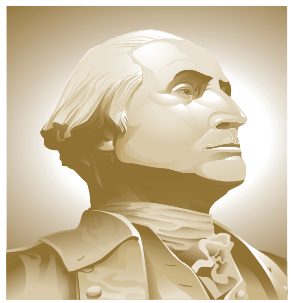
Introduction to Natural Language Processing

Natural Language Processing
Lecture 1-a



THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

What is natural language processing?

- The study of language and linguistic interactions from a computational perspective.
- Natural Language Processing (NLP) enable of us to develop following algorithms and models.

1- Natural Language Understanding (NLU)

2- Natural Language Generation (NLG)

Some Brief History

- 1940s and 1950s:
  - Markov Process and finite state machine for grammar.
- 1957–1970:
  - Symbolic and stochastic.
- 1970–1983
  - Stochastic paradigm, logic-based paradigm, natural language understanding and discourse modeling
- 1983–1993:
  - Finite state models
- 2000–2007:
  - Rise of Machine Learning

Symbolic Vs. Statistical

- **Symbolic Method**:
  - Many of early NLP methods involved symbolic methods. Basically these are rule based techniques.
  - The process of identify the meaning.
  - Symbolic process is not slow and does not perform well on known NLP tasks.
- **Statistical Method**:
  - Statistical methods inspired from speech recognition work and requires data.
  - It uses context and it performs pretty well on NLP task.
  - It handles a lot of variability seen in the language.

What is natural language understanding?

- Can computers or machines be intelligent?
- The following would be the examples of NLU.

1- IBM Watson
2- Alexa
3- Siri
4- Google search engine

Linguistic Analyses

- Levels of linguistic analyses
  - **Syntax**:
    - what is grammatical? *ProgrammingAnalogy* $\Rightarrow$ No Compiler Error
  - **Semantics**:
    - what does it mean? *ProgrammingAnalogy* $\Rightarrow$ No Bugs
  - **Pragmatics**:
    - what does it do? *ProgrammingAnalogy* $\Rightarrow$ No algorithmic error

NLP Used Areas

Search	Web Search	Documents Search	Autocomplete Words
Editing	Spelling Error	Grammar Correction	Style of Writing
Dialog	Chatbot	Speech Assistant	Question answering
Email	Spam	Classification	Ordering
Text mining	Summarization	Knowledge extraction	Topic modeling
News	Find Events	Fact checking	Headline detection
Sentiment analysis	Reviews	Product review	Customer care
Behavior prediction	Medical	Election forecasting	Business
Creative writing	Movie scripts	Transcriptions	Song lyrics

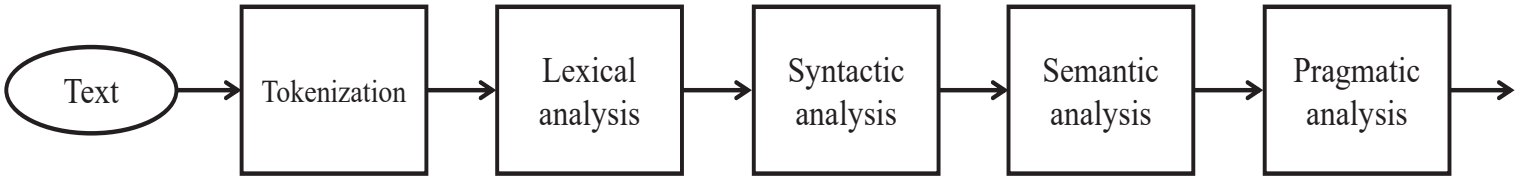
NLP Approaches with applications

- Classical
- Statistical
- Applications

Classical

The Classical Toolkit Process

- The process is :



- **Lexical analysis:**
  - is the process of converting a sequence of characters into a sequence of tokens or strings.
- **Syntactic analysis:**
  - is the analysis of basic unit of meaning like sentence.
- **Semantic analysis:**
  - is the process of relating syntactic structures, from the levels of phrases, clauses, sentences and paragraphs to the level of the writing as a whole.
- **Pragmatic analysis:**
  - refers to a set of linguistic and logical tools

Simple View

- The simple view is that

- the sentences of a text are first analyzed in terms of their syntax

- this provides an order and structure that is more amenable to an analysis in terms of semantics, or literal meaning

- then this is followed by a stage of pragmatic analysis whereby the meaning of the utterance or text in context is determined.

Tokenization

- The first stage of tokenization and sentence segmentation is a very crucial step.
- Text is generally not made up of the short, neat, well-formed, and well-delimited sentences.
- Languages such as Chinese, Japanese, or Thai, which do not share the apparently easy space-delimited tokenization.
- We may believe to be a property of languages like English but other langue can be very different.
- Lexical analysis would address the finer-grained decomposition of tokenization.

Challenges of Tokenization

- In both unsegmented and space-delimited languages, the specific challenges posed by tokenization are largely dependent on both thewriting system ((logographic, syllabic, or alphabetic).
- Tokenization in Space-Delimited Languages:
  - 76 cents a share, which on the surface consists of four tokens; a 76-cents-a-share dividend, it is normally hyphenated and appears as one.
  - $3.9 to $4million differently than if it had been written as 3.9 to 4 million dollars or $3,900,000 to $4,000,000.
- Tokenizing Punctuation:
  - Quotation marks and apostrophes (“ ” ‘ ’) are a major source of tokenization ambiguity
  - In English, ’s can serve as a contraction for the verb is, as in he’s, it’s, she’s, and Peter’s head and shoulders above the rest.
  - It also occurs in the plural form of some words, such as I.D.’s or 1980’s,
- Multi-Part Words
  - In French, for example, hyphenated compounds such as va-t-il (will it?), c’est-à-dire (that is to say), and celui-ci (it) need to be expanded during tokenization,

Challenges of Tokenization

Multiword Expressions

- For example, the three-word English expression in spite of is, for all intents and purposes, equivalent to the single word despite, and both could be treated as a single token.
- Dates and times, money expressions, and percents, can be treated as a single token.
- Several examples of such phrases can be seen: March 26, $3.9 to $4 million, and Sept. 24 could each be treated as a single token.
- For example, the English date November 18, 1989 could alternately appear in English texts as any number of variations, such as Nov. 18, 1989, 18 November 1989, 11/18/89 or 18/11/89.

Tokenization in Unsegmented Languages

- The nature of the tokenization task in unsegmented languages like Chinese, Japanese, and Thai is fundamentally different from English.
- The Chinese writing system consists of several thousand characters known as Hanzi, with a word consisting of one or more characters.
- Common unsegmented alphabetic and syllabic languages are Thai, Balinese, Javanese, and Khmer.
- Such writing systems have fewer characters than Chinese and Japanese, they also have longer words.

Text Preprocessing

- As we mentioned not all languages deliver text in the form of words neatly delimited by spaces.
- First we need to perform segmentation process.
- Identify the words that make up an utterance.
- What constitutes a word?
- Next we need to do sentence segmentation. Break the text into sentence-sized pieces.
- This turns out not to be so trivial either.

Challenges of Text Preprocessing

- Character-Set Dependence: Eight-bit character sets, UTF-8 variable-length character encoding
  - In this system, the German word über would be written as u”ber or ueber, and the French word déjà would be written as de’ja’ or de1ja2. Languages that do not use the roman alphabet, such as Russian and Arabic, required much more elaborate romanization systems
- Character Encoding Identification and Its Impact on Tokenization
  - Tokenization rules would then be required to handle each symbol (¿ ¡ Ł ç)
- Language Identification
  - For languages with a unique alphabet not used by any other languages, such as Greek or Hebrew, language identification is determined by character set identification
- Application Dependence
  - In languages such as Chinese, which do not contain white space between any words, a wide range of word segmentation conventions are currently in use.

Sentence Boundary Punctuation

- In NLP, sentence boundary punctuation marks considered are the period, question mark, and exclamation point, and the definition of sentence is limited to the text sentence which begins with a capital letter and ends in a full stop.
- Grammatical sentences can be delimited by many other punctuation marks.

  - Here is a sentence. Here is another.
  - Here is a sentence; here is another.
  - There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, 'Oh dear! Oh dear! I shall be late!' (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually TOOK AWATCHOUT OF ITS WAISTCOATPOCKET, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge.
  - 'Oh dear! Oh dear! I shall be late!'

- The Importance of Context: Many contextual factors have been shown to assist sentence segmentation in difficult cases. These contextual factors include

  - Case distinctions: In languages and corpora where both uppercase and lowercase letters are consistently.
  - Part of speech: Parts of speech of the words within three tokens of the punctuation mark can assist in sentence segmentation.
  - Word length: Used the length of the words before and after a period as one contextual feature.
  - Lexical endings: Used morphological analysis to recognize suffixes and thereby filter out words which were not likely to be abbreviations.
  - Prefixes and suffixes: Used both prefixes and suffixes of the words surrounding the punctuation mark as one contextual feature.
  - Abbreviation classes: divided abbreviations into categories such as titles and corporate designators.
  - Internal punctuation: Used the presence of periods within a token as a feature.

Lexical Analysis

- The words, of course, are not atomic, and are themselves open to further analysis.
- By taking words apart, we can uncover information that will be useful at later stages of processing.
- This mean that decomposing words into their parts, maintaining rules for how combinations are formed.
- This is much more efficient in terms of storage space than would be the case if we simply listed every word as an atomic element in a huge inventory.
- Also, there would be words always missing from any such inventory.

Syntactic Parsing

- The basic unit of meaning analysis is the sentence.
- A sentence expresses a proposition, an idea, or a thought, and says something about some real or imaginary world.
- Extracting the meaning from a sentence is thus a key issue.
- Sentences are not, however, just linear sequences of words.
- Analysis of each sentence, which determines its structure in one way or another.
- In NLP approaches based involve the determining of the syntactic or grammatical structure of each sentence.

Challenges Syntactic Parsing

- One of the strongest cases for expressive power has been the occurrence of long-distance dependencies, as in English wh-questions: There is no clear limit as to how much material may be embedded between the two ends.
  - Who did you sell the car to -?
  - Who do you think that you sold the car to -?
  - Who do you think that he suspects that you sold the car to -?
  - Oh dear! Oh dear! I shall be late!
- Second difference concerns the extreme structural ambiguity of natural language.
  - Put the block in the box on the table
  - Put the block [in the box on the table]
  - Put [the block in the box] on the table
- If we add another prepositional phrase (“in the kitchen”), we get five analyses; if we add yet another, we get 14, and so on.

Semantic Analysis

- Identifying the underlying syntactic structure of a sequence of words is only one step in determining the meaning of a sentence.
- Meaning of a sentence provides a structured object that is more amenable to further manipulation.
- It is these subsequent steps that derive a meaning for the sentence in question.
- Mostly, the semantics of natural language have been less studied than syntactic issues.

Challenges Semantic Analysis

- Logical Approaches:
  - Some politicians are mortal. [There is an x (at least one) so that x is a politician and x is mortal.]
  - All Australian students like Kevin Rudd. [For all x with x being a student and Australian, x likes Kevin Rudd.]
  - Conrad is tired. Whenever Conrad is tired, he sleeps. Conrad sleeps.
- Discourse Representation Theory
  - A man sleeps. He snores.
  - man (x)
  - sleep (x)
  - y=x
  - snore (y)
- In the DRT representation, the quantification in the first sentence results in an if-then condition: if x is a man, then x sleeps. This condition is expressed through a conditional ($A \rightarrow B$) involving two DRSs.

Natural Language Generation

- Determining the meaning of an utterance is only really one-half of the story of natural language processing.
- In many situations, a response then needs to be generated.
- For many of today's applications, what is required here is rather trivial and can be handled by means of canned responses.
- Increasingly, however, we are seeing natural language generation techniques applied in the context of more sophisticated back-end systems.
- The need to be able to custom-create fluent multi-sentential texts on demand becomes a priority.

Statistical

Approaches

- **Corpus Creation**:
  - Corpus Size, Balance, Representativeness, Sampling, Data Capture, Copyright and Multilingual Corpora
- **Treebank Annotation**:
  - Morphosyntactic Annotation, The Penn Treebank, Semantic, and Discourse Annotation
- **Fundamental Statistical Techniques**:
  - Binary Linear Classification, Multi-Category Classification, MaximumLikelihood Estimation, Generative and Discriminative models , Naive Bayes and Logistic Regression,Hidden Markov Model
- **Part-of-Speech Tagging**:
  - Parts of Speech, Rule-Based Approaches, Markov Model Approaches, Maximum Entropy Approaches, POS Tagging in Languages Other Than English
- **Statistical Parsing**:
  - Probabilistic Context-Free Grammars
- **Multiword Expressions**:
  - Idiomaticity, Collocations, A Word on Terminology and Related Fields

Applications

Approaches

- **Chinese Machine Translation**
- **Information Retrieval**
- **Question Answering**
- **Information Extraction**
- **Report Generation**
- **Sentiment Analysis and Subjectivity**
- **Text Mining**
- **Language Models**
- **Text Classification and Text Clustering**