Natural Language Processing

Natural Language Processing is a subfield of linguistics, computer science and artificial intelligence, it is concerned with the interactions between machines and natural language

NLP is concerned with the ability of machines to analyze, understand and process natural language, hence deriving meaning to it. NLP involves following tasks

>> Summarization
>> Translation
>> Entity Recognition
>> Relationship Extraction
>> Sentiment Analysis
>> Speech Recognition
>> Topic Segmentation

Natural Language Processing is characterized as a difficult problem; human language/natural language is rarely precise. Hence to understand natural language is to understand not only the words, but the concepts and how they're linked together to create meaning. Despite language being one of the easiest things for the human mind to learn, the ambiguity of language is what makes the NLP a difficult problem to master.

NLP combines computational linguistics-rule-based modeling of natural language -- with statistical, machine learning and deep learning models. These models enable machines to process natural language in the form of text or voice data and to 'Understand' its meaning, complete with the speaker or writer's intent and sentiment.

Challenges in NLP : Ambiguities

- Lexical Ambiguity (Semantic ambiguity or Homonymy) : Lexical ambiguity is the presence of two or more possible meanings for a single word. The lexicon contains entries that are homophonous or even co-spelled, but differ in meanings and even syntactic categories. Example, 'Duck' is both a verb and a noun as in cover.
- This sort of ambiguity is often very easy to detect by simple linguistic reflection, especially when the meanings are widely distinct such as in the case of 'Bat'
- Lexical ambiguity is sometimes used deliberately to create puns and other types of wordplay
- According to the editors of the MIT Encyclopedia of the Cognitive Sciences "True Lexical ambiguity is typically distinguished from polysemy or from vagueness, though the boundaries can be fuzzy."

  Example: **Will Will will Will's will?**
  Here, Will -> Modal verb(like Should/Would/Can)
  Will -> Name of a person
  will -> Verb
  Will's -> Name of a person (Second person)
  will? -> Noun

  Rose rose to put rose roes on her rows of roses
  Here, Rose -> Noun (Name of the person)
  rose -> Verb (Action)
  rose -> Adjective
  roes -> Noun (Seafood)
  rows -> queue
  roses -> Noun (Name of a flower)

  Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo
  City Animal City Animal Verb Verb City Animal
  Here the word "Buffalo" is used in three different sense
  Buffalo - City in the U. S.
  buffalo - Animal
  Buffalo - bully(Verb)

The meaning of the sentences comes out to be “Buffaloes from Buffalo, New York, whom buffaloes from Buffalo bully, bully buffaloes from Buffalo.”

Sometimes co-spelled words are distinct sounding, such as ‘refuse’ which is distinct sounding (in my dialect at least) between ‘ree-fuze’ and ‘reh-fuse’. Fortunately, we have the relevant categories to describe these differences and we can talk about ambiguity in sound or in notation (or in sign)

Structural ambiguity or Syntactic ambiguity : Structural ambiguity is the potential of multiple interpretations for a piece of written or spoken language because of the way words or phrases are organized i.e. the sentences are ambiguous because of the organisation of the words or phrases. Some structural ambiguity is the result of writing errors, such as misplaced modifiers.

- The man saw the boy with the binoculars : This sentences interprets different meanings i.e. The man saw the boy who has binoculars or the man with the binoculars saw the boy.
- Flying planes can be dangers : Here, the sentences can be interpreted as, Flying a plane is dangerous or Flying plane itself is dangerous
- Hole found in the room wall; police are looking into it : The last part of the sentence is found to be ambiguous , What is police looking into? Hole or something else

Language imprecision and vagueness :

Examples of Ambiguity

1. Hospitals are sued by **7 foot Doctors** :
   Ambiguity : 7 foot Doctors
   Actual meaning : The Hospitals are sued by 7 “Foot Doctors”

2. Stolen painting **found by Tree** :
   Ambiguity : found by Tree
   Actual meaning : Stolen paintings were found near the tree.

3. Teacher Strikes Idle Kids :
   Ambiguity : Teacher Strikes Idle Kids
   Actual meaning : Due to teachers strike, kids are Idle

4. I made her Duck
   Possible interpretations : covered her; Duck refers to covering
   cooked Duck for her (Dative pronoun, Ditransitive)
   cooked a duck of hers (Possessive pronoun, Ditransitive)
   made her like Duck
   made a duck for her
   made her lower her head
   Type of Ambiguity : Syntactic ambiguity
   Explanation : The word ‘Duck’ here can be a noun or verb and ‘her’ can be a possessive(of her) or dative(for her) pronoun. ‘Make’ can mean ‘create’ or ‘cook. ‘Make’ can be Transitive or Ditransitive or Action-Transitive

Empirical Laws

Function Words and Content words

A particular language and its vocabulary is made up of some kinds of words called function words and content words.Content words are the words including information and meaning while Function words are the words consisting of necessary words for grammar. In other words, content words give us the most important information while function words are used to stitch those words together.

Content words are usually nouns, verbs, adjectives and adverbs. As these gives us some information about the following

Noun = person, place or thing
Verb = action, state
Adjective = describes an object, person, place or thing
Adverb = tells us how, where or when something happens

Function words help us connect information, function words are important for understanding, but they add little meaning beyond defining the relationship between two words. Function words include auxiliary verbs, prepositions, articles, conjunctions, and pronouns. Auxiliary verbs are used to establish the tense, prepositions show relationships in time and space, articles show us something that is specific or one of many, and pronouns refer to other nouns.

Auxiliary verbs = do, be, have (help with conjugation of tense)
Prepositions = show relationships in time and space
Articles = used to indicate specific or non-specific nouns
Conjunctions = words that connect
Pronouns = refer to other nouns

Lets see some examples

1. Mary has lived in England for ten years
   Here,
   Mary has lived in England for ten years

Noun : Mary, England
Verb : lived
Adverbs : ten years
Auxiliary verbs : has
Preposition : in, for

2. He's going to fly to Chicago next week.
   Here,
   He's going to fly to Chicago next week.

Noun : Chicago
Verb : fly
Adverbs : next week

Preposition : to

Corpus of Tom Sawyer

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal in□nitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Type and Token

In the study of texts; the ratio of the number of different words, called types while the total number of words in a certain text is called token.The distinction between a *type* and its *tokens* is a useful metaphysical distinction. Type is considered to be concept and takens are instances of concept

Will will will -----> { Type = 1, Token = 3

(Unique words) (Total number of words)

Type-Token distinction

The distinction between a *type* and its *tokens* is an ontological one between a general sort of thing and its particular concrete instances (to put it in an intuitive and preliminary way). The **type–token distinction** is the difference between naming a *class* (type) of objects and naming the individual *instances* (tokens) of that class.

Type/Token Ratio

The type/token ratio (TTR) is the ratio of the number of different words(types) to the number of running words(tokens) in a given text. This index indicates how often, on average, a new ‘word form’ appears in the text or corpus.

Case : Mark Twain’s Tom Sawyer
Tokens : 71370
Type : 8018
TTR : Type/Token = 8010/71370 = 0.112

Complete Shakespeare work
Token : 884647
Type : 29066
TTR : 0.032

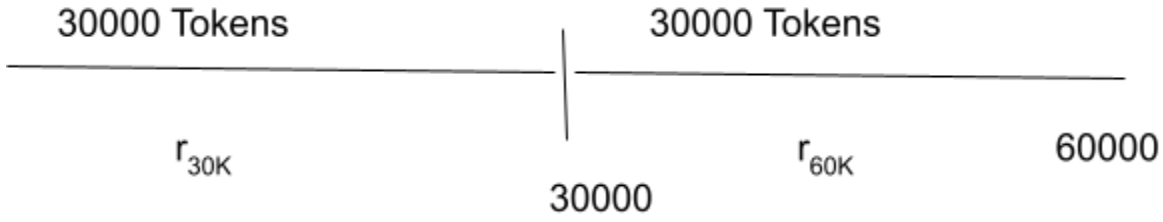
High TTR → Tendency to use new words
Low TTR → Tendency to use less new words (repetition of words)

Comparing Conversations, Academics, News, Fiction
Conversations → Lowest TTR … as we tend to repeat words in normal conversation
Fiction → Low TTR
Academic → High
News → Highest

News > Academic > Fiction > Conversations

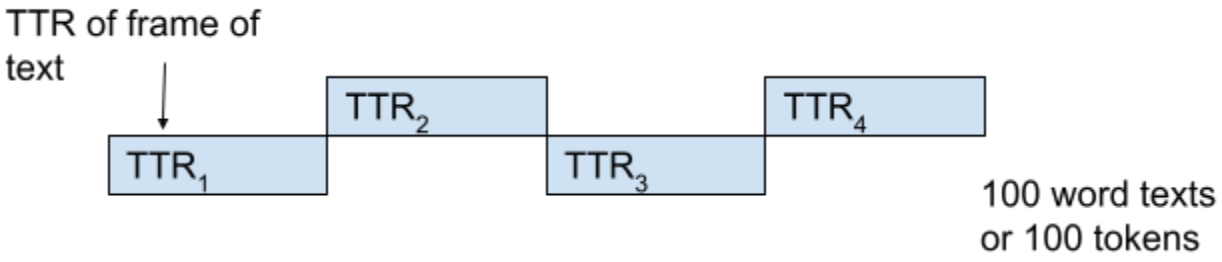
**TTR in not a valid measure of text complexity as the values varies with size of the text and for valid measure, a running average is computed on consecutive 1000-words chucks of the text.

Consider a text, containing 60000 tokens. The TTR for the first 30000 tokens is equal to r_{30K} and for other 30000 tokens is r_{60K}. Therefore, we can imply that $r_{60K} > r_{30K}$. Since most of the words are considered to be covered in the first 30000 tokens.



Calculation of Running TTR

Suppose we have a text with a total N words/tokens, in order to calculate an accurate Type-token ratio we calculate the TTR of a specific frame and then calculate the TTR of the exact next frame then take an average of the calculated TTR.



$$\text{Running TTR} = \frac{TTR_1 + TTR_2 + TTR_3 + TTR_4}{\text{No. of frames}} = \frac{TTR_1 + TTR_2 + TTR_3 + TTR_4}{4}$$

Zipf's Law