

A STUDY OF CURRENT TRENDS IN NATURAL LANGUAGE PROCESSING OF INDIAN LANGUAGE

Sapna Mittal ¹, Hema Arora ²

^{1,2}Computer Science , RSD College , Ferozepur

THEORITICAL FRAME WORK OF THE STUDY

Natural Language Processing (NLP) is the computerized approach to analyzing 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 analyzing 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 analyzed should not be specifically constructed for the purpose of the analysis, but rather that the text is gathered from actual usage. The notion of 'levels of linguistic analyses refer' 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.

'For a range of tasks or applications' points out that NLP is not usually considered a goal in and of itself, except perhaps for AI researchers. For others, NLP is the means for

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.

Origins

As most modern disciplines, the lineage of NLP is indeed mixed, and still today has strong emphases by different groups whose backgrounds are more influenced by one or another of the disciplines. Key among the contributors to the discipline and practice of NLP are: Linguistics - focuses on formal, structural models of language and the discovery of language universals - in fact the field of NLP was originally referred to as Computational Linguistics; Computer Science - is concerned with developing internal representations of data and efficient processing of these structures, and; Cognitive Psychology - looks at language usage as a window into human cognitive processes, and has the goal of modeling the use of language in a psychologically plausible way.

Divisions

While the entire field is referred to as Natural Language Processing, there are in fact two distinct focuses – language processing and language generation. The first of these refers to the analysis of language for the purpose of producing a meaningful representation, while the latter refers to the production of language from a representation. The task of Natural Language Processing is equivalent to the role of reader/listener, while the task of Natural Language Generation is that of the writer/speaker. While much of the theory and technology are shared by these two divisions, Natural Language Generation also requires a planning capability. That is, the generation system requires a plan or model of the goal of the interaction in order to decide what the system should generate at each point in an interaction. We will focus on the task of natural language analysis, as this is most relevant to Library and Information Science.

Another distinction is traditionally made between language understanding and speech understanding. Speech understanding starts with, and speech generation ends with, oral language and therefore relies on the additional fields of acoustics and phonology. Speech understanding focuses on how the 'sounds' of language as picked up by the system in the form of acoustical waves are transcribed into recognizable morphemes and words. Once in this form, the same levels of processing which are utilized on written text are utilized. All of these levels, including the phonology level, will be covered in Section 2; however, the emphasis throughout will be on language in the written form.

BRIEF HISTORY OF NATURAL LANGUAGE PROCESSING

Research in natural language processing has been going on for several decades dating back to the late 1940s. Machine translation (MT) was the first computer-based application related to natural language. While Weaver and Booth (1); (2) started one of the earliest MT projects in 1946 on computer translation based on expertise in breaking enemy codes during World War II, it was generally agreed that it was Weaver's memorandum of 1949 that brought the idea of MT to general notice and inspired many projects (3). He suggested using ideas from cryptography and information theory for language translation. Research began at various research institutions in the United States within a few years.

LEVELS OF NATURAL LANGUAGE PROCESSING

The most explanatory method for presenting what actually happens within a Natural Language processing system is by means of the 'levels of language' approach. This is also referred to as the synchronic model of language and is distinguished from the earlier sequential model, which hypothesizes that the levels of human language processing follow one another in a strictly sequential manner. Psycholinguistic research suggests that language processing is much more dynamic, as the levels can interact in a variety of orders. Introspection reveals that we frequently use information we gain from what is typically thought of as a higher level of processing to assist in a lower level of analysis. For example, the pragmatic knowledge that the document you are reading is about biology will be used when a particular word that has several possible senses (or meanings) is encountered, and the word will be interpreted as having the biology sense. Of necessity, the following description of levels will be presented sequentially. The key point here is that meaning is conveyed by each and every level of language and that since humans have been shown to use all levels of language to gain understanding, the more capable an NLP system is, the more levels of language it will utilize.

- **Phonology**

This level deals with the interpretation of speech sounds within and across words. There are, in fact, three types of rules used in phonological analysis: 1) phonetic rules – for sounds within words; 2) phonemic rules – for variations of pronunciation when words are spoken together, and; 3) prosodic rules – for fluctuation in stress and intonation across a sentence. In an NLP system that accepts spoken input, the sound waves are analyzed and encoded into a digitized signal for interpretation by various rules or by comparison to the particular language model being utilized.

- **Morphology**

This level deals with the componential nature of words, which are composed of morphemes – the smallest units of meaning. For example, the word preregistration can be morphologically analyzed into three separate morphemes: the prefix pre, the root registrar, and the suffixation. Since the meaning of each morpheme remains the same across words, humans can break down an unknown word into its constituent morphemes in order to understand its meaning. Similarly, an NLP system can recognize the meaning conveyed by each morpheme in order to gain and represent meaning. For example, adding the suffix–ed to a verb, conveys that the action of the

verb took place in the past. This is a key piece of meaning, and in fact, is frequently only evidenced in a text by the use of the -ed morpheme.

- **Lexical**

At this level, humans, as well as NLP systems, interpret the meaning of individual words. Several types of processing contribute to word-level understanding – the first of these being assignment of a single part-of-speech tag to each word. In this processing, words that can function as more than one part-of-speech are assigned the most probable part-of-speech tag based on the context in which they occur.

Additionally at the lexical level, those words that have only one possible sense or meaning can be replaced by a semantic representation of that meaning. The nature of the representation varies according to the semantic theory utilized in the NLP system. The following representation of the meaning of the word launch is in the form of logical predicates. As can be observed, a single lexical unit is decomposed into its more basic properties. Given that there is a set of semantic primitives used across all words; these simplified lexical representations make it possible to unify meaning across words and to produce complex interpretations, much the same as humans do.

Launch (a large boat used for carrying people on rivers, lakes harbors, etc.) ((CLASS BOAT) (PROPERTIES (LARGE) (PURPOSE (PREDICATION (CLASS CARRY) (OBJECT PEOPLE))))))
--

The lexical level may require a lexicon, and the particular approach taken by an NLP system will determine whether a lexicon will be utilized, as well as the nature and extent of information that is encoded in the lexicon. Lexicons may be quite simple, with only the words and their part(s)-of-speech, or may be increasingly complex and contain information on the semantic class of the word, what arguments it takes, and the semantic limitations on these arguments, definitions of the sense(s) in the semantic representation utilized in the particular system, and even the semantic field in which each sense of a polysemous word is used.

- **Syntactic**

This level focuses on analyzing the words in a sentence so as to uncover the grammatical structure of the sentence. This requires both a grammar and a parser. The output of this level of processing is a (possibly delinearized) representation of the sentence that reveals the structural dependency relationships between the words. There are various grammars that can be utilized, and which will, in turn, impact the choice of a parser. Not all NLP applications require a full parse of sentences, therefore the remaining challenges in parsing of prepositional phrase attachment and conjunction scoping no longer stymie those applications for which phrasal and clausal dependencies are sufficient. Syntax conveys meaning in most languages because order and dependency contribute to meaning. For example the two sentences: ‘The dog chased the cat.’ and ‘The cat chased the dog.’ differ only in terms of syntax, yet convey quite different meanings.

- **Semantic**

This is the level at which most people think meaning is determined, however, as we can see in the above defining of the levels, it is all the levels that contribute to meaning. Semantic processing determines the possible meanings of a sentence by focusing on the interactions among word-level meanings in the sentence. This level of processing can include the semantic disambiguation of words with multiple senses; in an analogous way to

how syntactic disambiguation of words that can function as multiple parts-of-speech is accomplished at the syntactic level. Semantic disambiguation permits one and only one sense of polysemous words to be selected and included in the semantic representation of the sentence. For example, amongst other meanings, 'file' as a noun can mean either a folder for storing papers, or a tool to shape one's fingernails, or a line of individuals in a queue. If information from the rest of the sentence were required for the disambiguation, the semantic, not the lexical level, would do the disambiguation. A wide range of methods can be implemented to accomplish the disambiguation, some which require information as to the frequency with which each sense occurs in a particular corpus of interest, or in general usage, some which require consideration of the local context, and others which utilize pragmatic knowledge of the domain of the document.

- **Discourse**

While syntax and semantics work with sentence-length units, the discourse level of NLP works with units of text longer than a sentence. That is, it does not interpret multisentence texts as just concatenated sentences, each of which can be interpreted singly. Rather, discourse focuses on the properties of the text as a whole that convey meaning by making connections between component sentences. Several types of discourse processing can occur at this level, two of the most common being anaphora resolution and discourse/text structure recognition. Anaphora resolution is the replacing of words such as pronouns, which are semantically vacant, with the appropriate entity to which they refer (30). Discourse/text structure recognition determines the functions of sentences in the text, which, in turn, adds to the meaningful representation of the text. For example, newspaper articles can be deconstructed into discourse components such as: Lead, Main Story, Previous Events, Evaluation, Attributed Quotes, and Expectation (31).

- **Pragmatic**

This level is concerned with the purposeful use of language in situations and utilizes context over and above the contents of the text for understanding the goal is to explain how extra meaning is read into texts without actually being encoded in them. This requires much world knowledge, including the understanding of intentions, plans, and goals. Some NLP applications may utilize knowledge bases and inference modules. For example, the following

two sentences require resolution of the anaphoric term 'they', but this resolution requires Pragmatic or world knowledge.

- **Summary of Levels**

Current NLP systems tend to implement modules to accomplish mainly the lower levels of processing. This is for several reasons. First, the application may not require interpretation at the higher levels. Secondly, the lower levels have been more thoroughly researched and implemented. Thirdly, the lower levels deal with smaller units of analysis, e.g. morphemes, words, and sentences, which are rule-governed, versus the higher levels of language processing which deal with texts and world knowledge, and which are only regularity-governed. As will be seen in the following section on Approaches, the statistical approaches have, to date, been validated on the lower levels of analysis, while the symbolic approaches have dealt with all levels, although there are still few working systems which incorporate the higher levels.