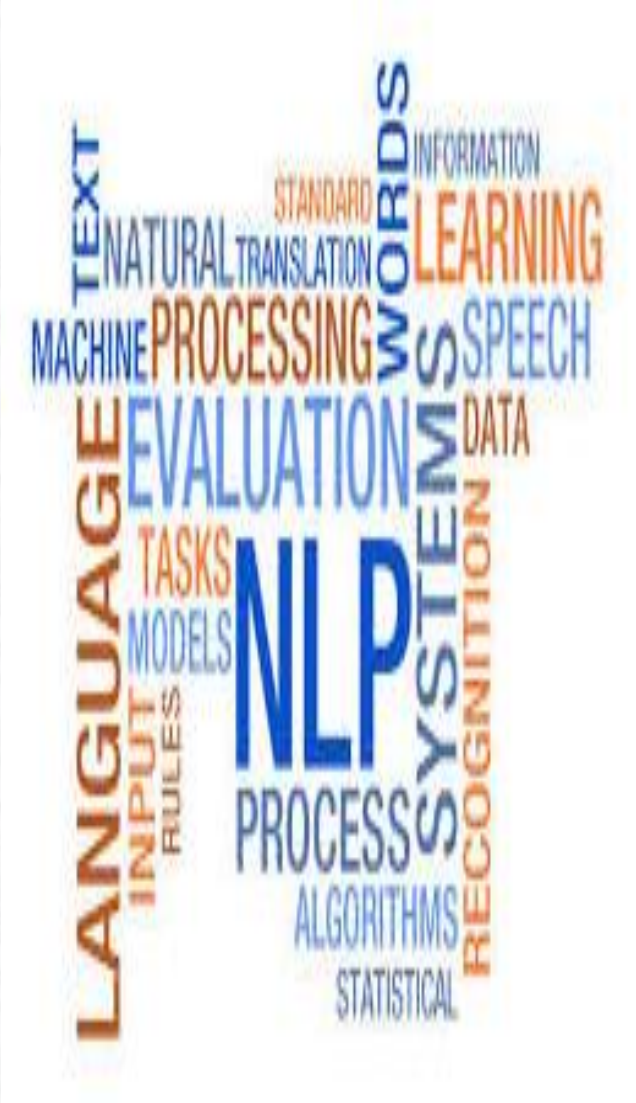
Introduction to NLP



What is Natural Language Processing?

Instructor : Dr. Hanaa Bayomi Ali
Mail : h.mobarz @ fci-cu.edu.eg

Outlines

- Course overview
- Introduction to NLP
- Applications of NLP
- NLP System
- Why NLP is hard?

Course Description

- introduction to NLP system
- Words
  - -Regular expression and Finite state automata
  - - Morphology and finite state transducer
- Syntax
  - -Word classes and Part-of-speech Tagging
  - -Context free grammar (for English)
  - - Parsing with context free grammar
- Semantics
  - - Lexical Semantics
  - - Word Sense Disambiguation

Recommended Textbooks

- ***Speech and Language Processing***, Daniel Jurafsky and James Martin, Prentice-Hall (second edition).

- ***Natural Language Processing with Java*** eBook: Richard M Reese, 2015-07-05

- ***Statistical NLP.*** By Michael Collins, Columbia University

Chapter will be covered

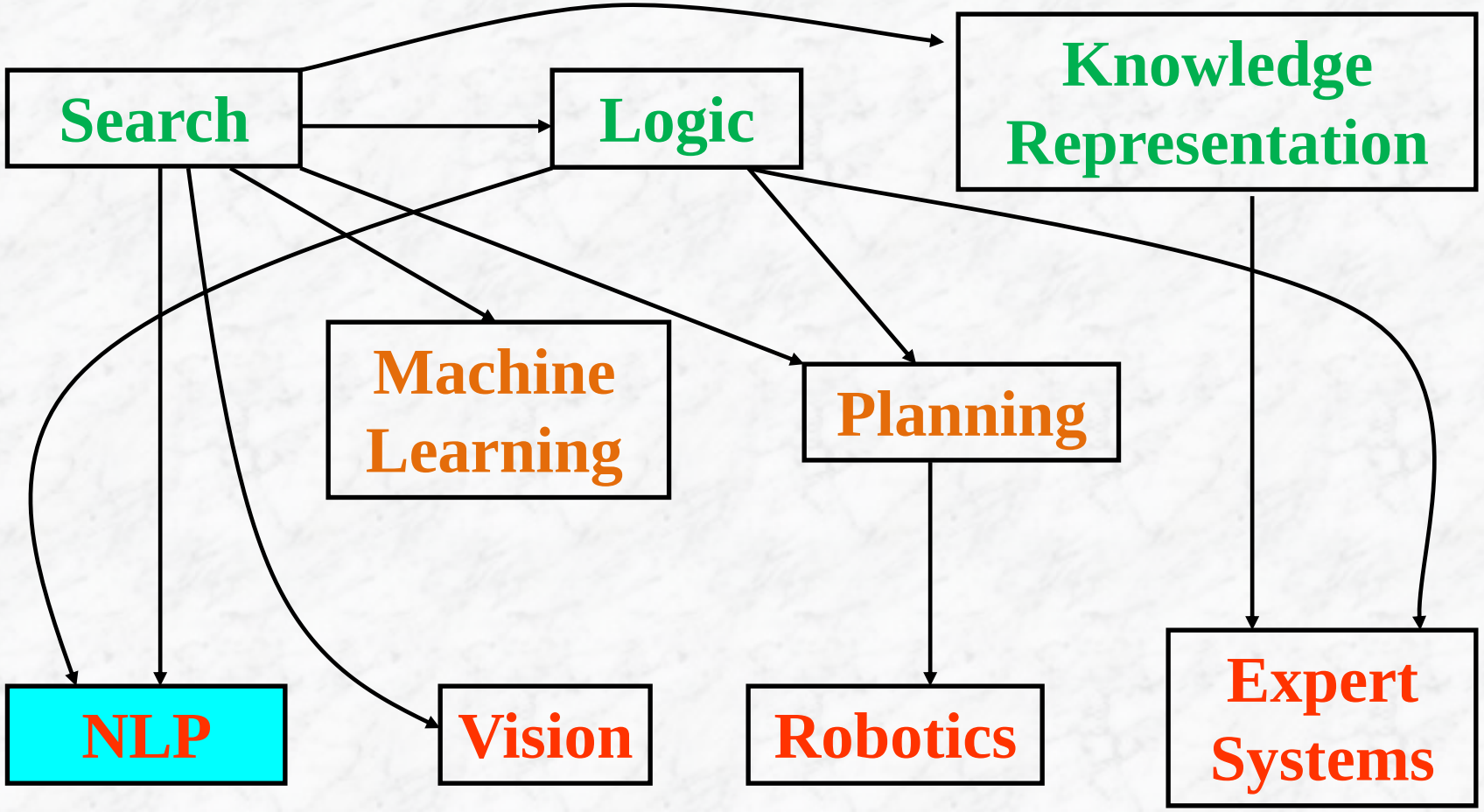
- introduction to NLP system — chapter 1
- Regular expression and Finite state automata — chapter 2
- Morphology and finite state transducer — chapter 3
- Word classes and Part-of-speech Tagging — chapter 8
- Context free grammar (for English) — chapter 9
- Parsing with context free grammar — chapter 10
- Lexical Semantics — chapter 16

Grading

- Assignments 10%
- Midterm Exam 20%
- Project 10%
- Final 60%

Enroll_Access_Code	Course_ID	Course_Name
155730	202102.FCI.CS462	Natural Languages Processing

Perspective of NLP: Areas of AI and their inter-dependencies



What's the difference between Machine Learning (ML), AI, and NLP?

- AI = building systems that can **do intelligent things**

- NLP = building systems that **can understand language** $\subsetneq$ AI

- ML = building systems that can **learn from experience** $\subsetneq$ AI

- NLP $\cap$ ML = building systems that can **learn how to understand language**

What is NLP?

- NLP is a field of computer science, artificial intelligence, and computational linguistics .

- Concerned with the interactions between computers and human (**natural**) languages.

- NLP has 2 Goals

1. **Science Goal** : Understand the way language operates
2. **Engineering Goal**: Build systems that analyze and generate language; reduce the man machine gap

What is NLP?

- NLP is a field of computer science, artificial intelligence, and computational linguistics .

Example

"I went to the bank to withdraw some money"

- why did you go to the bank?

1. **Science Goal** : Understand the way language operates
2. **Engineering Goal**: Build systems that analyze and generate language; reduce the man machine gap

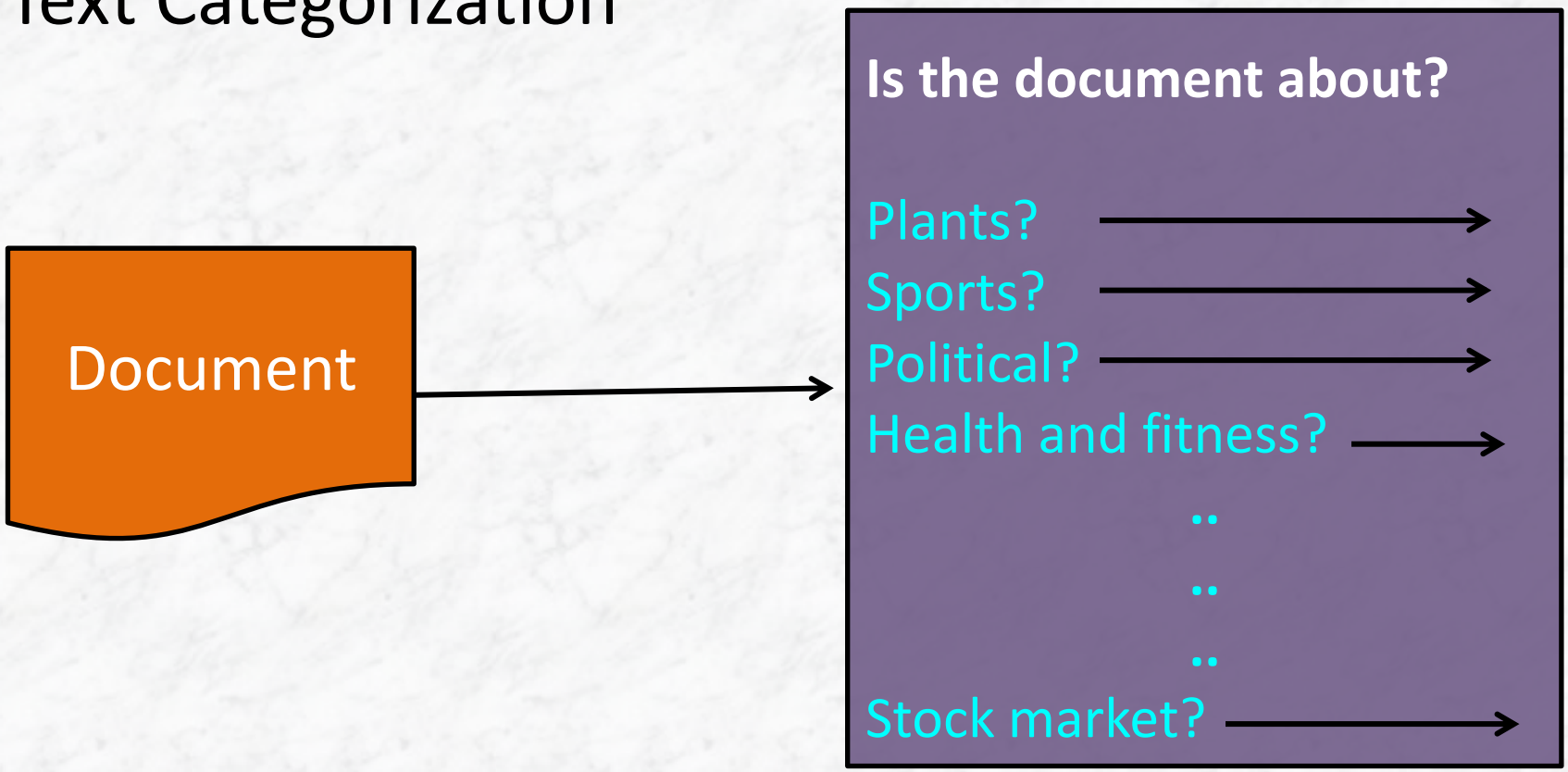
Why Should You Care?

Trends

1. An enormous amount of knowledge is now available in **machine readable** form as natural language text(emails, news articles, web pages, IM, scientific articles, insurance claims, customer complaint letters, transcripts of phone calls, technical documents, government documents, patent portfolios, court decisions, contracts, ……)
2. **Conversational agents** are becoming an important form of human-computer communication
3. Much of human-human communication is now mediated by computers

Application of NLP

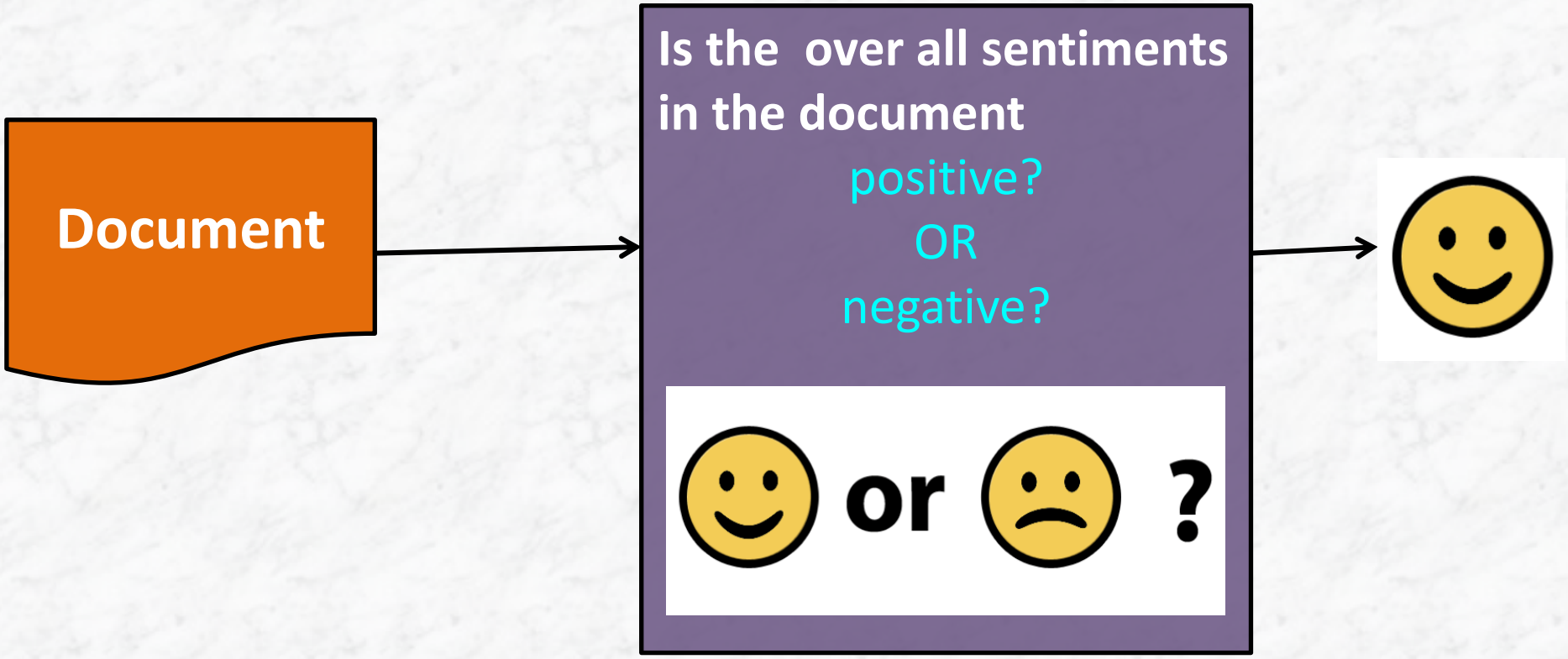
1- Text Categorization



Ex. Uclassify, Weka

Application of NLP (cont)

2- Sentiment classification

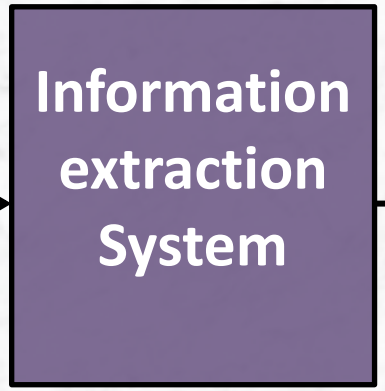


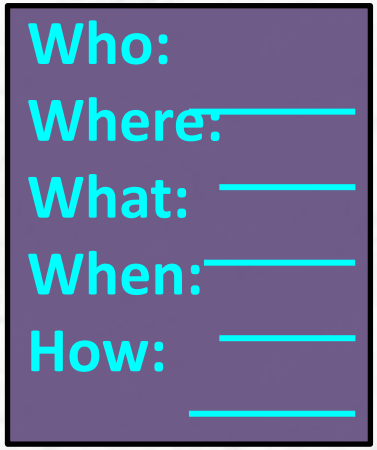
In general, sentiment classification appears to be harder than categorizing by topic. (EX. “Epinion” “consumer review”)

Application of NLP (cont)

3- Information Extraction (IE)

Text collection





Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jur

Hi Dan, we've now scheduled the curriculur

It will be in Gates 159 tomorrow from 10:00

-Chris

Event:	Curriculum mtg
Date:	Jan-16-2012
Start:	10:00am
End:	1:30am
Where:	Gates 159

Application of NLP (cont)

3- Information Extraction (IE) cont.

- Recognition, tagging, and extraction into a structured representation, certain key elements of information, e.g. persons, companies, locations, organizations, from large collections of text.

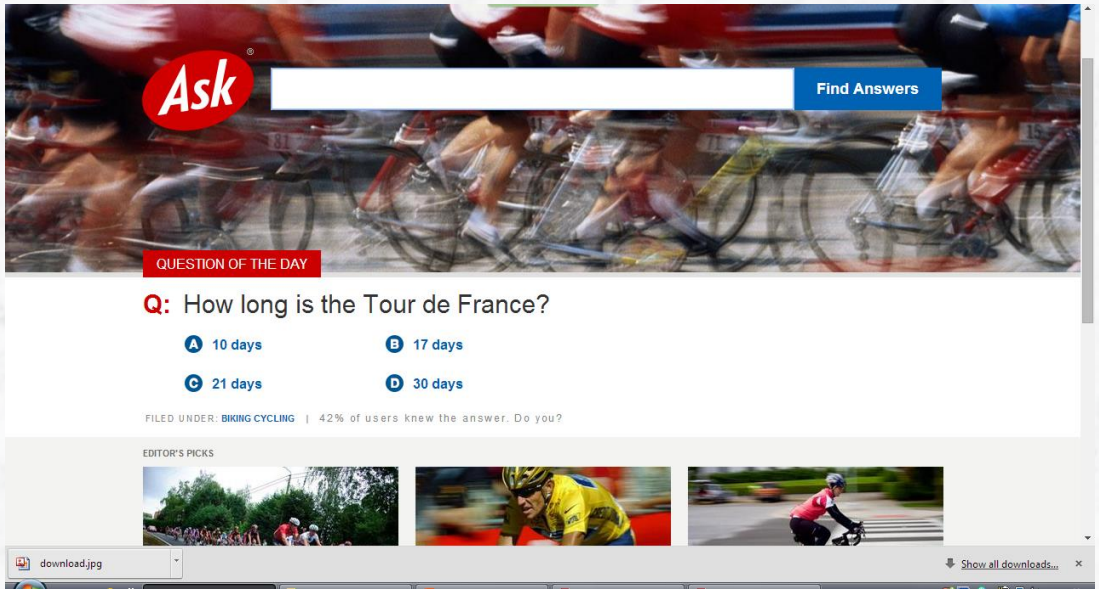
- These extractions can then be utilized for a range of applications including question-answering, visualization, and data mining.
- Ex, Monster.com, HotJobs.com (Job finders) .

Application of NLP (cont)

4- Question-Answering

- In contrast to Information Retrieval, which provides a list of potentially relevant documents in response to a user's query.
- provides the user with either just the text of the answer itself or answer-providing passages.

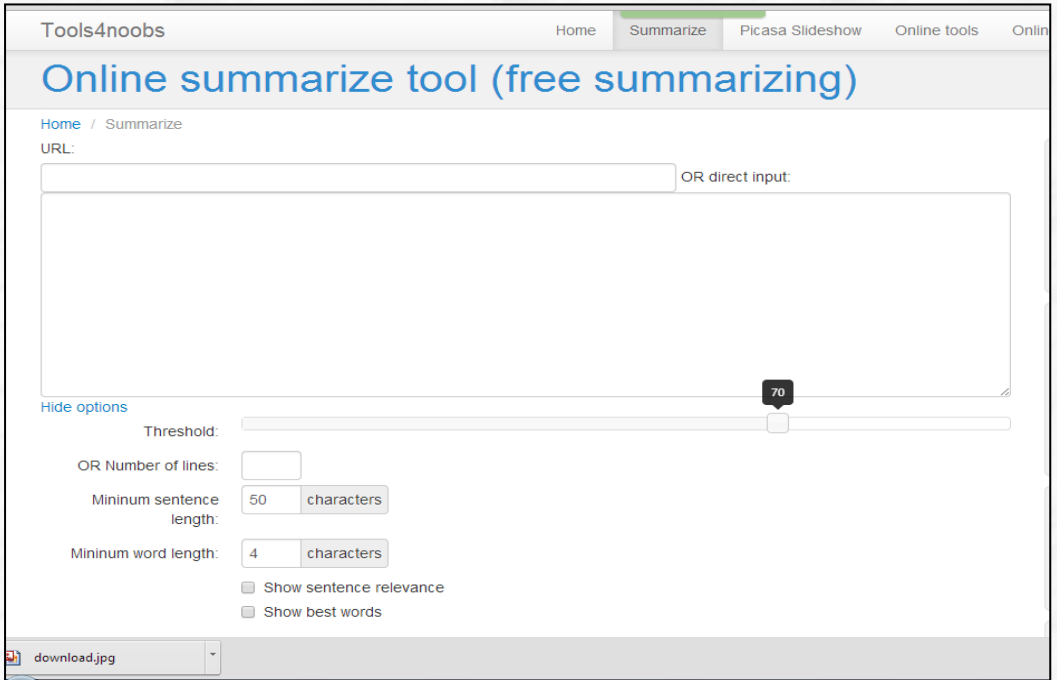
- Ex. Ask Jeeves



Application of NLP (cont)

5- Summarization

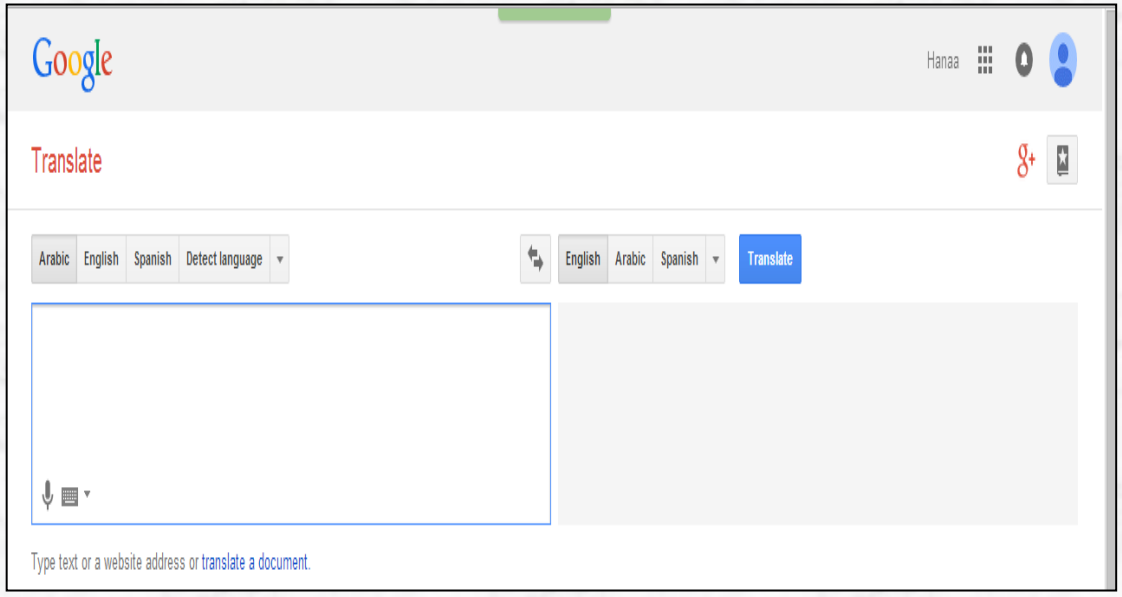
- reduces a larger text into a shorter, yet richly constituted abbreviated narrative representation of the original document.
- Very context-dependent!
- Ex. Tools for noobs.



Application of NLP (cont)

6- Machine translation

- perhaps the oldest of all NLP applications, various levels of NLP have been utilized in MT systems, ranging from the 'word-based' approach to applications that include higher levels of analysis.
- EX, Google, SysTtran



Level of difficulties

Mostly Solved	Good progress	Still Hard
Easy	**intermediate**	**Hard**
Cleanup, Tokenization	Information Retrieval and Extraction (IR)	Machine Translation
Stemming	Relationship Extraction	Automatic Summarization/ **Paraphrasing**
Lemmatization	Named Entity Recognation (NER)	Natural Language Generation
Part of Speech Tagging	Sentiment Analysis/Sentance Boundary Dismbiguation	**Automatic short answer scoring**
Query Expansion	World sense and Dismbiguation	Quation Answering System
Parsing	Text Similarity	Dialog System
Topic Segmentationand Recognation	Coreference Resolution	Image Captioning & other Multimodel Tasks
Morphological Degmentation (Word/Sentences)	Discourse Analysis	

The Problem of NLP

- When people see text, they understand its meaning (by and large)
- When computers see text, they get only character strings (and perhaps HTML tags)
- We'd like computer agents to see meanings and be able to intelligently process text
- These desires have led to many proposals for structured, semantically marked up formats
- But often human beings still resolutely make use of text in human languages
- This problem isn’t likely to just go away.
- Ambiguities (Syntactic , Semantic)

General NLP—Too Difficult!

- Word-level ambiguity
  - **"design" can be a noun or a verb** (Ambiguous POS)
  - **"root" has multiple meanings** (Ambiguous sense)
- Syntactic ambiguity
  - **"natural language processing"** (Modification)
  - **"A man saw a boy *with a telescope*."** (PP Attachment)
- Anaphora resolution
  - **"John persuaded Bill to buy a TV for *himself*."**

    (*himself* = John or Bill?)
- Presupposition
  - **"He has quit smoking." implies that he smoked before.**

Humans rely on context to interpret (when possible).
This context may extend beyond a given document!

Language Processing Tasks

- Processing **spoken language** involves all NLP stages, plus speech recognition

- Processing **written text** using lexical, syntactic and semantic knowledge about the language, as well as the required real world information

- Another dimension understanding (**analysis,Parsing**) vs. generation (**synthesis**)

Understanding VS. Generation

- ***Natural language understanding (NLU) :*** mapping the given input (i.e. text) *into a useful representation*:

- By "understand" we do not mean that the computer has humanlike thoughts, feelings, and knowledge.
- But can **recognize** and **use information expressed** in a human language.
- the system needs to disambiguate the input sentence to produce the machine representation language (appropriate syntactic and semantic schema)

- ***NLU faces the challenge of*** *understanding a text without ambiguity.*

EX. Automatically tagging part of speech of words (easy), automatic grading of student essays (hard)

Understanding VS. Generation (cont)

- ***Natural language generation(NLG) :*** starts from the data to *product a text which is the result of the interpretation and analysis of this data*. our goal is much *more complex*: we must, from data placed here and there, product text – but in what order, about what subject and in what form?

  Ex. Automatic Summarization.

Stages of language processing

1- Phonetics and phonology	Speech sound
2- Lexical Analysis	Dividing the whole chunk of txt into paragraphs, sentences, and words
3- Morphology & Lexicon	Words & their forms
4- Syntactic Analysis	Structure of sentences
5- Semantic Analysis	Meaning of words & sentences
6- Pragmatics	Meaning in context & for a purpose
7- Discourse	Connected sentence processing in a larger body of text

Stages of language processing (Cont.)

Phonetics and phonology

- How words are related to their sound
- Every language has an “alphabet” of sound called *phonemes*
- ***Phoneme*** is the smallest unit of sound
- Sound waves are continuous but phonemes are discrete.
- In order to understand a speech, a computer must segment the continuous stream of speech into discrete sounds, then classify each sound as a particular phoneme.

Stages of language processing (Cont.)

Phonetics and phonology

Human Speech

- Difficult medium
  - Background noise
  - Words can be pronounced very differently
    - different people: accents, age, sex
    - same person: emotional state, illness
  - Words maybe pronounced alike with different meaning
    - Week → weak
    - To → two
    - Sandwich → sand which
- Computer speech relies heavily on waveform analysis and pattern recognition

Stages of language processing (Cont.)

Lexical analysis

Tokenization

- A **sentence** is a sequence of tokens ended by a period, a colon, a semicolon, an exclamation point, or a question mark
- The process of segmenting a string of characters into words is known as **tokenization**, and maybe assign part of speech (**POS)** to each word
- A sequence of tokens separated by blanks. Blank characters are white spaces, carriage returns, tabulations, etc.

Stages of language processing (Cont.)

Lexical analysis

Tokenization

How to use sentence tokenize in NLTK?

After installing nltk and nltk_data , you can launch python and import sent_tokenize tool from nltk:

```
>>> text = “this’s a sent tokenize test. this is sent two. is this sent three? sent 4 is cool! Now it’s your turn.”
>>> from nltk.tokenize import sent_tokenize
>>> sent_tokenize_list = sent_tokenize(text)
>>> len(sent_tokenize_list)
5
>>> sent_tokenize_list
[“this’s a sent tokenize test.”, ‘this is sent two.’, ‘is this sent three?’, ‘sent 4 is cool!’, “Now it’s your turn.”]
```

Stages of language processing (Cont.)

Lexical analysis

Tokenization

Tokenizing text into words

Tokenizing text into words in NLTK is very simple, just called word_tokenize from nltk.tokenize module:

```
>>> from nltk.tokenize import word_tokenize

>>> word_tokenize('Hello World.')
['Hello', 'World', '.']

>>> word_tokenize("this's a test")
['this', "'s", 'a', 'test']
```

Stages of language processing (Cont.)

Morphological Analysis

- **Purpose** determine meanings of individual word. is the study of how root words and affixes – the **morphemes** – are composed to form words- **Morpheme** – It is primitive unit of meaning in a language.
- Analyzing words into their linguistic components
  - Replace original word by root+affixes
    - *unbreakable* → *un + break + able* *('under')*
- Lookup the root in a database of meanings : a **lexicon**
- **Problem** word level ambiguity words may have several meanings, the **correct** one cannot be chosen
  - Example : the word "bank", the word "mean"
  - Further problem domain specialized meanings

Stages of language processing (Cont.)

Syntactic Analysis

- **Parsing** : It involves analysis of words in the sentence for **grammar** and arranging words in a manner that shows the relationship among the words.

- **Parsing**: given a sentence and a grammar
  - Checks that the sentence is correct according with the grammar and if so returns a **parse tree** representing the structure of the sentence.

Stages of language processing (Cont.)

Semantic Analysis

- It draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the task domain.

- The semantic analyzer disregards sentence such as

  “hot ice-cream”.

Stages of language processing (Cont.)

Pragmatic Analysis

- During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge.

EX. Backward & forward references – Coreference resolution

"The man went near the dog. It hits him."

Often co reference & ambiguity go together as in –

"The dog went near the cat. It hits it."

Stages of language processing (Cont.)

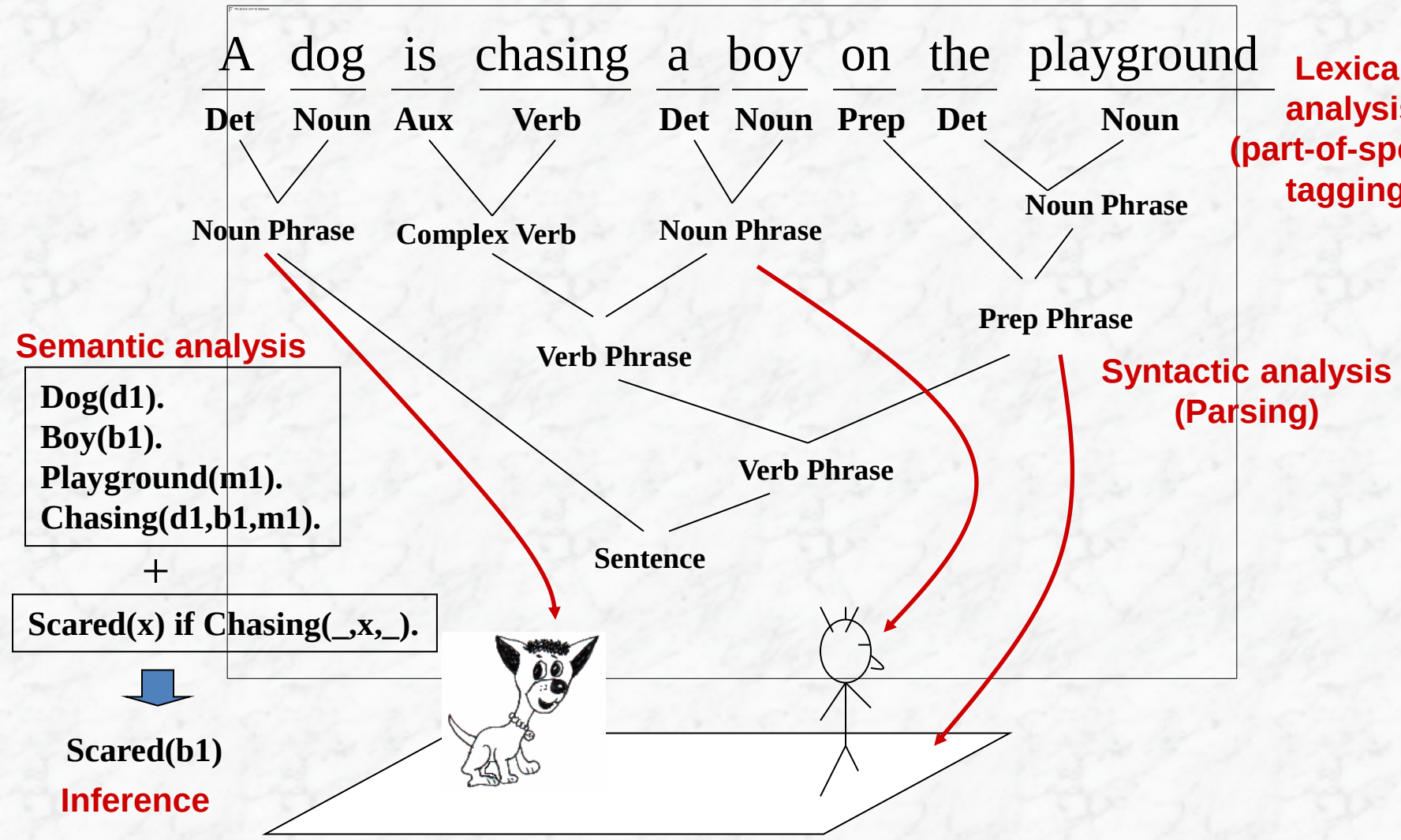
Discourse

- The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.

EX. User situation & context

"*Is that water?*" – the action to be performed is different in a *chemistry lab* and on a *dining table*.

Natural Language Processing



(Taken from ChengXiang Zhai, CS 397cxz – Fall 2003)