Natural language processing: Introduction

Lê Anh Cường

Reading

- Chapter 1 [1]
- Chapter 1,3,4 ([2])

Outline

- What is NLP?
- Rationalist and Empiricist Approaches to Language
- The ambiguity of Language: why NLP is difficulty?
- Linguistic essentials
- Corpus based work

Why Natural Language Processing?

- Huge amounts of data
  - At least 2.5 billion pages
- Applications for processing large amounts of texts
  - ➢ Require NLP expertise

- Classify text into categories
- Index and search large texts
- Automatic translation
- Information Extraction
- Automatic summarization
- Question answering
- Knowledge acquisition
- Text generations/dialogs

Natural?

- Natural language?
  - Refer to the languages spoken by people (English, Vietnamese,…), as opposed to artificial languages (C++, Java,…)
- Natural language processing
  - Applications that deal with natural language processing
- Computational Linguistics
  - Doing linguistics on computers
  - More on the linguistic side than NLP, but closely related

Do you understand?

- Sdfsdf;sldjfsdf
- Iuerpnlc;lwke;rkef
- Klskdfsjkdpwoierpmcfs;df
- Sdjkf ;lkjewkewo s;dlmf;sdfm;

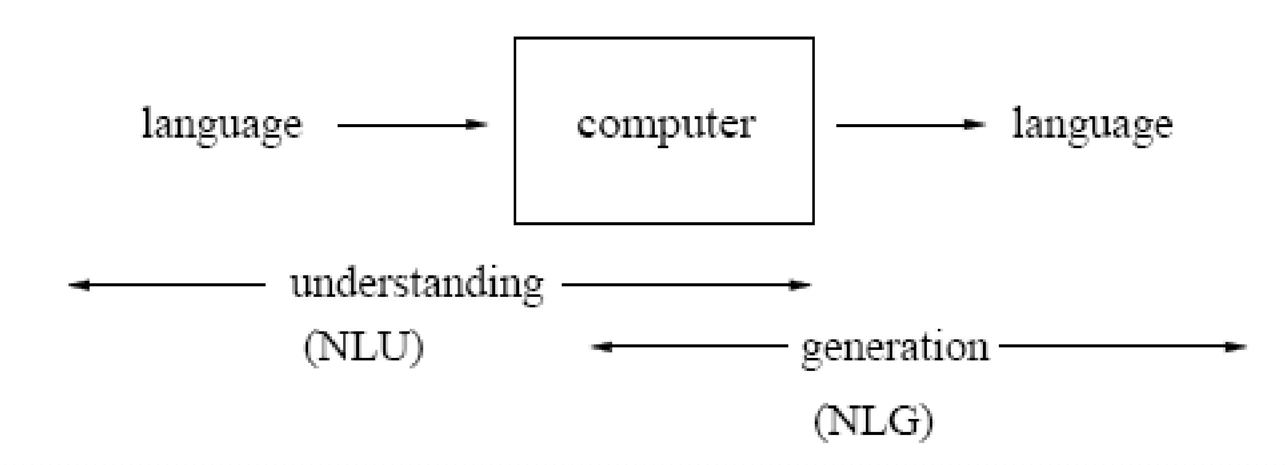
Computer lack knowledge

- Computer 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

Unless we teach them!

What is Natural Language Processing?

- Automatically process natural language
- Computer using natural language as input and/or output



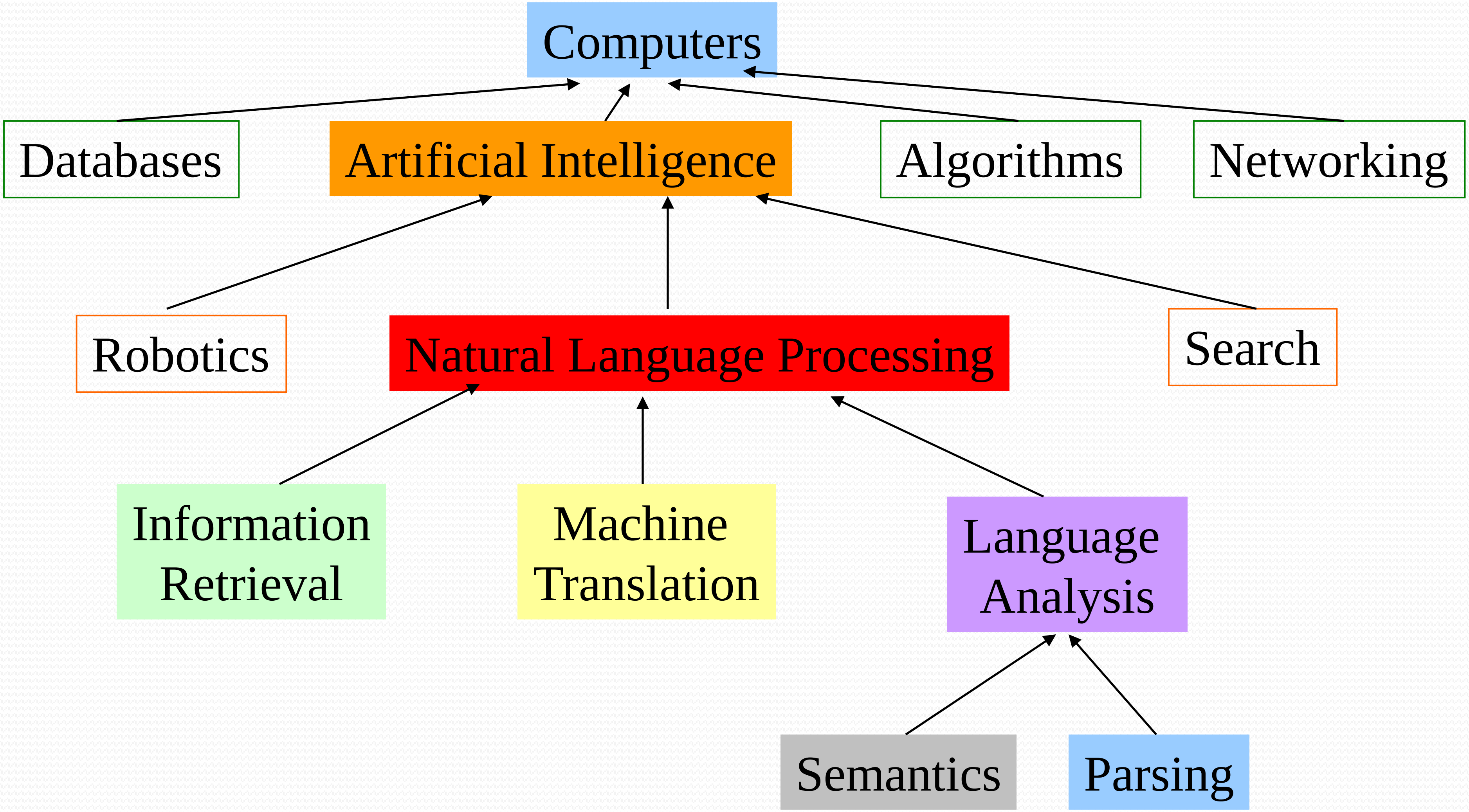
NLP- Definitions

- **Natural language processing** (**NLP**) is a subfield of artificial intelligence and computational linguistics. It studies the problems of automated generation and understanding of natural human languages.

- Natural-language-generation systems convert information from computer databases into normal-sounding human language.

- Natural-language-understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate.

NLP in Computer Science



Computers

Databases

Artificial Intelligence

Algorithms

Networking

Robotics

Natural Language Processing

Search

Information Retrieval

Machine Translation

Language Analysis

Semantics

Parsing

Major tasks in NLP

- Language model
- Morphology analysis
- Part of Speech tagging
- Syntactic parsing
- Word sense disambiguation
- Semantic representation
- Collocation/multi-word expression extraction
- Anaphora resolution
- Preposition attachment
- Word Net

- Text categorization
- Information extraction
- Information retrieval
- Machine translation
- Named entity recognition
- Text generation
- Question answering
- Sentiment analysis & Opinion mining

Basic problems in NLP

- Language model
- Morphology analysis
- Part of Speech tagging
- Syntactic parsing
- Word sense disambiguation
- Semantic representation
- Collocation/multi-word expression extraction
- Corellation
- Preposition attachment
- Word Net

Language Model

- Estimate the probability of a sequence of words (sentence) in a language.

P("I like it") = ?

P("I lay it") = ?

Morphology analysis

- To analyze the structure of a word.

For example:

going -> go[V] + ing

dogs -> dog [N]+ s

computerization -> ?

preprocessing -> ?

Part-Of-Speech tagging

INPUT:

Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT:

Profits/N soared/V at/P Boeing/N Co./N ,/, easily/ADV topping/V forecasts/N on/P Wall/N Street/N ,/, as/P their/POSS CEO/N Alan/N Mulally/N announced/V first/ADJ quarter/N results/N ./.

N = Noun

V = Verb

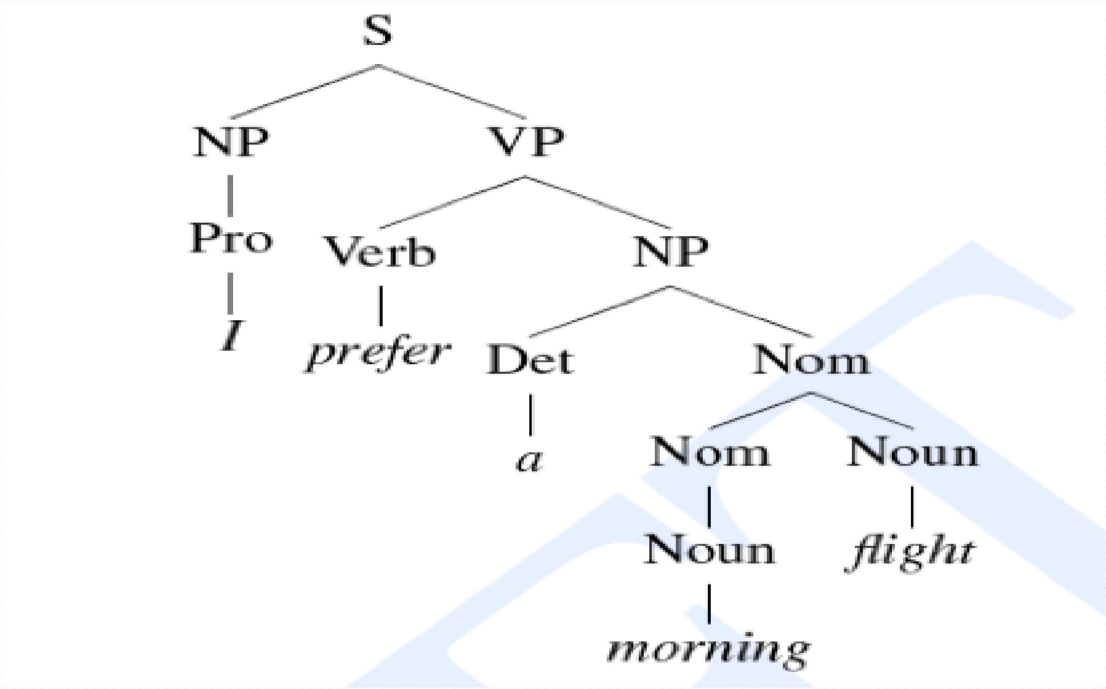
P = Preposition

Adv = Adverb

Adj = Adjective

Syntactic parsing

- To determine grammatical structure of a sentence



Word Sense Disambiguation

- Task of automatically selecting the correct sense for a word given a context

Go to the bank to deposite money.

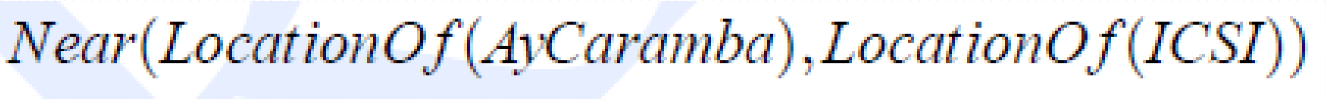
Go along the bank.

Representation meaning

- For example

“Ay Caramba is near ICSI”

Presenting by First Order Logic



$$Near(LocationOf(AyCaramba), LocationOf(ICSI))$$

Collocation extraction

- In corpus linguistics, collocation defines a sequence of words or terms that co-occur more often than would be expected by chance.

  Strong tea -> not powerful tea

  Powerful computer -> not strong computer

- Collocation extraction is a task that extracts collocations automatically from a corpus, using computational linguistics.

Anaphora resolution

- The problem of resolving what a pronoun, or a noun phrase refers to
- For example

  “The dog entered my room. It scared me”

  Find the connection between dog and it, not room and it

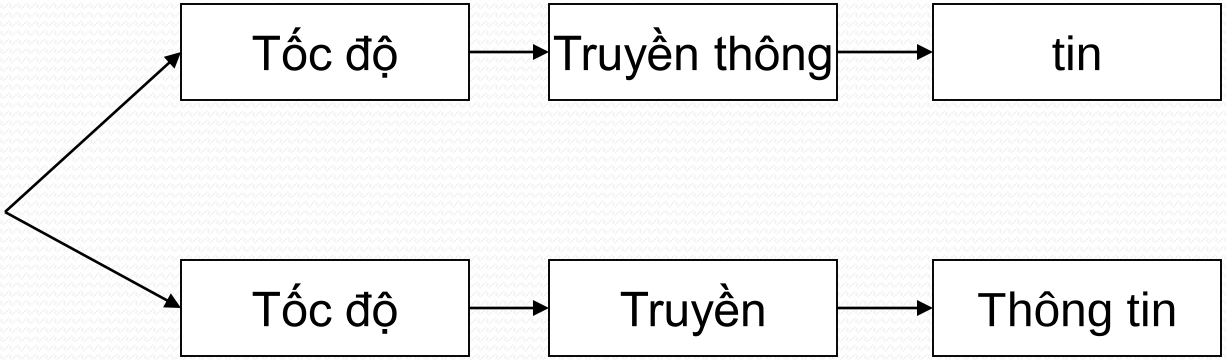
Preposition attachment

- Prepositional phrase attachment is a common cause of structural ambiguity in natural language
- For example

  “I saw the man in the park with a telescope”

Word Segmentation

- Tốc độ truyền thông tin



Applications NLP

- Text categorization
- Information extraction
- Information retrieval
- Machine translation
- Named entity recognition
- Natural language generation
- Question answering
- Sentiment analysis
- Opinion mining

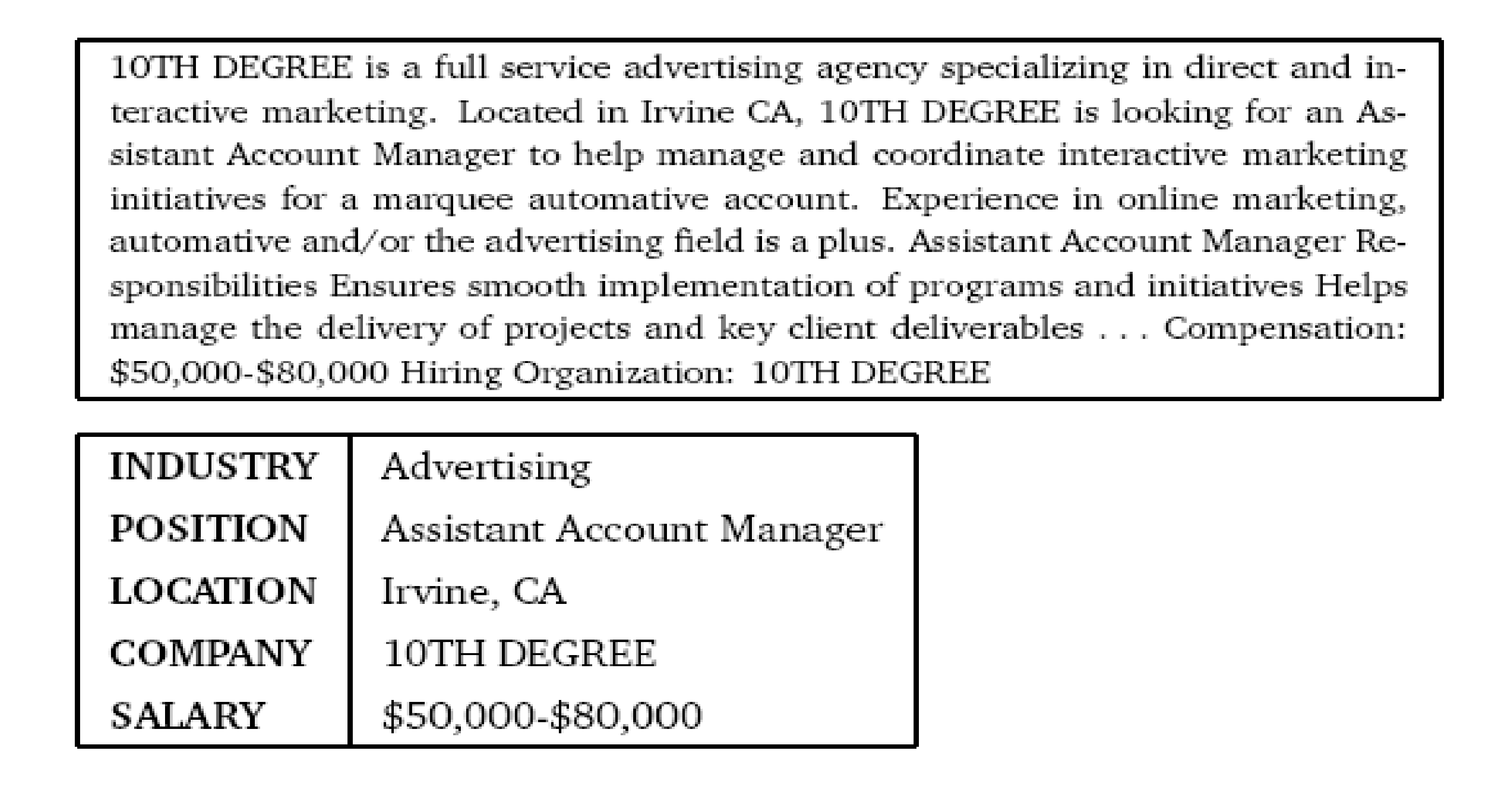
Text Categorization

- Classify documents by: topics, language, author, spam filtering, information retrieval (relevant, not relevant), sentiment classification (positive, negative)

Information Extraction

- Motivation:
  - Complex searches ("Find me all the job in advertising paying at least $50,000 in Boston")
  - Statistical queries ("Does the number of jobs in accounting increases over the years?")
- Goal: map a document collection to structured database

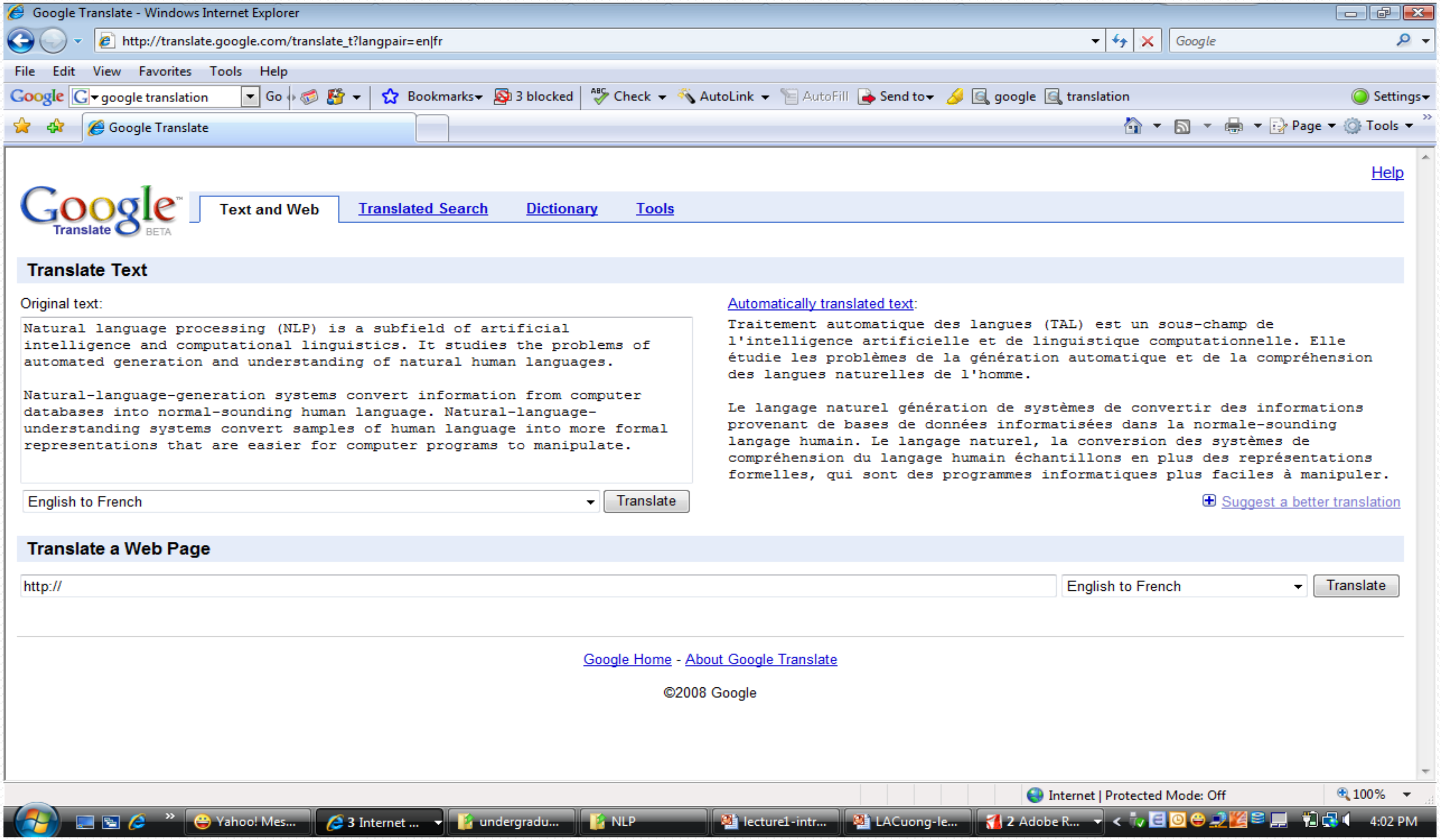
Information Extraction



10TH DEGREE is a full service advertising agency specializing in direct and interactive marketing. Located in Irvine CA, 10TH DEGREE is looking for an Assistant Account Manager to help manage and coordinate interactive marketing initiatives for a marquee automative account. Experience in online marketing, automative and/or the advertising field is a plus. Assistant Account Manager Responsibilities Ensures smooth implementation of programs and initiatives Helps manage the delivery of projects and key client deliverables . . . Compensation: $50,000-$80,000 Hiring Organization: 10TH DEGREE

INDUSTRY	Advertising
POSITION	Assistant Account Manager
LOCATION	Irvine, CA
COMPANY	10TH DEGREE
SALARY	$50,000-$80,000

Machine Translation


Google Translate - Windows Internet Explorer
http://translate.google.com/translate_t?langpair=en|fr
Google
File Edit View Favorites Tools Help
Google google translation Go Bookmarks 3 blocked Check AutoLink AutoFill Send to google translation Settings
Google Translate
Page Tools
Help
Google Translate BETA
Text and Web
Translated Search
Dictionary
Tools

Translate Text

Original text:

```
Natural language processing (NLP) is a subfield of artificial
intelligence and computational linguistics. It studies the problems of
automated generation and understanding of natural human languages.

Natural-language-generation systems convert information from computer
databases into normal-sounding human language. Natural-language-
understanding systems convert samples of human language into more formal
representations that are easier for computer programs to manipulate.
```

English to French

Translate

Automatically translated text:

```
Traitement automatique des langues (TAL) est un sous-champ de
l'intelligence artificielle et de linguistique computationnelle. Elle
étudie les problèmes de la génération automatique et de la compréhension
des langues naturelles de l'homme.

Le langage naturel génération de systèmes de convertir des informations
provenant de bases de données informatisées dans la normale-sounding
langage humain. Le langage naturel, la conversion des systèmes de
compréhension du langage humain échantillons en plus des représentations
formelles, qui sont des programmes informatiques plus faciles à manipuler.
```

Suggest a better translation

Translate a Web Page

http://

English to French

Translate

Google Home - About Google Translate

©2008 Google

Text summarization

- Automatic summarization is the creation of a shortened version of a text by a computer program.

- The product of this procedure still contains the most important points of the original text.

Name Entity Recognition

INPUT: Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT: Profits soared at [Company Boeing Co.], easily topping forecasts on [Location Wall Street], as their CEO [Person Alan Mulally] announced first quarter results.

Dialog System (Conversation agent)

- User: I need a flight from Boston to Washington, arriving by 10 pm.
- System: What day are you flying on?
- User: Tomorrow
- System: Returns a list of flights

Information retrieval

- General model:

  -A huge collection of texts

  -A query

  Task: find documents that are relevant to the given query

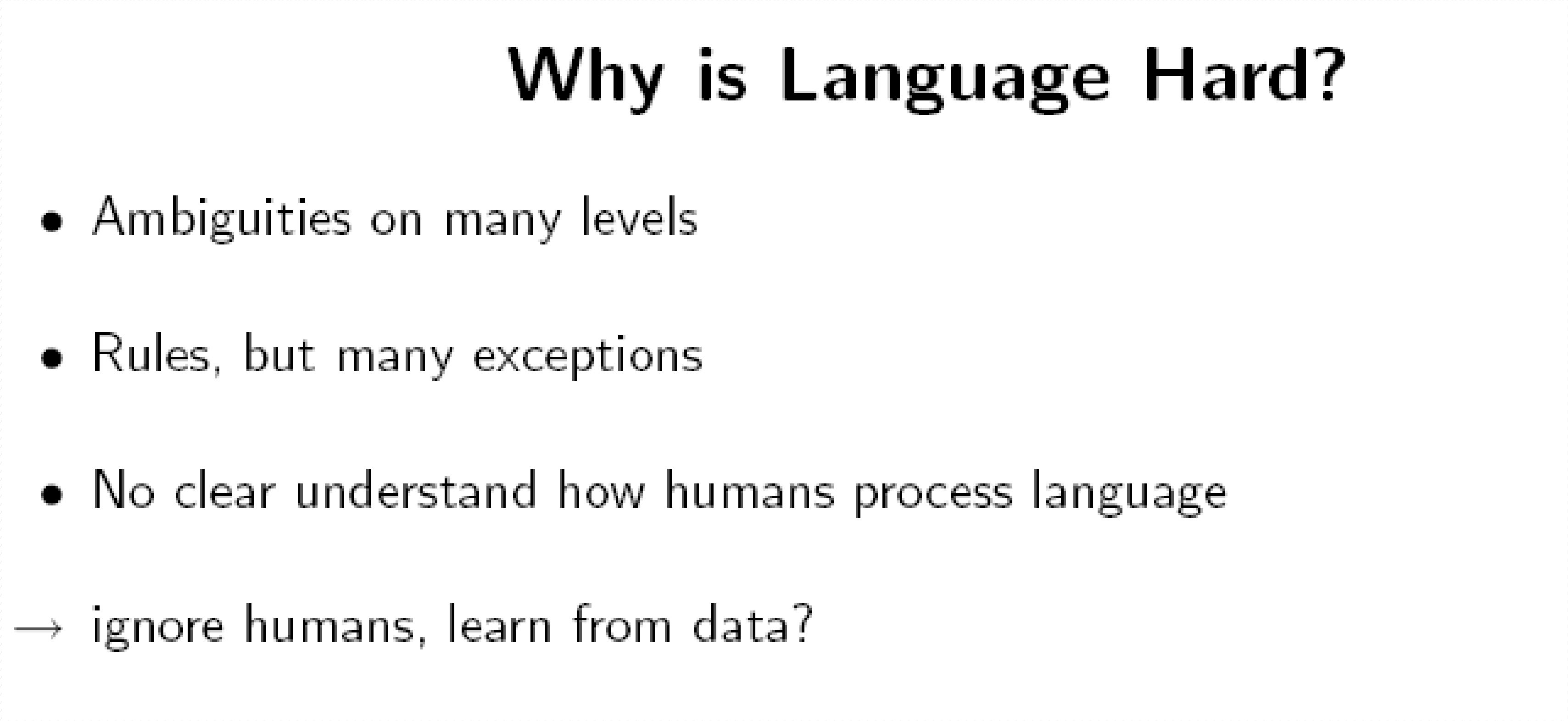
  Examples: google, yahoo, bing,…

Natural language generation

- Task of generating natural language from a machine representation system such as a knowledge base or a logical form.
- Text generation:
  - Generate natural sentence
  - For summarization
  - For changing the way of sentence

New problems in NLP

- Sentiment analysis and Opninion mining
- New languages: young people, chatting language, social networks, …
- Extracting knowledge from Texts in many resources (news, social network, ….)
  - Mining stock (forex) trend from twitter



Why is Language Hard?

- Ambiguities on many levels
- Rules, but many exceptions
- No clear understand how humans process language

→ ignore humans, learn from data?

Why NLP is difficult

- **Language is ambiguous**
  - At all levels: lexical, phrase, semantic
  - Iraqi Head Seeks Arms
    - Word sense is ambiguous (head, arms)
  - Stolen Painting Found by Tree
    - Thematic role is ambiguous: tree is agent or location?
  - Ban on Nude Dancing on Governor's Desk
    - Syntactic structure (attachment) is ambiguous: is the ban or the dancing on the desk?
  - Hospitals Are Sued by 7 Foot Doctors
    - Semantics is ambiguous : what is 7 foot?

Why NLP is difficult

- Language is flexible
  - New words, new meanings
  - Different meanings in different contexts
- Language is complex!

Why NLP is difficult

- MANY hidden variables
  - Knowledge about the world
  - Knowledge about the context
  - Knowledge about human communication techniques
    - *Can you tell me the time?*
- Problem of scale
  - Many (infinite?) possible words, meanings, context
- Problem of sparsity
  - Very difficult to do statistical analysis, most things (words, concepts) are never seen before
- Long range correlations

Why NLP is difficult

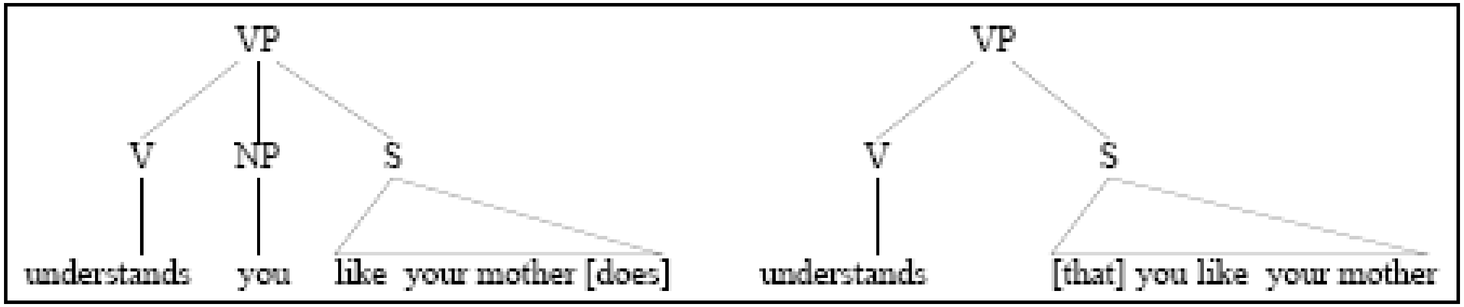
- Key problems:
  - Representation of *meaning*
  - Language presupposes knowledge about the world
  - Language only reflects the surface of meaning
  - Language presupposes communication between people

Why is NLP hard?

“At last, a computer that understands you like your mother”

Ambiguity at many levels

At the syntactic level:



Different structures lead to different interpretations.

Approaches in NLP

- Knowledge based approach
- Statistical approach

Corpus-based statistical approaches to tackle NLP problem

- How can a machine understand these differences?
  - Decorate the cake with the frosting
  - Decorate the cake with the kids
- Rules based approaches, i.e. hand coded syntactic constraints and preference rules:
  - The verb *decorate* require an animate being as agent
  - The object *cake* is formed by any of the following, inanimate entities (cream, dough, frosting.....)
- Such approaches have been showed to be time consuming to build, do not scale up well and are very brittle to new, unusual, metaphorical use of language

Corpus-based statistical approaches to tackle NLP problem

- A Statistical NLP approach seeks to solve these problems by automatically learning lexical and structural preferences from text collections (corpora)
- Statistical models are robust, generalize well and behave gracefully in the presence of errors and new data.
- So:
  - Get large text collections
  - Compute statistics over those collections
  - (The bigger the collections, the better the statistics)

Corpus-based statistical approaches to tackle NLP problem

- Decorate the cake with the frosting
- Decorate the cake with the kids

- From (labeled) corpora we can learn that:

  #(kids are subject/agent of decorate) > #(frosting is subject/agent of decorate)

- From (UN-labeled) corpora we can learn that:

  #(“*the kids decorate the cake*”) >> #(“*the frosting decorates the cake*”)

  #(“*cake with frosting*”) >> #(“*cake with kids*”)

  etc..

- Given these “facts” we then need a statistical model for the attachment decision

Corpus-based statistical approaches to tackle NLP problem

- Topic categorization: classify the document into semantics topics

Document 1

The U.S. swept into the Davis Cup final on Saturday when twins Bob and Mike Bryan defeated Belarus's Max Mirnyi and Vladimir Voltchkov to give the Americans an unsurmountable 3-0 lead in the best-of-five semi-final tie.

Topic = sport

Document 2

One of the strangest, most relentless hurricane seasons on record reached new bizarre heights yesterday as the plodding approach of Hurricane Jeanne prompted evacuation orders for hundreds of thousands of Floridians and high wind warnings that stretched 350 miles from the swamp towns south of Miami to the historic city of St. Augustine.

Topic = disaster

Corpus-based statistical approaches to tackle NLP problem

- Topic categorization: classify the document into semantics topics

Document 1 (sport)

The U.S. swept into the Davis Cup final on Saturday when twins Bob and Mike Bryan …

Document 2 (disasters)

One of the strangest, most relentless hurricane seasons on record reached new bizarre heights yesterday as….

- From (labeled) corpora we can learn that:

  #(sport documents containing word *Cup*) > #(disaster documents containing word *Cup*) **-- feature**

- We then need a statistical model for the topic assignment

Corpus-based statistical approaches to tackle NLP problem

- Feature extractions (usually linguistics motivated)
- Statistical models
- Data (corpora, labels, linguistic resources)

Summarization

- Learn about the problems and possibilities of natural language analysis:
  - What are the major issues?
  - What are the major solutions?
- At the end you should:
  - Agree that language is difficult, interesting and important
  - Be able to assess language problems
    - Know which solutions to apply when, and how
    - Feel some ownership over the algorithms
  - Be able to use software to tackle some NLP language tasks
  - Know language resources
  - Be able to read papers in the field