**Natural Language Processing**

The term Natural Language Processing encompasses a broad set of techniques for automated generation, manipulation and analysis of natural or human languages. Although most NLP techniques inherit largely from Linguistics and Artificial Intelligence, they are also influenced by relatively newer areas such as Machine Learning, Computational Statistics and Cognitive Science. Before we see some examples of NLP techniques, it will be useful to introduce some very basic terminology. Please note that as a side effect of keeping things simple, these definitions may not stand up to strict linguistic scrutiny.

* **Token**: Before any real processing can be done on the input text, it needs to be segmented into linguistic units such as words, punctuation, and numbers or alphanumeric. These units are known as tokens.
* **Sentence**: An ordered sequence of tokens.
* **Tokenization**: The process of splitting a sentence into its constituent tokens. For segmented languages such as English, the existence of whitespace makes tokenization relatively easier and uninteresting. However, for languages such as Chinese and Arabic, the task is more difficult since there are no explicit boundaries. Furthermore, almost all characters in such non-segmented languages can exist as one-character words by themselves but can also join together to form multi-character words.
* **Corpus**: A body of text, usually containing a large number of sentences.
* **Part-of-speech (POS) Tag**: A word can be classified into one or more of a set of lexical or part-of-speech categories such as Nouns, Verbs, Adjectives and Articles, to name a few. A POS tag is a symbol representing such a lexical category - NN(Noun), VB(Verb), JJ(Adjective), AT(Article). One of the oldest and most commonly used tag sets is the Brown Corpus tag set. We will discuss the Brown Corpus in more detail below.
* **Parse Tree**: A tree defined over a given sentence that represents the syntactic structure of the sentence as defined by a formal grammar. Now that we have introduced the basic terminology, let’s look at some common NLP tasks:
* **POS Tagging**: Given a sentence and a set of POS tags, a common language processing task is to automatically assign POS tags to each word in the sentences. For example, given the sentence The ball is red, the output of a POS tagger would be The/AT ball/NN is/VB red/JJ. State-of-the-art POS taggers [9] can achieve accuracy as high as 96%. Tagging text with parts-of-speech turns out to be extremely useful for more complicated NLP tasks such as parsing and machine translation, which are discussed below.
* **Computational Morphology**: Natural languages consist of a very large number of words that are built upon basic building blocks known as morphemes (or stems), the smallest linguistic units possessing meaning. Computational morphology is concerned with the discovery and analysis of the internal structure of words using computers.
* **Parsing**: In the parsing task, a parser constructs the parse tree given a sentence. Some parsers assume the existence of a set of grammar rules in order to parse but recent parsers are smart enough to deduce the parse trees directly from the given data using complex statistical models. Most parsers also operate in a supervised setting and require the sentence to be POS-tagged before it can be parsed. Statistical parsing is an area of active research in NLP.