CSE528

Natural Language Processing

Venue:ADB-405 SLOTS: A2+TA2 Topic: Introduction

Prof. Tulasi Prasad Sariki,

SCSE, VIT Chennai Campus

www.learnersdesk.info

Contents

- Introduction to NLP
- Ambiguity
- Need for Natural Language Processing
- Natural Languages vs. Computer Languages
- Why Natural Language Processing ?
- Linguistics Levels of Analysis
- Basic terms / terminology in NLP
- Different Tasks in NLP

Introduction

- Why do we need a language?
- Computers would be a lot more useful if they could handle our email, do our library research, talk to us …
- But they are fazed by natural human language (ambiguity).
- How can we tell computers about language? (Or help them learn it as kids do?)

Natural Language Processing

Ambiguity

John said, "I saw the man on the hill with a telescope.“

List the reasonable interpretations?

Past tense to See

- I saw the man. The man was on the hill. The hill had a telescope.
- I saw the man. The man was on the hill. The man had a telescope.
- I saw the man. I was on the hill. The hill had a telescope.
- I saw the man. I was on the hill. I saw him using a telescope.

Need for 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
- Speech understanding
  - Understand phone conversations (IVR)
- Information extraction
  - Extract useful information from resumes
- Automatic summarization
  - Condense 1 book into 1 page (Summary)
- Question answering
- Knowledge acquisition
- Text generations / dialogues
- Automatic translation

Natural?

- Natural Language?
  - Refers to the language spoken by people, e.g. English, Telugu, Tamil, 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

Natural Languages vs. Computer Languages

- Ambiguity is the primary difference between natural and computer languages.
- Formal programming languages are designed to be unambiguous, i.e. they can be defined by a

grammar that produces a unique parse for each sentence in the language.

- Programming languages are also designed for efficient (deterministic) parsing, i.e. they are deterministic context-free languages (DCLFs).
  - A sentence in a DCFL can be parsed in O(*n*) time where *n* is the length of the string.

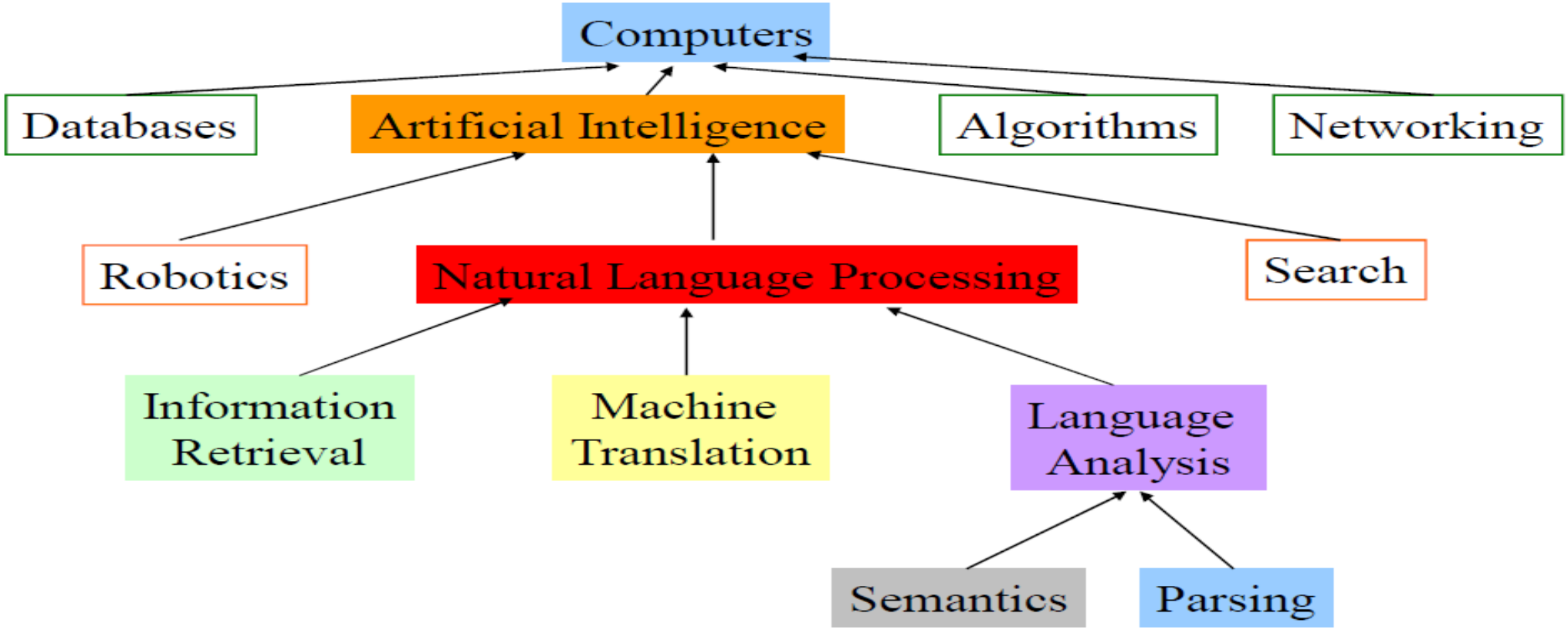
Why Natural Language Processing?

- 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

Basic terms / terminology in NLP

Token: Before any real processing can be done on the input text, it needs to be segmented into linguistic units such as words, punctuation, numbers or alphanumerics. These units are known as tokens.

Sentence: An ordered sequence of tokens.

Tokenization: The process of splitting a sentence into its constituent tokens. For segmented languages such as English, the existence of whitespace makes tokenization relatively easier and uninteresting. However, for languages such as Chinese and Arabic, the task is more difficult since there are no explicit boundaries.

Basic terms / terminology in NLP

Corpus: A body of text, usually containing a large number of sentences.

Corpora: Collection of texts. Plural form of corpus.

Bilingual corpus: A collection of texts in which each text appears in two languages.

Dialogue: Communicative linguistic activity in which at least two speakers or agents participate.

n-gram : A sequence of n tokens.

Semantics: The study of linguistic meaning.

Basic terms / terminology in NLP

Part-of-speech (POS) Tag: A word can be classified into one or more of a set of lexical or part-of-speech categories such as Nouns, Verbs, Adjectives and Articles etc.,

A POS tag is a symbol representing such a lexical category - NN(Noun), VB(Verb), JJ(Adjective), AT(Article).

POS Tagging: Given a sentence and a set of POS tags, a common language processing task is to automatically assign POS tags to each word in the sentences.

For example, given the sentence The ball is red, the output of a POS tagger would be The/AT ball/NN is/VB red/JJ.

Basic terms / terminology in NLP

Parse Tree: A tree defined over a given sentence that represents the syntactic structure of the sentence as defined by a formal grammar.

Computational Morphology: Natural languages consist of a very large number of words that are built upon basic building blocks known as morphemes (or stems), the smallest linguistic units possessing meaning.

Parsing: In the parsing task, a parser constructs the parse tree given a sentence. Some parsers assume the existence of a set of grammar rules in order to parse but recent parsers are smart enough to deduce the parse trees directly from the given data using complex statistical models.

Why is Language Ambiguous?

Having a unique linguistic expression for every possible conceptualization that could be conveyed would make language overly complex and linguistic expressions unnecessarily long.

Allowing resolvable ambiguity permits shorter linguistic expressions, i.e. data compression.

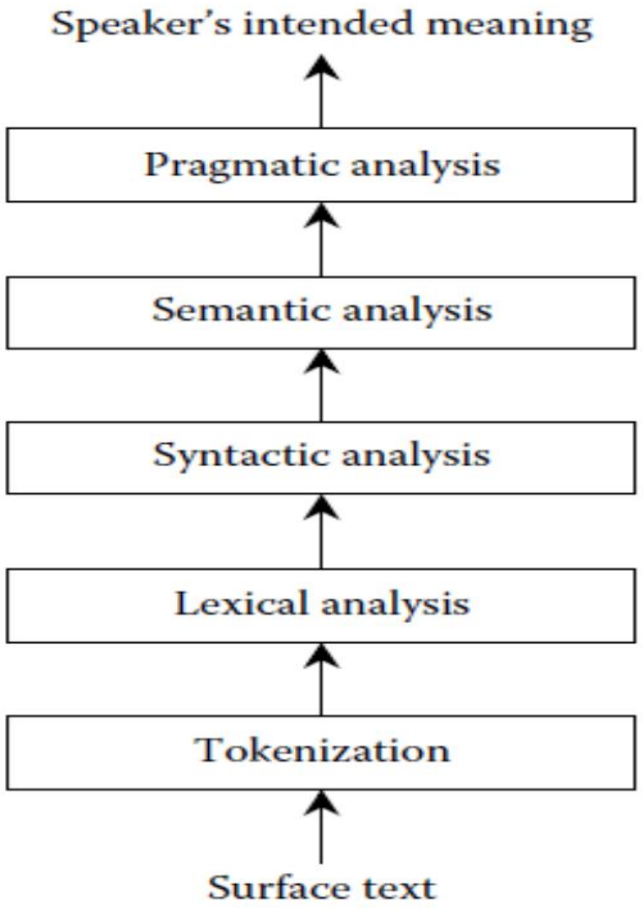
Language relies on people's ability to use their knowledge and inference abilities to properly resolve ambiguities.

Infrequently, disambiguation fails, i.e. the compression is lossy.

Simple View of NLP

Traditionally, work in natural language processing has tended to view the process of language analysis as being decomposable into a number of stages, mirroring the theoretical linguistic distinctions drawn between SYNTAX, SEMANTICS, and PRAGMATICS.

Sentences of a text are first analyzed in terms of their syntax; this provides an order and structure that is more amenable to an analysis in terms of semantics, or literal meaning;



Simple View of NLP

And this is followed by a stage of pragmatic analysis whereby the meaning of the utterance or text in context is determined. This last stage is often seen as being concerned with DISCOURSE, whereas the previous two are generally concerned with sentential matters.

It is widely recognized that in real terms it is not so easy to separate the processing of language neatly into boxes corresponding to each of the layers.

However, such a separation serves as a useful pedagogic aid, and also constitutes the basis for architectural models that make the task of natural language analysis more manageable from a software engineering point of view.

Simple View of NLP

Natural language analysis is only one-half of the story. We also have to consider natural language generation, where we are concerned with mapping from some (typically nonlinguistic) internal representation to a surface text.

Syntax, Semantic, Pragmatics

Syntax concerns the proper ordering of words and its affect on meaning.

- The dog bit the boy.
- The boy bit the dog.
- * Bit boy dog the the.

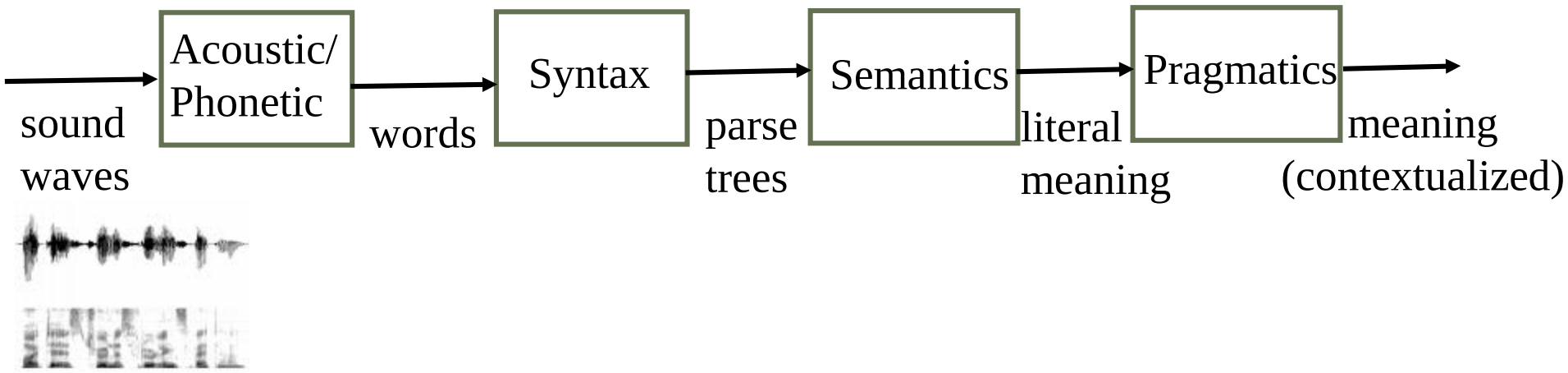
Semantics concerns the (literal) meaning of words, phrases, and sentences.

- “plant” as a photosynthetic organism
- “plant” as a manufacturing facility

Syntax, Semantic, Pragmatics

Pragmatics concerns the overall communicative and social context and its effect on interpretation.

- The ham sandwich wants another beer. (co-reference, anaphora)
- John thinks vanilla. (ellipsis)



Syntactic Tasks

Word Segmentation

Breaking a string of characters (graphemes) into a sequence of words.

In some written languages (e.g. Chinese) words are not separated by spaces.

Even in English, characters other than white-space can be used to separate words [e.g. , ; . - : ( )]

Examples from English URLs:

- jumptheshark.com ⇒ jump the shark .com

Syntactic Tasks

Morphological Analysis

Morphology is the field of linguistics that studies the internal structure of words.

A ***morpheme*** is the smallest linguistic unit that has semantic meaning e.g. “carry”, “pre”, “ed”, “ly”, “s”

Morphological analysis is the task of segmenting a word into its morphemes:

- carried ⇒ carry + ed (past tense)
- independently ⇒ in + (depend + ent) + ly
- Googlers ⇒ (Google + er) + s (plural)

Syntactic Tasks

Part Of Speech (POS) Tagging

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

I ate the spaghetti with meatballs.
Pro V Det N Prep N

John saw the saw and decided to take it to the table.
PN V Det N Con V Part V Pro Prep Det N

Useful for subsequent syntactic parsing and word sense disambiguation.

Syntactic Tasks

Phrase Chunking

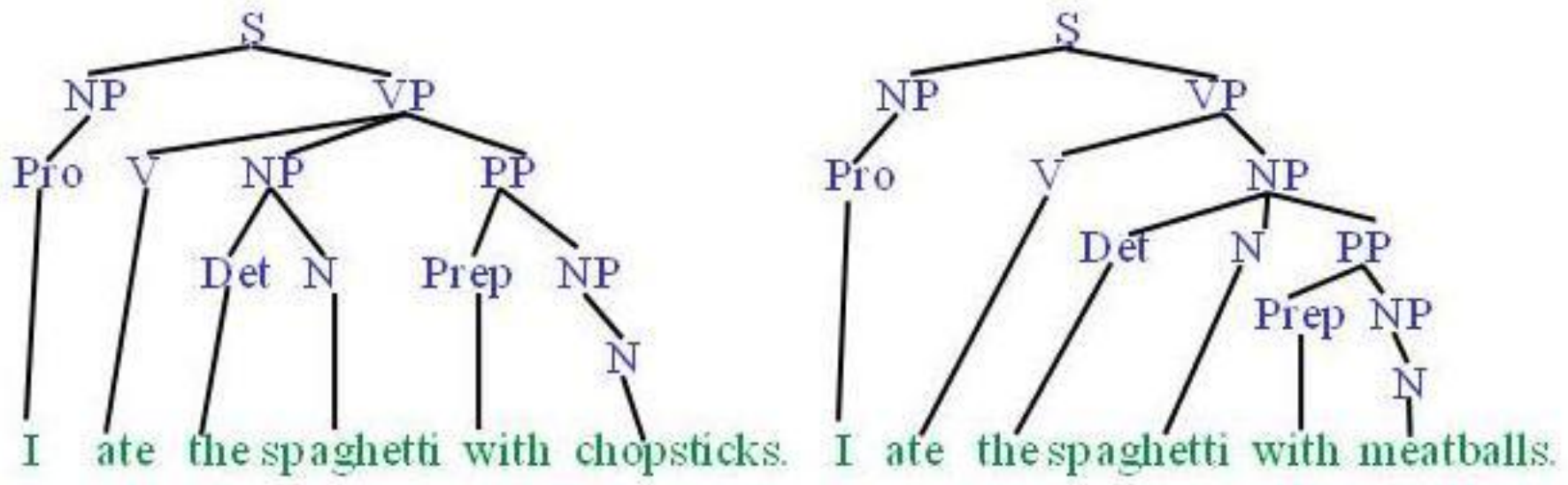
Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.

- [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
- [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

Syntactic Tasks

Syntactic Parsing

Produce the correct syntactic parse tree for a sentence.



Semantic Tasks

Word Sense Disambiguation (WSD)

Words in natural language usually have a fair number of different possible meanings.

- Ellen has a strong interest in computational linguistics.
- Ellen pays a large amount of interest on her credit card.

For many tasks (question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.

Semantic Tasks

Semantic Role Labeling (SRL)

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

agent patient source destination instrument

- John drove Mary from Austin to Dallas in his Toyota Prius.
- The hammer broke the window.

Also referred to a “case role analysis,” “thematic analysis,” and “shallow semantic parsing”

Semantic Tasks

Semantic Parsing

A ***semantic parser*** maps a natural-language sentence to a complete, detailed semantic representation (***logical form***).

For many applications, the desired output is immediately executable by another program.

Example: Mapping an English database query to Prolog:

How many cities are there in the US?

answer(A, count(B, (city(B), loc(B, C), const(C, countryid(USA))),A))

Pragmatics/Discourse Tasks

Anaphora Resolution/ Co-Reference

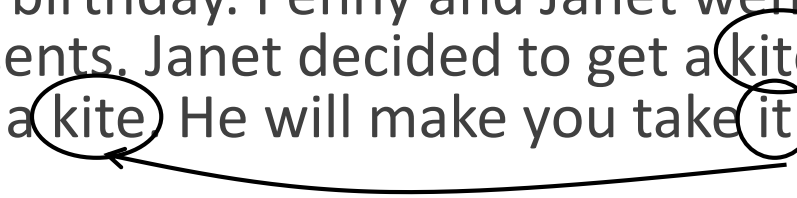
Determine which phrases in a document refer to the same underlying entity.

- John put the carrot on the plate and ate it.

- Bush started the war in Iraq. But the president needed the consent of Congress.

Some cases require difficult reasoning.

- Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."



Pragmatics/Discourse Tasks

Anaphora (right toward/ antecedent)

a. Susan dropped the plate. **It** shattered loudly. It points to the left toward its antecedent the plate.

b. The music stopped, and **that** upset everyone. It points to the left toward its antecedent The music stopped.

c. Fred was angry, and **so** was I. - The adverb so is an anaphor; it points to the left toward its antecedent angry.

d. If Sam buys a new bike, I will **do it** as well. - The verb phrase do it is anaphor; it points to the left toward its antecedent buys a new bike.

Pragmatics/Discourse Tasks

Cataphora (right toward / postcedent)

a. Because **he** was very cold, David put on his coat. It points to the right toward its postcedent David.

b. **His** friends have been criticizing Jim for exaggerating. It points to the right toward its postcedent Jim.

c. Although Sam might **do** so, I will not buy a new bike. It points to the right toward its postcedent buy a new bike.

d. In **their** free time, the kids play video games. It points to the right toward its postcedent the kids.

Pragmatics/Discourse Tasks

Exophora (Something not directly present)

a. **This** garden hose is better than **that** one. - The demonstrative adjectives this and that are exophors; they point to entities in the situational context.

b. Jerry is standing over **there**. The adverb there is an exophor; it points to a location in the situational context.

Ellipsis Resolution

Frequently words and phrases are omitted from sentences when they can be inferred from context.

Applications of NLP

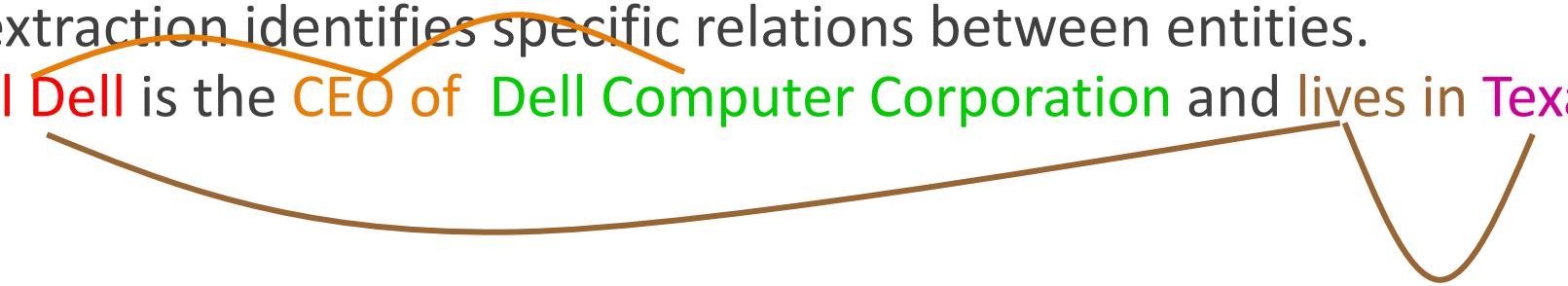
Information Extraction (IE)

Identify phrases in language that refer to specific types of entities and relations in text.

Named entity recognition is task of identifying names of people, places, organizations, etc. in text.

people organizations places
- Michael Dell is the CEO of Dell Computer Corporation and lives in Texas.

Relation extraction identifies specific relations between entities.
- Michael Dell is the CEO of Dell Computer Corporation and lives in Texas.



Applications of NLP

Question Answering

Directly answer natural language questions based on information presented in a corpora of textual documents (e.g. the web).

When was Barack Obama born? (*factoid*)

- August 4, 1961

Who was president when Barack Obama was born?

- John F. Kennedy

How many presidents have there been since Barack Obama was born?

- 9

Applications of NLP

Text Summarization

Produce a short summary of a longer document or article.

- Article: With a split decision in the final two primaries and a flurry of superdelegate endorsements, Sen. Barack Obama sealed the Democratic presidential nomination last night after a grueling and history-making campaign against Sen. Hillary Rodham Clinton that will make him the first African American to head a major-party ticket. Before a chanting and cheering audience in St. Paul, Minn., the first-term senator from Illinois savored what once seemed an unlikely outcome to the Democratic race with a nod to the marathon that was ending and to what will be another hard-fought battle, against Sen. John McCain, the presumptive Republican nominee….
- Summary: Senator Barack Obama was declared the presumptive Democratic presidential nominee.

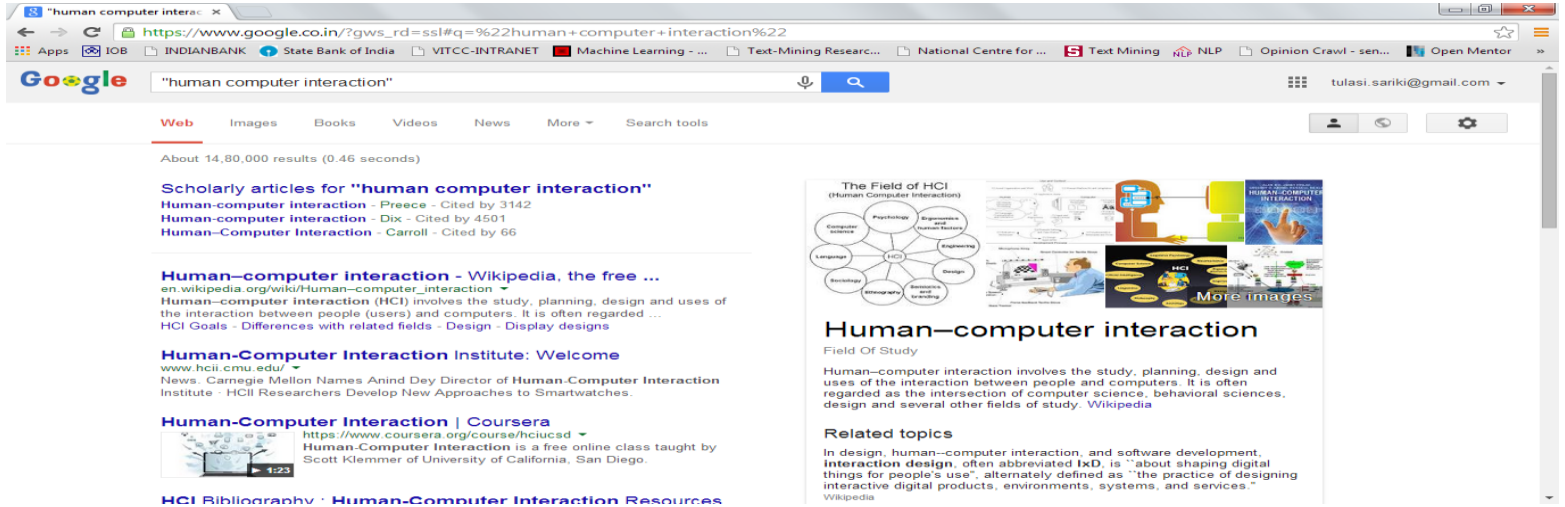
Applications of NLP

Machine Translation (MT)

Translate a sentence from one natural language to another.

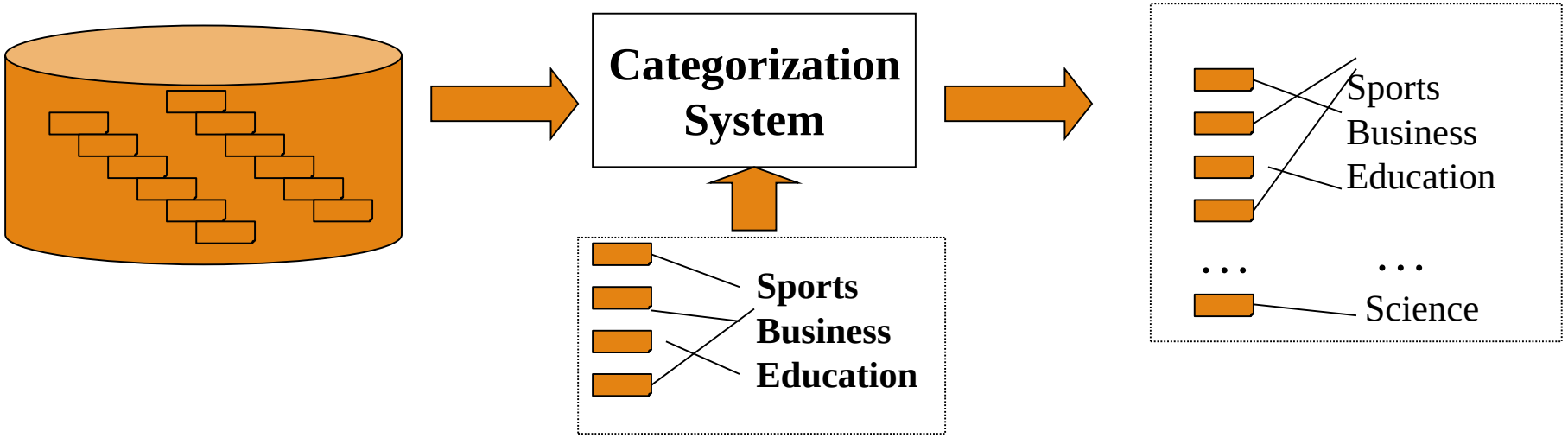
- Hasta la vista, bebé ⇒ See you later, baby.

Information Retrieval



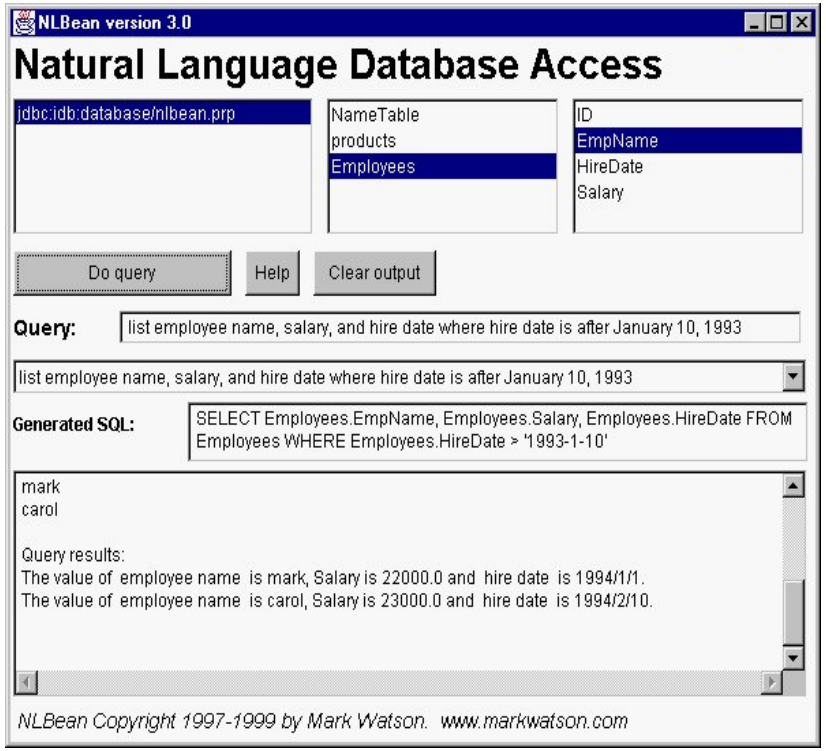
Applications of NLP

Text Categorization

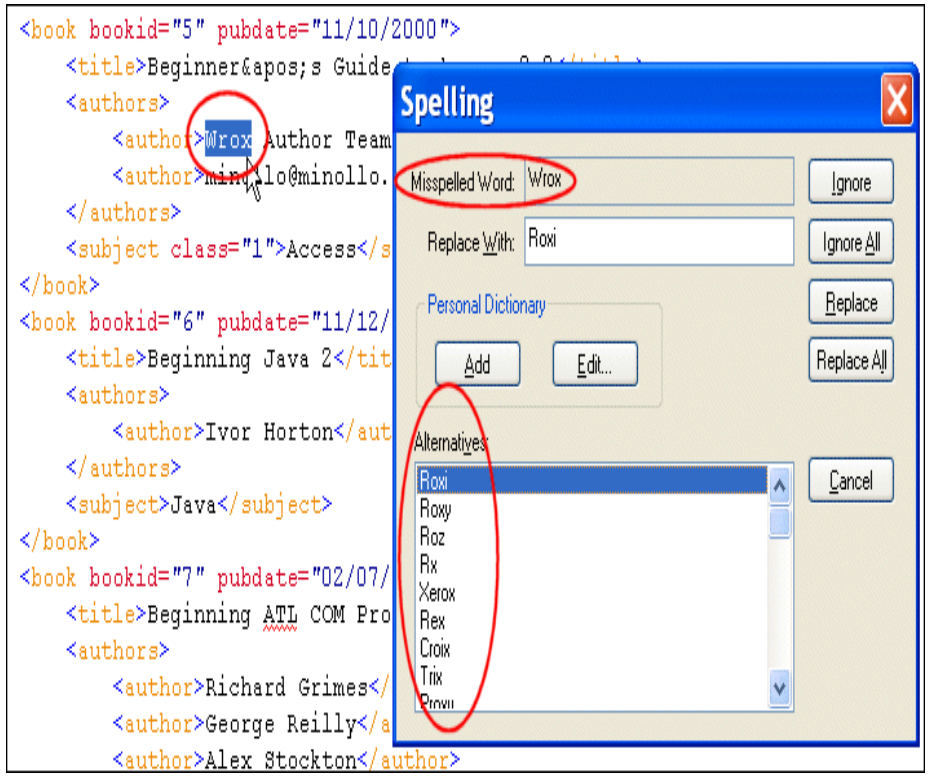


Applications of NLP

Natural Language Interfaces

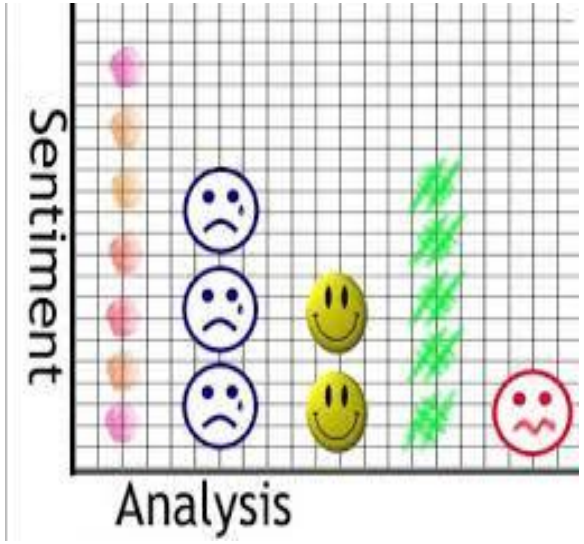


Spell Checking

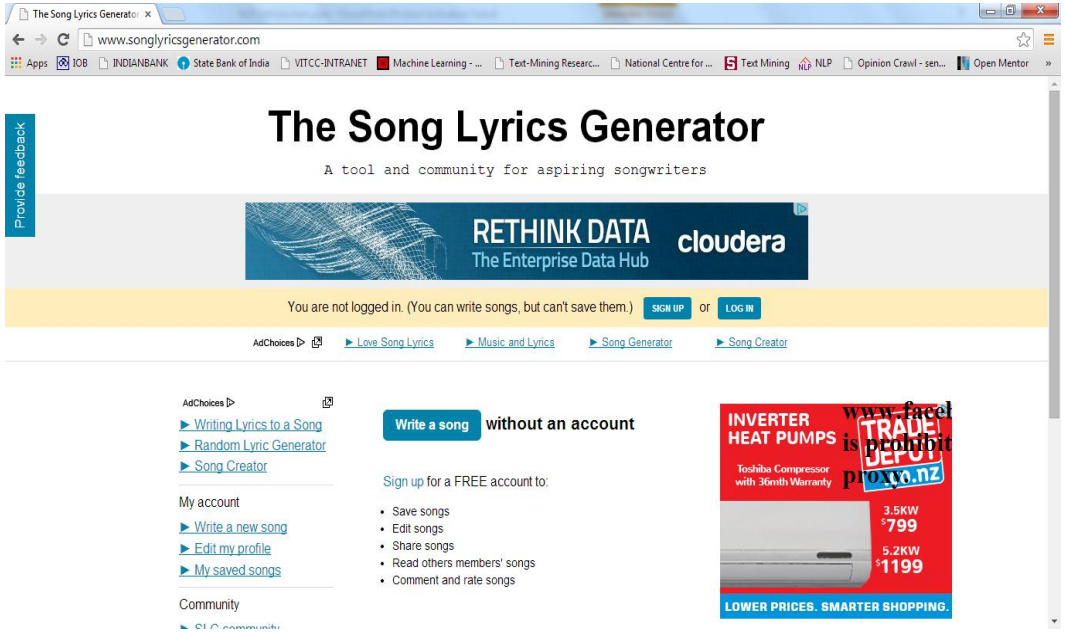


Applications of NLP

Sentiment Analysis

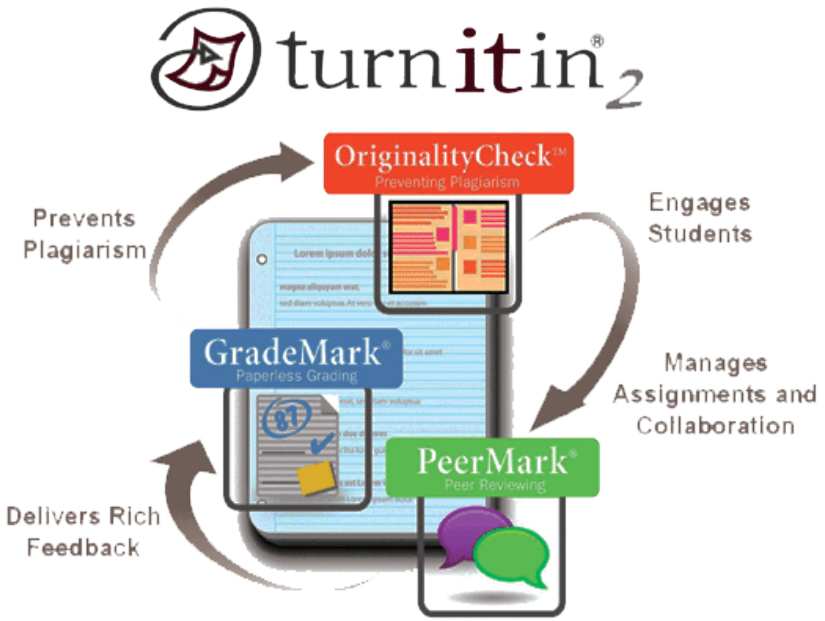


Automatic Lyrics Generation



Applications of NLP

Plagiarism Detection



Speech Recognition

END