ISSN No. 0976-5697

Volume 6, No. 3, May 2015 (Special Issue)

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Word Reordering using Rule Based Parser in Machine Translation

Kanusu Srinivas Rao
Assistant Professor,
Dept. of Computer Science,
Yogi Vemana University, Kadapa ,
AndhraPradesh, India. Pin: 516003,
Email: kanususrinivas@yahoo.co.in

Ratnakumari challa
Assistant Professor,
Dept. of Computer Science & Engg,
JNTUK, Kakinada,
AndhraPradesh, India.
Email: ratnamala3784@gmail.com

Abstract: Word Reordering is an intermediary stage in Machine Translation, where the source language sentences with disordered words are oriented into the grammatical structure of target language. The main focus is imposed on the reordering of given words into the basic structure of target language. The approach is based on target side reordering, also referred as post-ordering, where one to one word mapping is done before the reordering process. Initially, the source sentence is tokenized and each word is mapped to its corresponding word in target language. It results in target words with the source language structure. This disordered sentence obtained from the direct translation of the source sentence is taken as input and each word is assigned with Part Of Speech (POS) tagging. A parser is employed on the POS tagged words, which uses the rule bank to orient the words into correct form. The parser with the set of rules, reorders the source language structured sentence into target language structure.

Keywords: parsering, source sentence, target sentence, post processing, word reordering.

I. Introduction

Machine translation (MT) is the task of translating the text in source language to target language. It can be said to doing Natural language Processing (NLP). Even though Machine Translation was envisioned as a computer application in the 1950's and research has been made for 60 years, machine translation is still considered to be an open problem.

In a linguistically diverse country like India, Machine Translation is an important and most appropriate technology for globalization. Human translation in India can be found since the ancient times which are being evident from the various works of philosophy, arts, mythology, religion and science which have been translated among ancient and modern Indian languages. As of now, human translation in India finds application mainly in the administration, media and education, and to a lesser extent, in business, arts and science and technology.

In such a situation, there is a big market for translation between English and the various Indian languages. Human translation is slow and also consumes more time and cost compared to machine translation. The reason for choosing automatic machine translation rather than human translation is that machine translation is better, faster and cheaper than human translation.

I. Machine Translation Process

The process involves a series of steps such as tokenization, tagging, translation, reordering and post processing.

A. *Text input*

This is the first phase in the machine translation process. The sentence categories can be classified based on the degree of difficulty of translation. Sentences that have relations, expectations, assumptions, and conditions make the Machine Translation relatively difficult. Speaker's intentions and mental status expressed in the sentences require discourse analysis for interpretation. This is due to the inter-relationship among adjacent sentences. World knowledge and common sense could be required for interpreting some sentences.

B. *Tokenization*

The next step involves tokenization where the sentence is divided into tokens. Each token constitutes of a word. They can be divided basing on the space character or punctuation marks. The tokens are further grouped based on the Parts-Of-Speech (POS) tagging. For example, if space character is used as a separator, it can be identified by the ASCII character 32.

C. *Tagging*

Tagging means the identification of linguistic properties of the individual words. Each word in the sentence is assigned with Parts-Of-Speech (POS) tagging. These tags depend on the context and meaning of the word. Basing on the POS tags, the rules are applied for reordering the sentence.

D. *Translation*

The words are directly translated into the target language basing on word to word mapping. It particularly deals with recognition, analysis and generation of words. Some of the morphological processes are inflection, derivation, affixes and combining forms. Inflection is the most regular and productive morphological process across languages. Inflection alters the form of the word in number, gender, mood, tense, aspect, person and case. This results in target words but in the order of source language which leaves no meaning.

E. Parsing

Parsing is the assessment of the functions of the words in relation to each other. The parsing is done using a rule based parser which makes use of rule bank. The rule bank consists of the syntax of the target language. As words are foundation of speech and language processing, syntax can be considered as the skeleton.

Syntactic analysis concerns with how words are grouped into classes called Parts of Speech, how they group their neighbours into phrases and the way in which words depends on other words in a sentence.

F. Post processing

Post processing constitutes the final step of Machine Translation which takes the reordered sentence as input. The reordering using a parser, only results in the basic structure of target language. It still needs further processing such as inclusion of auxiliary verbs and other necessary modifications to produce the best output that grammatically satisfies the target language.

II. WORD REORDERING

This is an important step and penultimate step in Machine Translation. A good word reordering mechanism produces an effective output. Word reordering can be done at source side, called pre-ordering or at target side, called post-ordering. In pre-ordering, the words are arranged in target language structure and are then translated into target language. The post-ordering employs an opposite mechanism where the words are first translated into target language and are then ordered as per the structure of target language. Both mechanisms have proven to be effective and the post-ordering is recently gaining importance.

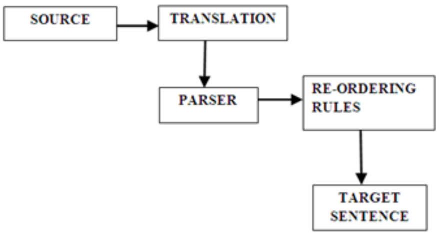

Figure 1.Reordering in Machine Translation System

This is an intermediary phase before post-processing where words are oriented in target structure i.e. in the basic structure of the target language. The reordered sentences can be used for training and testing. A good reordering mechanism improves the performance and efficiency of machine translation system.

The following flowchart well explains the process implemented in reordering the given input.

III. RULE-BASED MACHINE TRANSLATION

A Rule-Based Machine Translation (RBMT) system consists of collection of rules, called grammar rules and software programs to process the rules. RBMT involves more information about the linguistics of the source and target languages, using the morphological and syntactic rules and semantic analysis of both languages.

The basic approach involves linking the structure of the input sentence with the structure of the output sentence using a parser and an analyzer for the source language, a generator for the target language, and a transfer lexicon for the actual translation. Nevertheless, building RBMT systems entails human effort to code all of the linguistic resources, such as source side part-of-speech taggers and syntactic parsers.

A RBMT system always is extensible and maintainable. Rules play a major role in various stages of translation, such as syntactic processing, semantic interpretation, and contextual processing of language. Generally, rules are written with linguistic knowledge gathered from linguists.

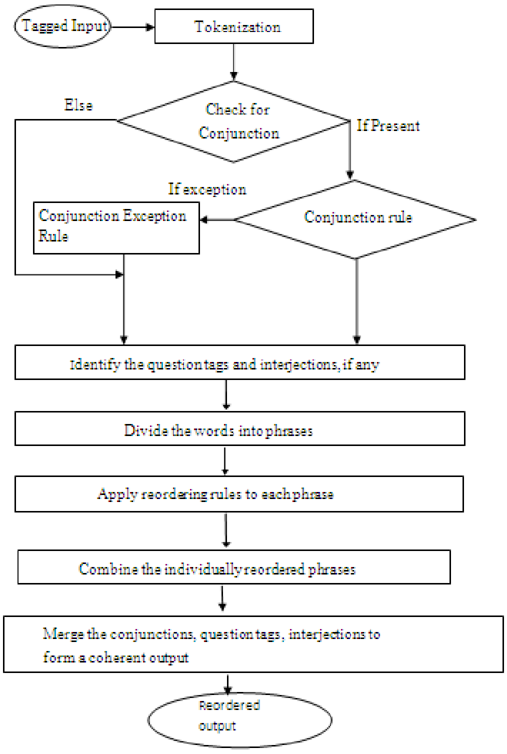
Tagged Input
Tokenization
Else
Check for Conjunction
If Present
If exception
Conjunction Exception Rule
Conjunction rule
Identify the question tags and interjections, if any
Divide the words into phrases
Apply reordering rules to each phrase
Combine the individually reordered phrases
Merge the conjunctions, question tags, interjections to form a coherent output
Reordered output

Figure 2. Flowchart to describe the process involved in Reordering.

In addition, there are many stylistic differences. For example, it is common to see very long sentences in English, using abstract concepts as the subjects of sentences, and stringing several clauses. Such constructions are not natural in Indian languages, and present major difficulties in producing good translations. With the current state of art in MT, it is not possible to have Fully Automatic, High Quality, and General-Purpose Machine Translation. Practical systems need to handle ambiguity and the other complexities of natural language processing.

Therefore, to have a good translation system, reordering the source sentence in accordance to target sentence is needed. So reordering system for source sentences can make significant improvements over Indian languages.

IV. CONCLUSION

The word reordering system gives efficient results as a part of Machine Translation. Reordering is done through the implementation of a rule based parser. To improve efficiency, the reordering is performed at the lowest level of phrase and these constituent phrases of a main phrase are logically combined through bottom-up approach. The output is clearly displayed in the form of parse tree so that a naive user can also understand and get the clear picture of how parsing is done.

V. REFERENCES

[1] R.Gangadharaiah & N.Balakrishnan, "Application of Linguistic Rules to Generalized Example Based Machine Translation for Indian Languages", Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages, India, 2006.

[2] Mustafa Abusalah, John Tait & Michael Oakes, "Literature Review of Cross Language Information Retrieval", World Academy of Science, Engineering and Technology, 2005.

[3] Maja Popovic & Hermann Ney, "POS-based Word Reorderings for Statistical Machine Translation", in Proceedings of the Fifth International conference on Language Resources and Evaluation, 2006.

[4] Sethuramalingam S, "Effective Query Translation Techniques for Cross-Language Information Retrieval", MS Thesis submitted at IIIT Hyderabad, India, 2009.

[5] Isao Goto, Masao, Utiyama, "Post-ordering by Parsing for Japanese-English StatisticalMachine". Brown, C.P., "The Grammar of the Telugu Language". New Delhi: Laurier Books Ltd, 2001.

[6] Gwynn and Krishnamurti: "A Grammar of Modern Telugu", volume 11, Oxford University Press, Delhi, 1987.

[7] W.John Hutchins and Halord L. Somers, "An Introduction To Machine Translation", Academic Press Ltd.,1992.

[8] Jurafsky, Daniel and Martin, James.H, "Speech and Language Processing-An Introduction to Natural Language processing, Computational Linguistics and Speech Recognition", 2002.