LANGUAGE IN INDIA

Strength for Today and Bright Hope for Tomorrow

Volume 9 : 8 August 2009

ISSN 1930-2940

Managing Editor: M. S. Thirumalai, Ph.D.
Editors: B. Mallikarjun, Ph.D.
Sam Mohanlal, Ph.D.
B. A. Sharada, Ph.D.
A. R. Fatihi, Ph.D.
Lakhan Gusain, Ph.D.
K. Karunakaran, Ph.D.
Jennifer Marie Bayer, Ph.D.

Computational Linguistics as a Curriculum for Engineering Students in India

G. Bhuvaneswari, Ph. D. and K. P. Soman, Ph.D.

Computational Linguistics as a Curriculum for Engineering Students in India

G. Bhuvaneswari, Ph. D. and K. P. Soman, Ph.D.

Abstract

NLP and Machine Translation tools are upcoming areas of study in the field of Computational Linguistics. The development of Language technology and its growth leads to the need for the detailed study of computational aspect of Language and especially for those who mastered the field of Technology. This paper tends to create a basic reference tools for those students and researchers who are interested in Computational Linguistics and also give some basic idea of every tool used to develop such software tools.

Introduction

Computational Linguistics is an Interdisciplinary of Linguistics. The very term infer that it is a bridge for Language and Technology while these two fields in the earlier stage were defined as different areas of study. Computational Linguistics is the advanced step for using computers in the Language field. While Language is a form of communication that reaches even the ordinary people, computers were meant only for those who dealt with technology and sciences. Bridging these two is a great challenging task which can be studied under Computational Linguistics and hence it is necessary to include Computational Linguistics as a Special and Independent subject for Engineering graduates.

Though Computational Linguistics sounds more of to be a practical subject, it also has theoretical components. Theoretical Linguistics deals with issues involved in Theoretical Linguistics and Cognitive science. Many Theories are to be dealt in order to understand and generate Natural

Language and this is handled by the knowledge of Theoretical Linguistics. Various theories done by researchers and pioneers in the field of Computational Linguistics come under Theoretical Issues. Further improvement in this field can be undertaken based on the issues involved in Theoretical Linguistics.

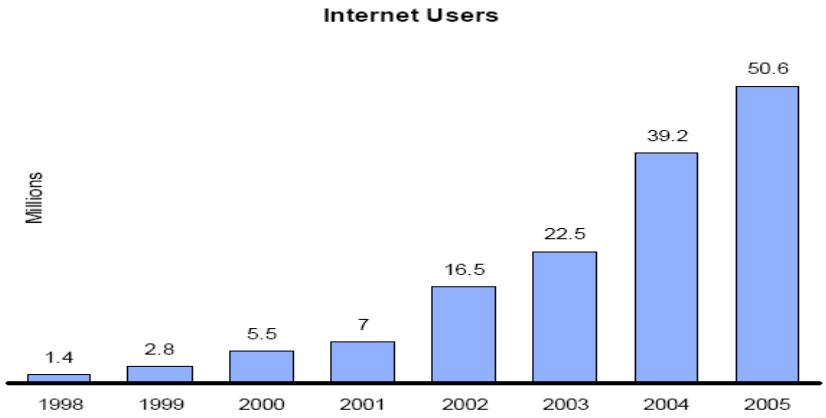
Applied Linguistics is more of practical side of Computational Linguistics. It involves the practical outcome of modeling human language use. Developing tools, techniques and methodology to develop such tools and techniques are handled in Applied Linguistics. The purpose of developing these tools are to create software products that acts more interactive with human knowledge and in turn attain some knowledge of human language. That does not mean that Computers either learn human language or replaces the human brain. It aims actually at training computers to understand the logic and the mathematical ability that deals with the language generation which is also knows as Machine Learning.

Though Language is a set of rules, it is not always easy to make a Machine to understand all set of rules since it needs to understand human logic which is very complicated. This becomes a great challenge in this field.

Although we use language for communication in everyday life, still it remains a query that how people acquire language. Some language acquisition researchers believe that, language acquisition is based primarily on general learning mechanisms. Humans have the understanding of complex mechanisms which keeps the language growing. The process of making the computer learn such mechanisms help us to learn better about the language acquisition theory by humans and also to learn the concealed characteristics of human language. This characteristics may in turn be common to other set of languages also which come under same family of languages.

Growth of Internet and Need for Machine Translation Tools

There is a rapid development of use of Internet which actually narrows down the world under a small tree. But due to language barrier the knowledge is not reachable to everyone as it is not available in Regional Languages in India.



(Cf. Developing a Speech Translator for Indian Market, Soman K.P, .A.G. Menon)

Though English is the main stream of medium of Communication in the area of Internet and scientific communication, it stops the growth of regional language. The need for the internet in Regional Languages is beyond debate. This would narrower the world into one and in a Multilingual country, this cheap mode of information transfer and retrieval is the need of the hour. This will not only bring the far corners of the country close to one another, but it will also provide a window to the outside world.

Need for Computational llinguistics as Individual Subject in Engineering Departments

Computational Linguistics is the cup of tea for many researches in NLP field and they mostly are interested in various disciplines that become the combination of humanities, sciences and Engineering. This approach of interlinking the fields attracts the researchers. When this subject is taught as a course in engineering departments it would lead to more developments in the language

field with the help of technology since both technology and language shine when linked together and that would lead to developing more software tools with the help of linguistic applications.

Computational Linguistics and Engineering Students

When Computational Linguistics becomes one of the courses for engineering students, there are certain issues that would pop up. The major issue is that an engineering student is totally unaware of the basic knowledge of Linguistics and they would need to learn the detailed study of the subject. The students will also not know in detail what kind of tools they certainly can develop. This paper would serve as a resource material for those Engineering students who are interested in the research area of Computational Linguistics. The major work in Computational Linguistics can be done in Natural Language Processing, Machine Translation and the tools developed for Machine Translation.

Translation is always in demand in a society where the country like India is a multilingual one and which is involved with more than one language in all the fields of study. Manual translation, , it depends on one particular individual and mostly relied on that particular person's efficiency in those languages. Thus manual translation is rather slow, time taking and expensive. When translation process is mechanized, it almost resolves the problem of tediousness in manual translation. Moreover, as in any field of study, use of computers plays a vital role because using computers in the Manual work saves lot of time and money.

Translation by Machine is also a difficult process but not a tedious one because it becomes a challenging one since the task involved is a multidisciplinary task. When a Machine translation system is developed effectively it would be helpful for any kind of translation involved in any field. Development of a machine translation system would be one time task which would be used for further translations.

Machine Translation – Rule based and Statistical and Tools

Translation is always in demand in a society where the country is a multilingual one and which is involved with more than one language in all the fields of study. India is a multilingual country where each state in India is separated based on language used by people. There is a growing demand for translations of scientific and technical documents, commercial and business transactions, administrative memoranda, legal documentation, instruction manuals, agricultural and medical information, industrial documents, publicity leaflets, newspaper reports, etc. When translation is involved in these fields, it becomes not only challenging but also difficult for an individual translator to cater to the needs.

The demand for translations in these fields is becoming more and more such that there will be a sharp fall in the availability of translators in the society.

Though it is an oldest dream to mechanize translation, in the twentieth century, it has almost become a reality. This involves both linguistic as well as technical knowledge and they have to be interlinked efficiently for the development of a successful Machine Translation system. A Machine Translation system has to have lot of linguistic and computational resources equally and approximately balanced to realize the goal.

It has been claimed that building a Machine Translation System between a pair of genetically related languages is comparatively an easier task than developing one that involves genetically unrelated pair. The more the similarities between the languages, the less are the problems in the development of Machine Translation System.

Tools needed for Machine Translation

Morph analyzer

Part of Speech Tagger

Parser

Transliteration Tool

Morph synthesizer

Word net

Chunking

Dictionaries

Morphological Analyser

Morphological Analyzer is the main tool in Machine Translation. This contains the data of Root words and their inflections. Developing this tool differs for different types of Machine Translation. In Rule based MT, the task of Morphological Analyser is to identify the root and Morphological features of the word. Words in the input text are first processed by the morphological analyzer. Its task is to identify the root, lexical category, and other morphological features of the given word. For Statistical MT, machine learning methodology is used for creating Morphological analyzer. In Statistical machine learning approach, linguistic knowledge is automatically extracted from an annotated corpus. It does not require manual encoding of linguistic rules. Morphological analysis is redefined as a classification task. Van den Bosch et al. (1996) discern two types of classification tasks:

Identification task: Choose the correct classification for a given input. Example: Choose the correct part-of-speech of a word.

Segmentation task: Decide for the current position in the input whether it is associated with a boundary. Example: Decide for each phoneme in a word whether it is at a morpheme boundary or not.

Part of Speech Tagger

Part of Speech Tagging is the process of assigning a part of speech or lexical category uniquely to each word in a sentence. It essentially involves the task of marking each word in a sentence with its appropriate part of speech.

Words are often ambiguous; consequently they are analyzed in more than one way by the morphological analyzer. Words listed in the dictionary also often marked for more than one

lexical category. Here, POS tagging helps to find out the exact grammatical category of the word to avoid the ambiguity.

The POS annotation of a text is the procedure for categorizing words in terms of various parts of speech. A collection of POS tags for a given language is called a POS tag set.

Parser

Parser is a tool that has to be developed for the purpose of syntactic analysis. It is the process to determine the grammatical structure of words with respect to the given formal grammar. The parser basically is developed to resolve the ambiguity in the structure of the Language. The parsers mostly rely on corpora which has trained and annotated data. This approach allows the system to gather information about the frequency with which various constructions occur in specific contexts.

Machine Transliteration

Machine Transliteration is to transcribe a word written in a script with approximate phonetic equivalence in another language. It is useful for machine translation, cross-lingual information retrieval, multilingual text and speech processing. It is the practice of transcribing a word or text written in one writing system into another writing system, such that a reader should be able to reconstruct the original spelling of unknown transliterated words.

Morphological Synthesis

Morphological synthesis is the process of producing the inflectional forms given the base form. It is the inverse form of Morphological Analyser. It generates the root words and their inflections to a word form.

Wordnet

The purpose of word net is to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. It groups words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. This tool is basically used to disambiguating the meanings of words depending upon the context.

Chunking

The analysis of a sentence which identifies the category of words is known as chunking. It neither specifies the internal structure of the words nor their role in the main sentence. It is a technique widely used in natural language processing. It is similar to the concept of lexical analysis for computer languages.

Programming Languages

Many Engineering students are really not aware of programming languages other than Java or C++. For Machine translation system, it would rather be better if students are exposed to learn string based programming languages such as Perl or Python. This would create more interest in students in language oriented field with which they can really play with the languages.

Conclusion

Promotion of mother tongue literacy and education with all the information made available, especially, that in science and technology in mother tongue through machine assisted translation is the concept of developing translation tools. Information literacy is crucial to the competitive advantage of individuals, enterprises, regions and nations. (cf. Anne Marie Perrault, 2006) For example, Wikipedia may be made available in all languages by introducing easy to use domain specific high precision translation tools. Scientists, teachers and advanced level student communities should be educated and motivated to volunteer the translation process. Academic programs and policies should encourage the development of information literacy skills.

This paper attempts to give an idea to develop course material or tools for NLP in order to teach the Engineering students with no basic knowledge of Linguistics in order to develop software tools for the societal needs and even making those computer illiterates to benefit the use of these software. Further developments such as developing speech translation tool would rectify the language barrier that is more common in the linguistic diversity Indian society.

References

Anne Marie Perrault American Competitivenessin the Internet Age Report Information Literacy Summit 2006

Arnold, D.J. Lorna Balkan, Siety Meijer, R.Lee Humphreys and Louisa Sadler Machine Translation: an Introductory Guide, Blackwells-NCC, London, 1994

Chris Brew, Markus Dickinson W. Detmar Meurers "Language and Computers, Creating an Introduction for a General Undergraduate AudienceProceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching NLP and CL, pages 15–22, Ann Arbor, June 2005.

Fei xia, "The Evolution of a statistical NLP course," Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics (TeachCL-08), pages 45-53,Columbus, Ohio, USA, June 2008.

Heike Zinmeister, "Freshmen's CL curriculum: the benefits of redundancy" Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics (TeachCL-08), pages 19-26,Columbus, Ohio, USA, June 2008.

Reva Freedman, "Teaching NLP to Computer Science Majors via Applications andExperiments," Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics (TeachCL-08), pages 114–119,Columbus, Ohio, USA, June 2008.

Robert Dale "What's the Future for Computational Linguistics?" Computational Linguistics Volume 34, Number 4 pages 622-624

Soman K.P, A.G. Menon, "Developing a Speech Translator for Indian Market - Research Proposal to IBM"

Vijaya M.S, Loganathan R, Shivapratap G, Ajith V.P, Soman K.P,"English to Tamil Transliteration using Sequence Labeling Approach",In proceedings of International Conference on Asian LanguageProcessing, Chiang Mai, Thailand, pp 169-173, 2008.

G. Bhuvaneswari , Ph. D.
Computational Engineering and Networking (CEN)
Amrita Vishwa Vidya Peetham
Coimbatore
Tamilnadu, India
bhuvaneswari.sb@gmail.com

K. P. Soman, Ph.D.
Computational Engineering and Networking (CEN)
Amrita Vishwa Vidya Peetham
Coimbatore
Tamilnadu, India
kp_soman@amrita.edu