Mathematics and Big Data

Natural Language Processing - Part 1

Sundus Zafar

Department of Mathematics
Autonomous University of Barcelona

1. Natural Language Processing - Basics
   - Introduction
   - Why NLP ?
   - History
   - Applications
   - Challenges
   - NLP Techniques

Introduction

- Natural Language Processing, or NLP for short, is broadly defined as the automatic manipulation of natural language, like speech and text, by software.
- It is a branch of AI that helps computers understand, interpret and manipulate human language.
- NLP draws from many disciplines, including computer science and computational linguistics.
- Its pursuit to fill the gap between human communication and computer understanding.
- Its study has been around for more than 50 years to retrieve high quality information from text.

Why NLP ?

Natural language refers to the way we, humans, communicate with each other, namely, speech and text.
We are surrounded by text. Think about how much text you see each day:

- Emails
- SMS
- Web Pages
- Blogs
- Endless list of text...

History

A brief history of NLP

- Early enthusiasm (1950's): Machine Translation
  - Too ambitious
  - Bar-Hillel report (1960) concluded that fully-automatic high-quality translation could not be accomplished without knowledge (Dictionary + Encyclopedia)
- Less ambitious applications (late 1960's & early 1970's): Limited success, failed to scale up
  - Speech recognition
  - Dialogue (Eliza) **Shallow understanding**
  - Inference and domain knowledge (SHRDLU="block world")

Deep understanding in limited domain

- Real world evaluation (late 1970's – now)
  - Story understanding (late 1970's & early 1980's) **Knowledge representation**
  - Large scale evaluation of speech recognition, text retrieval, information extraction (1980 – now) **Robust component techniques**
  - Statistical approaches enjoy more success (first in speech recognition & retrieval, later others) **Statistical language models**
- Current trend:
  - Boundary between statistical and symbolic approaches is disappearing.
  - We need to use all the available knowledge **Applications**

Let's build an artificial system
that understands human language

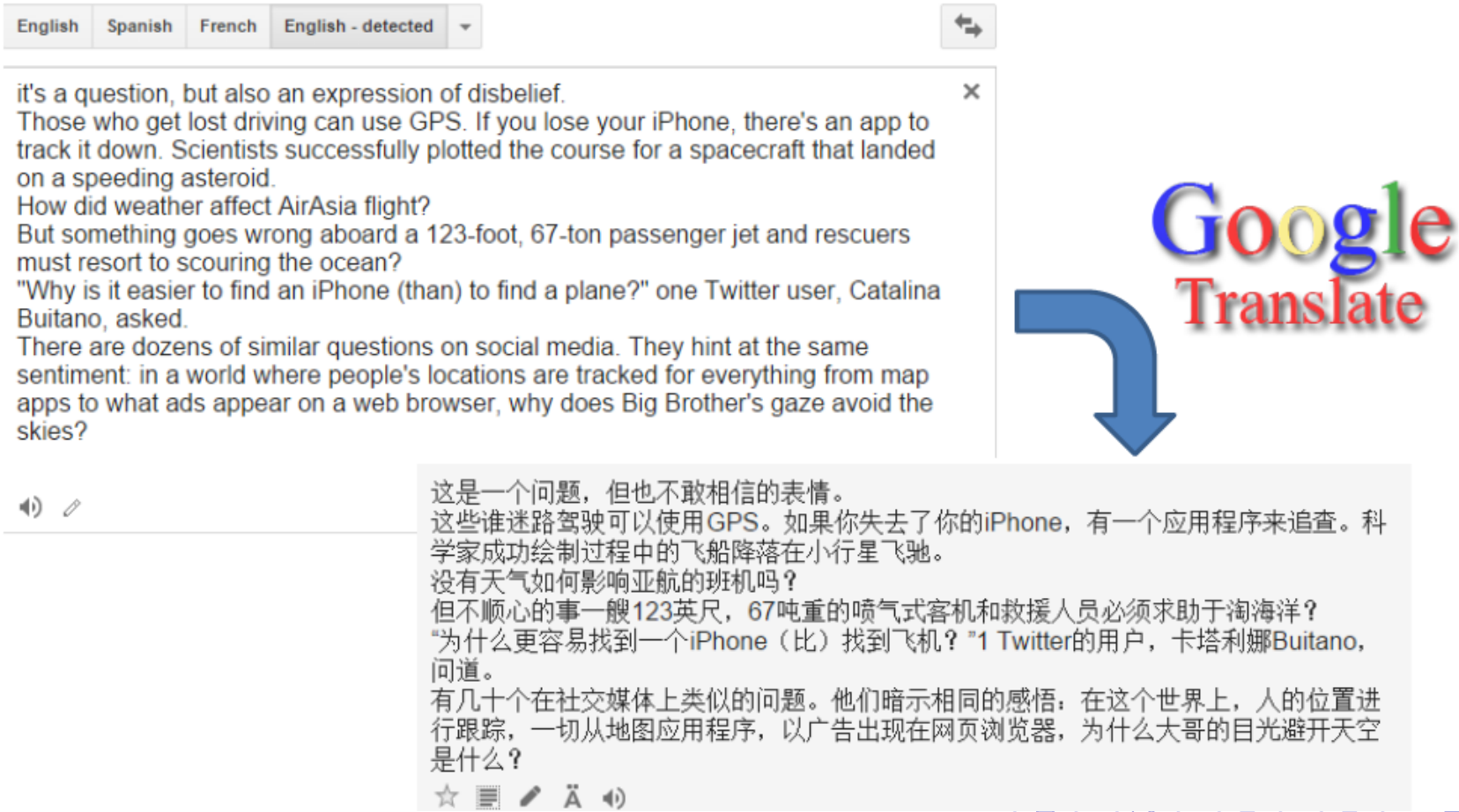
Applications

Which NLP application do we use every day?

- Google translate
- Facebook
- Twitter.
- Blogspot.
- Job seeking.
- Google search
- Yahoo search.
- Many more ...

Applications

Machine translation



Applications

The vast applications of NLP can be viewed in the following categories:

- Automatically answer our emails.
- Translate languages accurately.
- Manage, summarize and aggregate information.
- Use speech as a UI when needed.
- Talk to us.
- Litsen to us.

Challenges

Following challenges are commonly faced while working with text mining:

- Word level ambiguity
  - Words have multiple meaning e.g "Root".
  - Words can be noun or verb e.g "Design".
- Context sensitivity
  - Presupposition e.g He has quit school implies he attended school before.
- Syntactic ambiguity
- Others

Challenges

Issues in Syntax

Syntax does not deal with the meaning of a sentence, but it may help?!

"the dog ate my homework"

Who ate? →dog

The important thing when we analyze a syntax is to identify the part of speech (POS): Dog = noun ; ate = verb ; homework = noun

There are programs that do this automatically, called: Part of Speech Taggers. (also called grammatical tagging)

Accuracy of English POS tagging: 99%.

Identify collocations

mother in law, hot dog

Compositional versus non-compositional collocates

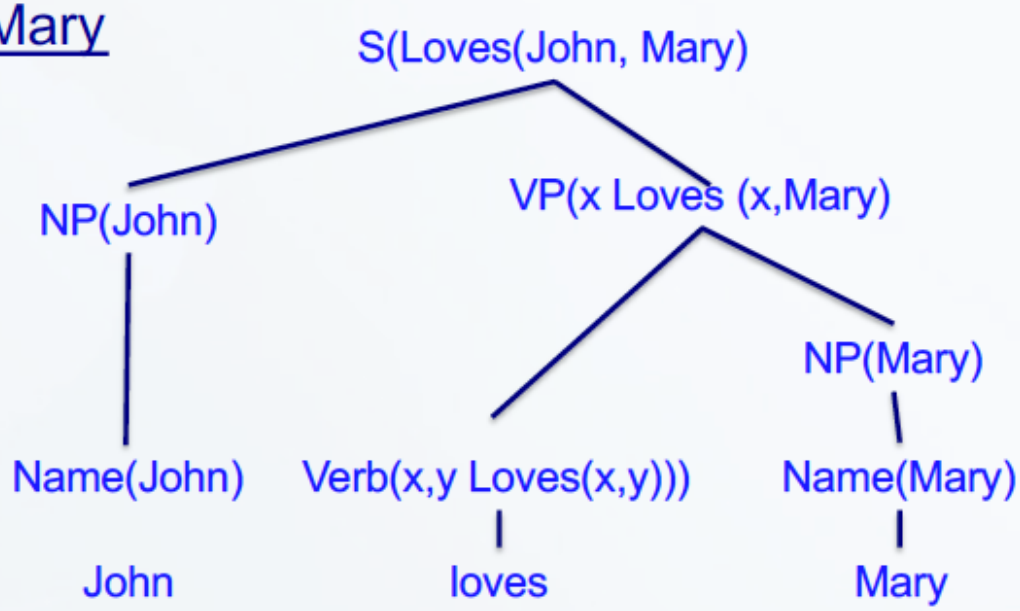
Challenges

Issues in Syntax (Part of Speech Tagging)

Based on [1]

Assume input sentence **S** in natural language **L**. Assume you have rules (*grammar* **G**) that describe syntactic regularities (patterns or structures). Given **S** & **G**, find syntactic structure of **S**. Such a structure is called a Parse Tree

Pars tree: John loves Mary



Helps a computer to automatically answer questions like -Who did what and when?

Tokenization

Tokenization/Segmentation

- Split text into words and sentences
  - Task: what is the most likely segmentation / tokenization?

There was an earthquake near Barcelona. I've even felt it in Terrassa, Sabadell, etc.

There + was + an + earthquake + near + Barcelona.

I + ve + even + felt + it + in + Terrassa, + Sabadell + etc.

Named Entity Recognition (NER)

Named entity recognition

- Determine text mapping to proper names
  - Task: what is the most likely mapping

Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

Organization, Location, Person

Parts of Speech (POS)

Part-of-Speech tagging

- Marking up a word in a text (corpus) as corresponding to a particular part of speech
  - Task: what is the most likely tag sequence

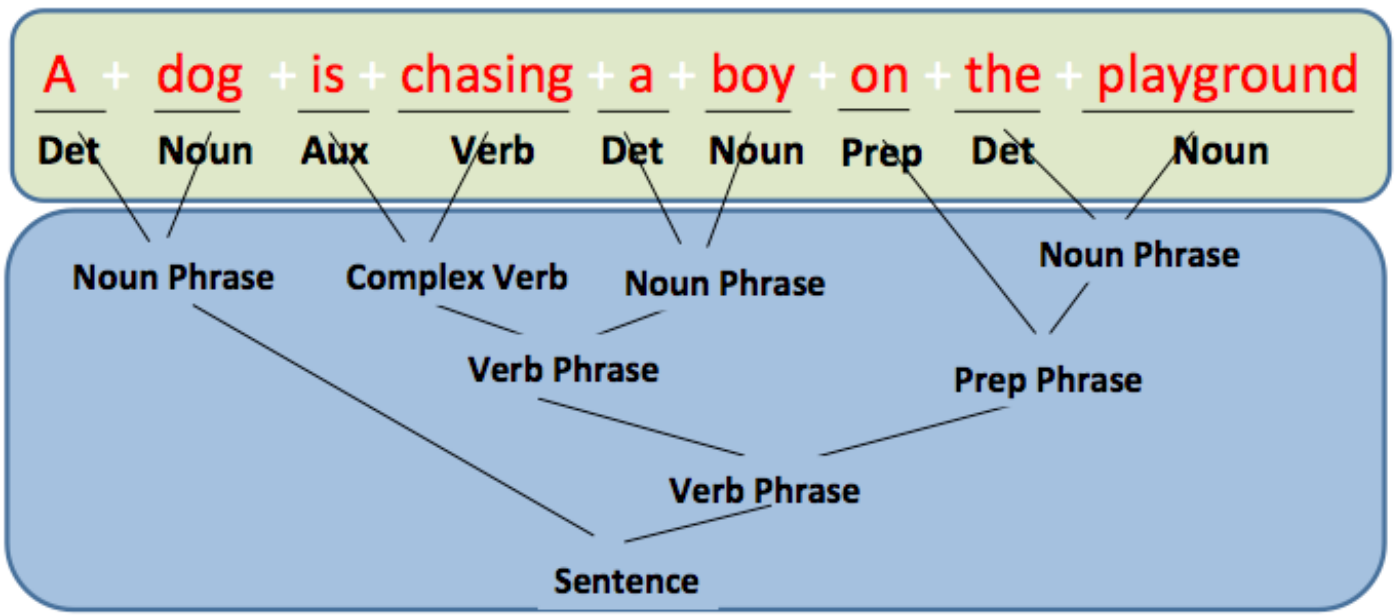
A + dog + is + chasing + a + boy + on + the + playground

A	dog	is	chasing	a	boy	on	the	playground
Det	Noun	Aux	Verb	Det	Noun	Prep	Det	Noun

Syntactic Parsing

Syntactic parsing

- Grammatical analysis of a given sentence, conforming to the rules of a formal grammar
  - Task: what is the most likely grammatical



References

References

Some of the slides in this lecture are based on the following resources , but with many additions and revision:

[1] Rada Mihalcea: Natural Language Processing, 2008
www.cs.odu.edu/~mukka/cs480f09/Lecturenotes/.../Intro1.ppt

[2] Markus Dickinson: Introduction to Natural Language Processing (NLP), Linguistics 362 course, 2006 http://www9.georgetown.edu/faculty/mad87/06/362/syllabus.html