Natural Language Processing: 101

Natural Language Processing (NLP) is a field of Artificial Intelligence (AI) that deals with computers perceive, process and understand natural languages.

One of the greatest challenges in PLN is ambiguity, the same expression can have different meanings.

Different levels of analysis

Morphological Analysis

Focused on the study and classification of isolated words. Splits text into tokens, identifying part-of-speech (PoS) tags, lemmas, and radicals.

The PoS tag identifies what is the function of each token (e.g. verb, punctuation, proper noun). It can also identify gender (male or female), singular or plural, and degree.

Ambiguity at this level is related with words (tokens) that can belong to different PoS tags.

Tasks related with this analysis include: sentence splitting and *tokenization*.

Syntactic Analysis

Syntactic analysis, or parsing, studies the relations between words (tokens) in a sentence.

Tasks related with this analysis consist mainly on creating trees (e.g. dependency, constituency) that capture relations between tokens.

Semantic Analysis

Semantic analysis is focused with extracting the concepts and ideas (the meaning) of a sentence. This usually entails representing the meaning of the text using a unambiguous formal representation.

Extracting structured triples of information from unstructured text is a common approach.

Pragmatic Analysis

Pragmatic analysis is focused on studying how the sentence/text is related with the context, how it is related with real world concepts and ideas, and is usually

also related with some previous knowledge on the topic.

Phonetic Analysis

The study of sounds. Ambiguity problems related with different words that are pronounced (have the same sounds) the same.

References

- Jurafsky, D. and Martin, J. H. (2008). Speech and Language Processing. Prentice Hall, 2nd edition.
- Jackson, P. and Moulinier, I. (2007). Natural Language Processing for Online Applications: Text retrieval, extraction and categorization. John Benjamins Publishing, 2nd edition.