International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

NLP-The Future of Machine Learning

Sameer*

Shah Satnam Ji P.G Boys' College, Sirsa, India

Accepted 12 April 2017, Available online 29 April 2017, **Vol.7, No.2 (April 2017)**

Abstract

Natural Language Processing is the hot cake of research scholars today due to its myriad applications areas especially Computational Linguistics, and Speech Recognition and Synthesis. NLP allows computers to not only interact with humans, but also interact with fellow computers. This is the earliest area of Artificial Intelligence which got great attention as it is helpful in spelling correction, grammar checking, information retrieval and machine translation. Although computers lack common sense and logic, which cause herculean problems, however this paper deals with a solution-based approach for NLP.

Keywords*: Artificial Intelligence, Natural Language Processing, Accent, Speech Recognition*

Introduction

Natural Language Processing is one of the major areas, which is currently being achieved by research and development teams of AI all over the world. Natural language processing is a big paradigm broadly it is defined in either of the following two ways.

1) *Voice recognition*- This study describes us about the ways towards the understanding of a human language not by any commands, but with the help of vibrations and sounds which a human produce during any conversation or at the time when he express himself. This approach deals with verbal communication between humans and computers/ humans.
2) *Text manipulation*- This definition concentrates on understanding a written paper of natural language (like a book or newspaper). This approach is quite dissimilar from voice recognition and more practical as grammar is strictly followed in written form instead of verbal field.

These two ways of human expression scenario is developed in thousands of years, with many variations and environmental effects. There are several languages in the world and several ways to write them; making a general AI application to understand even the slightest of any language is a big challenge. However a lot of progress has been done in this field with the rise of technology and sciences. When we compare the two fields of NLP, the theory of text manipulation is at better progress than voice evolution methods. But still there are lot of problems in both areas of NLP. One such problem is "Intonation", which deals with the focus on each and every word of a written sentence. But sometime it is hard to understand the meaning due to complex hierarchy of words; this process is called Intonation. Other problems in understanding of NLP are discussed in the following section.

Problems encountered in understanding human language

Understanding a language means knowing what concepts a word or phrase stands for and knowing how to link those concepts together in a meaningful way. It is a big irony that natural language, the symbol system that is easiest for humans to learn and use, is hardest for a computer to master. Long after machines have proven capable of inverting large matrices with speed and grace, they are still incompetent to master the basics of human spoken and written languages.

1) *Different way to say the same thing*- Humans generally speak according to their comfort as well as requirement of the situation. Different people say the same statement differently to convey the message as everyone has their own way to express it. This causes a chaotic situation where the pattern of words does not match although the meaning is same. Hence interpretation of the words and in-turn deduction of obvious meaning becomes unfeasible. Moreover a large database is to be maintained to match each case to find exact meaning of the sentence.
2) Example-"Please bring a glass of water". "Bring me a glass of water". "I need a glass of water".

*Corresponding author: Sameer

3) Furthermore a lot of rules need to be defined and followed to match spoken sentences with the defined database.
4) *Spoken English is far different from written English*- Due to the excess change in the nature of spoken language; spoken languages got too much variation. Generally syntax of grammar is not properly followed by people while speaking. Additionally, there are silent words and letters in English; and everyday new words are introduced without any special need to make the language more complex and chaotic. These real-world changes make it tricky for a computer to understand natural language because it only follows rules of grammar and know only the dictionary words.
5) *Unlike words may have same phonetics*- There exist some words with different meaning but having same speaking pattern. For example: Words "*their"* and *"there"* have same representation phonetically. These two words have different meaning but when used in a sentence it is pretty complicated for a computer to determine which one has been used in the sentence.
6) *Distinct parse tree for same sentence*- from beginning in 1970, the researches used the technique of parse trees to assemble and understand a sentence in pieces (subjects, object, verb, adjective etc. But same sentence can give multiple parse trees. Like in example provided. Some words can act as verb and as well as noun. So it creates problems in understanding sentence structure for a machine.

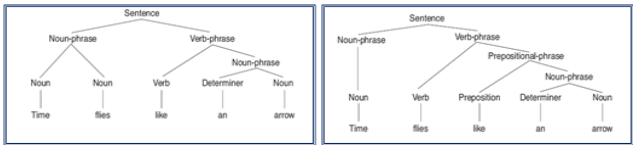


Figure 1: Two different ways to parse a given sentence

7) *Usage of disparate accents*- Accent can be defined as specific way or dialect of speaking a language according to geographical area. There are many variation of a speaking language according to diverse regions of the world. For example: English has several popular accents namely British, American, Australian, Indian, etc. However, people in America also have northern and southern accents. But when a Frenchman try to speak English there is another variation present of the English language with some leg and foots. This creates problems for the machine to understand the words in different accents.
8) *Use of Idioms*- Idioms is described as the sentences which are used to refer something indirectly. The words which form sentences are not actually contain the literal meaning. Instead they always have a hidden meaning and they refer to something else. When a machine tries to process idiom it will not understand what the user actually wanted to say, unless and until it infers the real meaning of sentence rather than literal meaning.
9) *Same word conveys distinctive meanings at different times*- Every language has some words which express disparate meaning according to the sentences in which they are used. For example -
   - "This is the season of cold."
   - "He has been caught by cold."

   The first sentence describe the autumn season while second one illustrates an illness. This dilemma for a machine can cause turmoil in understanding and figuring out gist of the language. Here is another example to prove the point.
   - "He is rough while talk."
   - "The Paper should be a bit rough to get a better hold on ink."
10) Complex to comprehend the expression- The language especially in spoken form is more often affected by human expressions than any other thing. Individuals use many expressions in place of words and sometimes leave the sentence incomplete by laughing or any other activity. Humans can understand this type of hidden language but it is not easy for a computer to estimate what the speaker is up to.
11) Ambiguous sentence structure- English and several other languages sometimes don't specify which word an adjective applies to. For example, in the string "pretty little girls' school".
    a) Does the school look little?
    b) Do the girls look little?
    c) Do the girls look pretty?
    d) Does the school look pretty?
12) Problems with pronouns- Pronouns are used for refer nouns or objects indirectly. The pronoun issues must be resolved effectively for better understanding of language. This will be clear by an example- -"We give apples to boys because they were good". Now the word "they" may refer either to apples or to boys. Both are perfect replacement for the word "they". As this situation is ambiguous even for humans; interpretation of sentence by a machine will be a lost challenge.

The above mentioned points are some major criteria on behalf of which it can be easily realized that human making a machine understanding natural language is a herculean task. Still over the fifty years of hard work and the use of ultimate power of computer to infer exact meaning of written or spoken language is not a daydream. Scholars are at the stack of success (at least in text manipulation) and have developed some predefined methods like Parse tree generation to understand a NLP. Researchers have already fabricated a 6- step based method to do a text manipulation based NLP.

Current enhancement in NLP

Today many softwares have been build which are valuable for Natural Language Processing and understanding. These softwares are build on the central idea of disintegrating problems into phasis,

then use solutions to remove problems at each phase and make parse trees to understand a particular sentence. Some of the solutions to basic problems of NLP already identified by research scholars are as follows -

1) *Word level structure*- Existing NLP softwares are developed with fundamental plan to break up every sentence into individual words, so that each word can be processed and understand separately. By doing so each word will be judged on the basic of dictionary meaning as well as sentence orientation.
2) *Pronouns and reference resolving (Discourse Integration)*- In general practice, persons take help of the previous line to resolve out the references due to a pronoun. For example in the sentence - "Sam got an apple. He ate it."
3) Here, "He" will be replaced by "Sam". But this method is not proven to be correct for each complex sentence.
4) *Semantic structure analysis*- This field of NLP is totally concerned with the meaning for a particular sentence. If the machine is able to find out a particular meaning, it is ok, otherwise it will generate an error message. It checks the integrity of the sentence by parse tree analysis.
5) *Pragmatic Analysis*- In this analysis, the software checks what type of sentence it is, whether a command, a question, a statement or a request. On this basis, the machine decides its execution sequence to get the correct meaning of the sentence.

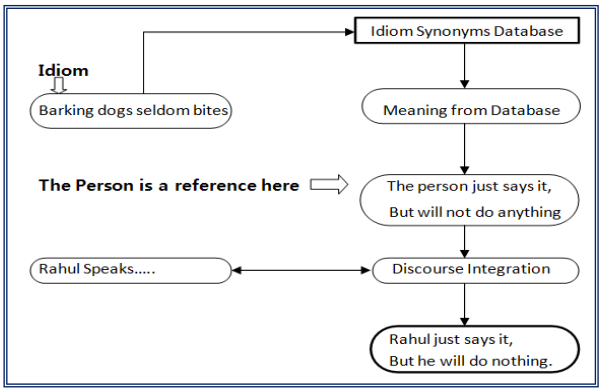
Solutions for NLP problems

Before the classification is done, people must understand that as human is not perfect as god; similarly machines are not ideal as humans in understanding human language. If academics want to make a machine to think like a man, they must also accept the facts which lead to imperfect nature of a human being.

As per the neuron network is concerned as the back propagation and generic algorithms are used to establish the human ways to manipulate things in machine there are some outsource way to make the execution and understanding easy. These ways are on conceptual basis not any programming basis. Some novel solutions for solving primary NLP problems are discussed as below:-

1) *The power to guess*- An underlying hard fact about human nature is that they are not familiar to every word of a natural language; especially in case of a foreign language (Example: English is foreign language for Indians. In case when humans don't know meaning of any particular word, then they simply guess it according to structure of the sentence. The same mechanism can be applied for a machine. There is no need to make a complete database; instead there should be references of some lines. Through a match–guess process, machines can achieve the "power to guess" and can help in case where computer can't understand a spoken word.
2) *The ability to ask and cross questioning*- If a computer does not find a good match for a word or sentence; then a problem solving strategy to employ is to simply ask the speaker what does he really wants to speak. The motto behind this point is make computer more interactive. There should be special interrupts for calling a procedure to ask queries. This functionality will also be of great use in case of idioms; as computer can query the speaker if computer is unable to determine the real meaning of an idiom.
3) *Resolving Idiom issue*- The idioms are the special meaning phrases, which are different from usual sentences. Idioms always have a dissimilar meaning than their literal meaning. As humans know idioms always have a different meaning other than their literal meaning and they are having references to different persons at diverse time and place. As Idioms are extensively used in any spoken language, hence there is a gigantic need to make them understood by machines. Solving idiom issues to understand them fully can be achieved by using following two techniques: *Synonyms database* and *Discourse integration.* Let us take an example -

"Rahul always praise himself but u know barking dogs seldom bite"



4) ***Heuristic comparison of different accents-*** As different people speak in different tone and manner, human can't neglect the point that they are speaking the same thing. There is only small variation in terms of vibration of voice that is due to change in vocal abilities. In place of checking exactly the same vibration, a range of vibration should be checked to uncover the heuristic pattern. For text, this mechanism is already established. Another technique that is being developed now-a-days is to create a font in human writing. Amplifying a sin wave is the simplest way to do it. The given input should be treated as a distorted one; a "nearby Similarity check" for signal can be

applied to correct it. Signal is going to be checked against the predefined signals of voice sample and will be converted to the wave which is most resemblance to the predefined one.

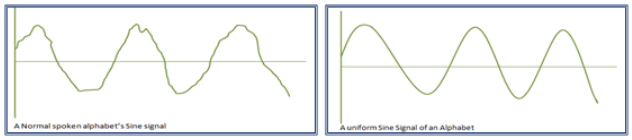


Figure 4: Correcting Sine signal of an alphabet

5) *Multi- synonyms/ instance based Database-* Every word that is being used at more than one instance should have an unclassified, incomplete multi-synonym definition in the knowledgebase. The instance at which the word is used should be saved with special keywords (if possible) so that machine can indentify its behaviour next time. Also before apply a particular solution to a multi-meaning word, every instance should be checked. It will also help in the situation where two words have same phonetics.

Table 1: An example of multi- synonyms database

Word	Meaning	Situational Reference	Tags
Key	Solution to a problem	Work is the key to success.	Work, Adjective sentence structure
	Object to open locks	He lost his key.	Subject oriented, physical object, have a shape
	Text to unlock a software	This software requires activation key.	Virtual object, must be written somewhere, Software

6) *Learning the internet language (abbreviations)-* Beyond the knowledge base and updating in database due to new occurrences, the machines should also consider learning of first letters for Internet language or SMS language or longer names. For such situations, the place of such words in the sentence is very vital. There must be some protocol for a robot which defines special meaning according to this place. For example: Dr. in a hospital refers to doctor working as a medical professional, while Dr. in a college will denote a professor working in academics. Also a name database can be created to store name with initials.

Conclusion

Results of some NLP experiments reported in this paper show encouraging results. This paper provides a solution based approach to NLP rather than problem based. First of all, some common problems that occur in NLP are discussed and then some novice and practical solutions are provided. The paper also proposes new concepts like solving Idiom issues and finds out what possible resolution could be applied to these problems. The paper also gives insight into the structure of the knowledge-based for natural language processing. Though machine translations are not always perfect and do not produce as good translations as human translators would produce, the results, and evidences of interests in improving the performance level of NLP systems, are very encouraging. Finally it can be concluded that machines are not perfect as human, but still it can be made intelligent by novice solutions provided in the paper.

References

Allen, (1987) *Natural language understanding.* Menlo Park, CA: The Benjamin Cummings Publishing Company, Inc.

Joshi, (1999), Supertagging: an approach to almost parsing. Computational Linguistics, 25, 237-265.

Barahona ; Alferes September (1999), Progress in Artificial Intelligence. 9th Portuguese Conference on Artificial Intelligence, EPIA'99. Proceedings, Evora, Portugal. Berlin: Springer-Verlag.

Bates ; Weischedel, (1993) *Challenges in natural language processing.* Cambridge: Cambridge University Press

Carballo ; Strzalkowski (2000) Natural language information retrieval: progress report. Information Processing & Management, 36, 155-178.

Charniak, (1995) Natural language learning. ACM Computing Surveys, 27, 317-3319.

Jurafsky ; Martin, (2000) Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition. Upper Saddle River, NJ: Prentice Hall.

Kay (1976), *Webster's Collegiate Thesaurus. G.* & C. Merriam Co.

King (1996) Evaluating natural language processing systems. Communications of the ACM, 39, 73-80.

Kupiec, (1993), MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In R. Korfhage, E. Rasmussen and P. Willett (eds.) *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval.* New York: Association for Computing Machinery. 181-190.

McRoy (1992) Using multiple knowledge sources for word sense discrimination. *Computational Linguistics* 18.1.1-30.

Rosner ; Johnson (1992) *Computational linguistics and formal semantics*. Cambridge: Cambridge University Press.