ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

Applications of Computational Linguistics to Language Studies: An Overview

Sakthi Vel S.
University of Kerala, Thiruvananthapuram, Kerala, India

***Abstract* - Computational Linguistics is the scientific study of human languages. It is one of the most relevant interdisciplinary field of Linguistics and Computer science. Intelligent Natural Language Processing (NLP) is based on the science called Computational Linguistics (CL). It is closely connected with General Linguistics and Applied Linguistics. Computational linguistics might be considered as a synonym of automatic processing of natural language, since the main task of computational linguistics is to construct computer programs to process Text/Speech in natural languages [4]. The paper attempts to provide an introductory view about computational linguistics and its various applications to language studies. The paper has three sections. The first section deals with introduction to computational linguistics, its major divisions (applied and theoretical) and some fundamental ideas about the topic. Second section deals with structure of computational linguistics -a tree diagram representation. The third section describes the various applications of computational linguistics in close association with language studies. Finally the paper summed up with conclusion and list of references.**

***Index Terms:*— Linguistics, Computer Science, Computational Linguistics, Human Languages, Text, Speech.**

I. INTRODUCTION

Computational linguistics is the scientific study of human languages from a computational perspective. Computational linguists provide computational models of various types of linguistic phenomena [1]. Computer oriented various studies on languages have emerged into a hybrid type called computational linguistics [6]. As an inter disciplinary field computational linguistics has a history of more than forty years [6]. The ultimate aim of computational linguistics is to describe the basic techniques that are used in building computer models for natural languages production and comprehension.

As the power, potentiality and sophistication of computers increased, the use of computers have been extended to the non-numerical data like graphics, symbols, etc., and one such application is the study of texts and speech in natural languages. The study has grown into an extent of establishing itself as a separate discipline called Computational Linguistics or Natural Language Processing [8]. We live in the age of digital information. The various sources of its access are pages of newspapers, magazines, radio, TV and computer screens. The major portions of these information are in the form of natural language text or speech [4]. It all describes the work of the inter-disciplinary field called computational linguistics that arises from the research in Artificial Intelligence (AI) [9].

The goal of AI is to create software products that have some knowledge of human language. The major task of the computational linguist is to develop a computational theory of language, using the notion of algorithms and data structures from computer science [8]. Language can be studied from the point of view of *structure* and *use*. The Study of language structure is called *structural* or *formal linguistics* and language use is called as *functional linguistics* [7]. Computational Linguistics is the product of the inter relation between linguistics and computer science, which is concerned with the computational form of the human language. The main aim of computational linguistics is to use computers as a tool to interface, understand, generate or implement linguistics theories. Today's computers do not understand human language more over computer languages are difficult to learn and do not correspond to the structure of human thought.

Computational linguistics is a field of vital importance in the information age. Computational linguists create tools for important practical tasks such as machine translation, speech recognition, speech synthesis, information extraction from text, grammar checking, text mining and more [1]. Computational linguistics is the study of computer systems for understanding and generating natural languages [3]. The tools that work in computational linguistics make use of artificial intelligence: algorithms, data structures, formal models for representing knowledge, models of reasoning process and so on [6].

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

II. THE STRUCTURE OF COMPUTATIONAL LINGUISTICS- A TREE DIAGRAM

Computational Linguistics is an interdisciplinary field concerned with the statistical or rule-based modeling of natural languages from a computational perspective. The prime aim of computational linguistics is to create software/ computer systems that understand and speak any human languages or natural languages. Intelligent natural language processing is also based on the science called computational linguistics. Computational linguistics is closely connected with applied linguistics and linguistics in general [4]. Computational linguistics might be considered as a synonym of automatic processing of natural languages, since the main task of computational linguistics is to construct computer programs to process speech and texts in natural languages [2]. Computational Linguistics can be divided into two major areas depending upon the medium of the language being processed:

1) Text processing and
2) Speech processing

Computational linguistics has two components:

A. Theoretical Computational Linguistics
It takes up issues in theoretical linguistics and cognitive science. It deals with formal theories about the linguistic knowledge that human needs for generating and understanding languages. The essential components of language or the building blocks of languages are sounds, words, sentences and meanings. All these elements are equally important. They give rise to understand the building blocks of language at different levels of structuring: phonology, morphology, syntax, semantics and discourse [5].

B. Applied Computational Linguistics
It focuses on the practical outcome of modeling human language use. The methods, techniques, tools and applications used in the applied area is also termed as Human Language Engineering (HLE)/ Human Language Technology (HLT). Applied computational linguistics is now widely used in language studies, business, and scientific research & development domains for many purposes [10, 11]. The following tree diagram shows the different branches and levels of computational linguistics in tune with the study of languages.

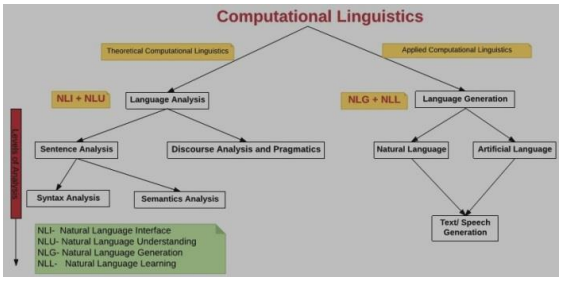

Fig. 1: Different branches of Computational Linguistics

III. MOTIVATIONS OF COMPUTATIONAL LINGUISTICS

A. First, the linguistics or cognitive motivation
A linguistics motivation is to gain a better understanding of how humans communicate by using natural languages. This motivation is shared with theoretical linguistics and psycholinguistics [6].

B. Second, the technological motivation
Technological motivation is to build intelligent computer systems, such as from Natural Language Interface (NLI) to databases, automatic machine translation systems, text analysis systems, speech understanding systems and computer- aided instructions systems [6].

IV. APPLICATIONS OF COMPUTATIONAL LINGUISTICS

A. Primary Applications
The primary applications concerned about the development of specific practical systems which involves the use of natural languages. Three classes of application which have been central in the development of computational linguistics are:

***1. Machine Translation*:**
The Work in relation with Machine Translation began in the late 1950s. Machine translation is the application of computers to the translation of texts from one natural language into another language equivalent. Machine Translation is an important area of Computational Linguistics. The task of machine translation is highly complex in the sense that it involves, the identification of Parts-Of-Speech (POS), an understanding of grammar, it requires an extensively elaborate task of translating a Source Language (SL) into Target Language (TL) equivalent.

***2. Information Retrieval (IR)*:**
Information Retrieval Systems (IRS) are designed to search for relevant information in large documentary databases [4]. It is to retrieve scientific, technical, business

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

document from a corpus or database. Information Retrieval Systems (IRS) are designed to search for relevant information in large documentary databases.

3. Human Computer Interaction (HCI):
It enables the users to communicate with the computer or any other electronic devices through English, Tamil, Malayalam or any other human languages.

B. Secondary Applications
The secondary applications can be divided into two major classes: Text-based applications and Dialogue-based applications [6, 8].

a) Text- based applications:
Involve the processing of written text, such as books, newspapers, reports, manuals, e-mail message, and so on [6, 8]. Text-based natural language research is ongoing in application such as
1. Finding appropriate documents on certain topics from a data-base of texts (for example, finding relevant books in a library).
2. Extracting information from messages or articles on certain topics (for example, building a database of all stock transactions described in the news on a given day).
3. Translating documents from one language to another (for example, producing automobile repair manuals in many different languages).
4. Summarizing texts for certain purposes (for example, producing a 10-page summary of a 100-page government report) [8].

b) Dialogue-based applications:
Involve human-machine communication. More commonly it is about spoken language, but it also includes interaction using keyboards. Its typical potential applications include: [6, 8].
1) Question-answering systems (QA), where natural language is used to query a database (for example, a query system to a personal database).
2) Automated customer service over the telephone (for example, to perform business transactions or order items from a catalogue).
3) Tutoring systems, where the machine interacts with a student (for example, an automated languages tutoring system).
4) Spoken language control the machine (for example, voice control of a VCR or computer).
5) General cooperative problem-solving systems (for example, a system that helps a person plan and scheduled flight shipments) [6].

V. COMPUTATIONAL LINGUISTICS TRIES TO SOLVE PROBLEMS IN THE FOLLOWING AREAS

- ***Automatic hyphenation***: Hyphenation is intended for the proper splitting of words in natural language texts. When a word occurring at the end of a line is too long to fit on that line within the accepted margins, a part of it is moved to the next line [4].

- ***Spell checking***: The objective of spell checking is the detection and correction of typographic and orthographic errors in the text [4].

- ***Grammar checking***: Detection and correction of grammatical errors by taking into account adjacent words in the sentence or even the whole sentence are much more difficult tasks for computational linguistics and software developers than just checking orthography [3].

- ***Style checking***: The stylistics errors are those violating the laws of use of correct words and word combinations in language, in general or in a given literary genre [4].

- ***References to words and word combinations***: The references from any specific word give access to the set of words semantically related to the former, or to words, which can form combinations with the former in the text. This is one of the most important applications and nowadays it is performed with linguistic tools of two different kinds: autonomous on-line dictionaries and built-in dictionaries of synonyms [8].

- ***Text summarization***: It produces a coherent summary of a set of text. It is used to provide summaries or detailed information of texts of a known type.

- ***Computational Lexicography:*** The objective of computational lexicography is to investigate the design, construction and use of electronic dictionaries in natural language processing. It covers computational methods and tools designed to assist the various lexicographical tasks, including the preparation of lexicographical evidence from many sources such as the recording of relevant linguistic information from the database, the editing of lexicographic and entitles, and the dissemination of lexicographical products [6].

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

- ***Machine Readable Corpus (MRC):*** is useful in automatic lexical analysis. Lexical analysis includes the following procedures, searching for words sequences of words and parts of speech in the texts etc. It is an important procedure in lexical analysis [7]. MRC provides easy ways of searching and manipulating the data and it can be enriched with additional information whenever known information is collected [7]. It has specific linguistic application in domains of linguistics studies such as language teaching, contrastive linguistics and translations [7].

- ***Parts Of Speech (POS)****:* annotation is the process of assigning codes for indicating the parts of speech to the lexical and textual units found in the corpus. It enables easy retrieval of items from the corpus and further analysis of the texts like the parsing analysis and semantic annotation. Parts of speech annotation is needed for certain reasons. It is used by the computers to predict the parts of speech of textual elements from the context [7].

- ***Natural Language Understanding (NLU)***: Natural Language Understanding is a hard task because it requires formulating not only a grammar for the language but also using background knowledge including common sense knowledge [12]. NLU involves the following tasks such as, Mapping the given input in natural language into useful representation and analyzing different aspects of the language [5]. NLU system is organized around the three levels of representation such as Syntactic, logical form and meaning representation.Natural language understanding systems are the most general and complex systems involving NLP [4].

- ***Natural Language Interface (NLI)*:** Natural Language Interface enables the users to communicate with the computer or any other electronic devices in English, Tamil, Malayalam or any other human languages. Some applications of such interfaces are database queries, Information Retrieval from texts, so-called Expert system, and robot control. The task performed by a natural language interface to a database is to understand questions entered by a user in natural language and to provide answers-usually in natural language [5].

- ***Natural Language Generation (NLG)*:** Convert information from computer databases into readable human language.

- ***Natural Language Processing (NLP)*:** NLP gives the machine the ability to read and understand the languages that humans speak. A sufficiently powerful NLP system would enable natural language user interfaces and the acquisition of knowledge directly from human-written sources, such as newswire texts. Some straightforward applications of NLP include Information Retrieval (or text mining), Question Answering (QA) and Machine Translation.

- ***Optical Character Recognition (OCR)*:** Given image representing printed text, helps in determine the corresponding or related text [5].

- ***Multilingual Multimedia content development***: In our communication we mix language with other modes of communication and other information media. We combine speech with gesture and facial expressions. Digital texts are combined with pictures and sounds. Movies may contain language in spoken and written form. Thus speech and text technologies overlap and interact with many other technologies that facilitate processing of multimodal communication and multimedia documents [11, 12].

- ***Word Sense Disambiguation (WSD)*:** Many words have more than one meaning. We have to select the meaning which makes the most sense in context. For this problem, we are given a list of words and associated word senses. E.g. form a dictionary from an online resources such as Word Net.

- ***Natural Language Interface to Data Bases (NLI-DB)*:** Computers have been widely used to store and manage large amounts of data. The data might pertain to railway reservation, library, banking, management, information, and so on. Normally, to use these systems, specialized computer knowledge is necessary. The goal of NLI is to remove this barrier. The user is expected to interact in natural language (by means of a keyboard, and a screen). In Natural language interface to database, users communicate their information need by means of natural language query [14].

- ***Cross Language Information Retrieval* (CLIR):** Cross Language Information Retrieval System aims to break language barrier and make domain information accessible to all users irrespective of language and region. In this system, a user can submit a Natural Language query in a source

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

language and user will be able to access documents available in the language of the query as well as the target language by using a Machine Translation System e.g. MANTRA [7].

- ***Morphological segmentation*:** Separate words into individual morphemes and identify the class of the morpheme.

- ***Sentiment analysis***: Extract subjective information usually from a set of documents, for the purpose of marketing and business [12].

- ***Named Entity Recognition (NER)*:** NER is a process to determine which items in the text relates to proper names, such as people or places, and what type such name or place we are referring belongs to.

- ***Audio-Video Search***: This system aims to focus on extraction and retrieval of information from audio and video sources, not just by their metadata, but by performing search on its content [12]. Like textual information, audio and video files also contain multiple types of audio and visual information which are difficult to extract. For extracting information from such sources the process like automatically recognized Speech Transcripts, image similarity matching etc. are used. This application works on the transcribed text of the audio/video files. [14].

- ***Parsing***: It refers to the parse tree such as grammatical analysis or evaluation of a given sentence [8].

- ***Discourse analysis***: The task is identifying the discourse structure of a connected text, i.e. the nature of the discourse relationships between sentences E.g., elaboration, explanation, contrast.

- ***Topic segmentation and recognition***: Given a chunk of text, separate it into segments each of which is devoted to a topic, and identify the topic of the segment.

- ***Sentence breaking***: Find the sentence boundaries. Sentence boundaries are often marked by periods or other punctuation marks, but these same characters can serve other purposes.

VI. CONCLUSION

To conclude, the paper aimed to discuss the basic idea about Computational Linguistics. The twenty- first century is the century of the total information revolution. Computational linguistics is also the product of this digital revolution. Even though, Computational Linguistics is not attained complete perfection because of the complex nature of human languages and it is difficult to capture the entire linguistics knowledge with hundreds percent accuracy in processing. The development of the tools for the automatic processing of the natural languages have vital significance in the overall development of any country. And it is equally inevitable to compete with each other globally.

REFERENCES

[1] http://www.thebestschools.org/rankings/best-computational-linguistics-graduate-programs/.

[2] http://www.aclweb.org/aclwiki/index.php?title=Frequently_asked_questions_about_Computational Linguistics.

[3] Ralph Grishman. (1994). Computational linguistics an introduction. Cambridge University Press. Pp. (17,33, 63).

[4] Igor Bolshakov and Alexander Gelbukh. (2004). Computational Linguistics Models, Resources, Applications. INSTITUTO POLITĔCNICO NACIONAL Publication. Pp. (17, 25, 54-55, 58, 60-63, 73, 77).

[5] Akshar Bharati, Vineet Chaitanya, Rajeev Sangal. (2001). Natural Language Processing A Paninian Perspective. Prentice- Hall of India, New Delhi. Pp. (1-2, 7).

[6] Directorate of Distance Education. Annamalai University. (2005). P.G. Diploma in NLP. Computer Application to Language Studies. Annamalai University Publication. Pp. (1, 5, 31, 89, 134-135).

[7] Directorate of Distance Education. Annamalai University. (2005). P.G. Diploma in NLP. Corpus Linguistics. Annamalai University Publication. Pp. (1, 4, 14, 27, 31).

[8] Directorate of Distance Education. Annamalai University. (2005). P.G. Diploma in NLP. Natural Language Parsing. Annamalai University Publication. Pp. (1-2, 11).

5

[9] Directorate of Distance Education. Annamalai University. (2005). P.G. Diploma in NLP. Artificial Intelligence. Annamalai University Publication. Pp. (54-55).

[10] http://www.ldcil.org/areasOfWorkSpeech.aspx.

[11] http://www.dfki.de/hansu/HLT-Survey.pdf.

[12] https://cdac.in/index.aspx?id=mlingual.

[13] http://bosslinux.in.

[14] http://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm.