Natural Language Processing[1]

INTRODUCTION

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. The definition I offer is:

> **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 be gathered from actual usage.

The notion of *'levels of linguistic analysis'* (to be further explained in Section 2) 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.

'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

---

[1] Liddy, E. D. In Encyclopedia of Library and Information Science, 2nd Ed. Marcel Decker, Inc.

accomplishing a particular task. Therefore, you have Information Retrieval (IR) systems that utilize NLP, as well as Machine Translation (MT), Question-Answering, etc.

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 rely 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.

Early work in MT took the simplistic view that the only differences between languages resided in their vocabularies and the permitted word orders. Systems developed from this perspective simply used dictionary-lookup for appropriate words for translation and reordered the words after translation to fit the word-order rules of the target language, without taking into account the lexical ambiguity inherent in natural language. This produced poor results. The apparent failure made researchers realize that the task was a lot harder than anticipated, and they needed a more adequate theory of language. However, it was not until 1957 when Chomsky (4) published *Syntactic Structures* introducing the idea of generative grammar, did the field gain better insight into whether or how mainstream linguistics could help MT.

During this period, other NLP application areas began to emerge, such as speech recognition. The language processing community and the speech community then was split into two camps with the language processing community dominated by the

theoretical perspective of generative grammar and hostile to statistical methods, and the speech community dominated by statistical information theory (5) and hostile to theoretical linguistics (6).

Due to the developments of the syntactic theory of language and parsing algorithms, there was over-enthusiasm in the 1950s that people believed that fully automatic high quality translation systems (2) would be able to produce results indistinguishable from those of human translators, and such systems should be in operation within a few years. It was not only unrealistic given the then-available linguistic knowledge and computer systems, but also impossible in principle (3).

The inadequacies of then-existing systems, and perhaps accompanied by the over-enthusiasm, led to the ALPAC (Automatic Language Processing Advisory Committee of the National Academy of Science - National Research Council) report of 1966. (7) The report concluded that MT was not immediately achievable and recommended it not be funded. This had the effect of halting MT and most work in other applications of NLP at least within the United States.

Although there was a substantial decrease in NLP work during the years after the ALPAC report, there were some significant developments, both in theoretical issues and in construction of prototype systems. Theoretical work in the late 1960's and early 1970's focused on the issue of how to represent meaning and developing computationally tractable solutions that the then-existing theories of grammar were not able to produce. In 1965, Chomsky (8) introduced the transformational model of linguistic competence. However, the transformational generative grammars were too syntactically oriented to allow for semantic concerns. They also did not lend themselves easily to computational implementation. As a reaction to Chomsky's theories and the work of other transformational generativists, case grammar of Fillmore, (9), semantic networks of Quillian, (10), and conceptual dependency theory of Schank, (11) were developed to explain syntactic anomalies, and provide semantic representations. Augmented transition networks of Woods, (12) extended the power of phrase-structure grammar by incorporating mechanisms from programming languages such as LISP. Other representation formalisms included Wilks' preference semantics (13), and Kay's functional grammar (14).

Alongside theoretical development, many prototype systems were developed to demonstrate the effectiveness of particular principles. Weizenbaum's ELIZA (15) was built to replicate the conversation between a psychologist and a patient, simply by permuting or echoing the user input. Winograd's SHRDLU (16) simulated a robot that manipulated blocks on a tabletop. Despite its limitations, it showed that natural language understanding was indeed possible for the computer (17). PARRY (18) attempted to embody a theory of paranoia in a system. Instead of single keywords, it used groups of keywords, and used synonyms if keywords were not found. LUNAR was developed by Woods (19) as an interface system to a database that consisted of information about lunar rock samples using augmented transition network and procedural semantics (20).

In the late 1970's, attention shifted to semantic issues, discourse phenomena, and communicative goals and plans (21). Grosz (22) analyzed task-oriented dialogues and proposed a theory to partition the discourse into units based on her findings about the relation between the structure of a task and the structure of the task-oriented dialogue. Mann and Thompson (23) developed Rhetorical Structure Theory, attributing hierarchical structure to discourse. Other researchers have also made significant contributions, including Hobbs and Rosenschein (24), Polanyi and Scha (25), and Reichman (26).

This period also saw considerable work on natural language generation. McKeown's discourse planner TEXT (27) and McDonald's response generator MUMMBLE (28) used rhetorical predicates to produce declarative descriptions in the form of short texts, usually paragraphs. TEXT's ability to generate coherent responses online was considered a major achievement.

In the early 1980s, motivated by the availability of critical computational resources, the growing awareness within each community of the limitations of isolated solutions to NLP problems (21), and a general push toward applications that worked with language in a broad, real-world context (6), researchers started re-examining non-symbolic approaches that had lost popularity in early days. By the end of 1980s, symbolic approaches had been used to address many significant problems in NLP and statistical approaches were shown to be complementary in many respects to symbolic approaches (21).

In the last ten years of the millennium, the field was growing rapidly. This can be attributed to: a) increased availability of large amounts of electronic text; b) availability of computers with increased speed and memory; and c) the advent of the Internet. Statistical approaches succeeded in dealing with many generic problems in computational linguistics such as part-of-speech identification, word sense disambiguation, etc., and have become standard throughout NLP (29). NLP researchers are now developing next generation NLP systems that deal reasonably well with general text and account for a good portion of the variability and ambiguity of language.

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.

(Figure 1: Synchronized Model of Language Processing)

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 *registra,* and the suffix *tion*. 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 multi-sentence 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 inferencing modules. For example, the following two sentences require resolution of the anaphoric term ‘they’, but this resolution requires pragmatic or world knowledge.

> *The city councilors refused the demonstrators a permit because **they** feared violence.*
>
> *The city councilors refused the demonstrators a permit because **they** advocated revolution.*

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.

APPROACHES TO NATURAL LANGUAGE PROCESSING

Natural language processing approaches fall roughly into four categories: symbolic, statistical, connectionist, and hybrid. Symbolic and statistical approaches have coexisted since the early days of this field. Connectionist NLP work first appeared in the 1960's. For a long time, symbolic approaches dominated the field. In the 1980's, statistical approaches regained popularity as a result of the availability of critical computational resources and the need to deal with broad, real-world contexts. Connectionist approaches also recovered from earlier criticism by demonstrating the utility of neural networks in NLP. This section examines each of these approaches in terms of their foundations, typical techniques, differences in processing and system aspects, and their robustness, flexibility, and suitability for various tasks.

Symbolic Approach

Symbolic approaches perform deep analysis of linguistic phenomena and are based on explicit representation of facts about language through well-understood knowledge representation schemes and associated algorithms (21). In fact, the description of the levels of language analysis in the preceding section is given from a symbolic perspective. The primary source of evidence in symbolic systems comes from human-developed rules and lexicons.

A good example of symbolic approaches is seen in logic or rule-based systems. In logic-based systems, the symbolic structure is usually in the form of logic propositions. Manipulations of such structures are defined by inference procedures that are generally truth preserving. Rule-based systems usually consist of a set of rules, an inference engine, and a workspace or working memory. Knowledge is represented as facts or rules in the rule-base. The inference engine repeatedly selects a rule whose condition is satisfied and executes the rule.

Another example of symbolic approaches is semantic networks. First proposed by Quillian (10) to model associative memory in psychology, semantic networks represent knowledge through a set of nodes that represent objects or concepts and the labeled links that represent relations between nodes. The pattern of connectivity reflects semantic organization, that is; highly associated concepts are directly linked whereas moderately or weakly related concepts are linked through intervening concepts. Semantic networks are widely used to represent structured knowledge and have the most connectionist flavor of the symbolic models (32).

Symbolic approaches have been used for a few decades in a variety of research areas and applications such as information extraction, text categorization, ambiguity resolution, and lexical acquisition. Typical techniques include: explanation-based learning, rule-based learning, inductive logic programming, decision trees, conceptual clustering, and K nearest neighbor algorithms (6; 33).

Statistical Approach

Statistical approaches employ various mathematical techniques and often use large text corpora to develop approximate generalized models of linguistic phenomena based on actual examples of these phenomena provided by the text corpora without adding significant linguistic or world knowledge. In contrast to symbolic approaches, statistical approaches use observable data as the primary source of evidence.

A frequently used statistical model is the Hidden Markov Model (HMM) inherited from the speech community. HMM is a finite state automaton that has a set of states with probabilities attached to transitions between states (34). Although outputs are visible, states themselves are not directly observable, thus "hidden" from external observations. Each state produces one of the observable outputs with a certain probability.

Statistical approaches have typically been used in tasks such as speech recognition, lexical acquisition, parsing, part-of-speech tagging, collocations, statistical machine translation, statistical grammar learning, and so on.

Connectionist Approach

Similar to the statistical approaches, connectionist approaches also develop generalized models from examples of linguistic phenomena. What separates connectionism from other statistical methods is that connectionist models combine statistical learning with various theories of representation - thus the connectionist representations allow transformation, inference, and manipulation of logic formulae (33). In addition, in connectionist systems, linguistic models are harder to observe due to the fact that connectionist architectures are less constrained than statistical ones (35); (21).

Generally speaking, a connectionist model is a network of interconnected simple processing units with knowledge stored in the weights of the connections between units (32). Local interactions among units can result in dynamic global behavior, which, in turn, leads to computation.

Some connectionist models are called localist models, assuming that each unit represents a particular concept. For example, one unit might represent the concept "mammal" while another unit might represent the concept "whale". Relations between concepts are encoded by the weights of connections between those concepts. Knowledge in such models is spread across the network, and the connectivity between units reflects their structural relationship. Localist models are quite similar to semantic networks, but the links between units are not usually labeled as they are in semantic nets. They perform

well at tasks such as word-sense disambiguation, language generation, and limited inference (36).

Other connectionist models are called distributed models. Unlike that in localist models, a concept in distributed models is represented as a function of simultaneous activation of multiple units. An individual unit only participates in a concept representation. These models are well suited for natural language processing tasks such as syntactic parsing, limited domain translation tasks, and associative retrieval.

Comparison Among Approaches

From the above section, we have seen that similarities and differences exist between approaches in terms of their assumptions, philosophical foundations, and source of evidence. In addition to that, the similarities and differences can also be reflected in the processes each approach follows, as well as in system aspects, robustness, flexibility, and suitable tasks.

Process: Research using these different approaches follows a general set of steps, namely, data collection, data analysis/model building, rule/data construction, and application of rules/data in system. The data collection stage is critical to all three approaches although statistical and connectionist approaches typically require much more data than symbolic approaches. In the data analysis/model building stage, symbolic approaches rely on human analysis of the data in order to form a theory while statistical approaches manually define a statistical model that is an approximate generalization of the collected data. Connectionist approaches build a connectionist model from the data. In the rule / data construction stage, manual efforts are typical for symbolic approaches and the theory formed in the previous step may evolve when new cases are encountered. In contrast, statistical and connectionist approaches use the statistical or connectionist model as guidance and build rules or data items automatically, usually in relatively large quantity. After building rules or data items, all approaches then automatically apply them to specific tasks in the system. For instance, connectionist approaches may apply the rules to train the weights of links between units.

System aspects: By system aspects, we mean source of data, theory or model formed from data analysis, rules, and basis for evaluation.

- *Data:* As mentioned earlier, symbolic approaches use human introspective data, which are usually not directly observable. Statistical and connectionist approaches are built on the basis of machine observable facets of data, usually from text corpora.

- *Theory or model based on data analysis:* As the outcome of data analysis, a theory is formed for symbolic approaches whereas a parametric model is formed for statistical approaches and a connectionist model is formed for connectionist approaches.

- *Rules:* For symbolic approaches, the rule construction stage usually results in rules with detailed criteria of rule application. For statistical approaches, the criteria of rule

application are usually at the surface level or under-specified. For connectionist approaches, individual rules typically cannot be recognized.

- *Basis for Evaluation:* Evaluation of symbolic systems is typically based on intuitive judgments of unaffiliated subjects and may use system-internal measures of growth such as the number of new rules. In contrast, the basis for evaluation of statistical and connectionist systems are usually in the form of scores computed from some evaluation function. However, if all approaches are utilized for the same task, then the results of the task can be evaluated both quantitatively and qualitatively and compared.

Robustness: Symbolic systems may be fragile when presented with unusual, or noisy input. To deal with anomalies, they can anticipate them by making the grammar more general to accommodate them. Compared to symbolic systems, statistical systems may be more robust in the face of unexpected input provided that training data is sufficient, which may be difficult to be assured of. Connectionist systems may also be robust and fault tolerant because knowledge in such systems is stored across the network. When presented with noisy input, they degrade gradually.

Flexibility: Since symbolic models are built by human analysis of well-formulated examples, symbolic systems may lack the flexibility to adapt dynamically to experience. In contrast, statistical systems allow broad coverage, and may be better able to deal with unrestricted text (21) for more effective handling of the task at hand. Connectionist systems exhibit flexibility by dynamically acquiring appropriate behavior based on the given input. For example, the weights of a connectionist network can be adapted in real-time to improve performance. However, such systems may have difficulty with the representation of structures needed to handle complex conceptual relationships, thus limiting their abilities to handle high-level NLP (36).

Suitable tasks: Symbolic approaches seem to be suited for phenomena that exhibit identifiable linguistic behavior. They can be used to model phenomena at all the various linguistic levels described in earlier sections. Statistical approaches have proven to be effective in modeling language phenomena based on frequent use of language as reflected in text corpora. Linguistic phenomena that are not well understood or do not exhibit clear regularity are candidates for statistical approaches. Similar to statistical approaches, connectionist approaches can also deal with linguistic phenomena that are not well understood. They are useful for low-level NLP tasks that are usually subtasks in a larger problem.

To summarize, symbolic, statistical, and connectionist approaches have exhibited different characteristics, thus some problems may be better tackled with one approach while other problems by another. In some cases, for some specific tasks, one approach may prove adequate, while in other cases, the tasks can get so complex that it might not be possible to choose a single best approach. In addition, as Klavans and Resnik (6) pointed out, there is no such thing as a “purely statistical” method. Every use of statistics is based upon a symbolic model and statistics alone is not adequate for NLP. Toward this end, statistical approaches are not at odds with symbolic approaches. In fact, they are

rather complementary. As a result, researchers have begun developing hybrid techniques that utilize the strengths of each approach in an attempt to address NLP problems more effectively and in a more flexible manner.

NATURAL LANGUAGE PROCESSING APPLICATIONS

Natural language processing provides both theory and implementations for a range of applications. In fact, any application that utilizes text is a candidate for NLP. The most frequent applications utilizing NLP include the following:

- Information Retrieval – given the significant presence of text in this application, it is surprising that so few implementations utilize NLP. Recently, statistical approaches for accomplishing NLP have seen more utilization, but few systems other than those by Liddy (37) and Strzalkowski (38) have developed significant systems based on NLP

.

- Information Extraction (IE) – a more recent application area, IE focuses on the recognition, tagging, and extraction into a structured representation, certain key elements of information, e.g. persons, companies, locations, organizations, from large collections of text. These extractions can then be utilized for a range of applications including question-answering, visualization, and data mining.

- Question-Answering – in contrast to Information Retrieval, which provides a list of potentially relevant documents in response to a user's query, question-answering provides the user with either just the text of the answer itself or answer-providing passages.

- Summarization – the higher levels of NLP, particularly the discourse level, can empower an implementation that reduces a larger text into a shorter, yet richly-constituted abbreviated narrative representation of the original document.

- Machine Translation – perhaps the oldest of all NLP applications, various levels of NLP have been utilized in MT systems, ranging from the 'word-based' approach to applications that include higher levels of analysis.

- Dialogue Systems – perhaps the omnipresent application of the future, in the systems envisioned by large providers of end-user applications. Dialogue systems, which usually focus on a narrowly defined application (e.g. your refrigerator or home sound system), currently utilize the phonetic and lexical levels of language. It is believed that utilization of all the levels of language processing explained above offer the potential for truly habitable dialogue systems.

CONCLUSIONS

While NLP is a relatively recent area of research and application, as compared to other information technology approaches, there have been sufficient successes to date that suggest that NLP-based information access technologies will continue to be a major area of research and development in information systems now and far into the future.

Acknowledgement

Grateful appreciation to Xiaoyong Liu who contributed to this entry while she was a Ph.D. student and a Research Assistant in the Center for Natural Language Processing in the School of Information Studies at Syracuse University.