IMPLEMENTATION OF NATURAL LANGUAGE TRANSLATION SYSTEM BASED ON NLP: MYANMAR TO JAPAN TRANSLATION PROCESS

Nwe Oo Khin
Computer University (Maubin)
nweookhin22@gmail.com

ABSTRACT

The translation of natural language is one of the most important in NLP interface. Therefore language translation becomes a major concern in the present day.This paper is based on the natural language processing concept. It also translates Myanmar sentence into the equivalent Japanese sentence. Myanmar sentences are usually long and sentence constructions are complex but this paper concentrates to the short or simple sentences.There are so many system Language Translation which base on NLP, but the most significant fact of our translation is that the famous pagodas in Myanmar and the name of well- known Images in Shwedagon Pagoda have already been stored in the lexicon. If a user wants to add new words, he needs to enter into the lexicon entry to use them and this paper, he can use not only the limited grammar rules, but also the new sentence patterns that he wants to use.It can translate both simple sentences and short paragraphs. The goal of this system is to support the Myanmar students who want to write the Japanese sentence in accordance with thinking in Myanmar.This paper is implemented by using Java and Microsoft Access Database.

1. INTRODUCTION

Natural Language Processing (NLP) opens the door for direct human communication. NLP tries to make the computer to be capable of understanding commands written in standard human languages. NLP is one of the most prominent research areas carry out in artificial intelligence. NLP can offer possible solution to the problems of communication between humans and computers. There are two common kinds of languages such as natural language and formal language. Formal language is artificial language deliberately developed for a special design and purpose. A good example is a computer language. Natural language is not deliberately invented or designed.They have evolved over the centuries as human beings learned to communicate with one another. For example, Myanmar, English, French and Japanese,etc., are natural languages [8]. The goal of the Natural Language Processing (NLP) group is to design and build software that will analyze, understand and generate languages that humans use naturally [5]. Human language consists of morphology (the way words are built up from small meaning bearing units, syntax (sentence structure), sematics(meaning),and countless ambiguities. A language translation system analyze the grammars of sentence. The rules of this system are to put words together to form complete sentence called grammar.Grammar is composed of two basic parts : syntax and sematics [6].

This paper explains how the parser and analyzer combine to tokenize, how to built parse tree and how the transfer process uses lexicon. And then it also illustrates Myanmar sentence generation.There are five sections in this paper, where the first section refers to introduction.Section2 describes the system of Natural Language Processing. Section 3 shows the detailed description of the overview of our processing system, mention the Myanmar parser and syntax analyzer, explain source to target transfer which is being support by the lexicon and understarder and elaborates Japanese generation. Section4 shows the Exprimental Result and Description.The last section expresses the section 5 conclusion.

2 . NATURAL LANGUAGE PROCESSING

Natural Language Processing is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies. Natural Language Processing includes five major elements: the parser, the lexicon, the

understander, the knowledge base and the generator .

Parser : Parser is the section of software that reads the input sentence, word by word. Each word is isolated and its part of speech is identified. Then the parser maps the words into a structure called the parser tree .

Lexicon : Lexicon is a computer - held dictionary. The lexicon contains all of the words that the program is capable of recognizing and contains the correct spelling of each word and also designates its parts of speech.

Understander : The understander works in conjunction with the knowledge base to determine what the sentence means . A knowledge base is a special kind of database for knowledge management .

Knowledge : The knowledge base is the primary means of understanding what has been said. Production rules, semantic networks, frames and scripts are all common methods for implementing natural language knowledge base.

Generator : The generator uses the understood input to create a usable output.The understander creates another data structure that represents the meaning and understanding of the sentence and stores it in memory.

3. OVERVIEW OF THE PROPOSE SYSTEM

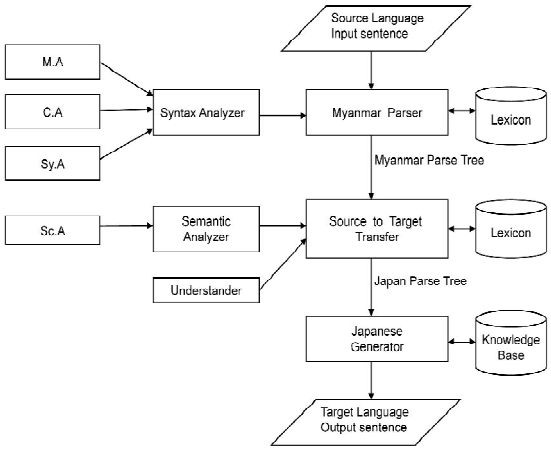


Figure (1) Myanmar to JapaneseProcessing System

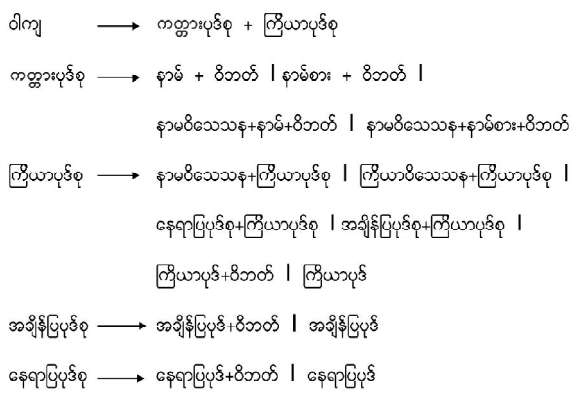
M.A → Morpholexical Analysis
C.A → Constituent Analysis
Sy.A → Syntatic Analysis
Se.A → Semantic Analysis

Figure (1) Shows Myanmar to Japanese Processin System.The source language of this system is in Myanmar in which target language is Japanese. As soon as the Myanmar sentence is being inputted, it has to pass parser first. Parser and Syntax Analyzer need to combine to analysis the language. The first stage of syntax analyzer which is known as morpholexical analysis tokenizes the words from input Myanmar sentence and their part of speech. Tokenizer tokenizes each space, and when a user inputs Myanmar sentences, it categorises them by using myanmar rules, and put spaces after each and every particles such as adjective, adverb, preposition and conjunctions between the words. After that the constituent analysis brakes them into chunk by using Context Free Grammar rules.Syntatic analysis is checking the grammar rule of input Myanmar sentence. Myanmar grammar rules store in the knowledge base. The words which are tokenized by the parser are transferred into Source to Target transfer process. In Source to Target transfer, the semantic analyzer uses bilingual dictionary in the lexicon.The understander supports at using conjunctions in sentences in this process. The third component of Japannese generator rearranges the sentence by using Japanese grammar rules which are based on Myanmar grammar. The Myanmar sentence from the previous process is generated to the targed language, Japanese.

3.1 Myanmar Parser and Syntax Analyzer

The key element in a natural language system is the parser. The parser is a piece of software that analyzes the input sentence syntactically. A parser determines in a sentence what part -of- speech to assign to each of the words and combines these part-of-speech tagged words into larger and larger grammar [9]. The parser then maps the words into a structured called the parser tree. The parser tree shows the meaning of all of the words and how they are assembled. The entire

parsing processing analogous to the process of diagramming sentences. The job of the parser is to examine each word in a sentence and create the parser tree that identifies all of the word and puts them together in the right way [7]. By using Myanmar grammer rules the parser can help sentence to pase in this system. The grammar rules which is produced by CFG are following;



Parser and Syntax Analyzer are linked to analysis the sentence. Syntax analyzer shows the grammatical function of each constituent of a sentence. Syntax Analyzer has three steps: morpholexical analysis, constituent analysis and syntatic analysis.

Morpholexical analysis morphemically can be tokenized the input Myanmar sentence. The tokenizer breaks input Myanmar sentence into it components. The sentence is split into tokens. The tokenizer tokenizes the sentence using inter word space into token.

Constituent analysis can be considered as chunking. The chunker is a non-deterministic version of the left to right (LR) parser. That uses a hand-written context-free grammar (CFG) and employs a best-first search. The parser shifts words matching the right-hand side for a rule from the grammar. Syntatic analysis is based on link grammar and identifies the grammatical relations between separate chunks. The parse in this system generate a parse tree of Myanmar sentence. The input Myanmar sentence is after parsing process, the result is shown the parse tree

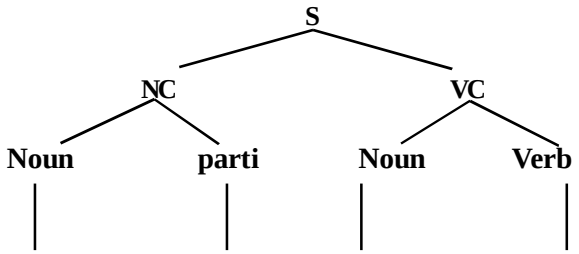


Figure (2) Example of a Myanmar Parse Tree

3.2 Source to Target Transfer

In the function of source to target transfer process the lexicon is used to describe the connection between meanings of each and every word which is tokenized by parser. This transfer process is supported by semantic analyzer and understander. The semantic analyzer analyzes the semantic and it refers to the meaning of the language. Semantic analysis helps to find the works by using lexicon.

3.2.1 Lexicon

The lexicon can be seen as a dictionary that lists the words of a language alphabetically. It contains all of the words that the program is capable of recognizing and contain the correct spelling of each word and designates its part of speech. For words that can have more than one meaning the lexicon lists all of the various meanings permitted by the system. The lexicon is a bridge between a language and the knowledge base expressed in that language[3].The lexicon can be used to automatically translate on a word-by-word basis, and in some cases phrase-by-phrase. In a monolingual lexicon the meanings are expressed as paraphrases in the same language, in a bilingual lexicon as a translation into another language[2].

In this system, Myanmar sentence is being translated and bilingual lexicon is used to make Japanese sentence. The lexicon can retrieve the correct words from the sentence which is being inputted by the user. In our translation system, the user can contribute the words in the lexicon table.When we input the new words, we need to put both meaning-bearing and their parts of speech.

3.2.2 Understander

The purpose of the understander is to use the parser tree to refer to the knowledge base. Thus, the understander may answer a question using the knowledge base. If the input sentence is a correct Myanmar sentence, the undrestander determine correct Janpanese pattern by looking up Japanese grammar rules in the knowledge base. An example sentence is shown in the following:

In the above two sentences, preposition]wGif} has two meanings ' ni ' and ' de ' according to the form of Japanese sentences. For the first sentence, it chooses the meaning ' ni ' in accordance with the prior word and the sentence, " mise no mae ni kuruma ga tomarimashita", is produced. In the second sentence, it chooses the meaning ' de ' looking for the prior word and it produces the sentence, " shingoo de kuruma ga tomatte imasu".

3.3 Japanese Sentence Generator

For Japanese sentence generation we use Japanese grammar rules associated with input Myanmar sentence.We base this grammar on the Myanmar grammar. Japanese sentence structure is a S+O+V (Subject, Object, Verb) , S+P+V (Subject, Place, Verb) and S+T+V (Subject, Time, Verb)Language like Myanmar. But they both have particle before and after the noun. So Subject and Object have noun+particle, Place has particle and Time has time+particle. As Myanmar and Japanese rules are similar with each other, the number of Japanese grammar rules could be same with Myanmar.

S → NC + VC

NC → Noun + parti / Pronoun + Parti / adj + noun + parti / adj + pronoun + parti

VC → adj + VC / adv + VC / NC + VC / PC + VC / Verb+ parti / verb

TC → Time + parti / Time

PC → Place + Parti / Place

As they are similar, we need not reorder the words in the sentence. Our specification grammar rules are stored in knowledge base. The parse tree (one of the output of parser) is matched with Japanese grammar rules. If grammar rules are found, the generator generates the target language,Japanese. Example Japanese parse tree is shown in figure.

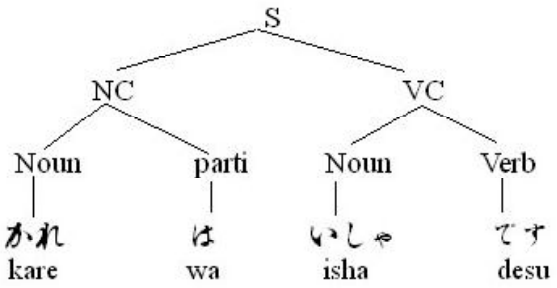
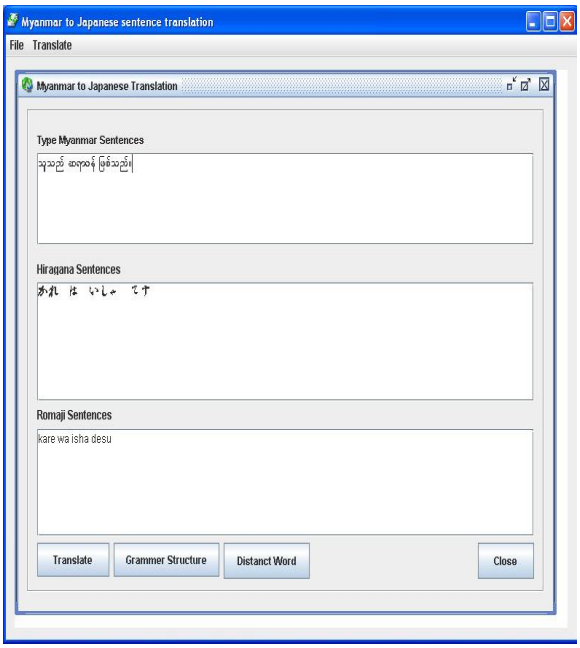


Figure (3) Example of Japanese Parse Tree

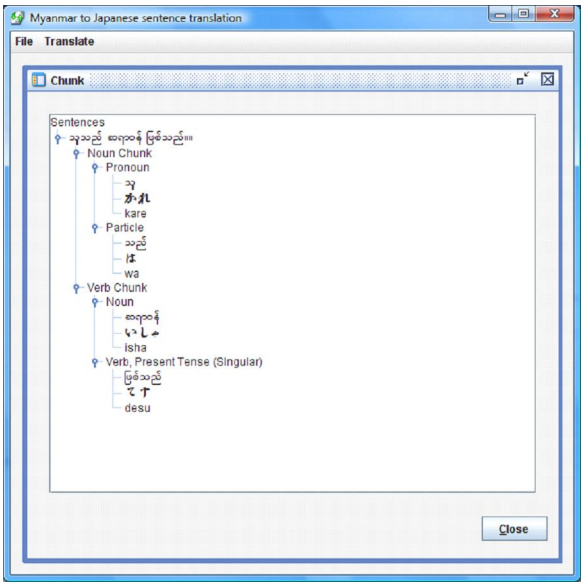
4. EXPERIMENTAL RESULT AND DESCRIPTION

When user starts to travel the Myanmar to Japanese translation program, menu bar will appear. It consists of two buttons on menu bar, file and translate. Part of speech, lexicon, sentences pattern and exit are shown in the file menu.



For translation, as soon as we enter it; the translation box comes out. There are all together four buttons in the translation box: translate, grammar strature, distance word and close. If user wants to translate Myanmar sentence to Japanese, he has to

type the Myanmar sentence in the Type Myanmar Sentence column. It is obey that when user type the sentences, he has to use space after particle;adjective, adverb,preposition, particle and so also “ ” end of the sentence. When the translate button is pressed , it shows error message box it means the sentences is incorrect of input sentence is successful, Japanese sentence will appear.



If user chooses the grammar structure button, grammar relation between input Myanmar sentence and output Japanese sentence is shown by parse tree in it.

5. CONCLUSION

This system can translate the sentences and it is capable of covering the sentence patterns in the knowledge base system. The system accepts the Myanmar Language with Myanmar input sentence and generates the equivalent Japanese sentence. This sytem allows only the simple sentence. This system does not allow all of the sentence patterns and all of the complete tense. The commonly used words form Myanmar-Japan Dictionary and meaningful names, places and times have already been stored in the Lexicon. The user should enter to the lexicon according to the guideline when the user wants to put more words.This system can help to overcome the language barriers between the Myanmar and Japanese. It can be applied as Myanmar-Japan Dictionary for Myanmar students who are learning Japanese Language. While the user types the Myanmar sentece, the user should separate one space after particle, Adverb and Adjective.

REFERENCES

[1] Bilingual Parallel Corpora and Language Engineering Carl, M.999 : Inducing translation templates for example- based machine translation, Machine Translation Summit, Singapore.

[2] D.Hiemastra, “ Using statistical methods to create a bilingual dictionary”, Master’s thesis Parlevink Group Section Software Engineering and Theoretical Informatics Department of Computer Science.

[3] john F. Sowa,
“Concepts in the lexicon: Introduction”,
email : sowa@bestweb.net.

[4] Manning and H.Schutze,
“Foundations of Statistical Natural Language Processing”, MIT Press, Cambridge, Massachusetts.

[5] M.S.C, Thomas and K.Plunkett
“Representing the bilingual’s two lexicons”, University of Oxford, UK, URL www.psyc.bbk.ac.uk/ people/academic/thomas-m/Thomas-cogsci95.pdf.

[6] Myat Thuzar Tun, Ma
“English Syntax Analyzer for English-to-Myanmar Machine Translation”
Ph.D Thesis, 2007.

[7] Thiri Haymar Kyaw, Ma
“Implementation of Natural Language Translation system from Myanmar to English
M.C.Sc (Thesis), 2004.

[8] Turban, Efraim,
“Expert Systems and Applied Artificial Intelligence”, ISBN 0-02-946565-6, Macmillan Publishing Company.

[9] http://donelaitis.vdult/publikacijos/SDOCL1.htm

[10] http://en.wikipedia.org/wiki/Machine-Translation.