**MSDS-7337-Natural Language Processing-Glossary**

Rashmi Patel1

1Master of Science in Data Science

Southern Methodist University

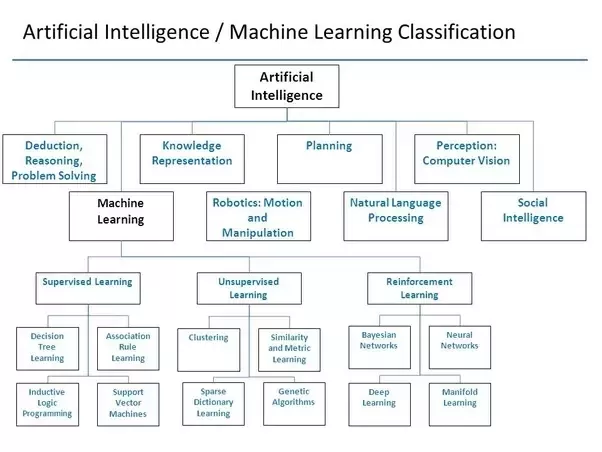
Dallas, Texas, USA

[rashmip@smu.edu](mailto:rashmip@smu.edu)

**Abstract: This document contains the important terms and concepts that are used and implemented in Natural Language Processing (NLP).**

**Unit1: Introduction**

**What is Natural Language Processing (NLP)?**



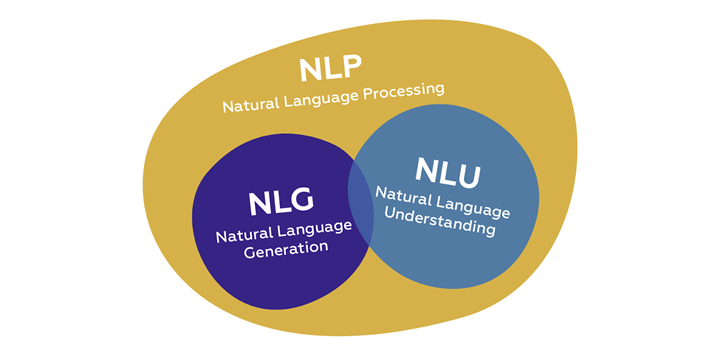
Natural Language Processing (NLP), a branch of artificial intelligence (AI), is nowadays most frequently used technology which concerns about how the computer/ machine can understand text and spoken words like humans. Natural Language Processing (NLP) is a combination of computational linguistics, machine learning and deep learning modelling techniques.

Natural Language vs. Artificial Language:

Natural languages are those that evolved or emerged gradually over time, largely unconsciously. Examples of Natural Languages are: English, Spanish, Japanese, sign language, etc.

Artificial Language are those that were designed, crafted, or invented with conscious purpose, largely all at once and not gradually. Examples of Artificial Languages are: LISP, C, C++, Python, etc

**Two sides of NLP:**



There are two sides of Natural Language Processing (NLP):

1. Natural Language Understanding (NLU)
2. Natural Language Generation (NLG)

Natural Language Understanding (NLU): The branch of NLP that uses computer software to understand input made in the form of natural language and produce a useful representation of some inputted natural language.

Natural Language Generation (NLG): The branch of NLP that uses computer software to generate usable, natural language output that is not just identical to its input.

**Applications of NLP:**

As we know that there are 2 sides of NLP, we will be looking at applications of both NLU and NLG.

Applications of NLU:

1. Text annotation
   1. Tagging
   2. Metadata extraction/generation
   3. Classification
   4. Document summarization
2. Corpus analytics
   1. Theme extraction
   2. Clustering
   3. Taxonomy mapping
   4. Sentiment analysis
3. Search applications
   1. Query repair
   2. Query refinement
   3. Results postprocessing
4. Advanced applications
   1. Machine translation
   2. Knowledge discovery
   3. Question handling
      1. Question typing/routing
      2. Question-FAQ approximate matching

Applications of NLG:

1. Text annotation
   1. Document summarization
   2. Generation of callouts/headlines
2. Corpus analytics
   1. Labelling of clusters
   2. Synopsizing of corpus-wide topic and/or sentiment trends
3. Search applications
   1. Advanced capsule generation (summarization modified to fit the query)
   2. Advanced query refinement (next-gen version of “did you mean?” for disambiguation of query’s meaning)
4. Advanced applications
   1. Machine translation (construction in target language)
   2. Knowledge discovery (human-friendly presentation)
   3. Question handling
      1. Question refinement
      2. Question answering

**Unit2: Levels of Analysis in NLP**

**A picture containing timeline

Description automatically generated**

**Lexical analysis:**

1. Fundamental foundation of all analyses that analyze words.
2. Analysis goes up and down through these levels to help correct errors and fill-in gaps for all other levels of analysis.

Terminology and Concepts in Lexical Analysis:

1. **Lexicon** - dictionary, vocabulary
2. **Morphology**-study of morphemes-units of which words are made
3. **Stemmer** - algorithm that provides root morpheme
4. **Metadata** - corpus derived meta-data supports developing : word frequency score, common collocation, and commonly co-occurring words and context words
5. **Collocations** - words commonly occurring together
6. **Head word** - specific listing in the lexicon
7. **Polysemous** - more than one meaning
8. **Text and Corpus Analytics** - includes spell correction, terminology extraction, lexical diversity measurement
9. **WordNet** - the Bible of word senses

**Syntactic analysis:**

* + 1. Grammar like analysis. Putting full sentences through a program that can draw structure of grammar.
    2. Analysis goes up and down through these levels to help correct errors and fill-in gaps for all other levels of analysis.
    3. Second level of NLP that focuses on grammatical analysis of documents

Terminology and Concepts in Syntactic Analysis:

1. **Sentence boundary detection** - needed since syntax analysis is sentence by sentence analysis. grammar parsing relies on sentence boundary detection. non-trivial problem.
2. **Part of speech (POS)** - each word in a sentence can be assigned to a part of speech. most words have more than one part of speech; therefore, POS tagging is non-trivial. needs contextual clues - words preceding and succeeding,
3. **Penn Treebank Tag set** - commonly used tag sets for parts of speech, approximately 40 different POS with this set, support eventual grammar parsing
4. **Parsing** - break the sentence into grammar parts
5. **Lemmatization** - canonical (conventional) form that represents a set of related word forms
6. **Discrete text field analysis** - unitizing, normalizing, smart ETL.

**Semantic analysis:**

1. Entire meanings based on usage in a full like sentence way. 90% of data science projects stop here.
2. Analysis goes up and down through these levels to help correct errors.

Terminology and Concepts in Semantic Analysis:

1. **Named entity extraction (NEE a.k.a. NER)** - recognizes the entities (persons organizations, places, events) without typing.
2. **Relationship extraction (between NEs)** - What relation does one NE have to another? Sometimes we can answer this during the NEE process.
3. **Word sense disambiguation (WSD)** - still unsolved problem in AI - how to distinguish the meaning of a word in context, from all of the possible meanings of the word
4. **Classification** - we use a tree-structured graph to place documents into categories. Often, we employ machine learning methods, such as SVM (support vector machine).
5. Additional types of semantic analysis include:
   1. Tagging
   2. Topic segmentation
   3. Sentiment analysis

**Discourse analysis:**

1. Entire meanings that can be inferred based on another semantic meaning
2. Nuances of human conversation like metaphorical, ironic, comical, etc.

Terminology and Concepts in Discourse Analysis:

1. **Anaphora resolution** - the problem of resolving what a pronoun, or a noun phrase refers to.
2. **Discourse modelling** - A discourse model is a mental object that constitutes an. individual's knowledge of a discourse. It is constructed based on what has occurred in the discourse supplemented by general and specific knowledge.
3. **Question answering** – Most straightforward approach is matching pre-existing Q-A materials. The matching is often not straightforward. We can build models that adapt question formats.
4. **Textual entailment** -Inferences generated follows a strict logic. In textual entailment framework, the entailing and entailed texts are termed text (t) and hypothesis (h), respectively.
5. **Pragmatic analysis** – Inferences generated does not follow strict logic.

The below table summarizes typical levels of analysis employed in NLP. Generally, we use lower levels also when we move to “higher” levels of analysis.

Table

Description automatically generated

**Unit3: Trade-Offs in NLP**

**Shallow vs. Deep:**

1. When there is a need to answer a question like- Shall I build a very robust, exhaustive representation of text meaning, or shall I just pick out a few elements needed for my application? then the best approach is Shallow vs. Deep.
2. Shallow NLP works on breaking sentences into chunks whereas deep NLP consists of 3 or more hidden layers.

**Statistical vs. Symbolic:**

1. When there is a need to answer a question like-Shall I leverage powerful, complex statistical methods, or use rules that can be operated on by strict logic? then the best approach is Statistical vs. Symbolic.
2. In symbolic NLP we have a rule-based system whereby discrete tests are performed to apply each rule whereas a combination of signals or nodes of varying strengths is used in statistical NLP.

**Feature engineering vs. Feature learning:**

1. When there is a need to answer a question like- Shall I involve human experts to engineer features, or shall I use them only to establish a training set? the best approach is Feature Engineering vs. Feature learning.
2. In feature learning, deficiencies in precision or recall are addressed by expanding the training set whereas, deficiencies in precision or recall are addressed by examining counterexamples with the SMEs in feature engineering.

**Top-down vs. bottom- up:**

1. When there is a need to answer a question like- Shall I start with high-level classifications and text characteristics, then gradually break them down into more detail, or shall I do the reverse? then the best approach is Top-down vs. Bottom-Up.
2. Top-down approach involves building your NLP framework around a set of high-level concepts and categories into which everything fits whereas bottom-up approach involves building up a huge list of stemmed keywords from raw text, then try to roll them up into phrases, tag clouds, hypernym trees, and the like.
3. Non-technical people prefer top-down approach whereas technical people use bottom-up approach as this sets you up to have rich context information for every feature of each document.

**Transparent vs. opaque (AI vs. XAI):**

1. When there is a need to answer a question like- Shall I engineer a system that is explainable to an intelligent SME or is it okay to build something only my fellow data scientists and engineers could understand? then the best approach is Transparent vs. Opaque.
2. Transparent means it is easy enough to see what the algorithm is doing whereas opaque is result of more complex machine learning and produces results not easily explainable

**Unit4: Working in NLP**

**Job roles that utilize NLP:**

1. Software engineer
2. Knowledge engineer
3. Data scientist
4. DBA
5. Applied linguistics researcher
6. Cognitive scientist
7. Marketing technologist

**Sectors that utilize NLP:**

Public and private sectors that use NLP include:

1. Information services
2. eCommerce
3. Customer service desks
4. Law enforcement or military
5. Legal
6. Business intelligence
7. Consumer devices
8. Embedded technologies
9. Publishing
10. Research institutes

**Organizations that relate to NLP:**

Associations supporting NLP include:

1. ACM- Association for Computing Machinery
2. IEEE-Institute of Electrical and Electronics Engineers (IEEE Computer Society)
3. AAAI- Association for the Advancement of Artificial Intelligence
4. IJCAI- International Joint Conference on Artificial Intelligence Organization
5. AAAL- American Association for Applied Linguistics
6. ICLA- International Cognitive Linguistics Association

**Unit5: Low-Level Analysis**

**Text pre-processing:**

**To perform this step we need to do following:**

1. Sentence segmentation (sentence tokenization)
2. Lexical analysis (word tokenization)

**Diagram

Description automatically generated**

**Text normalization:**

To normalize a word-tokenized text source, we often address the following: •

1. Contractions, expansion
2. Stop words, removal
   1. **Content words**: Content words are an open class, meaning they are an unfinished list to which new words are readily added. Examples are dog, cat, angrily, run, clever, democracy, green, God, wisdom.
   2. **Function words**: are a closed class, meaning they are a fixed list to which no new words are added. Examples are is, am, are, was, were, he, she, you, we, they, if, then, therefore, possibly.
   3. **Stop words:**
3. Misspellings, resolving of
4. Stemming, if called for

**Low-level document feature extraction:**

**Unit6: Lexical Knowledge Bases**

Lexical knowledge bases

Resources for creating or extending lexical knowledge bases

Applications of lexical knowledge bases

**Unit7: Syntactic Analysis: POS Tagging**

POS-tagging

Using POS tags

**Unit8: Syntactic Analysis: Parsing**

Shallow parsing Using chunks

Full grammar parsing

Uses for full parse trees

**Unit9:**

**Unit10: Semantic Analysis: Semantic Relatedness**

**Unit11: Semantic Analysis: Document Clustering**

**Unit12: Semantic Analysis: Text Classification**

**Unit13: Semantic Analysis: Topic Modelling**

**Unit14: Semantic Analysis: Sentiment and Rhetoric**