Role of NLP in Indian Regional Languages

Prof. Langote Manojkumar S[1], Miss Kulkarni Sweta[2], Miss Mansuri Shabnam[3], Miss Pawar Ankita[4] and Miss Bhoknal Kishor[5]

[1]Asst. Prof., PIRENS Technical Campus, Loni

[2,3,4,5]Student, PIRENS Technical Campus, Loni

ABSTRACT

The NLP is closer for interfacing among the peoples knowing different languages. If we consider an example of India there are various peoples talking in various languages. Huge literature is available in different local languages which is not understandable to others in India itself. So we can use Information technology for Natural Language Processing.

Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages; it began as a branch of artificial intelligence. In theory, natural language processing is a very attractive method of human–computer interaction. Natural language understanding is sometimes referred to as an AI-complete problem because it seems to require extensive knowledge about the outside world and the ability to manipulate it. Modern NLP algorithms are grounded in machine learning, especially statistical machine learning. Research into modern statistical NLP algorithms requires an understanding of a number of disparate fields, including linguistics, computer science, and statistics. In this paper we want to study on Role of NLP for Indian Language conversions, like Marathi to Hindi, Hindi to Gujarati etc. If we observe the different languages in India they look similar in different aspects like Grammar, Words, and Alphabets. This paper will discuss the solutions available, problems and challenges in Indian Language conversions.

Introduction

As far as digitization for India is concern one major issue is that India is one of the big countries having heterogeneous nature of culture, Languages, Geography and History. For the digitization languages becomes more significant barrier. It is a big problem we could not think in all aspects. In this paper we want to focus on Natural Language Processing for Indian Regional Languages.

We will explain the concept of NLP for Indian Regional Languages in 5 steps

1. Concept of NLP
2. Nature of Indian Languages
3. Inter-language (Indian Languages) conversion

4. Existing Solutions
5. Problems in Conversion
6. Probable Solutions

Concept of NLP

Basically we will start to think over concept of NLP and some common methods available so far. Various schemes for the machine-based translation of natural language have been proposed. Typically, the system used for translation includes a computer which receives input in one language and performs operations on the received input to supply output in another language. This type of translation has been an inexact one, and the resulting output can require significant editing by a skilled operator. The translation operation performed by known systems generally includes a structural conversion operation. The objective of structural conversion is to transform a given parse tree (i.e., a syntactic structure tree) of the source language sentence to the corresponding tree in the target language. Two types of structural conversion have been tried, grammar-rule-based and template-to-template. In grammar-rule-based structural conversion, the domain of structural conversion is limited to the domain of grammar rules that have been used to obtain the source-language parse tree (i.e., to a set of sub nodes that are immediate daughters of a given node).

[Koichi Takeda, *"Pattern-Based Machine Translation"*, Tokyo Research Laboratory, IBM Research, Proc. of COLING-96, Copenhagen, Denmark]

A method and apparatus for parsing in a spoken language translation system are provided, wherein an input is received comprising at least one input sentence or expression. A parsing table is accessed and consulted for a next action, wherein the parser looks up in the next action in the parsing table. During parsing operations, the parser may perform shift actions and reduce actions. In performing a shift action, a next item of the input string is shifted onto a stack or intermediate data structure of the parser. A new parse node is generated, and a feature structure or lexical feature structure of the shifted input item is obtained from a morphological analyzer and associated with the new parse node. The new node is placed on the stack or intermediate data structure. In performing a reduce action, a grammar rule and an associated compiled feature structure manipulation are applied. When the manipulations succeed, a new parse node is generated comprising the new feature structures resulting from the successful feature structure manipulations. When the entire input is successfully parsed then an accept action is performed, a rebuilding procedure is performed, and a structural analysis of the input is provided comprising a number of parse trees and sentential feature structures.

[Inventors: Duan; Lei (Cupertino, CA), Franz; Alexander M. (Palo Alto, CA) Assignee: Sony Corporation (Tokyo, JP) Sony Electronics, Inc. (Park Ridge, NJ) Appl. No.: 09/240,896 Filed: January 29, 1999]

The present invention is a system for translating text from a first source language into a second target language. The system assigns probabilities or scores to various target-language translations and then displays or makes otherwise available the highest scoring translations. The source text is first transuded into one or more intermediate structural representations. From these intermediate source structures a set of intermediate target-structure hypotheses is generated. These hypotheses are scored by two different models: a language model which assigns a probability or score to an intermediate target structure, and a translation model which assigns a probability or score to the event that an intermediate target structure is translated into an intermediate source structure. Scores from the translation model and language model are combined into a combined score for each intermediate target-structure hypothesis. Finally, a set of target-text hypotheses is produced by transuding the highest scoring target-structure hypotheses into portions of text in the target language. The system can either run in batch mode, in which case it translates source-language text into a target language without human assistance, or it can function as an aid to a human translator. When functioning as an aid to a human translator, the human may simply select from the various translation hypotheses provided by the system, or he may optionally provide hints or constraints on how to perform one or more of the stages of source transduction, hypothesis generation and target transduction.

[Inventors: Brown; Peter Fitzhugh (New York, NY), Cocke; John (Bedford, NY), Della Pietra; Stephen Andrew (Pearl River, NY), Della Pietra; Vincent Joseph (Blauvelt, NY), Jelinek; Frederick (Briarcliff Manor, NY), Lai; Jennifer Ceil (Garrison, NY), Mercer; Robert Leroy (Yorktown Heights, NY) Assignee: International Business Machines Corporation (Armonk, NY) Appl. No.: 08/459,454 Filed: June 2, 1995]

Nature of Indian Languages

India is one of the fastest growing nations today, making it an attractive market for businesses of all types. Creating documents for India, however, can present a number of difficulties. India’s population speaks many languages, which vary from region to region. Standard Hindi and English are the two primary official languages of India, with Hindi being read and understood by a large portion of the population. However, each Indian state may also have another official language. At present, we have extensive translation resources for Indian regional translation. Whether we need Indian regional language translations or documents translated into other Indian languages, we can handle the needs. Some of the issues for Indian-language translations include:

The Language of Government: As the official language of India, many documents destined for government review may need to be translated into formal Standard Indian language, using the Devanagari script. Although English is a second official language in India, Hindi translations may be required in many cases. We have translators and can format document in the proper script.

For Business Use: Indian language is also widely used in commerce, both in Devanagari script and in Romanized transcription. Translators will translate our documents, including contracts, legal documents, user manuals, IFU's, labels, product catalogs, and decals, into the style and format needed for your business documents, and will adapt the translation to the targeted audience.

Other Indian Languages: If regional requirements for the documents require translation into other Indian languages, such as Tamil, Urdu, Punjabi, Nepali, or others, we can provide skilled translators for these languages as well, for either single language translation or multilingual documents. In our paper translators specialize in multilingual translations. If any international company needs to have the documents translated to other languages, then this paper is useful.

International Language Services – Indian Language Translation for Business: We can bring our vast resources to handle any Indian language, always with the skilled document, content, and quality management we apply to every project. Add to that our formatting capabilities, and are assured of high quality, properly formatted translations of documents that present content to readers with the same tone and quality of the original.

Inter-language (Indian Languages) conversion

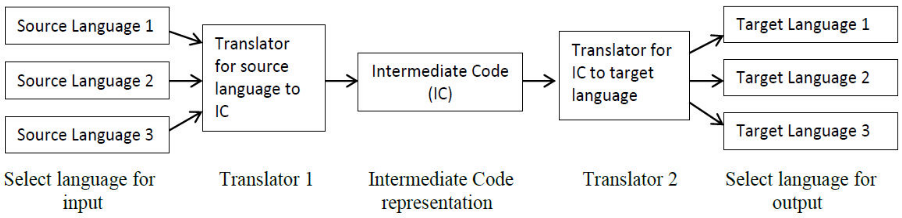
Source Language 1
Source Language 2
Source Language 3
Translator for source language to IC
Intermediate Code (IC)
Translator for IC to target language
Target Language 1
Target Language 2
Target Language 3
Select language for input
Translator 1
Intermediate Code representation
Translator 2
Select language for output

Figure 1: Overall Architecture

Sentence is formed using collection of tokens as discussed above that satisfies certain rules in production rules in P set. Thus, different activities performed to translate source language to target language are explained below.

Translator 1: Activities performed by translator 1 are explained as below,

1. Token separation (Lexical analysis) -The sentence contains differ types of tokens. These tokens may be of any type discussed in V set. These tokens are separated in this phase. The tokens are called as ‘lex’, so this phase is also called as “Lexical analysis”.
2. Token identification (Noun/ verb / adverb etc.) - After tokens get separated their type identification is performed in this step. Tokens separated may be of type noun, verb, adverb, article, etc.
3. Type Determination of sentence (Exclamation/ Question / Different tenses etc.) After completing token identification, the type of the sentence is determined in this step. The sentence may be in one of the following form, phrase, sentence, question,etc.
4. Phrase determination – Depending upon the type determined of the sentence by the above step, certain decisions are need to be taken. Suppose if the sentence is phrase, question, or exclamation then different type of subtasks are needed to be followed accordingly for proper translation.
5. Intermediate language generation after completing all the above steps finally intermediate representation of the source language is generated.

*Translator 2:*I/C to target language generation – Using intermediate representation of source language, final translation into target language is generated as a final output.

Problems in Conversion

1. Ambiguity in translation
2. Unmatched words
3. Lack of knowledge or vocabulary of all Indian languages
4. Hard to check the proper translations

Probable Solutions

1. Removal of Ambiguity: In Indian languages like Marathi/Hindi there are some words that have dual and multiple meanings. So it becomes to much complicated to consider the right meaning of the word while translation from one language to another. We can get the users opinion by giving all alternatives while translation
2. Unmatched Words: Can be converted using phonetic e.g. Marathi to Marathi
3. It is hardly impossible to improve and summarize the vocabulary of all Indian languages. So this problem can be overcome by using open source or free source of our Dictionary Database.
4. Open source is the best solution for black box testing. We can put our code on web so that many users of India can access and use it. Automatically they will give suggestions and opinions in our system.

Conclusion

As per my experience this field of NLP which is actually a branch of AI will need large efforts to put this work in functional state. This paper has tried to focus on new field of NLP for the bright future of digitization of India. There are many challenges in NLP for India.

References

Bharati,Akshar,Chaityanya Vineet and Sangal Rajeev,(1995),Natural Language Processing: A Paninian Perspective,Prentice-Hall of India.

Dhore M.L. and Dixit S.K.,Cross Language representation for commercial web applications in context of Indian languages using phonetic model,*CIIT international journal of Artifitial Inteligent systems and Machine Learning*,3(4).

DeNero John and Uszkoreit Jakob,(2011), Inducing Sentence Structure from Parallel Corpora for Reordering, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Gore Lata and Patil Nishigandha,(2002),English to Hindi-Translation System,*Proceedings of Symposium on translation systems strans.*

Padariya Nilesh,Chinnakotla Manoj,Nagesh Ajay and Dawant Om P.,(2008),Evaluation of Hindi to English, Marathi to English and English to Hindi.