Introduction to Text Processing

Sundharakumar KB

Department of Computer Science and Engineering
School of Engineering

Shiv Nadar University Chennai

NLP

- NLP is a branch of AI designed to enable machines to understand human language.

- The main intention of NLP is to build systems that are able to make sense of text and execute certain tasks like spell-check, text translation, topic classification, etc.

Components of NLP

- Natural language generation (NLG) – method of creating meaningful phrases and sentences from data.
- Text planning – retrieving applicable content
- Sentence planning – creating meaningful sentences and phrases.
- Text realization – mapping sentence plans to sentence structures.
- Eg: Chatbots, voice assistants, sentiment analysis platforms, etc.

Components of NLP

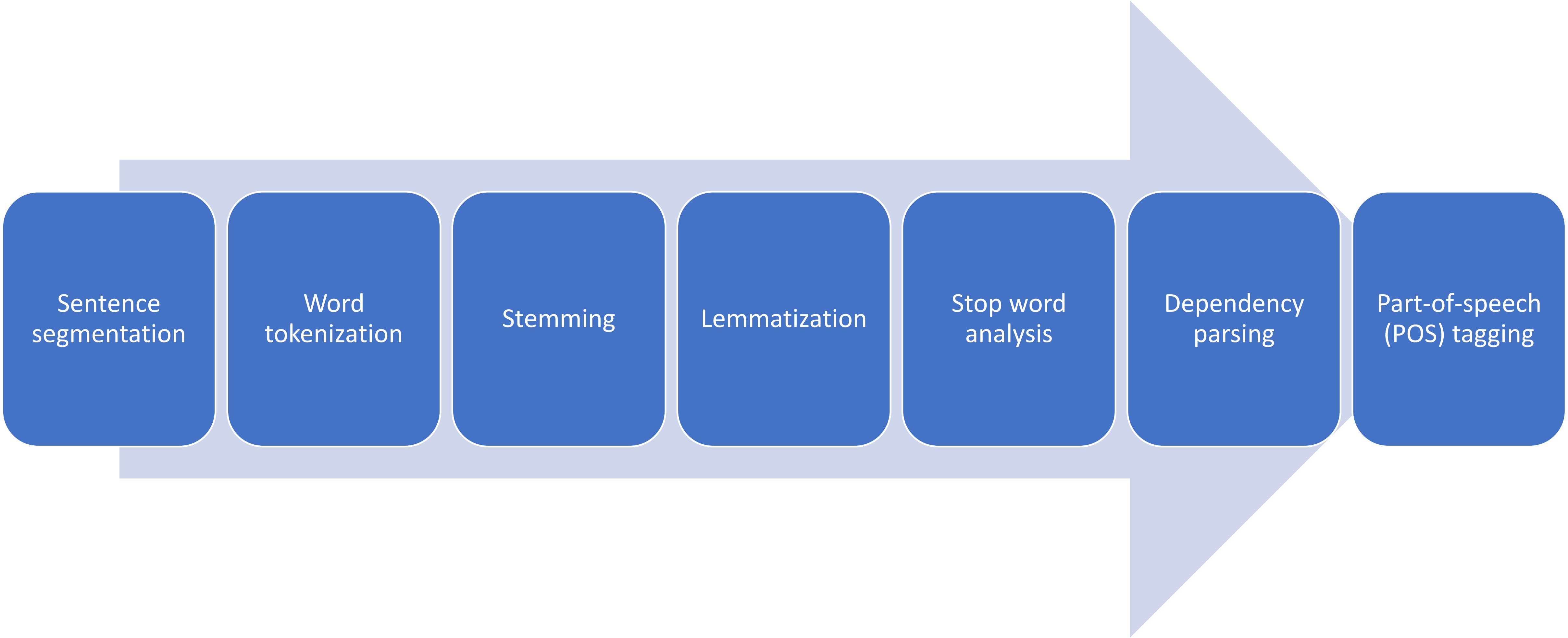
- Natural language understanding (NLU) - enable machines to understand and interpret human language by extracting metadata from content

- **Lexical ambiguity**: This means that one word holds several meanings. For example, "The man is looking for the match.“

- **Syntactic ambiguity**: This refers to a sequence of words with more than one meaning. For example, "The fish is ready to eat.”

- **Referential ambiguity**: This involves a word or a phrase that could refer to two or more properties. For example, Tom met Jerry and John. They went to the movies.

Pipeline of NLP



Regular expressions

- A RegEx, or Regular Expression, is a sequence of characters that forms a search pattern.

- RegEx can be used to check if a string contains the specified search pattern.

- Python has an in-built model "re" which can be used to work with Regular expressions.

Re functions

- match - string that is searched should be present in the beginning, else matching won't happen

- search - returns the span of first occurence of the string

- findall - returns list of all matchings

- split - splits the string based on a given character

- sub - replaces one or many matches

- compile - compiles a pattern in order to use with match or search

Re symbols

- [] - set of characters
- \ - special sequence
- ^ - startswith
- $ - endswith
- * - 0 or more
- + - 1 or more

Re symbols

- \d - digits
- \D - except digits
- \s - matches white character
- \S - returns a match that does not contain space
- \w - word (a-z, A-Z, 0-9 and _)
- \W - does not contain any words
- [a-n] - alphabetically between a and n
- [^ahl] - returns all matches except the characters mentioned