Building a Tool for Multi-word Expression Extraction in Kachin-Myanmar using Syntactic Parsing

Hkawn Ra Seng Bu
University of Computer Studies (Mandalay)
hkawnra.sengbu@gmail.com

Abstract

Syntactic analysis in Natural Language Processing (NLP) is one of the most important fields carried out in the world of artificial intelligence. NLP, a subfield of artificial intelligence and computational linguistics, can be applied as a productivity tool to translate from one language to another language. This thesis is based on Natural Language Processing concept and works for Kachin to Myanmar language translation by using syntactic analysis. Syntactic analysis is the Parsing-converting a flat input sentence into a hierarchal structure that corresponds to the units of with those stored in Bilingual Dictionary. If the computer understands the natural language, the computer and user communication will be smarter. Therefore language translation becomes a major role in the present day. The main purpose of this system is to translate Kachin Multi-word to Myanmar for document text. Moreover, this system can be applied as the Myanmar meaning for the Kachin students who are learning the Myanmar language.

Keywords: Syntactic analysis, Natural Language Processing, Kachin, Myanmar, Bi-lingual.

1. Introduction

Human realized the need for language translation at the dawn of civilization with the arising demand to communicate with neighboring people. Language translation is the system for communication. It embodies both verbal and written expression to help us communicate our thoughts and feelings. There are two kinds of language such as formal language and natural language. Formal languages are artificial languages deliberately developed for a special purpose such as computer language. Natural languages are human language to communicate with each other. For example: Myanmar, English, and Kachin.

Natural Language Processing (NLP) programs use artificial intelligence to allow the user to communicate with the computer in the user's natural language. The computer can both understand and respond to command given in a natural language.

Syntactic analysis of grammar rules in Natural Language Processing (NLP) plays a vital role in the world of artificial intelligence. The system accepts the natural language inputs in order to produce the natural language outputs. Firstly, the users input the Kachin sentence in the system and then the system tokenizes the input sentence and produces the words. The computer analyzes the input syntax using the grammar rules. If the input's rule corrects one of the rules in the database, syntactic analysis is completed.

The main purpose of this thesis is to translate Kachin multi-word to Myanmar using syntactic parsing in Natural Language Processing (NLP). Moreover, we explain 8 objectives of this thesis in section 1.1.

1.1 Objectives

The objective of this system is to understand the Natural Language Processing (NLP) concepts and theory in Artificial Intelligence (AI), to know how the Multi-word Expression in Syntactic Analysis of Natural Language Processing, to know the basic grammar rules of Kachin and Myanmar language, to improve the communication between Kachin and Myanmar languages, to study the relationship between Kachin and Myanmar languages, to implement a system that translates from Kachin to Myanmar languages, to develop the communication between the user and the computer using Natural Language Processing (NLP), to reduce the language barriers among people using Kachin language and Myanmar language

We briefly describe related works in Section 2. And then system architecture is described in Section 3. Finally, we conclude this system with the benefit.

2. Related Works

Natural Language Processing system becomes a major focus area within Artificial Intelligence (AI). One of the fundamental goals of NLP is to build computer systems that can understand natural language. The primary goal of NLP is to understand how exactly human beings understand, generate and learn languages. The purpose of an NLP program may be to 'understand' natural language sentences, but in order to do so, the NLP system is required to ' understand' the meaning descriptions given in the lexicon. The term lexicon is used in NLP to stand for a dictionary considered to be one of the components of a NLP system [1, 2, 5].

The term lexicon is used in a technical sense in linguistics. A lexicon contains a complete inventory of all words in a language along with relevant

information conforming to the specifications of a given theoretical framework. This information would be organized as required by the theory. The words of interest are usually open-class or content words such as nouns, verb and adjectives rather than closed-class or grammatical function words such as articles, pronouns, and prepositions, whose behavior is more tightly bound to the grammar of the language. A lexicon may also include multi-word expressions such as fixed phrases, phrasal verbs, and other common expressions. Each word or phrase in a lexicon is described in a lexical entry. To say exactly what is include in each entry depends of the purpose of the particular lexicon. The details that are given may include any of its properties of spelling or sound, grammatical behavior, meaning or use and the nature of its relationships with other words. A lexical entry is therefore a potentially large record specifying many aspects of the linguistic behavior and meaning of word. Therefore, a lexicon can be viewed as an index that maps from the written form of a word to information about that word. This is, however, not a one-to-one correspondence. Words that occur in more than one syntactic category will usually have a separate entry for each category [3, 4, 6].

3. System Architecture

3.1Kachin Language Parser

The core of any NLP system is the parser. The job of the parser is to examine each word in a sentence and create the parse tree that identifies all of the words and puts them together in the right way. The parser in this system generates a parse tree of Kachin Multi-word.

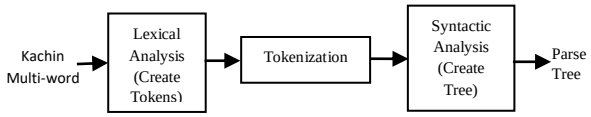

Figure 1. Parsing Process for Kachin Multi-word

The program splits the input Kachin multi-word looking for spaces to identify the individual words. It also identifies its Part of Speech of each word by working together with the lexicon. The parser works closer with the lexicon in doing syntactic analysis. The parser and lexicon works together to pick apart a sentence and then create the parse tree. The special characters in the sentence such as spaces are tokenized as word.

3.2 Token

A token is used to separate by morphological analysis. A token identifies a unit of information. Usually, tokens are the result of some processing part that has performed lexical analysis and divided a data set into the smallest units of information used for subsequent processing. A token is an occurrence of a term from the text of its source text. The type is an interned string, assinged by a lexical anlayzer naming the lexical or syntactic class that the token belongs to.

Kachin language is seperated by space between words. Therefore, this step is to divide words, these are Kachin tokens.

For example: Shi gaw jawng de sa ai

Shi	(Pronoun)
gaw	(Preposition)
jawng	(Noun)
de	(Noun)
sa ai	(Verb)

These words are seperated with using token of Kachin literature. It compares with the words in the lexicon.

3.3 Syntactic Analysis (Parsing)

The syntactic parsing is the first step towards trying to exact meaning from the meaning. To analyze an input sentence, the system first divides down the multi-word into individual words. The program scans the input multi-word for spaces to identify the individual words. And the program identifies the parts of Speech of each word in the sentence by working together with the lexicon.

Syntactic rules determine the correct order of words in a sentence. The words of a sentence can be divided into two or more groups, and within each group the words can be divided into subgroups, and so on, until only single words remain. It works out with the lexicon in doing syntactic analysis. Rules of syntax are specified by writing a grammar for the language.

Firstly, a parser checks if a multi-word is correct according to grammar and then its returns a representation of the multi-word structure.

For example:

The input sentence = Shi gaw jawng de sa ai
The Grammar Rule = <Subject>< Place>< Verb>
Sentence Rule = <Pronoun> <Preposition> <Noun> <Preposition> <Verb>
Syntactic Processing for multi-word,
Syntactic Analysis (Parsing)
Input: Shi gaw jawng de sa ai.
After parsing process, the result is
Shi (Pronoun), gaw (Preposition), jawng (Noun) de (preposition), sa ai (Verb).

Sentence	→	<Subject> <Place><Verb>
<Subject>	→	<Pronoun><Preposition>
<Place>	→	<Noun><Preposition>
<Pronoun>	→	shi
<Preposition>	→	gaw
<Noun>	→	jawng
<Preposition>	→	de
<Verb>	→	sa ai

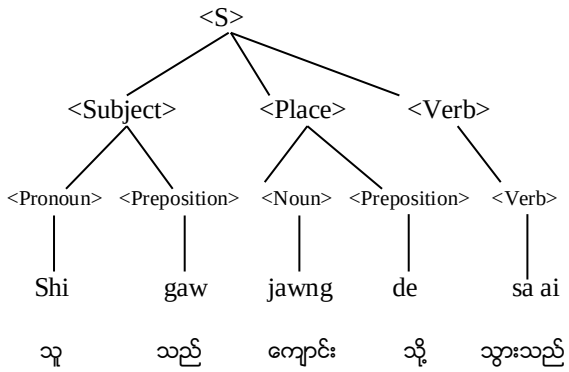

Figure 2. Syntactic Parsing in Kachin-Myanmar

The output sentence = သူသည် ကျောင်း သို့ သွားသည်
The result of the Myanmar
generation is = သူသည် ကျောင်း သို့ သွားသည်
Sentence Rule = Pronoun Preposition Noun Preposition Verb.
Grammar Rule = Subject Place Verb

3.4 Kachin-Myanmar Bilingual Lexicon

The system uses the bilingual lexicon. Bilingual lexicon is used to translate source (Kachin) multi-word to corresponding Myanmar sentences. Bilingual lexicons are important because they play a large role in translation of human languages.

The lexical database consists of the translation and contexts already presented in the human dictionary. The lexicon consists of a list of expressions in a given language. And some grammatical features are asssociated with each other usually in its meaning.

For example:

Shi (Pronoun) = သူ

gaw (Preposition) = သည်

jawng (Noun) = ကျောင်း

de (Preposition) = သို့

sa ai (Verb) = သွားသည်

3.5 Grammar Rules and Sentence Rules

In this system, there are 25 grammar rules and 65 sentences rules in knowledge Base. Grammar rule table has 2 fields: ID and Grammar Rules. Sentence rules table has 3 fields: ID, Sentence Rule and Grammar rules. The table of Grammar rules and Sentence rules are relationship.

Knowledge Base plays a vital role in translation from Kachin language to Myanmar language. A knowledge base is a dynamic resource that may itself have the capacity to learn, as part of an artificial intelligence (AI), contains facts and rules about objects in the specific domain.

By analyzing the structure of both languages, we found that the similarity is 80%.

For example (Grammar Rules):

- Subject Object Verb
- Subject Time Object Verb
  …
  …
- Subject Time Manner Object Verb
- Subject Time Place Object Manner Verb

For example (Sentences Rules):

- Noun Preposition Noun Preposition Verb
- Adverb Noun Preposition Noun Verb
  …
  …
  …
- Pronoun Preposition Noun Preposition Verb
- Preposition Noun Preposition Noun Verb

3.6 Grammar Analysis

A grammar is a finite set of productions, which contain both terminal and non-terminal symbols. One of the non-terminal symbols is the expression multi-word symbol of the grammar (the start symbol). A grammar is a description of a language. It consists largely of a set of production of rules to transform the non-terminal symbols on the left hand side and non-terminal symbols on the right hand side. Grammar analysis is the other application of NLP. This is done with a special NLP program that converts poor writing to acceptable writing.

For example:

Terminal: Shi gaw jawng de sa ai

Nonterminal: {Pronoun, Preposition, Noun, Preposition, Verb}
Sentence Rule = Pronoun, Preposition, Noun, Preposition, Verb
Grammar Rule = Subject Place Verb

3.7 Myanmar Sentence Generation

For Myanmar generation, this system use grammar rules associated with input Kachin multi-word. The grammar structure of the Kachin and Myanmar are the same. The parse tree (one of the output of parser) is matched with grammar rules. If grammar rules are found, this tree is produced.

For example:

Kachin multi-word (input): Shi gaw jawng de sa ai

Shi (Pronoun), gaw (Preposition), jawng (Noun), de (Preposition), sa ai (Verb)

The result of the Myanmar generation is

သူ သည် ကျောင်း သို့ သွားသည်

3.8 Lexicon

For the system, there must contain a lexicon databases. A lexicon database is like an electronic dictionary (a collection of words along with the information) with some additional information needed for the translation system or for some NLP applications. When the user typed the word, and then the word is matched in the lexicon. If all the word contain in the lexicon, translate output words. If the word is not contained in the lexicon, the user must contact Administrator to add new words in the lexicon. To enter the new word to the Lexicon, the Administrator type the Kachin word, choose the corresponding grammatical form for this word and type Myanmar meaning. If the entered word is "Verb" form, the "Verb" from the box is chosen.

Lexicon in this system refers to a table in a database which contains Kachin words and their Myanmar meanings and their Parts of Speech (POS).

There are (1388) words in the lexicon table. The words are stated according to their Parts of Speech type in the language: Noun, Pronoun, Adjective, Verb, Adverb, Preposition, and Conjunction. The system accepts the Kachin multi-word as an input and produces an equivalent Myanmar multi- word as an output. To prevent the interventation of the system, the Administrator account and password is used. The administrator can add the new words to the lexicon and update, delete the word.

Figure 3 describes the system architecture of language translation. The system has the following components:

1. Translation from Kachin multi-word to Myanmar – The system accepts the Kachin multi-word as an input text and translates Myanmar multi-word as an output.
2. Updating the Lexicon – The system allows the authorized user to update to the Lexicon.
3. Edit Lexicon – The system allows the users to edit, delete the lexicon.
4. Adding/ Updating Knowledge Base – The system allows the authorized users to add, update and delete to the Knowledge base.
5. Grammar Rule – The system shows the grammar rules of the multi-word.

4.1 Benefits of the System

By using this program and studying this system, the user can study the concept of Natural Language Processing (NLP) and multi-word expression in syntactic parsing. This system can handle the grammar rules and sentence rules of the Kachin- Myanmar language. This system can be applied as the Myanmar meaning for the Kachin students who are learning the Myanmar language.

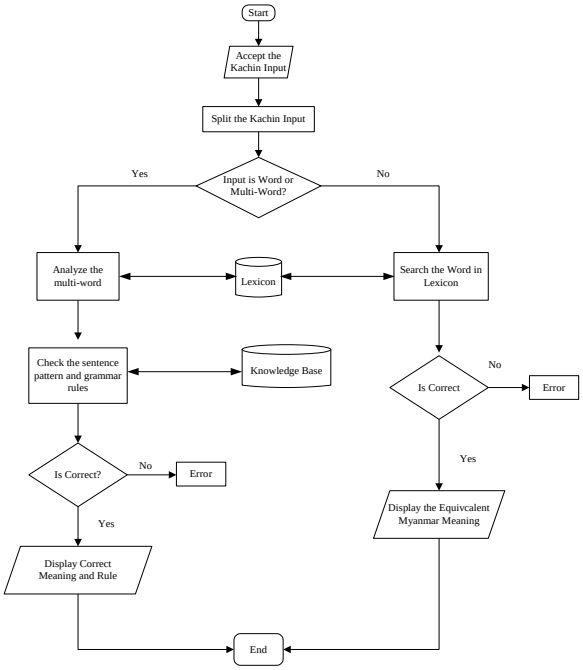

Figure 3. System Architecture of Language Translation

4.2 Limitation and Further Extension

This system can analyze only simple sentences. Paragraphs, question sentences and complex sentences will not be accepted. This system can only tokenize space. It cannot do any semantic works vice visa, from Myanmar to Kachin. This system can be used to give the general knowledge for anyone who interested in Myanmar and Kachin languages.

This system can be used on web page as translation application in the further. Furthermore, in order to perform highly reliable translation of excellent quality, it can be created to allow the computer to respond the voice output as speech recognition.

5. Conclusion

This system provides Natural Language Processing (NLP). This system especially emphasizes on the syntactic analysis of multi-word expression. This system accepts the Kachin language and generates the Myanmar language. This system is developed using parser, lexicon, knowledgebase and generator techniques and language translator.

The parser breaks the multi-word into its various parts of speech works together with the lexicon. In this

system, there are (1388) words in the lexicon but it allows adding the words by the authorized user.

The knowledge Base contains the sentences rules and grammar rules for translation from the Kachin language to Myanmar language.

The translation of natural language is one of the most important NLP interfaces. This system can help to overcome the language barrier between the Kachin and Myanmar languages. By studying this system, the users can more realized the syntactic analysis, building the lexicon, a tokenizer in natural language processing and knowledge Base. Moreover, this system can be applied as Kachin to Myanmar language dictionary for 1.5 million of Kachin people in Myanmar who are learning the Myanmar language.

6. References

[1] Efrain, Turbon
"Expert Systems and Applied Artificial Intelligence"
Macmillan Publishing Company, New York, 10022,
ISBN: 0-02-946565-6

[2] Eugence Charrid
"Introduction to AI"

[3] Haruyuki, Fujii
"Syntactic Analysis of Natural Language Instructions"
Showing a Way, 2001.Journal code: Y 0894A,
ISSN: 1340-4210

[4] Phyo, Myat Wut Yi
"Natural Language Translation Based On NLP: Japan-Myanmar Translation Process"
M.C.Sc (Thesis), January, 2009.

[5] Phyo, Wai Wai
"Syntactic Analysis of Grammar Rules in NLP"
M.C.Sc (Thesis), December, 2007.

[6] Thu, Khin
"Development of English to Myanmar Language Translator"
M.C.Sc (Thesis), June, 2008.