ISSN 2347 - 3983
Volume 4, No.10 October 2016
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter034102016.pdf

Insight to Computational Linguistics

Yogita Bansal
Assistant Professor (IT), JIMS, Sector-05, Rohini, New Delhi. INDIA.
yogita.sharma@jimsindia.org

ABSTRACT

The field of computational linguistics (CL), together with its engineering area of natural language processing (NLP), has burst out in recent years. It has emerged rapidly from a relatively unclear accessory of both AI and formal linguistics into a blooming scientific discipline. It has also become an important area of business development. The focus of research in CL and NLP has shifted over the past three decades from the study of small prototypes and theoretical models to robust learning and processing systems applied to large corpora [1]. For the last two centuries, human race has effectively coped with the computerization of many tasks using automatic and electrical devices, and these devices realistically help people in their everyday life. In the second half of the twentieth century, human consideration has turned to the automation of natural language processing. Community now wants assistance not only in automatic, but also in rational efforts. They would like the machine to read an unwary text, to test it for correctness, to carry out the instructions contained in the text, or even to realize it well enough to produce a reasonable reply based on its meaning. Intelligent natural language processing is based on the science called computational linguistics. Computational linguistics is closely connected with applied linguistics and linguistics in general [2].This paper intends to provide an introduction to the major areas of CL, and an impression of current work in this area.

Keywords: Computational linguistic, natural language processing, machine learning, corpora

1. INTRODUCTION

Computational linguistics is the application of linguistic theories and computational techniques to problems of natural language processing [3]. The goal of computational linguists is to write programs that can understand or generate as much natural language material as possible. These programs are good but only works in approximate ways; they cannot deal with all the matter in natural language, although they are capable of handling most common and interesting constructions. This fact, which is generally acceptable to a computational linguists, but may unacceptable for theoretical linguists, since it is part of their goal to find solution for all grammatical sentences of a language with their theory of grammar. The pragmatic work in Computational Linguistics views language understanding and generation as processes of symbol-manipulation in a rule-governed fashion. While theoretical Linguistics is interested in all characteristics of the language ability (the abstract knowledge of language and how it is used), its work should be testable by systems that are designed by computational linguists [4]. The task of constructing systems that understand or generate natural language is a complex one. It requires the incorporation of many kinds of data such as linguistic data (syntactic & semantic) and non-linguistic (knowledge of the domain of dialogue). It also requires an effective and efficient use of all data. In this sense, we may categorize the task of designing and building a natural language application as an engineering task. One general strategy to make design process easier is modularity, (i.e., dividing the problem into smaller sub problems). This notion is not unfamiliar to linguistics.

The language capacity can be represented and studied as hierarchy in levels of structure such as sounds, words and sentences. Linguists study the phonetics, the phonology, the morphology, the syntax and semantics of a language and they assume the existence of levels or modules in human competence (e.g., Chomsky's Autonomous Syntax Principle). This hierarchical view makes natural language systems flexible and easy to expand. Just how much knowledge is used in the understanding or generation process depends on the purpose of the application. For many applications the essential task is analyzing sentences, (i.e., determining what sentences mean). Some applications also require an analysis of corpora units, such as discourse and dialogue.

2. APPROACHES

The following section provides details for some of the text available across the entire field broken into four main approaches for area of discourse: developmental linguistics, structural linguistics, linguistic production, and linguistic comprehension. [5]

2.1 Development Approaches

The ability of infants to develop language is a marvelous example of how human can communicate and understand inherent ambiguity. This fact has also been modeled using robots in order to test linguistic theories. A robot is enabled to learn as children based on an affordance model in which mappings between actions, perceptions, and effects were created and linked to spoken words. These robots were able to perform word-to-meaning mappings without needing grammatical structure, vastly simplifying the learning process and shedding light on information which enhances the current understanding of linguistic development. It is important to note that this information can only be empirically tested using a computational approach.

Till now understanding is that the linguistic development of an individual is continually improving within a lifetime. This perception is also used in robots to continually improve their performance using neural networks. One major point to keep in mind is that languages themselves change and develop through time. Computational approaches to understanding this phenomenon have revealed very interesting information. This modeling effort achieved, through computational linguistics, what would otherwise have been impossible.

It is clear that the understanding of linguistic development in humans has been tremendously improved because of advances in computational linguistics. The ability to model and modify systems at will provides science principled method of testing hypotheses that would otherwise be intractable. [6, 7]

2.2 Structural Approaches

Language structure understanding is required to create better computational models of language. Till date the English language has been thoroughly studied using computational approaches to better recognize how the language works on structural level. Availability of large linguistic corpora is one of the most important pieces for being able to study linguistic structure. This provides computational linguists the raw data necessary to run their models and gain a better understanding of the underlying structures. These structures are present in the vast amount of data which is contained in any single language. One of the most cited English linguistic corpora is the Penn Treebank[1]. It contains over 4.5 million words of American English; this corpus has been tagged for part-of-speech information. This type of tagged corpus allows other researchers to apply hypotheses and measures that would otherwise be impossible to perform without the added information such as tagging part-of-speech.

Theoretical approaches are also developed to understand the structure of languages. These works allow computational linguistics to have a framework within which to work out assumptions made that will further enhance the understanding of the language in countless number of ways. One of the original theoretical theses on internalization of grammar and structure of language proposed two types of models. In these models, rules or patterns learned increases in number as the frequency of their usage increases. This work also created a question for computational linguists that how do an infant learn a specific and non-normal grammar (Chomsky Normal Form) without learning an over generalized version and getting stuck? [8] Theoretical efforts like these are crucial for the growth of field. They set the direction for research to go early in the lifetime of a field of study.

2.3 Production Approaches

Comprehension is only half the problem of communication. The other half is how a system produces language. Computational linguistics has made some very interesting discoveries in this part. ELIZA program is one of the earliest and best known examples designed to communicate with humans. It appears as if one is talking to another human only. It was developed by Joseph Weizenbaum at MIT in 1966. The program enacted a Rogerian psychotherapist in response to written statements and questions posed by a user. It looked talented of understanding what was said to it and responded intelligently, but in truth it simply followed a pattern matching routine. The machine routine relied on only understanding a few keywords in each sentence. Its responses were generated by recombining the unknown parts of the sentence around correctly translated versions of the known words. For example, in the phrase "It seems that you love me" ELIZA understands "you" and "me" which matches the general pattern "you [some words] me", allowing ELIZA to change the words "you" and "me" to "I" and "you" and replying "What makes you think I Love you?". In this example ELIZA has no understanding of the word "Love", but it is not required for a logical response in the context.

There are some problems which first started computational linguistics off as its own field. Some projects are still trying to solve these problems. Nowadays the methods have become more advanced and clever, and hence the results generated by computational linguists have become more informative. For example in an effort to improve computer translation, several models have been used and compared. This includes hidden Markov models, smoothing techniques, and the specific refinements of these such that they may be applied to verb translation. The model which was found to produce the most accepted translations of German and French words was a refined position model with first-order dependence and a fertility model. These models also provide efficient training algorithms, which gives other scientists the ability to enhance their results. This type of work could vastly improve understanding of how language is produced and comprehended by computers.

2.4 Comprehension Approaches

Much of the focus of modern computational linguistics is on language understanding. With the increase of the internet resources and the large quantity of easily accessible written human language material it is easy to write a program that may understand the inherent properties of human language. This ability to create a program capable of understanding human language would have many broad and exciting potential such as improved search engines, automated customer service, and online education.

[1] http://www.cis.upenn.edu/~treebank/home.html

Early work in comprehension included applying Bayesian statistics to the task of optical character recognition, as illustrated by Bledsoe and Browing in 1959 in which a large dictionary of possible letters were generated by "learning" from example letters and then the probability that any one of those learned examples matched the new input was combined to make a final decision [9]. Other attempts at applying Bayesian statistics to language analysis included the work of Mosteller and Wallace (1963) in which an analysis of the words used in *The Federalist* Papers was used to attempt to determine their authorship (concluding that Madison most likely authored the majority of the papers)[10].

In 1971 Terry Winograd developed an early natural language processing engine capable of interpreting naturally written commands within a simple rule governed environment. The primary language parsing program in this project was called SHRDLU, which was capable of carrying out a somewhat natural conversation with the user giving it commands, but only within the scope of the toy environment designed for the task. This environment consisted of different shaped and colored blocks, and SHRDLU was capable of interpreting commands such as "Find a block which is taller than the one you are holding and put it into the box." and asking questions such as "I don't understand which pyramid you mean." in response to the user's input. While impressive, this kind of natural language processing has proven much more difficult outside the limited scope of the toy environment [11]. Similarly a project developed by NASA called LUNAR was designed to provide answers to naturally written questions about the geological analysis of lunar rocks returned by the Apollo missions. These kinds of problems are referred to as question answering [12].

3. APPLICATIONS

Modern computational linguistics is often regarded as a combination of studies in computer science and programming, math, specially statistics, language structures, and natural language processing. Collectively, these fields most often lead to the development of systems that can recognize speech and perform some assignment based on that speech. Examples include softwares, such as Dragon, Apple's Siri feature, iListen, ViaVoice, spell-check tools and speech synthesis programs. These often help in working in smarter and more productive ways. With fast, accurate dictation and transcription, advanced customization, seamless integration across devices, and easy deployment for large enterprises, it uses ones voice—and put it to work. They are also used to help the disabled, and machine translation programs and websites, such as Google Translate and Word Reference [13].

Speech synthesis and recognition deals with how spoken language can be understood or created using computer programs. Parsing, lexical analysis, semantic analysis and generation are sub-divisions of computational linguistics. Machine translation remains the sub-division of computational linguistics dealing with handling computer translation between languages. The possibilities of automatic language translation still have many open problems yet to be realized and remain a notorious branch of computational linguistics.

Computational linguistics can be especially helpful in matters involving social media and the Internet. For example, filters in chat rooms or on website searches require computational linguistics. Chat operators often use filters to identify certain words or phrases so that users cannot submit them by marking them inappropriate. Another example of using filters is on websites. Schools and parents often use these feature to block some websites. They use filters so that certain keywords are matched against the content to restrict the usage. There are also many other programs available in which parents use Parental controls to put content filters in place. Computational linguists can also develop programs that may cluster and arrange content through Social media mining. An example of this is Twitter, in which programs can group tweets by subject or keywords [14].

Computational linguistics can be divided into major areas depending upon the medium of the language being processed, whether spoken or textual; and upon the task being performed, whether analyzing language (recognition) or synthesizing language (generation).[15]

Some of the areas of research that are studied by computational linguistics include [16]:

- Computational complexity of natural language largely modeled on automata theory.
- Computational semantics which consists of defining suitable logics. This comprises meaning representation, automatically constructing them and reasoning with them for linguistic purpose.
- Computer-aided corpus linguistics used for discourse analysis
- Design of parsers or chunkers for natural languages
- Design of taggers like POS-taggers (part-of-speech taggers)
- Machine translation as one of the earliest and most difficult applications of computational linguistics.
- Simulation and study of language evolution in historical linguistics.

4. CONCLUSION

Computational linguistics studies language production, comprehension and acquisition and ask questions such as what information is used, and how is it used? It may give insights into language disorders, and suggest possible Solutions for it. Natural language processing uses computers to process speech and texts for tasks such as information retrieval, extraction and summarization. Machine translation can be done using human-computer interface, statistical models. Machine learning plays a central role in both. Theory and practical applications interact in a productive way. The state of the art and the near term future can be explained with help of sample scenarios such as: – Generation of weather reports in several languages, Translation of Web pages/written text into different languages, Talking to our appliances such as “Speak to our phone for search” or “Find restaurants”, “Answer questions from clients on the phone”, “Grade GRE/SAT essays”.

ACKNOWLEDGEMENT

I would like to thank all the people who have helped me to give the knowledge about this research paper. Finally, I would like to thank all the websites and, International Journal of Emerging Trends in Engineering Research to create our research paper successfully.

REFERENCES

1. Clark A., Fox C., Lappin S. ***The Handbook of Computational Linguistics and Natural Language Processing***, Wiley-Blackwell publications,2010.
2. Igor A. Bolshakov and Alexander Gelbukh, **Computational linguistics models**, Resources, Applications, IPN-UNAM-FCE, 2004,
3. http://www.ba.umist.ac.uk/public/departments/registrars /academicoffice/ uga/lang.htm
4. Grishman, R. **Computational Linguistics**. Cambridge University Press. 1986.
5. https://en.wikipedia.org/wiki/Computational_linguistics
6. Salvi, G., Montesano, L., Bernardino, A., & Santos-Victor, J. **Language bootstrapping: learning word meanings from perception-action association**. IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society, 42(3), 660-71, 2012.
7. Gong, T.; Shuai, L.; Tamariz, M. & Jäger, G.. E. Scalas, ed. **Studying Language Change Using Price Equation and Pólya-urn Dynamics. PLoS ONE**. 7 (3): e33171.doi:10.1371/journal.pone.0033171,2012
8. Braine, M.D.S. **On two types of models of the internalization of grammars**. In D.I. Slobin (Ed.), The ontogenesis of grammar: A theoretical perspective. New York: Academic Press,1971
9. ledsoe, W. W. & Browning, I. **Pattern recognition and reading by machine**. Papers presented at the December 1–3, 1959, eastern joint IRE-AIEE-ACM computer conference on - IRE-AIEE-ACM ’59 (Eastern).1959
10. Mosteller, F. **Inference in an authorship problem**. Journal of the American Statistical Association, 1963.
11. Winograd, T. ***Procedures as a Representation for Data in a Computer Program for Understanding Natural Language***, 1971.
12. Woods, W.; Kaplan, R. & Nash-Webber, B. ***The lunar sciences natural language information system***, 1972.
13. Blei, D. & Ng, A. **Latent dirichlet allocation**. The Journal of Machine Learning.3:2003.
14. **Careers in Computational Linguistics**. California State University. Retrieved19 September 2016.
15. Oettinger, A. G. (1965). ***Computational Linguistics***. The American Mathematical Monthly, Vol. 72, No. 2, Part 2: Computers and Computing, 1965.
16. McEnery, Thomas (1996). ***Corpus Linguistics: An Introduction***. Edinburgh: Edinburgh University Press. p. 114, 1996.