Deep learning for natural language processing

Introduction to natural language processing

Benoit Favre <benoit.favre@univ-amu.fr>

Aix-Marseille Université, LIF/CNRS

03-2018

What is Natural Language Processing?

What is Natural Language Processing (NLP)?

- Allow computer to communicate with humans using everyday language
- Teach computers to reproduce human behavior regarding language manipulation
- Linked to the study of human language through computers (Computational Linguistics)

Why is it difficult?

- People do not follow rules strictly when they talk or write: "r u ready?"
- Language is ambiguous: "time flies like an arrow"
- Input can be noisy: speech recognition in the subway

NLP is everywhere

- Spell checker / grammar correction (Word)
- Information retrieval / search (Google)
- Machine translation (Google)
- Information extraction (Ask.com)
- Question answering (Jeopardy)
- Automatic summarization (Google news)
- Call routing (Telcos)
- Sentiment analysis (Amazon)
- Spam filtering (Email)
- Writing recognition (Cheque processing)
- Voice dictation (Dragon, Nuance)
- Speech synthesis (In-car GPS)
- Dialog systems (Siri/OK Google/Alexa...)

Domains reltated to NLP

- Artificial intelligence
- Formal language theory
- Machine learning
- Linguistics
- Psycholinguistics
- Cognitive Sciences
- Philosophy of language

Communication channel

From the point of view of the source (the speaker)

1. Intent: the message we want to communicate
2. Generation: the message in linguistic form
3. Production: the muscular action which leads to sound production

From a receiver point of view (listener)

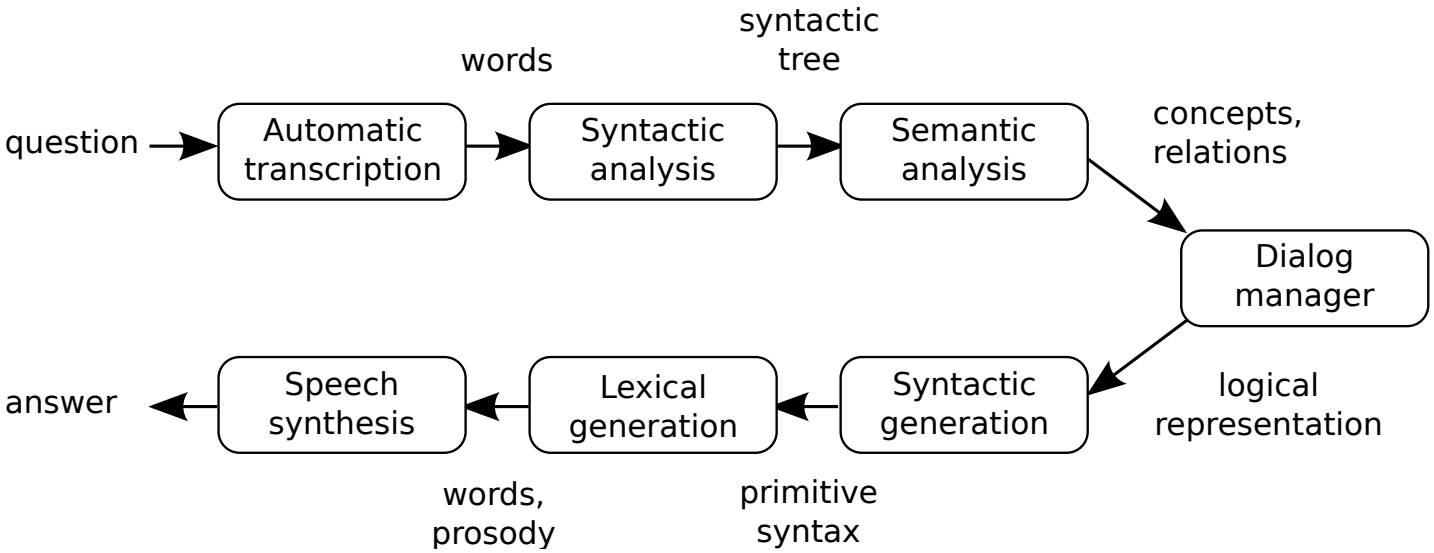
1. Perception: how the sound is transmitted to neurons
2. Analysis: interpretation of the linguistic message (syntactic, semantic...)
3. Integration: believe or not the information, reply...

Processing levels

"John loves Mary"

1. **Lexical** : segment character stream in words, identify linguistic units
   John/firstname-male *loves*/verb-love *Mary*/firstname-female
2. **Syntax** : identify grammatical structures
   (S (NP (NNP John)) (VP (VBZ loves) (NP (NNP Mary))) (. .))
3. **Semantic** : represent meaning
   love(person(*John*), person(*Mary*))
4. **Pragmatic** : what is the function of that sentence in context?
   Is it reciprocal ?
   Since when ?
   What does it entail ? know(*John*, *Mary*)

Modular approach



Language ambiguity

- Phonetic
  - I don't know! – I don't - no!
- Graphical
- Phonetic and graphical
  - I live by the bank (river bank or financial institution)
- Etymology
  - I met an Indian (from India or native American)
  - I love American wine (from USA or from the Americas)
- Syntactic
  - He looks at the man with a telescope
  - He gave her cat food
- Referential
  - She is gone. Who?
- Notational conventions
  - Birth date: 08/01/05

(wikipedia)

Basic NLP tasks

- Syntax
  - Word / sentence segmentation
  - Morphological analysis
  - Part-of-speech tagging
  - Syntactic chunking
  - Syntactic parsing
- Semantic
  - Word sense disambiguation
  - Semantic role labeling
  - Logical form creation
- Pragmatic
  - Coreference resolution
  - Discourse parsing

Word segmentation

Character sequence → word sequence (tokenization)

- Split according to delimiters [:,.!?']
- What about compounds? Multiword expressions?
- URLs (`http://www.google.com`), variable names (theMaximumInTheTable)
- In Chinese, no spaces between words:
  - 男孩喜歡冰淇淋。→ 男孩 (the boy) 喜歡 (likes) 冰淇淋 (ice cream) 。

Morphological analysis

Split words in relevant factors

- Gender and number
  - flower, flower+s, floppy, flopp+ies
- Verb tense
  - parse, pars+ing, pars+ed
- Prefixes, roots and suffixes
  - geo+caching
  - re+do, un+do, over+do
  - pre+fix, suf+fix
  - geo+local+ization
- Agglutinative languages
  - pronouns are glued to the verb (Arabic, spanish...)
- Rich morphology
  - Turkish, Finish
- $\rightarrow$ Lemmatization task: find canonical word form

Part-of-speech tagging

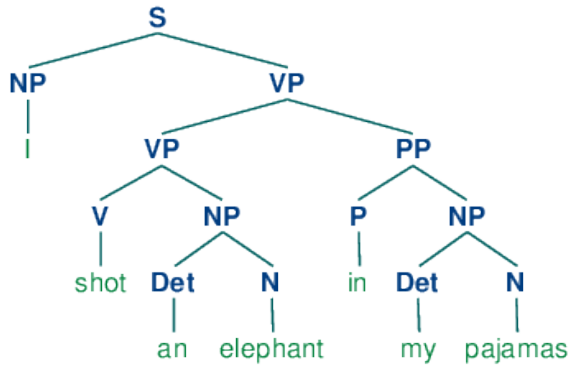
- Syntactic categories

Noun	Adverb	Discourse marker
Proper name	Determiner	Foreign words
Verb	Preposition	Punctuation
Adjective	Conjunctions	Pronouns

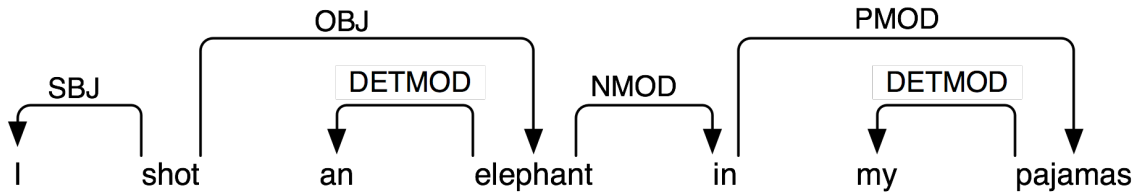
- Each word can have multiple categories
- Example : *time flies like an arrow*
  - flies: verb or noun?
  - like: preposition or verb?

Syntactic analysis

- Constituency parsing



- Dependency parsing



Word sense disambiguation (WSD)

What is the sense of each word in its context?

- **red**: color? wine? communist?
- **fly**: what birds do? insect?
- **bank**: river? financial institution?
- **book**: made of paper? make a reservation?

Word meaning highly depends on domain

- **apple**: fruit? company?
- **to pitch**: a ball? a product? a note?

Semantic parsing

Syntax is ambiguous

- The man **opens** the door
- The door **opens**
- The key **opens** the door

Semantic roles

- Who performed the action? **the agent**
- Who receives the action? **the patient**
- Who helps making the action? **the instrument**
- When, where, why?

John	sold	his car	to his brother	this morning
agent	predicate	instrument	patient	time

Reference resolution

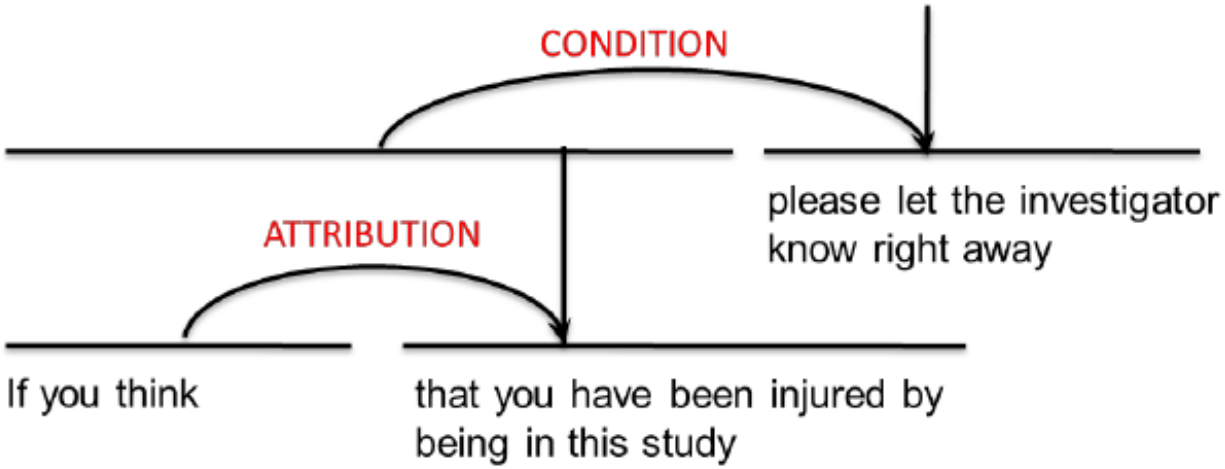
- Link all references to the same entity
  - “Alexander Graham Bell (March 3, 1847 –August 2, 1922)[4] was a Scottish-born[N 3] scientist, inventor, engineer, and innovator who is credited with patenting the first practical telephone.” (Wikipedia)

Ambiguity

- Pronouns (it, she, he, we, you, who, whose, both...)
- Noun phrases (the young man, the former president, the company...)
- Proper names (”Victoria”: South-African city, Canadian region, Queen, model...)

Discourse analysis

Relationship between sentences of a text, argument structure.



"Fully Automated Generation of Question-Answer Pairs for Scripted Virtual Instruction", Kuyten et al, 2012

Relation type (Rhetorical Structure Theory)

- Background
- Elaboration
- Preparation
- Contrast
- Objective
- Cause
- Circumstances
- Interpretation
- Justification
- Reformulation

Create a logical form

- Predicate representation
  - Can be used to infer new
- *John loves Mary but it is not reciprocal.*

$$\exists x, y, name(x, \text{“}John\text{”}) \wedge name(y, \text{“}Mary\text{”}) \wedge loves(x, y) \wedge not(loves(y, x))$$

- *John sold his car this morning to his brother.*

$$\exists x, y, z, name(x, \text{“}John\text{”}) \wedge brother(x, y) \wedge car(z)$$
$$\wedge owns(x, z) \wedge sell(x, y, z) \wedge time(\text{“}morning\text{”})$$

History of natural language processing

- 1950: Theory (test de Turing, grammaires de Chomsky)
  - Automatic translation during the cold war
- 1960: Toy systems
  - SHRDLU "place the red box next to the blue circle", ELIZA "the therapis"
- 1970:
  - Prolog (logic-base language for NLP), Dictionaries of semantic frames
- 1980: Dictation, Development of grammars
- 1990
  - Transition "introspection" $\rightarrow$ "corpus"
  - Evaluation campaigns
  - Neural networks are "forgotten"
- 2000
  - Machine learning
  - Applications: speech recognition, machine translation
- 2010...
  - Deep learning

Notion of corpus

- Language in the wild
  - Email
  - Forums
  - Chats
  - Speech recordings
  - Video
- **Manual Annotation** of all elements we want to predict
  - Text $\rightarrow$ topic
  - Sentence $\rightarrow$ parse tree
  - Review $\rightarrow$ sentiment

Methodology

Corpus-based natural language processing

1. Define a task
2. Write an annotation guide
3. Collect raw data
4. Ask people to annotate that data
5. Create a system to perform the task
6. Evaluate the output of the system

NLP Systems

- Input
  - Raw text, audio...
  - Sentences, contextualized words
  - Output of another system
- Output
  - n classes (ex: topics)
  - Structure (ex: syntactic parse)
  - Novel text (ex: translation, summary)
  - Commands for a system (ex: chatbots)
- Process
  - $\text{output} = f(\text{input})$
  - Deterministic vs random (evaluations need to be repeatable)
  - Parametrisable: $\text{output} = f(\text{input}, \text{parameters})$

What is the deep learning promise?

- "Classic" NLP system development
  - Requires a lot of feature engineering
  - Is affected by cascading errors
  - Hard to account for unlabeled data
  - Limited architectures and overly complex (ex: speech recognition...)
  - The curse of annotated data (you need linguists)
- Deep learning
  - Feature extraction is learned within the model
  - End-to-end training
  - Much more flexibility in model architectures
  - Can use tons of data
  - A step towards AI?
- Is this the end of linguistic expertise?