**Introduction to NLP**

Natural Language Processing (NLP) is the computerized approach to analysing text that is based on both a set of theories and a set of technologies. And, being a very active area of research and development, there is not a single agreed-upon definition that would satisfy everyone, but there are some aspects, which would be part of any knowledgeable person’s definition.

Natural Language Processing is a theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

Several elements of this definition can be further detailed. Firstly the imprecise notion of range of computational techniques’ is necessary because there are multiple methods or techniques from which to choose to accomplish a particular type of language analysis.

‘Naturally occurring texts’ can be of any language, mode, genre, etc. The texts can be oral or written. The only requirement is that they be in a language used by humans to communicate to one another. Also, the text being analysed should not be specifically constructed for the purpose of the analysis, but rather that the text be gathered from actual usage.

The notion of levels of linguistic analysis refers to the fact that there are multiple types of language processing known to be at work when humans produce or comprehend language. It is thought that humans normally utilize all of these levels since each level conveys different types of meaning. But various NLP systems utilize different levels, or combinations of levels of linguistic analysis, and this is seen in the differences amongst various NLP applications. This also leads to much confusion on the part of non-specialists as to what NLP really is, because a system that uses any subset of these levels of analysis can be said to be an NLP-based system. The difference between them, therefore, may actually be whether the system uses ‘weak’ NLP or ‘strong’ NLP.

‘Human-like language processing’ reveals that NLP is considered a discipline within Artificial Intelligence (AI). And while the full lineage of NLP does depend on a number of other disciplines, since NLP strives for human-like performance, it is appropriate to consider it an AI discipline.

**Goal:**

The goal of NLP as stated above is “to accomplish human-like language processing”.

The choice of the word ‘processing’ is very deliberate, and should not be replaced with ‘understanding’. For although the field of NLP was originally referred to as Natural Language Understanding (NLU) in the early days of AI, it is well agreed today that while the goal of NLP is true NLU, that goal has not yet been accomplished.

A full NLU System would be able to:

1. Paraphrase an input text

2. Translate the text into another language

3. Answer questions about the contents of the text

4. Draw inferences from the text

While NLP has made serious inroads into accomplishing goals 1 to 3, the fact that NLP

Systems cannot, of themselves, draw inferences from text. NLU still remains the goal of NLP. There are more practical goals for NLP, many related to the particular application for which it is being utilized. For example, an NLP-based IR system has the goal of providing more precise, complete information in response to a user’s real information need. The goal of the NLP system here is to represent the true meaning and intent of the user’s query, which can be expressed as naturally in everyday language as if they were speaking to a reference librarian. Also, the contents of the documents that are being searched will be represented at all their levels of meaning so that a true match between need and response can be found, no matter how either are expressed in their surface form.

For creating such a tremendous system, we will need some basic level functionality to be taken care of earlier. Let us assume English as the standard language for the world at this very moment. So we need to know and develop techniques which can do CFG parsing, tokenization, lemmatization and other.

So here is a brief look at some of these functionalities:

**1.1.1 STRUCTURE OF SENTENCE IN ENGLISH GRAMMER**:

|  |  |  |
| --- | --- | --- |
| **TENSE** | **STRUCTURE** | **EXAMPLE** |
| Simple Present Tense | S + V + O | He writes. |
| Simple Past Tense | S + V2 + O | He wrote. |
| Simple Future Tense | S + Will + V…. | He will write. |
| Present Continuous Tense | S + am/is /are + Ving…. | He is writing. |
| Past Continuous Tense | S + was/were + Ving…. | He was writing. |
| Future Continuous Tense | S + will be + Ving…. | He will be writing. |
| Present Perfect Tense | S + have/has + V3…. | He has written. |
| Past Perfect Tense | S + had + V3… | He had written. |
| Future Perfect Tense | S + will have + V3…. | He will have written. |
| Present Perfect Continuous Tense | S + have/has + been + Ving… | He has been teaching. |
| Past Perfect Continuous Tense | S + had + been + Ving… | He had been teaching. |
| Future Perfect Continuous Tense | S + will have + been + Ving…. | He will have been teaching. |

**1.1.2 PART OF SPEECH TAGGING:**

Parts-of-speech can be divided into two broad super categories: Closed class types and open class types. Closed classes are those that have relatively fixed membership. For example, prepositions are a closed class because there is a fixed set of them in English; new prepositions are rarely coined. By contrast nouns and verbs are open classes because new nouns and verbs are continually coined or borrowed from other languages (e.g., the new verb *to fax* or the borrowed noun *futon*). It is likely that any given speaker or corpus will have different open class words, but all speakers of a language, and corpora that are large enough, will likely share the set of closed class words. Closed class words are also generally function words like *of*, *it*, *and*, or *you*, which tend to be very short, occur frequently, and often have structuring uses in grammar. There are four major open classes that occur in the languages of the world; nouns, verbs, adjectives, and adverbs. It turns out that English has all four of these, although not every language does.

The closed classes differ more from language to language than do the open classes.

Here’s a quick overview of some of the more important closed classes in

English, with a few examples of each:

• **prepositions:** on, under, over, near, by, at, from, to, with

• **determiners:** a, an, the

• **pronouns:** she, who, I, others

• **conjunctions:** and, but, or, as, if, when

• **auxiliary verbs:** can, may, should, are

• **particles:** up, down, on, off, in, out, at, by,

• **numerals:** one, two, three, first, second, third

Apart from all these, standard define 8 parts of speech, which are as follows:

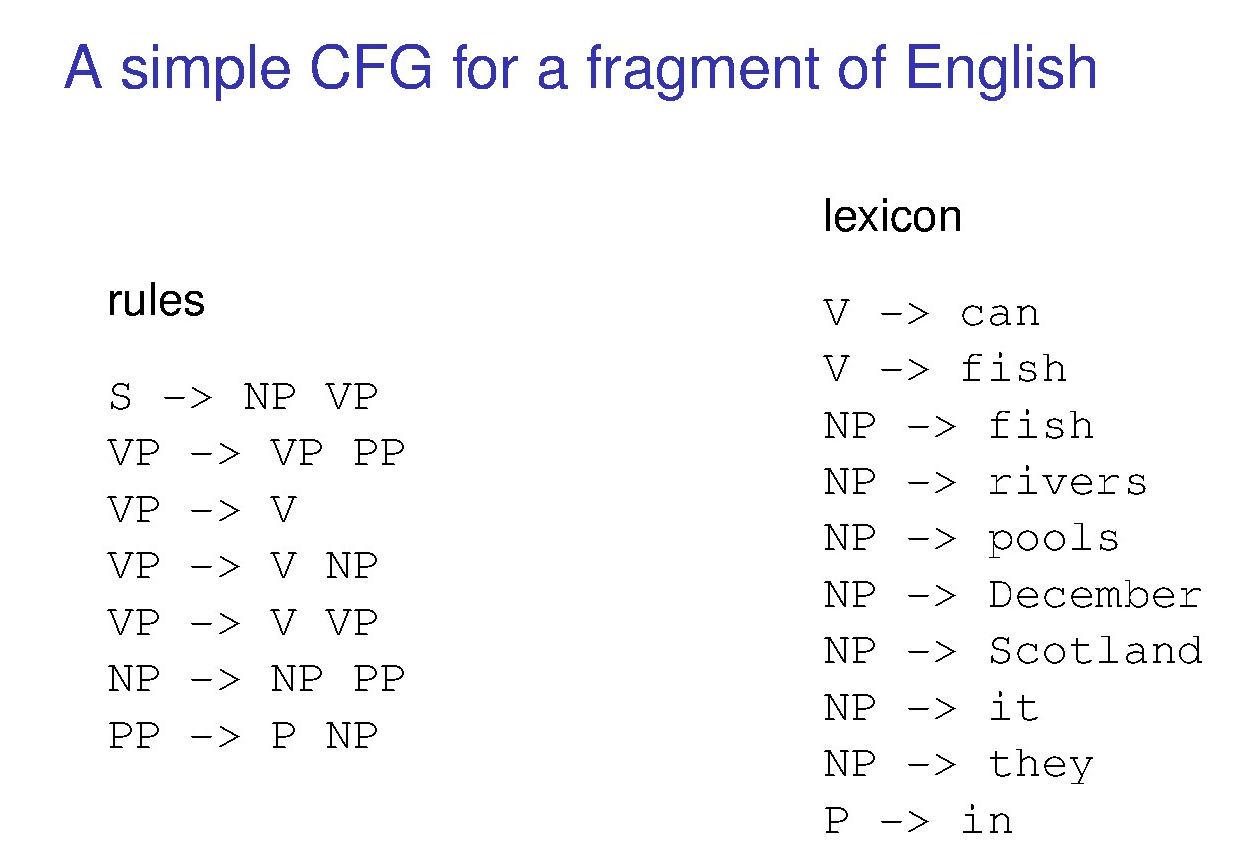
* **Noun**
* **Pronoun**
* **Verb**
* **Adverb**
* **Adjective**
* **Conjunction**
* **Disjunction**
* **Preposition**

Now in order to make tagging as well as later concepts a little better for the program functioning some tag-sets are provided by various research teams from time to time.

One of the most successful and popular tag-set called PENN TREEBANK is shown in following figure:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tag** | **Description** | **Example** | **Tag** | **Description** | **Example** |
| CC | Co-ordination, Conjunction | and, but, or | RP | Particle | up, off |
| CD | Cardinal number | one, two, three | SYM | Symbol | +,%, & |
| DT | Determiner | a, the | TO | “to” | To |
| DT | Existential ‘there’ | There | UH | Interjection | ah, oops |
| FX | Foreign word | mea culpa | VB | Verb, base form | Eat |
| IN | Preposition/sub-conjunction | of, in, by | VBD | Verb, past tense | Ate |
| JJ | Adjective | Yellow | VBG | Verb, gerund | Eating |
| JIR | Adj., comparative | Bigger | VBN | Verb, past participle | Eaten |
| JIS | Adj., superlative | Wildest | VBP | Verb, non-3sg pres | Eat |
| LS | List item marker | 1, 2, One | VBZ | Verb, 3sg pres | Eats |
| MD | Modal | can, should | WDT | Wh-determiner | which, that |
| NN | Noun, sing, mass | Llama | WP | Wh-pronoun | what, who |
| NNS | Noun, plural | Llamas | WP$ | Possessive wh- | Whose |
| NNP | Proper noun, singular | IBM | WRB | Wh-adverb | how, where |
| NNPS | Proper noun, plural | Carolinas | $ | Dollar sign | $ |
| PDT | Pre determiner | all, both | # | Pound sign | # |
| POS | Possessive ending | ’s | “ | Left quote | ‘ or “ |
| PRP | Personal pronoun | I, you, he | ” | Right quote | ’ or ” |
| PRPS | Possessive pronoun | your, one’s | ( | Left parenthesis | [, (, {, < |
| RB | Adverb | quickly, never | ) | Right parenthesis | ],), }, > |
| RBR | Adverb, comparative | Faster | , | Comma | , |
| RBS | Adverb, superlative | Fastest | . | Sentence-final punctuation | . ! ? |
|  |  |  | : | Mid-sentence punctuation | : ; ... – - |

**1.1.3 CONTEXT FREE GRAMMARS**

****

**1.1.4 PARSING**

The traditional approach to natural language processing takes as its basic assumption that a system must assign a complete constituent analysis to every sentence it encounters. The methods used to attempt this are drawn from mathematics, with context-free grammars playing a large role in assigning syntactic constituent structure.

Partee, ter Meulen and Wall (1993) provide an accessible introduction to the theoretical constructs underlying this approach, including set theory, logic, formal language theory, and automata theory, along with the application of these mechanisms to the syntax and semantics of natural language.

The program described in Alshawi 1992 is a very good example of a complete system built on these principles. For syntax, it uses a unification-based implementation of a generalized phrase structure grammar and handles an impressive number of syntactic structures which might be expected to appear in “interactive dialogues with information systems.

Although of course there is still a large residue even of this variety of English that the system fails to analyse properly.” (Alshawi 1992:61). In continuing research in this tradition, context-free grammars have been extended in various ways. The so-called “mildly context sensitive grammars,” such as tree adjoining grammars, have had considerable influence on recent work concerned with the formal aspects of parsing natural language. Several recent papers pursue non-traditional approaches to syntactic analysis. One such technique is partial, or underspecified, analysis.

For many applications such an analysis is entirely sufficient and can often be more reliably produced than a fully specified structure. Chen and Chen (1994), for example, employ statistical methods combined with a finite state mechanism to impose an analysis which consists only of noun phrase boundaries, without specifying their complete internal structure or their exact place in a complete tree structure. Agarwal and Boggess (1992) successfully rely on semantic features in a partially specified syntactic representation for the identification of coordinate structures.

In an innovative application of dependency grammar and dynamic programming techniques, Kurohashi and Nagao (1994) address the problem of analysing very complicated coordinate structures in Japanese.

A recent innovation in syntactic processing has been investigation into the use of statistical techniques.

The parsing techniques used yet are as follows:

1. **Top down parser**
2. **Bottom up parser**
3. **Bottom up left corner parser**
4. **Left corner parser with bottom up filter**
5. **Stepping parser (alternating top-down and bottom-up )**

Top-down parsing approach

