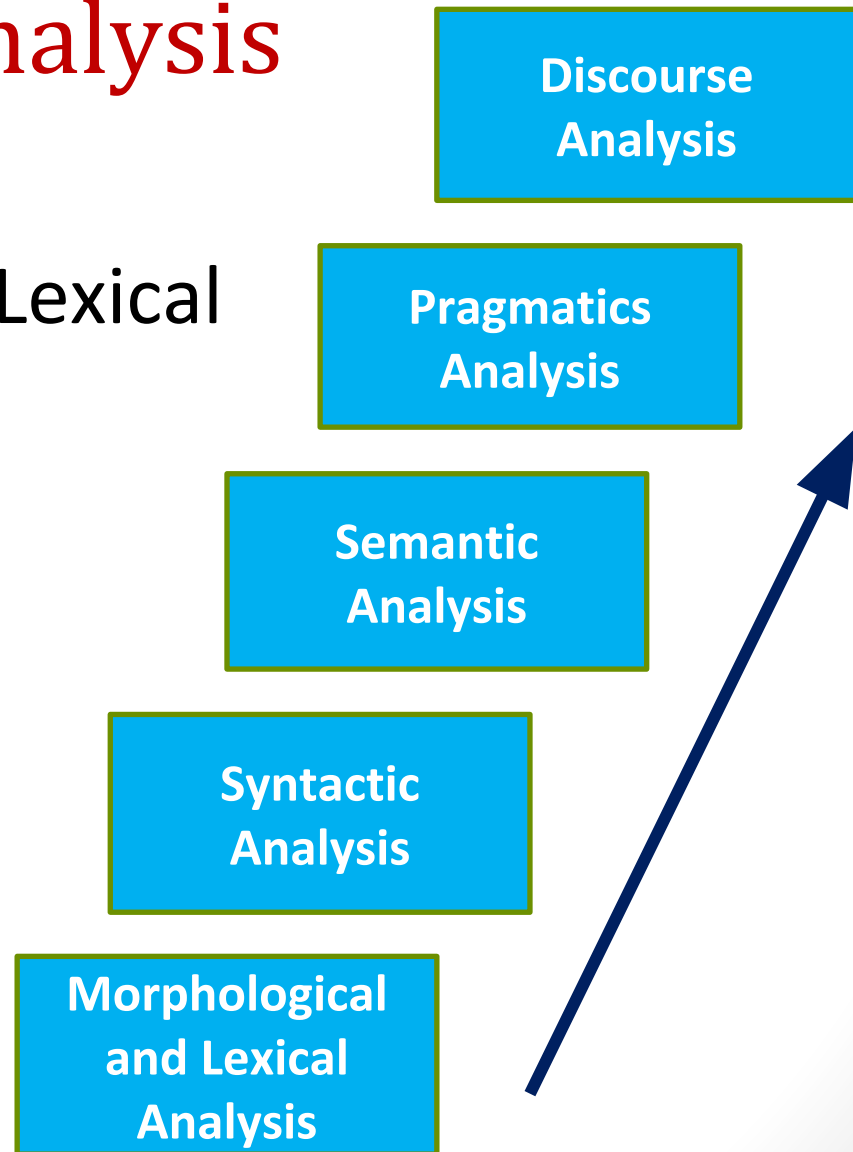
- *“There is a boy, who lost his toy, who jumped with joy, who drank the soy.”*

Goal of NLP

- Get computers to **perform useful tasks involving human language**
 - ✓ Enable **human-machine communication**
 - ✓ Improve **human-human communication**
 - ✓ Perform useful **text/speech processing**
- Example tasks
 - Conversational agents/dialogue system (HAL)
 - Machine translation
 - Question answering system
 - Information extraction
 - Word sense disambiguation

Levels of Language Processing & Analysis

- Morphological and Lexical Analysis
- Syntactic Analysis
- Semantic Analysis
- Pragmatics Analysis
- Discourse Analysis

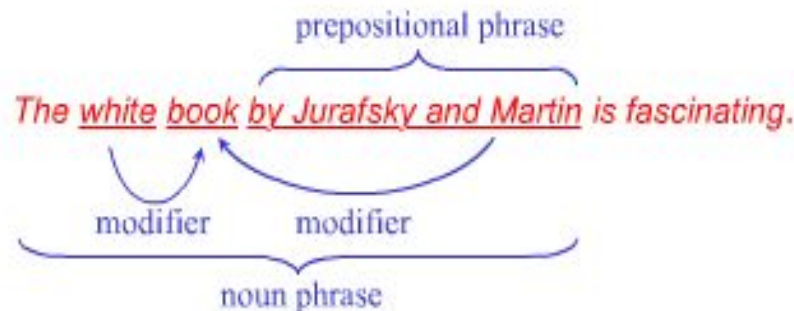


Phases of NLP

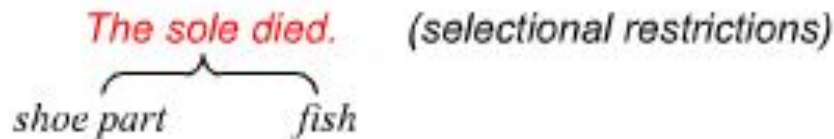
- Engaging in complex language behaviour requires various kinds of knowledge of language
 - Phonetics and Phonology - knowledge about linguistic sounds (t**a**p, bu**tt**er, **ch**ip, **s**heep)
 - Morphology - knowledge of the smallest meaningful units of words and their composition (cat**s**, child**ren**, check**ed**, buy**s**, friend**ly**)

Phases of NLP

- Syntax - knowledge of the structural relationships between words (i.e., in a sentence)



- Semantics - knowledge of meaning of words



– bass fishing, bass playing (word sense disambiguation)

Phases of NLP

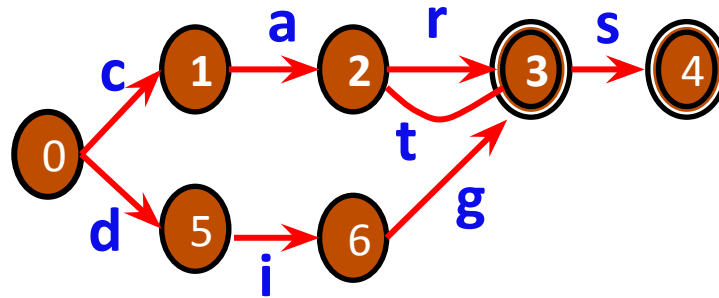
- Pragmatics - knowledge of the relationship of meaning to the goals/intentions of the speaker
- Discourse - knowledge about linguistic units larger than a single utterance

Common NLP Tasks

- Computational Morphology
- Part-of-Speech Tagging
- Text Summarization
- Topic Categorization
- Named Entity Recognition
- Word Sense Disambiguation
- Sentence Parsing
- Sentiment Analysis
- Co-reference Resolution
- Machine Translation

Computational Morphology

- Processing of words and word forms, in both their graphemic (written form) and their phonemic (spoken form)
- Example: finite state morphology



Part-of-Speech Tagging

- Assigning a **part-of-speech** (noun, verb, adjective, ...) to each word in a sentence

*“Malaysia/**N** has/**V** 25/**NUM** million/**N**
people/**N**”*

N – Noun

V – Verb

Num - Number

Text Summarization

- Text Summarization
 - Automatically reducing a text document to create a summary that preserves the most important points of the original document
 - Example : Given a single document, produce abstract, outline and headline

Topic Categorization

- Classifies documents according to their topics

“Serena and Nadal relieved after surviving tough opponents in Madrid” [Sports]

“Facebook eyes \$1billion deal for GPS app Waze” [Technology]

“Property, constructions to lead stock market” [Business]

“All eyes on cabinet lineups” [Politics]

Named Entity Recognition

- Identifies and labels sequences of words in a text that represents names of things (proper names), such as persons , locations and organizations.
- Classification into a set of predefined categories

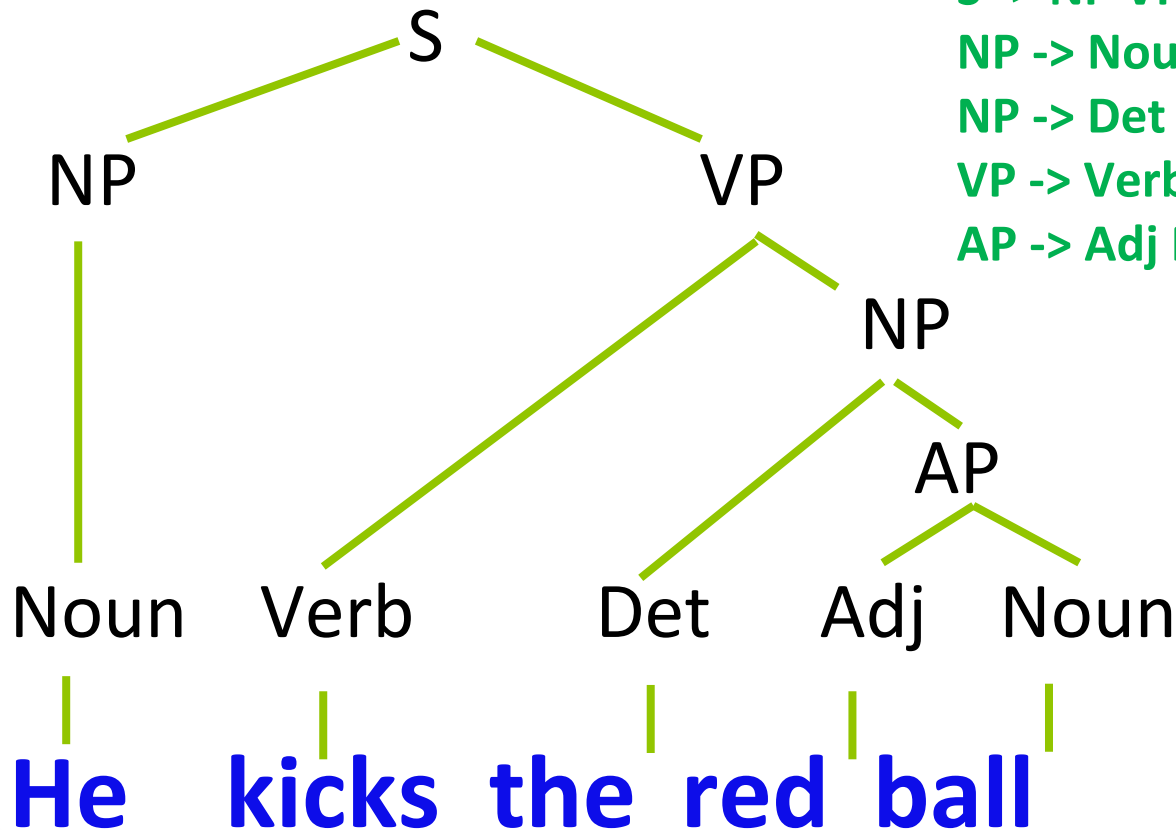
*“The **World Cup** (tournament) took place in **England** (country)”*

Word Sense Disambiguation

- Identify which **sense of a word** (i.e. meaning) is used in a sentence (in a context), when the word has **multiple meanings**.
- Classify an occurrence of the word in context into one or more of its **sense classes**.
- Examples:
 - bank (financial institution) vs bank (river)
 - book (reserve) vs book(reading material)
 - foot (body parts) vs foot (length/measurement)
 - fly ('take off' using wings) VS fly (insect)

Sentence Parsing

- Analysing a sentence into its **component categories** and **functions**



Rules

S -> NP VP

NP -> Noun

NP -> Det (Adj) Noun

VP -> Verb NP

AP -> Adj Noun

Sentiment Analysis

- Identify, analyze and classify **opinions in text** into categories such as "positive" , "negative" or "neutral"

*"I **love** Macintosh." (Positive)*

*"I **hate** Windows! " (Negative)*

*"What a **great** car, it **did not start** the first day"
(positive or negative???) → sarcasm*

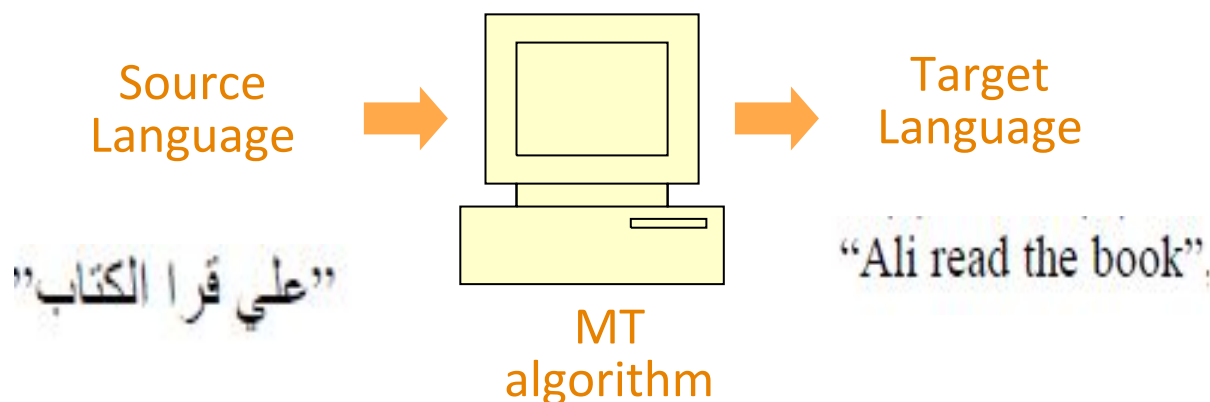
Co-reference Resolution

- Two textual entities that refers to the same object in the “real world” (Mitkov)

Saha Hisham Ismail₁, 45, said poor drainage₂ in the village₃ was the main cause of the problem₄. “We_{1,3} have reported it₂ to the authorities₅ and they₅ have promised to look into it₂, but nothing has been done to rectify the problem₂.”

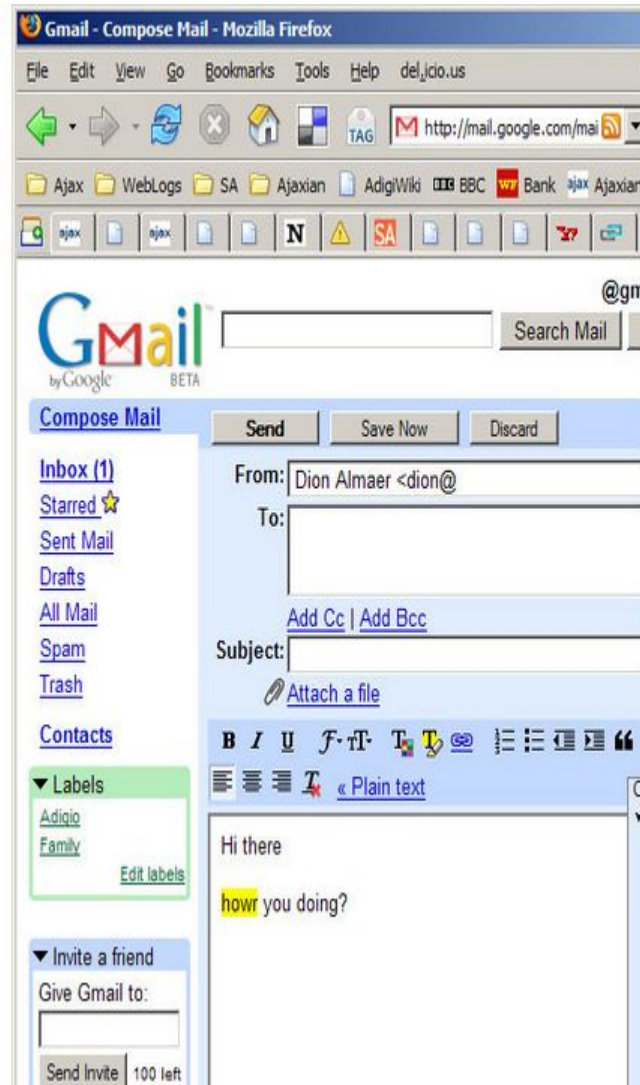
Machine Translation

- An automated system that analyzes text from **source language** and produces “equivalent” text in the **target language**



NLP Applications

- Gmail Spell Checker



NLP Applications

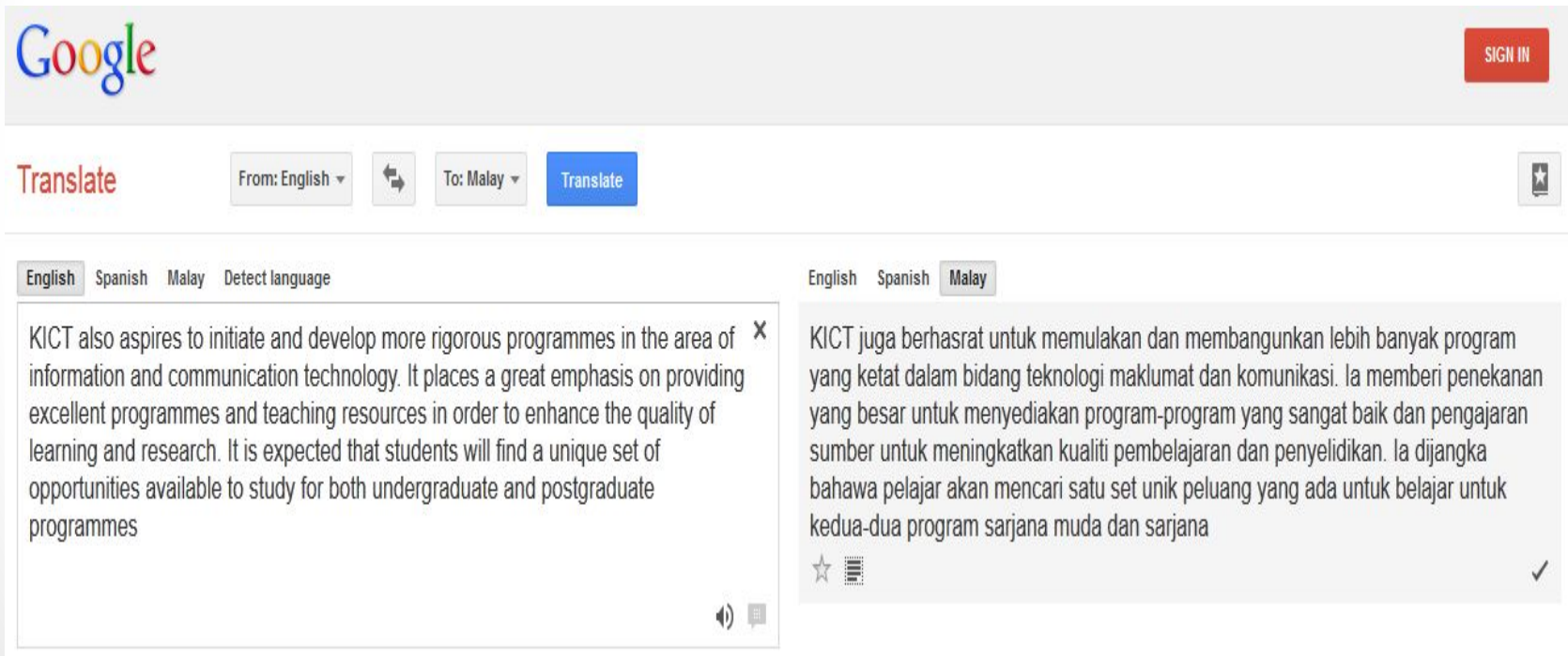
- Spam Classifier

The screenshot displays the Barracuda Spam Firewall 600 Quarantine Inbox. At the top, there is a logo for Barracuda Networks and a 'Log Off username' link. Below the logo, there are tabs for 'QUARANTINE INBOX' and 'PREFERENCES'. A language dropdown menu is set to 'English (US)'. The main area is titled 'Quarantine Inbox' and includes a 'Refresh' button, a 'Filter' dropdown set to 'None', and a 'Pattern' input field with an 'Apply Filter' button. A timeline slider shows dates from 10/20/2004 10:14 to 10/11/2004 10:22. Below the timeline, there are buttons for 'Deliver', 'Whitelist', 'Delete', 'Classify as Not Spam', and 'Classify as Spam'. A table of quarantined emails is listed below, with columns for 'Date Received', 'From', 'Subject', and 'Actions'. A mouse cursor is hovering over the 'Classify as Spam' button.

<input type="checkbox"/>	Date Received	From	Subject	Actions
<input checked="" type="checkbox"/>	10/20 08:20	Steven <aulpoo@erkapa.com>	Homeowners, Banks compete for your business	Deliver Whitelist Delete
<input type="checkbox"/>	10/20 19:15	WellsFargo <service@wellsfa...>	Security Alert on Microsoft Internet Expl...	Deliver Whitelist Delete
<input checked="" type="checkbox"/>	10/20 10:44	Bottom Line Secrets <bls@bo...>	Uncommon Cures for Common Ailments	Deliver Whitelist Delete
<input checked="" type="checkbox"/>	10/20 18:15	Bulls Eye Investing <eqnylt...>	Plain Facts Stock Newsletter	Deliver Whitelist Delete
<input checked="" type="checkbox"/>	10/20 14:03	"Kim A. Sandoval" <a_sandov...>	B*uy Vicodin online - next day shipping, ...	Deliver Whitelist Delete
<input checked="" type="checkbox"/>	10/19 23:15	hotcash@gamingexplore.com	Back to Basics: \$315 Player Bonus from Ve...	Deliver Whitelist Delete
<input checked="" type="checkbox"/>	10/19 14:57	Jack Taylor <jtaylor@joinho...>	Make up to \$10,000 working from home today!	Deliver Whitelist Delete
<input type="checkbox"/>	10/19 15:26	Discovery Shopper <Discover...>	Free Shipping for Early Bird Shoppers	Deliver Whitelist Delete
<input checked="" type="checkbox"/>	10/19 07:33	ruby <d236333sky9queen@vip.gr>	BEFORE AND AFTER PICTURES THAT WILL BLOW ...	Deliver Whitelist Delete
<input checked="" type="checkbox"/>	10/19 13:17	bonusound@gamingexplore.com	Back to Basics: \$315 Player Bonus from Ve...	Deliver Whitelist Delete
<input checked="" type="checkbox"/>	10/19 09:43	"Edgardo@greenwoodsports.us...	FREE FREE FREE FREE	Deliver Whitelist Delete
<input checked="" type="checkbox"/>	10/19 05:32	Daniel <Daniel@Poverful.Ace...>	Enjoy Unlimited PayPerView in Complete Fr...	Deliver Whitelist Delete
<input checked="" type="checkbox"/>	10/19 21:48	B R A D O X <Bradox0B90Supp...	Why did this happened to You, purccind?	Deliver Whitelist Delete

NLP Applications

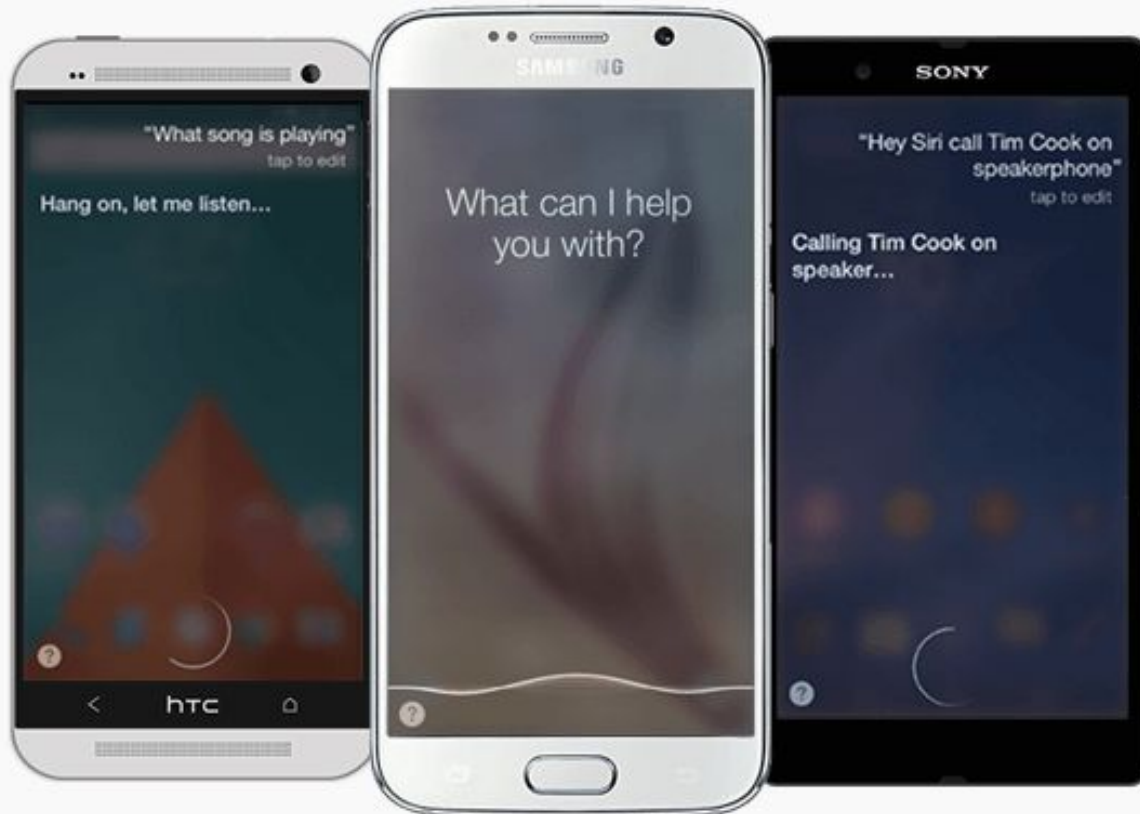
Google Translate



<https://translate.google.com/>

NLP Applications

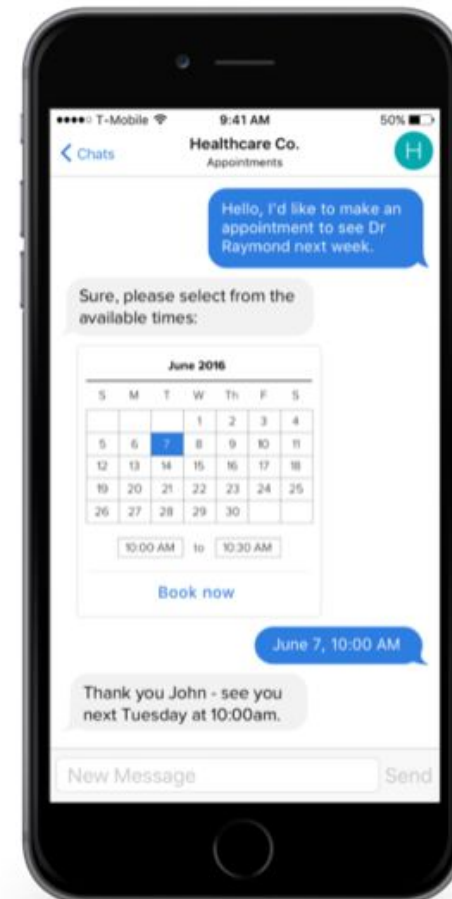
Personal Assistant (Speech-to-Text) :
SIRI ANDROID



<http://sirionandroid.com/>

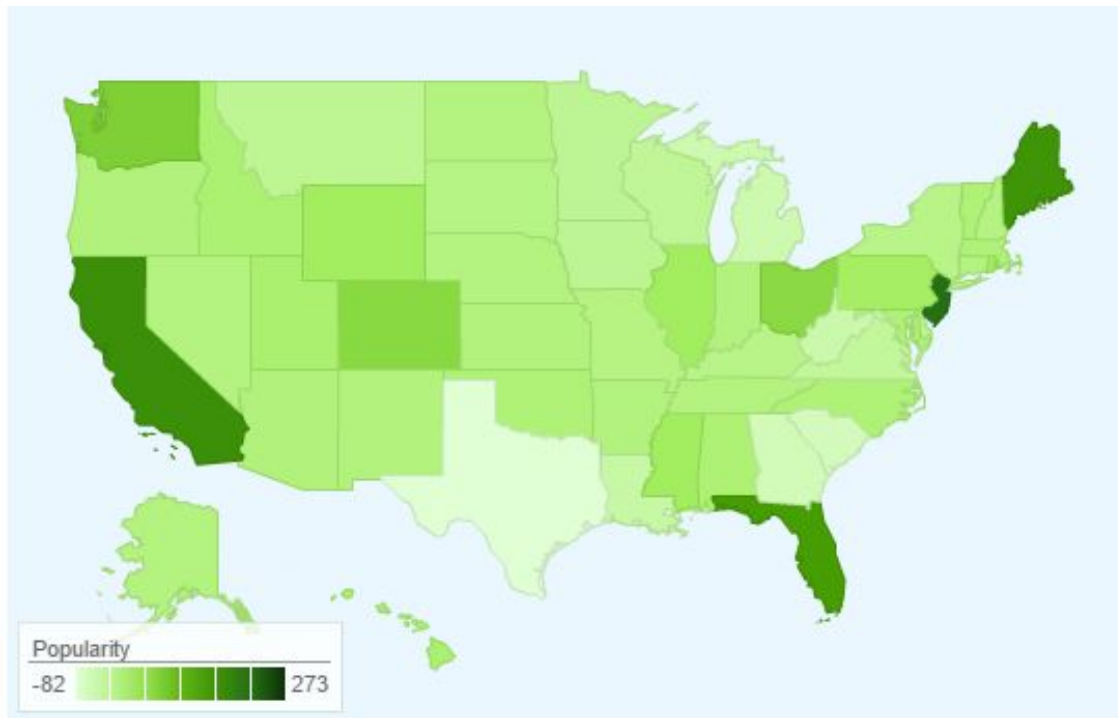
NLP Applications

Chatbot for Business: Helpdesk agent/Call Centers



NLP Applications

Text Analytics App: Prediction on US Presidential Election



The following Geomap shows Romney's popularity results:

<http://www.kazemjahanbakhsh.com/codes/election.html>

NLP Applications

IBM Watson Cognitive supercomputer : Speech-to-text service

Transcribe Audio

- Use your microphone to record audio.
- Upload pre-recorded audio (.mp3, .mpeg, .wav, .flac, or .opus only).
- Play one of the sample audio files.*

*Both US English broadband sample audio files are covered under the Creative Commons license.

The returned result includes the recognized text, word alternatives, and spotted keywords. Some models can detect multiple speakers; this may slow down performance.


Voice Model:


US English broadband model (16KHz) ▼


Keywords to spot:


IBM,admired,AI,transformations,cognitive,Artificial Intelligence,dz

☒ Detect multiple speakers

 Record Audio

 Upload Audio File

 Play Sample 1

 Play Sample 2

Text

Word Timings and Alternatives

Keywords (0/0)

JSON

<https://www.ibm.com/watson/services/speech-to-text/>