Introduction of Natural Language Processing

Lê Anh Cường
Faculty of Information Technology, TDTU

Content

1. What is NLP?
2. Applications of NLP
3. Why is NLP difficult?
4. Approaches in NLP

What is NLP?

Natural Language Processing (NLP) is “ability of machines to understand and interpret human language the way it is written or spoken”. The objective of NLP is to make computer/machines as intelligent as human beings in understanding language.

WHY DO WE NEED NLP?

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC
Einstein met with UN officials in Princeton		

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

⇨ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

PROCESS OF NLP

NATURAL LANGUAGE UNDERSTANDING

NLU or Natural Language Understanding tries to understand the meaning of given text. The nature and structure of each word inside text must be understood for NLU. For understanding structure, NLU tries to resolve following ambiguity present in natural language:

- **Lexical Analysis**
- **Syntactic Analysis**
- **Semantic Analysis**

PROCESS OF NLP

NATURAL LANGUAGE GENERATION

It is the process of automatically producing text from structured data in a readable format with meaningful phrases and sentences. The problem of natural language generation is hard to deal with. It is subset of NLP. Natural language generation divided into three proposed stages:

- **Text Planning** – Ordering of the basic content in structured data is done.
- **Sentence Planning** – The sentences are combined from structured data to represent the flow of information.
- **Realization** – Grammatically correct sentences are produced finally to represent text.

Names and disciplines

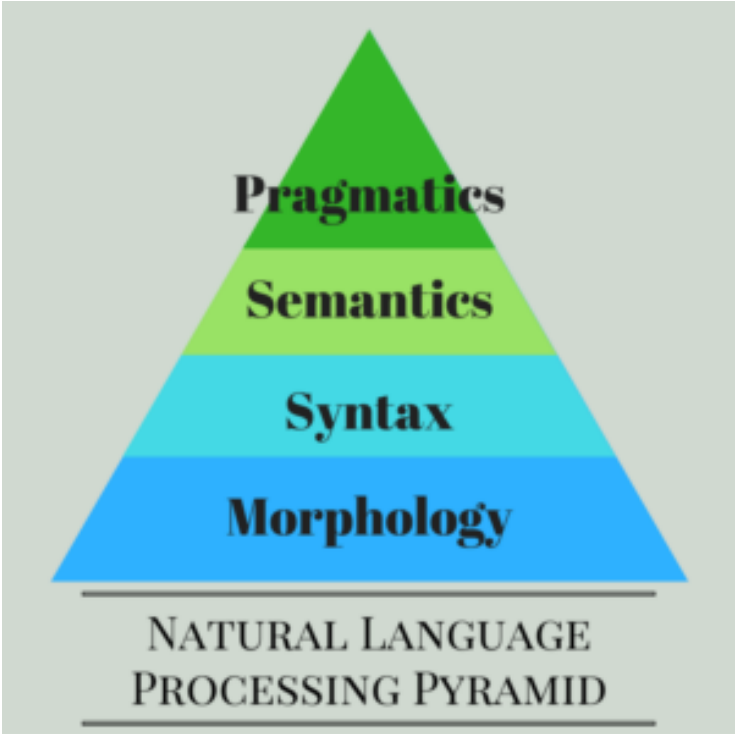
- **Natural Language Processing can take different names.** There may be some differences between them but the general direction of deriving meaning or understanding natural language is the same. Here are some alternative names:
- **Computational Linguistics** (nowadays used usually by people coming from a traditional linguistics background)
- **Text-Mining** (Usually when a lot of math/statistics are used)
- **Natural Language Understanding**

Names and disciplines

- Here are some disciplines related to modern natural language processing:
- Machine Learning
- Deep Learning
- Statistics
- Grammar
- Formal Languages, Regular Expressions

The NLP pyramid

- Natural Language Processing can also be viewed as a pyramid. The most common NLP tasks build one upon another.



Morphology

In traditional linguistics **morphology analyses how words are formed, what is their origin, how does their form change** depending on the context. In NLP you'll mostly deal with

- prefixes/suffixes
- singularization/pluralization
- gender detection
- word inflection
- lemmatization (the base form of a word).
- Spellchecking

In morphology, most of the operations are at a word level, where a word is viewed as a sequence of characters.

Syntax

Syntax cares about what proper word constructions are. **Determining the underlying structure of a sentence or building valid sentences is what syntax is all about.** In a way, syntax is what we usually refer to as grammar. Syntax is probably the most researched branch of computational linguistics. Here are only a few of the tasks:

- Part-of-speech tagging (assigning tags to words: Noun/Verb/Adjective/Adverb/Pronoun/Preposition/Conjunction etc …)
- Building Syntax Trees
- Building Dependency Trees

Syntax usually works on sentences, where a sentence is a sequence of words.

Semantics

Semantics derives meaning from text. This branch deals with the actual understanding of natural language. Here are some known problems:

- Named Entity Extraction
- Relation Extraction
- Semantic Role Labelling
- Word Sense Disambiguation

Semantics usually works on sentences, where a sentence is a sequence of words usually with some added semantics (like sense, role) attached.

Pragmatics

Pragmatics analyses the text as a whole. It's about determining underlying narrative threads, topics, references. Some discourse tasks are:

- Coreference / Anaphora resolution (find out what word refers what. Example: *John is fine. He**[John]**'s in no danger*.)
- Topic segmentation
- Lexical chains
- Summarization

Pragmatics usually works on a text represented as a sequence of sentences.

NLP Challenges

- Language is ambiguous.

I saw a man on a hill with a telescope.

NLP Challenges

- Language is ambiguous.

I saw a man on a hill with a telescope.

1. There's a man on a hill, and I'm watching him with my telescope.
2. There's a man on a hill, who I'm seeing, and he has a telescope.
3. There's a man, and he's on a hill that also has a telescope on it.
4. I'm on a hill, and I saw a man using a telescope.
5. There's a man on a hill, and I'm sawing him with a telescope.

Ambiguity in different levels

- **Lexical Ambiguity** – Words have multiple meanings
- **Syntactic Ambiguity** – Sentence having multiple parse trees.
- **Semantic Ambiguity** – Sentence having multiple meanings
- **Anaphoric Ambiguity** – Phrase or word which is previously mentioned but has a different meaning.

Why NLP?

- Text is everywhere, contains almost information
- Text is the way of communication

Why NLP?

- How does human can communicate with computer?
- How can we get knowledge from texts?

Question Answering: IBM's Watson

Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

Bram Stoker

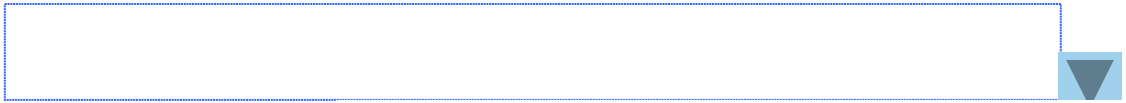
Information Extraction

Subject: **curriculum meeting**
Date: January 15, 2012

To: Dan Jurafsky

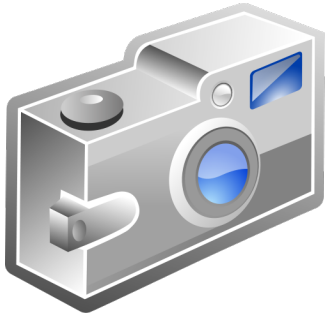
Event: Curriculum mtg
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

Hi Dan, we've now scheduled the curriculum meeting.
It will be in Gates 159 tomorrow from 10:00-11:30.
-Chris



Create new Calendar entry

Information Extraction & Sentiment Analysis



Attributes:

zoom

affordability

size and weight

flash

ease of use

Size and weight

✓ •nice and compact to carry!

•since the camera is small and light, I w

✓ around those heavy, bulky professional

•the camera feels flimsy, is plastic and v ...you

✗ have to be very delicate in the handling of this camera

Machine Translation

- Fully automatic

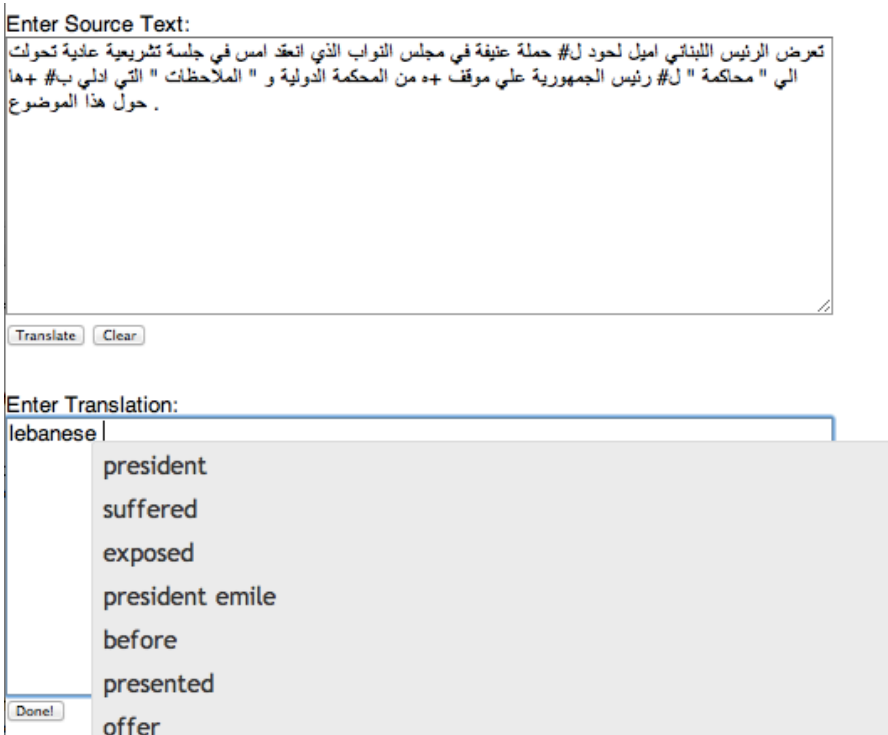
Enter Source Text:

这 不过 是 一 个 时间 的 问题 .

Translation from Stanford's *Phrasal*:

This is only a matter of time.

- Helping human translators



Language Technology

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC
Einstein met with UN officials in Princeton		

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

⇨ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Ambiguity makes NLP hard: "Crash blossoms"

Violinist Linked to JAL Crash Blossoms
Teacher Strikes Idle Kids
Red Tape Holds Up New Bridges
Hospitals Are Sued by 7 Foot Doctors
Juvenile Court to Try Shooting Defendant
Local High School Dropouts Cut in Half

Ambiguity is pervasive

New York Times headline (17 May 2000)

Fed raises interest rates

Fed raises interest rates

Fed raises interest rates 0.5%

In-video quizzes!

- Most lectures will include a little quiz
  - Just to check basic understanding
  - Simple, multiple-choice.
  - You can retake them if you get them wrong

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad
the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …
Let It Be was recorded …
… a mutation on the *for* gene …

But that's what makes it fun!

Making progress on this problem…

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources
- How we generally do this:
  - probabilistic models built from language data
    - P("maison" $\rightarrow$ "house") high
    - P("L'avocat général" $\rightarrow$ "the general avocado") low
  - Luckily, rough text features can often do half the job.

This class

- Teaches key theory and methods for statistical NLP:
  - Viterbi
  - Naïve Bayes, Maxent classifiers
  - N-gram language modeling
  - Statistical Parsing
  - Inverted index, tf-idf, vector models of meaning

- For practical, robust real-world applications
  - Text categorization
  - Information extraction
  - Spelling correction
  - Sentiment analysis
  - Machine translation
  - Question answering
  - Summarization

Skills you'll need

- Simple linear algebra (vectors, matrices)
- Basic probability theory
- Java or Python programming
  - Weekly programming assignments