An introduction to Natural Language Processing (NLP)

Houcine Senoussi

1. Introduction
2. Lexical Analysis
3. Syntactic Analysis
4. Semantic Analysis
5. Conclusion
6. References

What is it about ?

- Natural Language Processing(NLP) is a branch of Artificial Intelligence(AI).
- It is about communicating with machines using a natural (human) language.
- According to E.D. Liddy "Natural Language Processing is a theoretically motivated range of **computational techniques** for **analyzing** and representing **naturally** occurring **texts** at one or more **levels** of linguistic analysis for the purpose of achieving **human-like** language processing for a range of tasks or **applications**."[1]

What is it about ?

1. Which documents?
    - Natural languages (vs formal languages): those which humans naturally speak (Arabic, French, English, ...).
    - Written and spoken forms.
2. Which applications?
    - Machine translation.
    - Question answering.
    - Text summarization.
    - Many other applications.

What is it about?

3. Which methods/techniques?
   1. Automata theory.
   2. Machine learning.
   3. Statistics.
   4. Logics.

Approaches to NLP

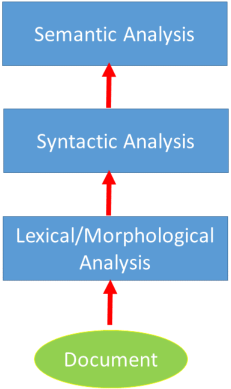
4. Which approaches?
   - Symbolic: Knowledge is (mainly) represented as facts or rules.
   - Statistical: uses large corpora to develop approximate generalized (statistic) models of linguistic phenomena based on actual examples of these phenomena.
   - Connectionist: using neural networks. Nowadays, the main NN used in NLP are LSTM (Long Short Term Memory) and ConvNet (convolutional neural networks).

Levels of language processing

5. Which levels of analysis?
    1. Phonology : studying sounds whose sequences define words.
    2. Morphology: studying words' forms.
    3. Lexical: studying words' role in sentences and their meanings.
    4. Syntactic: analyzing the words in a sentence.
    5. Semantic: understanding the meaning of the document.
    6. Discourse: focuses on the properties of the text as a whole rather than on sentences.
    7. Pragmatics: 'Studying of how knowledge about the world and language conventions interact with literal meaning' (ref 4)

NLP systems mainly deal with lower levels of processing and more particularly lexical (including morphological) /syntactic/semantic analyses.

Levels of language processing



Lexical Analysis

1. Lexical analysis is the analysis of texts at the level of the word.
2. For that we first need to transform a raw text into a sequence of words and other basic units. This is the role of **tokenization**. Then, we assign to each unit a category called its '**part-of-speech**" tag.
3. It requires a **lexicon**: a database containing the allowable words and grammatical and semantic information about them.

Tokenization

1. **Tokenization** is the first operation that must be applied to the text being processed. It is breaking up the document into **tokens**: words, punctuation marks, numbers and other discrete items.
2. For example, tokenization of the string 'Ernest Hemingway (July 21, 1899 - July 2, 1961) was an American novelist.' gives the tokens 'Ernest', 'Hemingway', '(', 'July', '21', ',', '1899', '-', 'July', '2', ',', '1961', ')', 'was', 'an', 'American', 'novelist'.

Parts of Speech

Several POS classifications have been developed. The following is a list of the nine POS that are often encountered:

1. Adjective: High, Nice, Ugly, Wide, ...
2. Adverb: Easily, Here, Now, Often, ...
3. Conjunction: And, But, For, Yet, ...
4. Determiner: A, An, That, The, ...
5. Noun: Bike, Car, Cat, Dog, ...
6. Preposition: At, In, From, On, ...
7. Pronoun: I, He, It, Us, ...
8. Proper noun: Albert, France, Python, ...
9. Verb: Does, Open, Read, Went, ...

Parts of Speech

- Assigning a part-of-speech tag to each word can be complex. This is the case, for example, when the same word has more than one possible part-of-speech (e.g. "He likes to **fish**." and "He caught a **fish**."). Words having more than one possible part-of-speech are assigned the most probable one based on the context in which they occur.
- Punctuation marks don't have POS tags. Depending on the application goal, they can be removed or not before POS tagging.

Stemming

- Stemming is an operation that identifies **morphologically complex** words, decomposes them into invariant **stems** (= basic form) and affixes, and deletes the affixes.
- Examples:
  - goes=go+es → *go*.
  - computers=computer+s → *computer*.
  - commanded=command+ed → *command*.
  - studies=studi+es → *studi*.
  - studying=study+ing → *study*.
  - meeting=meet+ing → *meet*.
- As we can see in the example of the word '*studies*', stems are not always valid words.

Lemmatization

- Lemmatization is another operation that aims at reducing words to bases forms: **lemmas**, also called dictionary forms. For that, a lemmatization algorithm needs to know the POS tag of the word and a set of rules explaining how different forms of a word are obtained.
  - better → *good*.
  - saw (the verb) → *see*.
  - saw (the noun) → *saw*.
  - meeting (the verb) → *meet*.
  - meeting (the noun) → *meeting*.

Syntactic Analysis

- Sentence is the basic unit of meaning in a text. Therefore, to extract the meaning of a text, we have to **identify** and **analyze** sentences.
- Sentences are not linear sequences of words. They are built according to some (**production**) **rules** which define sentences' **structure**.
- An example of rules(from reference [5]):
  1. Sentence → Noun Phrase, Verb Phrase.
  2. Noun Phrase → Determiner, Noun.
  3. Noun Phrase → Proper Noun.
  4. Noun Phrase → Noun Phrase, Conjunction, Noun Phrase.
  5. Verb Phrase → Verb, Noun Phrase.
  6. Verb Phrase → Verb, Preposition, Noun Phrase.

Grammars

- Such rules are part of the definition of a **grammar**. More generally a (formal) grammar is defined by:
  1. A set Σ of terminal symbols (words of the lexicon, punctuation signs, ..). These are the symbols that make up the sentences.
  2. A set N of non terminal symbols (Sentence, Noun Phrase, Verb Phrase, ...).
  3. A starting symbol S: the non terminal symbol which denotes the whole sentences (in our example: Sentence).
  4. A set of production (i.e. **rewrite**) rules of the form $LHS \rightarrow RHS$ in which LHS is a sequence of non terminal symbols and RHS a sequence of zero or more symbols (either terminal or non terminal).

Grammars-An example

- Σ=$\{Conj, Det, Nn, NP, PN, Prep, S, V, VP\}$
- N=$\{a, and, barked, cat, dog, fish, in, John,$
  $Mary, on, or, slept, swam, talked, the, to\}$
- Production rule:
  1. S → S Conj S.
  2. S → NP VP.
  3. NP → Det Nn| PN | NP Conj NP.
  4. VP → V NP|V Prep NP|V Conj VP.
  5. Det → *the*|*a*.
  6. Nn → *cat*|*dog*|*fish*.
  7. PN → *Mary*|*John*.
  8. V → *swam*|*talked*|*barked*|*slept*.
  9. Conj → *and*|*or*.
  10. Prep → *in*|*on*|*to*.

Grammars

- After Chomsky, we usually distinguish **four classes** of grammar depending on **the form of their production rules**. These classes can be arranged in a **hierarchy** in which a class describes all the languages described by the lower class, as well as additional languages.
- It is usually admitted that natural language grammars need at least the power of the so-called **context-free** class.
- Context-free grammars(CFG) are defined by rules having a single (non terminal) symbol on they left-hand side.

Parsing

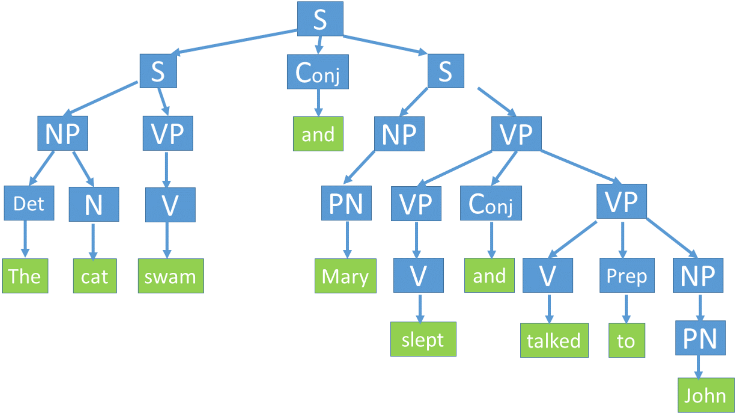
- Grammars/Set of rules give a specification of well-formed sentences. **Parsers** use them to analyze sentences.
- One way to show the result of syntactic analysis is with a **parse tree**: a tree in which the root is the sentence being analyzed, inner nodes are phrases (left-hand side of the rules), links represent application of rules and leaves terminal symbols (words).
- Example: the grammar given above recognizes the following sentence

The dog swam and Mary slept or talked to John.

Parse tree

- The parse tree of this sentence is given below.

Parse tree


S
S
Conj
S
NP
VP
and
NP
VP
Det
N
V
PN
VP
Conj
VP
The
cat
swam
Mary
V
and
V
Prep
NP
slept
talked
to
PN
John

Parse tree

- This parse tree can also be represented using the following labelled bracketed string:

Parse tree

```
(S
   (S
      (NP (Det the) (Nn dog)) (VP (V swam))
   )
   (Conj and)
   (S
      (NP (PN Mary))
      (VP
         (VP (V slept))
         (Conj or)
         (VP (V talked) (Prep to) (NP (PN John)))
      )
   )
)
```

Semantic Analysis

- Semantic analysis is about "**understanding** the **meaning**" of words, expressions, sentences and documents.
- All the levels of the analysis contribute to the semantics.
- We distinguish **lexical semantics** (meaning of individual words) and **supralexical semantics** (meaning of words' combinations such as phrases and sentences). The two parts/levels interact and interpenetrate in various ways.

Ambiguity

Many human/natural sentences are **ambiguous**: they are open to multiple interpretations. To understand the meaning of a sentence, an NLP system has to **resolve ambiguity**. The main types of ambiguity are the following:

- **Lexical** ambiguity: occurs when using a word that has more than one meaning.
  - Example: The priest married my sister.
- **Syntactic** ambiguity: occurs when a sentence has two or more different parses.
  - Example: Police help dog bite victim.
- **Referential** ambiguity: it occurs, for example, when there is a pronoun that can refer to several persons or things.
  - Example: He told the good news to his father. He was very happy.

Ambiguity-A famous example in French

We owe this example to the late Professor Daniel Kayser (cited in reference 7). The phrase 'Le prix Goncourt' designates:

1. an award in 'le Prix Goncourt a été attribué à X',
2. an amount in 'X a versé son Prix Goncourt à la Croix Rouge',
3. a person in 'le Prix Goncourt a été félicité par le Président',
4. a jury in 'le Prix Goncourt a admis un nouveau membre',
5. a book in 'peux-tu m'acheter le Prix Goncourt',
6. an event in 'depuis son Prix Goncourt, il est devenu arrogant',
7. and yet another meaning in 'le Prix Goncourt pervertit la vie littéraire'.

Lexical semantics

- It is about studying the meaning of individual words, and more generally the meaning of lexical units (lexical items) such as affixes or compound words.
- Analyzing the meaning of lexical items: analyzing their internal structure and content/representing their relations to other elements in the lexicon.

Lexical semantics

- Horizontal relations between lexical items:
  1. Synonymy: words having the same/close meaning(s). e.g. small/little.
  2. Opposition: antonymy (big/small), conversity (teach/learn, give/take, ...), ...
  3. Homonymy: different word with the same form (e.g. light (vs dark)/light(vs heavy)).
  4. Polysemy: different meaning for the same word (satellite: an electronic device that is sent into space and satellite: a country or state dependent on a more powerful one).

Lexical semantics

- Vertical relations between lexical items:
  1. Hyponymy/Hypernymy: the meaning of one lexical item, the hyponym, is more specific than the other, the hypernym (e.g. rose/flower, orange/fruit). Lexical items that are hyponyms of the same item are called co-hyponyms (e.g. cat and dog).
  2. Meronymy: a part-of relationship between an the meronym, and another, the holonym (e.g. nose/face, wheel/car).

Wordnet

- **Wordnet** is an electronic lexical database of English.
- It groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (**synsets**), each expressing a distinct concept.
- It also includes relations between synsets (hyperonymy/hyponymy, synonymy, ...).
- Wordnet can be used with Python thanks to the module NLTK.

Compositional semantics

- There exists several approaches for supralexical semantics. Compositional approach are based on the assumption that the meaning of a words' combination is determined by the meanings of its parts and the way in which those parts are combined (principle of compositionality).
- Logical approach belongs to this category of approaches. It uses first order logic (predicate logic) which represents semantic interpretation of sentences using variables, constants, predicates, functors, logical connectives and quantifiers (see the chapter 'Une introduction à la logique').
- In this approach, the meaning of a sentence is equated with the truth conditions of the formula(s) representing it, that is, the conditions under which these formulas are true.

Compositional semantics

- The following examples show two sentences and the interpretation of their meanings:
  1. *Some teachers are sick* → $\exists\ x\ teacher(x)\ \wedge\ sick(x)$
  2. *All French students learn Englich* → $\forall\ x\ (french(x)\ \wedge\ student(x))\ \Rightarrow learn(x, english)$
- Predicate rules also includes **inference rules** which determine which sentences are true given that some other sentences are true. The most famous of these rules is the **modus ponens** of which an example of application is the following:
  1. **Premise 1:** Pierre is sick ($P:\ sick(c)$).
  2. **Premise 1:** Whenever Pierre is sick he stays at home ($P\ \Rightarrow\ Q\ Q:\ stay\_at\_home(c)$).
  3. **Conclusion:** Pierre stays at home ($Q:\ stay\_at\_home(c)$).

Conclusion

- In this chapter we introduced Natural Language Processing (NLP).
- We briefly described the main levels of language processing.

References

1. Elizabeth D. Liddy.
2. A modern approach
3. pc2009
4. ref4 :
5. Introduction-to-natural-language-processing.pdf. R. Kibble.
6. ref6: handbook.
7. ref7: TALN