Critical Challenges in Natural Language Processing

Veena A Prakashe

Abstract

In this paper, the author attempts to enlist some of the basic bottlenecks that pose challenges while designing the automation of any natural language understanding system. In the beginning, some background material on the study of language and an overview of linguistics is presented for the benefit of the reader who might be new to the fields of artificial intelligence and cognitive science. Natural language systems are also discussed briefly so as to give a better insight into the processing of natural languages by computer systems. Then the three major threats or challenges of natural language processing, viz. knowledge acquisition from natural language; interaction with multiple underlying systems; and partial understanding of multi-sentence and fragments of language are discussed.

Keywords : Natural Language Processing.

0. Introduction

What is natural language? Natural language is any language that humans learn from their environment and use to communicate with each other. Whatever the form of the communication, natural languages are used to express our knowledge and emotions and to convey our responses to other people and to our surroundings. Natural languages are usually learned in early childhood from those around us. Children seem to recognize at a surprisingly early age the value of structure and uniformity in their utterances. Words, phrases, and sentences replace grunts, whines, and cries and better serve to convince others to recognize the child's needs. Natural languages can be acquired later in life through school, travel, or change in culture, but with very few exceptions, all humans in all cultures learn to communicate verbally in the language natural to their immediate environment.

In contrast to natural languages, artificial languages are languages created by humans to communicate with their technology, for example, computer programming languages.

Human process natural languages whenever they read Shakespeare, dictate a business letter, or tell a joke. Sign language is used by the hearing impaired to communicate thoughts and feelings with others and replaces the language they are unable to hear. Despite the different forms of language in each of these situations, aspects of the language used are similar. Whether language is spoken or written, message has a structure and the elements of language relate to each other in recognizable ways. Verbal communication or speech is characterized by the sounds which almost every human is capable of producing. Whether each person learns to produce a particular sound is determined by the languages learned rather than the anatomical speech production mechanisms. which are approximately the same for all normal humans. Speech is produced by stringing together individual human sounds in recognized patterns. The study of these patterns of sounds is called *phonology*. The study of the structure of language units and their relationships is called*syntax*. Phonology and syntax are both important parts of the field of linguistics.

Linguists are also concerned with semantics, the study of the relationship between the linguistics structures used and the meanings intended; in other words, how does what we say or write relate to what we mean? It is not enough for a sentence to be correct in form; it must also make sense. For example, the sentence.

The tree sang the chair.

Would not in ordinary discourse, be a meaningful sentence, even though it is grammatically reasonable. The noun phrase, The tree, can be the subject of a sentence; sang is obviously a verb; and the chair is a noun phrase which can serve as a direct object of a verb. But trees do not sing, and nothing that sings, sings chairs. So, how can this string of words be a sentence, even an unreasonable one?

We recognize that the following sentences have the same structure.

> The tree sang the chair.
>
> The students finished the exam.

The new, young vice-president in change of financial affairs in the company established extraordinary regulations concerning the procedures for reporting exceptional situations in the payroll department.

Thus, something in language makes us aware of similarities among sentences despite the variance in subject matter. The systems used by linguists to describe these similarities are called grammars. The term grammar is also used to refer to the methods taught in school such as diagramming sentences. These methods are designed to show the relationships among the various structures within sentences. Grammars consist of the elements allowed within sentences and the rules for putting these elements together. For example, the structure of some sentences can be described as a noun phrase followed by a verb phrase. In this case, the elements are the sentence, a noun phrase, and verb phrase. A rule to express their relationship could be written as:

SENTENCE ✍ NOUN PHRASE + VERB PHRASE

Obviously further definition of these elements would be required to describe noun phrases, verb phrases, and their components, and each of the components would have to be defined. This process of redefinition of the grammatical constructs would be continued until the elements were defined as specific words. The words in a grammar are called the vocabulary. The grammar is made up of the rules and the vocabulary along with the meanings associated with the vocabulary.

Besides linguists' use of grammars for describing language, grammars are used by logicians and formal philosophers to study formal languages. A formal grammar is essentially a set of rules and a list of elements upon which the rules can be applied.

Other logicians have sought to represent natural language by means of prepositional logic. All sentences in the language considered are written as propositions and can be manipulated according to the rules of formal logic. The statement, *All teenagers drive cars*, could be rewritten in logical notation as:

FOR-ALL (x) (EXISTS (y)(TEENAGER(x) (CAR(y) AND DRIVE(x,y))))

When a sentence has been thus transcribed, the rules of logic can be applied to test the validity of any references to the information contained in the sentence. However, prepositional logic can express only a subset of all sentences in any natural language. The method only applies to statements about which the truth or falsity can be known.

Cognitive psychologists, concerned with how humans think, approach language from a different perspective than linguists and logicians. They view language as a representational medium for thought rather than viewing language as an independent phenomenon.

Writing personal letters is a good example of social communication. Language used for social purposes follows the same rules as other language, but frequently is highly formulaic. The same phrases are used over and over in similar situations, often losing their meaning somewhat, yet still serving their basic function. For example,

> I love you.
>
> Hello, how are you? Fine, thank you, and you?
>
> Thank you very much for the gift. I like it a lot.
>
> Hey, bro. Wha's hap'nin'?

Dealing with language of this sort requires different analysis techniques from other language. The meaning behind the words seems to be of a different nature than the content of language used to convey information. Yet the notion of language as social interaction still fits the paradigm of the human information processing system.

1. Text Processing

Much of the information processed by computers is text, data of the type generally called character or alphanumeric. In natural language processing, all written material is text.

2. Characteristics of Text

Dealing with text is both simpler and harder than manipulating numeric data. In terms of the physical characteristics, it is simpler in that text is linear; the first character is handled, then the second, then the third, until the last is reached. At that point the data is processed. But in logical terms, text is quite slippery. Generally, in the computer, numeric data is represented in a specific form; a number is given a fixed quantity of bits and a set format. All integers occupy the same amount of storage in memory in a particular computer, as do real numbers. (Of course, extended precision may increase the amount, but it is still a fixed amount.) Text, on the other hand, is made up of words and names and other strings of characters, which are many different lengths, and thus require differing amounts of memory for storage.

Connected text, such as this paragraph which you are reading, can be handled as a linear string of characters, then broken up into words of varying length, which could then be processed. A word is defined in text processing as the string of characters, usually alphabetic, that fall between delimiters: blanks, commas, periods, parentheses, and any other allowable punctuation marks that indicate the end of a word. This definition covers some forms of text besides connected text, such as business letters and mailing lists containing names and addressed, or bibliographic data with various fields separated by specific punctuation marks. Much of this type of data is not referred to as natural language because it is not in sentence form. However, parts of some fields look very much like natural language, such as titles of books. This type of textual data has been the primary object of text editors and word processing systems without much concern for the language involved.

3. MARC Format and WEBMARC

A format designed to handle the various problems of dealing with text was developed by the U.S. Library of Congress MARC (MAchine Readable Catalog) Project in 1967. The MARC format has been used for a variety of library projects including communication and information exchange among the many libraries with machine readable information. Donald Sherman adapted the MARC format for recording dictionary data, specifically Webster's Seventh Collegiate Dictionary (known as W7), which was originally recorded in machine readable form by the Lexicographical Project at SDC. WEBMARC, Sherman's version of W7 in MARC format, contains 68,657 entries, each stored as a variable-length record representing the information about one word in the dictionary.

In WEBMARC, the leader is the first 24 characters in each record and contains fixed-length fields recording the record length, status (F for full record), source of data (W7), record extension number (usually, 0, for non-extended record, 1, 2,... for entry requiring more space than one physical record), an address pointer to the data part of the record, and a record identification number. The record directory follows the leader and is made up of a series of fixed-length (12 character) segments containing a tag identifying each part of the lexical entry, the address of the first character in that part, and the length of that part. The tag fields are three- digit numbers, the first digit of which identifies the type of the field.

The WEBMARC format illustrates several of the methods described for recording variable data, such as dictionary entries and bibliographical citations. It is not especially efficient in that no data compression is used and many of the record fields take up more space than required.

4. Design of Natural Language Systems

A natural language system designed to understand and manipulate language should be capable of accepting input in natural language text, storing knowledge related to the application domain, drawing inferences from that knowledge, answering questions based on the knowledge, and generating responses.

5. General Description of NLS

NLS is a knowledge based natural language understanding system. It processes natural language input and generates appropriate output. The knowledge base for this system is precompiled; in other words, a knowledge domain exists before execution begins. This knowledge base (KB) preserves both hierarchical and prepositional information about the data stored. The input accepted by the system includes statements to be paraphrased, i.e., restated to assure proper understanding; statements which represent knowledge to be learned, i.e., added to the KB; and questions to be answered by accessing the KB. The system outputs appropriate responses to input, as well as paraphrases of statements and answers to questions.

The major modules of NLS include the Parser, the Understander, and the Generator. The parser accepts the input string and maps in into an internal structure compatible with the KB. The generator maps from the internal structure to the output string. The understander module accesses the KB for various purposes; to obtain knowledge, to draw inferences, or to add knowledge to the KB. These three functions interact to accomplish various tasks.

6. Critical challenges for natural language processing

This paper identifies the problems that we believe must block widespread use of computational linguistics.

Knowledge acquisition from natural language (NL) texts of various kinds, from interactions with human beings, and from other sources. Language processing requires lexical, grammatical, semantic, and pragmatic knowledge.

Interaction with multiple underlying systems to give NL systems the utility and flexibility demanded by people using them. Single application systems are limited in both usefulness and the language that is necessary to communicate with them.

Partial understanding gleaned from multi-sentence language, or from fragments of language. Approaches to language understanding that require perfect input or that try to produce perfect output seem doomed to failure because novel language, incomplete language, and errorful language are the norm, not the exception.

7. State-of-the-art

The limitations of practical language processing technology have been summarized as follows:

Domains must be narrow enough so that the constraints on the relevant semantic concepts and relations can be expressed using current knowledge representation techniques, i.e. primarily in terms of types and sorts. Processing may be viewed abstractly as the application of recursive tree rewriting, including filtering out tree out matching a certain pattern.

Handcrafting is necessary, particularly in the grammatical components of systems (the component technology that exhibits least dependence on the application domain). Lexicons and axiomatizations of critical facts must be developed for each domain, and these remain time-consuming tasks.

The user must still adapt to the machine, but, as the products testify, the user can do so effectively.

Current systems have limited discourse capabilities that are almost exclusively handcrafted. Thus current systems are limited to viewing interaction, translation, and writing and reading text as processing a sequence of either isolated sentences or loosely related paragraphs. Consequently, the user must adapt to such limited discourse.

It is traditional to divide natural language phenomena (and components of systems designed to deal with them) into three classes:

Syntactic phenomena- those that pertain to the meaning of a sentence and the order of words in the sentence, based on the grammatical classes of words rather than their meaning.

Semantic phenomena- those that pertain to the meaning of a sentence relatively independent of the context in which that language occurs.

Pragmatic phenomena- those that relate the meaning of a sentence to the context in which it occurs. This context can be linguistic (such as the previous text or dialogue), or nonlinguistic (such as knowledge about the person who produced the language, about the goals of the communication, about the objects in the current visual field, etc.).

8. Knowledge acquisition for language processing

It goes without saying that any NLP system must know a fair amount about words, language, and some subject area before being able to understand language. Currently, virtually all NLP systems operate using fairly laboriously hand-built knowledge bases. The knowledge bases may include both linguistic knowledge (morphological, lexical, syntactic, semantic, and discourse) and nonlinguistic knowledge (semantic world knowledge, pragmatic, planning, inference), and the knowledge in them may be absolute or probabilistic. (Not all of these knowledge bases are necessary for every NLP system).

9. Types of knowledge acquisition

Just as there are many kinds of knowledge, there are a number of different ways of acquiring that knowledge:

Knowing by being pre-programmed - this includes such things as hand-built grammars and semantic interpretation rules.

Knowing by being told - this includes things that a human can “tell” the system using various user-interface tools, such as semantic interpretation rules that can be automatically built from examples, selectional restrictions, and various lexical and morphological features.

Knowing by looking in up - this means using references such as an online dictionary, where one can find exactly the information that is being sought.

Knowing by using source material – this means using references such as an encyclopedia or a corpus of domain-relevant material, from which one might be able to find or infer the information being sought; it may also mean using large volumes of material as the source of probabilistic knowledge (e.g., bank is more likely to mean a financial institution than the side of a river).

Knowing by figuring it out - this means using heuristics and the input itself (such as the part of speech of words surrounding an unknown word). Knowing by using a combination of the above techniques- this may or may not involve human intervention.

10. Interfacing to multiple underlying systems

Most current NL systems, whether accepting spoken or typed input, are designed to interface to a single homogeneous underlying system; they have a component geared to producing code for that single class of application systems, such as a relational database (Stallard, 1987; Parlance User Manual, Learner User Manual.) These systems take advantage of the simplicity of the semantics and the availability of a formal language (relational calculus and relational algebra) for the system’s output.

The challenge is to recreate a systematic, tractable procedure to translate from the logical expression of the user’s input to systems that are not fully relational, such as expert system functions, object-oriented and numerical simulation systems, calculation programs, and so on. Implicit in that challenge is the need to generate code for non-homogeneous software applications- those that have more than one application system.

The norm in the present generation of user environments is distributed, networked applications. A seamless, multi-model, NL interface should make use of a heterogeneous environment feasible for users and, if done well, transparent. Otherwise, the user will be limited by the complexity, idiosyncrasy, and diversity of the computing environment.

Such interfaces will be seamless in at least two senses:

The user can state information needs without specifying how to decompose those needs into a program calling the various underlying systems required to meet those needs. Therefore, no seams between the underlying systems will be visible.

The interface will use multiple input/output modalities (graphics, menus, tables, pointing, and natural language). Therefore, there should be no seams between input/output modalities.

Although acoustic and linguistic processing can determine what the user wants, the problem of translating that desire into an effective program to achieve the user's objective is a challenging, but solvable problem.

In order to deal with multiple underlying systems, not only must our NL interface be able to represent the meaning of the user's request, but it must also be capable of organizing the various application programs at its disposal, choosing which combination of resources to use, and supervising the transfer of data among them.

Partial understanding of fragments, novel language, and errorful language

It is time to move away from dependence on the sentence as the fundamental unit of language. Historically, input to NL systems has often had to consist of complete, well-formed sentences. The systems would take those sentences one at a time and process them. But language does not always naturally occur in precise sentence-sized chunks. Multi-sentence input is the norm for many systems that must deal with newspaper articles or similar chunks of text. Subsentence fragments are often produced naturally in spoken language and may occur as the output of some text processing. Even when a sentence is complete, it may not be perfectly formed; errors of all kinds, and new words, occur with great frequency in all applications.

11. Multi-sentence input

Historically, computational linguistics has been conducted under the assumption that the input to a NL system is complete sentences (or, in the case of speech, full utterances) and that the output should be a complete representation of the meaning of the input. This means that NL systems have traditionally been unable to deal well with unknown words, natural speech, language containing noise or errors, very long sentence (say, over 100 words), and certain kinds of constructions such as complex conjunctions.

12. Errorful language; including new words

Handling novel, incomplete, or errorful forms is still an area of research. In current interactive systems, new words are often handled by simply asking the user to define them. However, novel phrases or novel syntactic/ semantic constructions are also an area of research. Simple errors, such as spelling or typographical errors resulting in a form not in the dictionary, are handled in the state-of-the-art technology, but far more classes of errors require further research.

The state-of-the-art technology in message understanding systems is illustrative. It is impossible to build in all words and expressions ahead of time. As a consequence, approaches that try for full understanding appear brittle when encountering novel forms or errorful expressions.

The state of the art in spoken language understanding is similarly limited. New words, novel language, incomplete utterances, and errorful expressions are not generally handled. Including them poses a major roadblock, for they will decrease the constraint on the input set, increase the perplexity of the language model, and therefore decrease reliability in speech recognition.

The ability to deal with novel, incomplete, or errorful forms is fundamental to improving the performance users can expect from NLP systems.

13. Conclusion

We feel that knowledge acquisition, interaction with multiple underlying systems, and techniques for partial understanding are the three solvable problems that will have the most impact on the utility of natural language processing. The norm in the present generation of user environments is distributed, networked application. A seamless, multi-modal, natural language system should make use of a heterogeneous environment feasible for users, otherwise, the users may be limited by the complexity, idiosyncrasy, and diversity of the computing environments.

14. References

1. Bates, M, and Weischedel, R.M. (1993). Challenges in Natural Language Processing. Cambridge University Press. pp3-33.
2. Harris, Mary Dee. (1985). Introduction to Natural Language Processing. Reston Publishing. Company, Inc. pp55-66.
3. Bates, M., Boisen, S., and Makhoul, I (1991). "Developing an Evaluation Methodology for Spoken Language Systems", DARPA Speech and Natural Language Workshop, Hidden Valley, PA, Morgan Kaufmann Publishers, pp. 102-108.
4. Bobrow, R., Ingria, R., and Stallard, D. (1991). "Syntactic and Seminatic Knowledge in the DELPHI Unification Grammar," DARPA Speech and natural Language Workshop, Hidden Valley, PA, Morgan Kaufmann Publishers, pp. 230-236.
5. Neal, J., and Walter, S. (editors). (1991). Natural Language Processing Systems Evaluation. Workshop, Rome Laboratory.
6. Weischedel, R. M., Carbonell, J., Grosz, B., Marcus, M., Perrault, R., and Wilensky, R. (1990). Natural Language Processing, Annual Review of Computer Science, Vol.4, pp. 435-452.

About Author

Mrs. Veena A Prakashe is presently working as Information Scientist in Nagpur University Library. She holds M.Sc. (CSc.), MLISc, Diploma in German Language. She looking after the Computerization and Networking of Nagpur University Library, UGC- Info-Net Project. Research Experience : Worked on various Expert systems/ KBS projects of DoE and MEF, GoI, in CEERI, Pilani and NEERI, Nagpur. (Both CSIR Laboratories). Publications : a) A Book titled "DBASE III PLUS" in 1992, published by Pitamber Publishing House, New Delhi, financed by DoE, GoI. b) 9 papers published in the Conf. Proc. of various National Conferences. Membership : A life member of IWSA (Indian Women Scientist Association)

E-mail : sh_veena@hotmail.com