2022

NLP Dictionary

DS7337 NATURAL LANGUAGE PROCESSING

GOODWIN, BEN

Table of Contents

Unit 1 – Introduction to NLP

Natural Language Processing – A subfield of linguistics and computer science, with the objective of studying and understanding how to process interactions of human language by computers.

Natural Language – Languages developed over time, without a purpose in mind
Examples: English, French, Thai, Chinese
Artificial Language – Language developed with a specific purpose
Examples: Python, Klingon

NLP = NLU + NLG

Natural Language Understanding – Subfield within NLP that deals with machine reading comprehension. Considered an AI-hard problem.

Examples: Spell-checking, auto complete, and iPhone word suggestions

Natural Language Generation – A software process that produces natural language output. Generally, the output is text, but there is some debate as to whether the inputs of an NLG system need to be non-linguistic.

Examples: Chat bots

Unit 2 – Levels of Analysis in NLP

Lexical Analysis – This is the process of converting a sequence of characters into a sequence of tokens.

Syntactic Analysis – Tells us that logical meaning of certain given sentences or parts of those sentences.

Example: “The dog went away” “The dog” (noun phrase) “Went away” (Verb phrase)

Semantic Analysis – This is the process of drawing meaning from text. More formally, this is the process of relating syntactic structures from the levels of phrases, clauses, sentences, and paragraphs to the level of the writing to their language independent meanings.

Color is a hypernymy while “blue” “red” “yellow” are color’s hyponyms.

Discourse Analysis – An approach to the analysis of written, vocal, or sign language use, or any significant semiotic event. The objects of discourse analysis are variously defined in terms of coherent sequences of sentences.

Example: One may study the purposes and effects of different types of languages

Vocabulary - Number of words in a text or corpus

Vocabulary Size – Number of distinct words in a text

N-Gram – Grouping words around some whole number N

Example: Bigrams contain 2-word phrases, and a 3-gram would be 3-word phrases

Unit 3 – Trade-Offs in NLP

Shallow vs. Deep – At a high level this could refer to the computational resources available as these two terms are related to the amount of computation required for respective activities.

Deep: In the semantic sense this method looks at each word and its associated nuances. In the deep parsing sense, this breaks down all phrases and POS tags them.

Shallow: In the semantic sense this method looks at the document or text at a high level. This method typically breaks documents down at the chunk level.

Statistical vs. Symbolic – These two approaches differ in their approach to NLP. Currently statistical methods are favored, but in the past logical systems have proven effective.

Statistical: This approach is more math based and uses a handful of statistical approaches to perform NLP.

Example: Bayesian inference on words in corpus.

Symbolic: This approach is rule-based that applies criteria to inputs. The rules are clearly listed and explainable to anyone with basic familiarity with logic. The downside of this approach is that rules must be created for any possible situation, and situations can almost always be invented where the rules will fail.

Example: WordNet or POS Tagging

Feature Engineering vs. Feature Learning: In feature engineering, humans are involved in the process, defining, and developing features. Feature learning on the other hand follows the statistical approach where the features are learned based on a variety of statistical approaches.

Top-down vs. bottom up: In a top-down method, we start with a high level of classification and begin the process of slowly breaking down the details. In a bottom-up method, we start with the details and work our way up to the highest level of classification.

Transparent vs Opaque (AI vs. XAI): The difference between these terms relates to Artificial Intelligence versus explainable Artificial Intelligence. Transparent refers to the latter, and opaque, the former. Some AI methods are quite difficult to explain and get bucketed into “black boxes.”

Unit 4 – Working in NLP

NLP and Data Science: Natural Language processing is one of the most popular subfields in data science and is essential in progressing the field forward. Natural Language Processing is important because it helps to resolve ambiguity in language and adds useful numeric data for many downstream applications. Think of examples like chat bots or machine translation.

Job roles that utilize NLP: There is a laundry list of jobs that involve Natural Language processing some of those include: Email filters, Smart assistants (think Apple’s Siri, Amazon’s Alexa, or Google’s Assistant), predictive text, language translation, and text analytics.

Sectors that utilize NLP: Many different sectors utilize Natural Language Processing like healthcare, manufacturing, translation, and customer service.

Organizations that relate to NLP: Many businesses and organizations rely on NLP and may not be aware of it. Among these organizations: Google, InData Labs, Amazon, and IBM

Unit 5 – Low-Level Analysis

Text preprocessing: At a high-level text preprocessing in NLP is a method to clean the text data and make it ready to feed data to the model. This is a hugely important step in NLP as data in the wild is noisy and difficult for a computer to process. NLTK (Natural Language Tool Kit) can be used in this step for tasks like removing stop words, NEI (Named entity recognition), POS tagging (Part of speech tagging), and a handful of other useful tasks.

TF-IDF: Term frequency – inverse document frequency. This is a method of differential frequency analysis. This method looks at the frequency of the term in question and divides that by the phrases in the total document and determines its overall importance in each document.

Relative lexical diversity: Formula: Number of unique tokens/total number of tokens

Reading level: Examines syllable count, sentence length, and words relative to a compiled list.

Text Normalization: At a high-level text normalization is an attempt to reduce the randomness of the text. The goal is to reduce the amount of different information the computer must deal with. There are many details contained within this topic.

Stop words: A list of common words that serve no point in distinguishing between texts. These words are commonly removed and often are “a, an, the, on, of.”

Content and function words: Function words are best defined as words that do not contribute to the understanding of the text but serve purpose to the reader by defining answers to the subject. Content words are useful for determining meaning from a text.

Stemming: A stemmer reduces each word to its stem

Example: Running -> Run

Note: Stemmers don’t always produce a grammatically correct stem

Low-level document feature extraction: This is defined as the process of removing the smallest level of detail from a document such as stems.

Unit 6 – Lexical Knowledge Bases

Lexical knowledge bases: A repository of computational information about concepts intended to be generally useful in many application areas including computational linguistics (NLP). It contains information derived from machine-readable dictionaries, reference books, and results of statistical analyses of text usages. They’re to serve as a general repository of knowledge about lexical concepts.

Hyponym: These are generally subdivisions of more general words.

Example: Boston Terrier. Boston Terrier is a specific breed of dog.

Hypernym: A word meaning that encapsulates the meanings of other words. These are general words.

Example: Dog is more general towards an animal.

Meronym: A word that denotes a constituent part or a member of something.

Example: Peach tree

Resources for creating or extending lexical knowledge bases: There are a handful of extremely powerful lexical knowledge bases out there. IEEE contains many published articles on creating and extending lexical knowledge bases.

Applications of lexical knowledge bases: NLP is heavily reliant on these artifacts; they encode human knowledge of language in machine readable form.

WordNet: A series of lexical databases derived from the original Princeton WordNet which groups words into synonym sets and connects them using lexical and conceptual-semantic relations.

Unit 7 – Syntactic Analysis: POS-Tagging

POS-tagging: This task looks at text and determines the part of speech associated with it. A typical POS tagger will identify the different parts of a text or corpus and tag them (noun, adjective, verb). To implement POS tagging, investigate something like WordNet (Unit 6).

Users can also utilize the Penn Treebank which is defined at the bottom of this document.

Using POS tags: The Penn Treebank has greater than 90% accuracy, and most human experts only agree about 96% of the time, so this is still an unsolved problem. How can we teach computers if we don't even agree?

Unit 8 – Syntactic Analysis: Parsing

Parsing – The process of analyzing a string of symbols, either in natural language, computer languages, or data structures, conforming to the rules of grammar.

Shallow parsing: Shallow parsing in an analysis of a sentence which first identifies constituent parts of sentences and then links them to higher order units that have discrete grammatical meanings.

Chunks: Chunking is defined as the process of natural language processing used to identify parts of speech and short phrases present in each sentence. Chunks are what we're after in NLP (ex. Adjectives or noun phrases), whereas chinks are the unimportant parts.

Full grammar parsing: This is considered a complex and computationally intensive task to develop a full grammar tree.

Example: Useful for machine translation

Uses for full parse trees: Useful for machine translation

Penn Treebank Parts of Speech Tags

1) CC Coordinating conjunction – joins phrases
    (1) Ex. and, but, or
2) CD Cardinal number – Numbers spelled out
    (1) Ex. one, two, three, etc.
3) DT Determiner – Decides or selects something
    (1) Ex. a, the, every

4) EX Existential *there* – used to mark the existence of something.
    (1) Ex. There were many people at someplace
5) FW Foreign word - Word not in OED and some abbreviations
    (1) Ex. E.g., etc., i.e., foreign language words Latin, urban dictionary, etc…
6) IN Preposition or subordinating conjunction - A preposition usually indicates the temporal, spatial or logical relationship of its object to the rest of the sentence
    (1) Ex. Onto, since, above, along, and many more
7) JJ Adjective – Describes a noun
    (1) Red, large, old, good
8) JJR Adjective, comparative – an adjective which denotes some sort of comparison
    (1) Ex. More red, Larger, older, better
9) JJS Adjective, superlative – the "most" and the "-est"
    (1) Ex. Largest, Oldest, most red, best
10) LS List item marker –
    (1) Ex. These lists, 1., a., bullet points
11) MD Modal – indicate present operative and normative states
    (1) Ex. Shall Should, can could, may might, etc.
12) NN Noun, singular or mass – common person, place, thing, idea
    (1) Ex. Girl, car, tree
13) NNS Noun, plural – plural forms of NN
    (1) Ex. Girls, Cars, Trees
14) NNP Proper noun, singular – Specific names of NN
    (1) Jane, Thunderbird, Yggdrasil
15) NNPS Proper noun, plural – plural NNP
    (1) Cardinals (Team), Smiths (refers to lots of Smith)
16) PDT Predeterminer – Central, pre and Post determiners indicate things about the determiners
    (1) Ex. My, this, all, both, many, several
17) POS Possessive ending - Indicate ownership
    (1) Ex. He's, boy's, Mary's
18) PRP Personal pronoun – self references
    (1) Ex. I, me, we, us, our
19) PRP$ Possessive pronoun – possessive versions of PRP
    (1) Ex. Mine, ours, yours
20) RB Adverb – modifies verbs like adjectives with nouns
    (1) Ex. Slowly, never, rarely,
21) RBR Adverb, comparative – indicates some intensity order to the verbs
    (1) More slowly, more happily,
22) RBS Adverb, superlative – the most
    (1) Ex. Most slowly, most happily
23) RP Particle – unaccompanied prepositions
    (1) Ex. Along, away, off
24) SYM Symbol – non-alpha characters
    (1) Ex. &, -,
25) TO *to* – self explanatory

26) UH Interjection - a word used to convey emotion
    (1) Ex. Congratulations, peace, thanks, whoa
27) VB Verb, base form – action words
    (1) Ex. Run, swim, drive
28) VBD Verb, past tense – actions that happened in the past
    (1) Ex. Ran, swam, drove
29) VBG Verb, gerund, or present participle – indicate present states
    (1) Ex. Running, driving, dying
30) VBN Verb, past participle – past states of being
    (1) Ex. Studied, been living
31) VBP Verb, non-3rd person singular present – same as verb but convey habits, routines, thoughts
    (1) Ex. Same as verb forms run, swim, etc.
32) VBZ Verb, 3rd person singular present
    (1) Ex. Argues, Applies, tries
33) WDT Wh-determiner – determines selections not at beginning of sentence
    (1) Which, what, whatever – I don't know **what** you are talking about
34) WP Wh-pronoun – same as above but at beginning of sentence
    (1) **Which** bowl of soup
35) WP$ Possessive wh-pronoun – same as above but possessive
    (1) Whose, whom
36) WRB Wh-adverb – the events
    (1) When, where, why, how