Phrase Reordering Translation System in Myanmar-English

Aye Thida Win
University of Computer Studies, Mandalay
Ayethidawin.ucsm@gmail.com

Abstract

Machine Translation is the attempt to automate all or part of the process for translation from one human language to another. This definition involves accounting for the grammatical structure of each language and using rules and grammars to transfer the grammatical structure of the source language (SL) into the target language (TL). Myanmar to English machine translation can be used to facilitate learning in beginner of Myanmar-English language learner or vice versa and to help them to study of grammar, Myanmar and English languages are linguistically different language pairs. The aim of this paper is to reassemble from the not ordering set of English word into proper English sentence. In this paper proposed the position of English POS tag rules. The resulted raw sentence from translation process is reassembling to form the English sentence. In this system, especially we consider subject/verb agreement process, article checking process and tense adjustment process will also be performed according to the English grammar rules. The proposed system of this paper is reordering approach for English sentence. Our proposed system, we considered the position of English POS tags and then which are swapping to get the proper English sentences by using the reordering rules and mapping the English grammar patterns.

1. Introduction

The object of NLP is to design and build computer based systems that have ability to read, analyze, understand and generate required results and inferences from given language scenario in text form. A word is reordered when it and its translation occupy different positions within the corresponding sentence. Many factors contribute to the difficulty of machine translation, including words with multiple meanings sentences with multiple grammatical structures uncertainty about what a pronoun refers to, and other problems of grammar. We propose the target phrase reordering approach to incorporate machine Translation System for making translation easier. In this paper, we proposed reordering algorithm, using local reordering, for example swapping of adjective and noun in languages pairs like Myanmar and English. The raw English sentence from translation process can be reordered to get correct sentence by using English grammar rules. The problem of the Machine Translation can be viewed as consisting of three phrases (a) analysis of the source language to chose appropriate target language lexical item (word or phrase) for each source language lexical items, (b) reordering phrase where the chosen target language string, and (c) disambiguation of word sense where the correct meaning of words is chosen for translation.

Translation is the process of moving texts from one language (source language) to another (target language). To build a Natural Language Translation System, a lexical analyzer is required in the first step which is breaking input text into individual syllabic words or token and define the limit of word boundary. Our proposed algorithm is rule-based machine translation system, which is expensive in terms of formulating rules. It

easily introduces inconsistencies, and it is too rigid to be robust. However, rules are usually used in general domain and not for specific domain.

2. Related Work

Statistical Machine Translation moved from words to phrase as basic units of translation. A word is reordered when it and its translation occupy different positions within the corresponding sentence present in (D. Guta, et.al) which described reordering of nouns, verbs and adjectives by taking into account target-to-source words alignments and the distances between source as well as target words. They described additional feature function in the re-scoring stage of a SMT system using BTEC corpus for Japanese-to-English task and on the Europal corpus for the German-to-English task and comparison of BLEU scores [2]. The phrase-based SMT has achieved high translation quality, it still lacks of generalization ability to capture word order differences between langue described in (T. P. Nguyen et.al) which present general method for tree-to-string phrase based SMT and design syntactic transformation models using unlexicalized form of synchronous context-free-grammars. They shown result of English-Japanese and English-Vietnames translation showed a significant improvement over two baseline phrase-based SMT systems and reported statistic of CFG rules, phrase CFG rules, word-to-phrase tree transformation (W2PTT) rules and reordering rules [6]. The reordering rules are learned from the word aligned corpus and are integrated into decoding process by constructing a lattice, which contains all word reordering according to the reordering rules in (K. Rottmann, et. al) described that reorders the source side based on Part of Speech (POS) information. Phrase translation pairs are learned from the original corpus and from a reordered source corpus to better capture the reordered word sequence at decoding time. The resulted present English-Spanish and vise visa of German and English translations using the European Parliament Plenary Session Corpus. They described at decoding time all permutation need to be considered, which is impossible for any but very short sentences [4]. Comparison of word order, construction of sentence described Myanmar language and English language. This had described position of POS tag and the sentence feature of Myanmar and English languages [1]. Machine Translation System from English to Myanmar language presented in (E.E.Han, et. al) which described their rules based on Ontology and construction of Ontology [3]. An approach to reorder chunked phrases of the source language before full parsing described in (M. T. Tun, et. al). Which presented can be used to prevent parse steps that lead to parse trees that would be removed by the filter after parsing. Chunking divides text into segments which corresponds to a certain syntactic units such as a noun phrase, verb phrase and rule based one. They proposed to reduce parses disambiguation to reduce time for target language post editing and to get efficient translation by providing an appropriate sentence structure for translation to the target language [3].

3. Machine Translation Process

Machine Translation is the process of translation from source language text into the target language. When a computer translated an entire document automatically and then presents in some natural language to a human, the process is called machine translation. Machine translation has never measured up to the quality of human translation. Many factors contribute to the difficulty of machine translation, including words with multiple meaning, sentences with multiple grammatical structures, uncertainty about what a pronoun refers to and other problems of grammar. Computational morphology 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. Morphlogical analyser gives information concerning morphological properties of the words it analyses. 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.

4. Architecture of Machine Translation

The system based on the machine translation architecture. Machine translation engine can be classified into two classes: Direct or Transformer Architecture and Indirect or Linguistic Knowledge Architecture. Transformer engine is that input sentences can be transformed into output sentences by carrying out the simplest possible parse, replacing source word with their target language equivalents as specified in a bilingual dictionary, and then roughly re-arranging their order to suit the rules of the target language. Linguistic Knowledge Architecture (LKA) is translation from source language to target language based on linguistic knowledge base that can be classified into three steps analysis, transfer and synthesis. The fact that the target grammar is being used means that the output of the system, the target sentences, are far more likely to be grammatically correct that those of a Transformer system. Same grammars for each language are used regardless of the direction of the translation.

4.1. Interlingual Machine Translation

Interlingual machine translation is a methodology that employs interlingua for translation. Ideally the interlingual representation of the text should be sufficient to generate sentences in any language. Languages can have different parts of speech. In some cases two or more words in one language have a equivalent single word in another language. Interlingua approach addresses these structural differences between languages. The disadvantage is that the design of interlingua is too complex. This is due to the fact that there is no clear methodology developed so far to build a perfect interlingual representation. An interlingual lexicon is necessary to store information about the nature and behavior of each word in the language. The information includes events and actions.

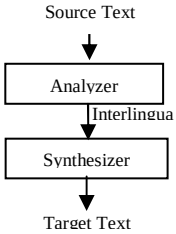

Figure 1: Interlingua Machine Translation

5. Background of Myanmar-English Translation

Statistical Natural Language processing methods are popular because one does not have to spend a long time learning and discovering all the rule of a language. In this system, the input Myanmar sentence is tokenized, segmented and tagged with POS. And then this sentence is translated word by word translation according to literal translation. We will get resulted raw sentence with POS tag which are not ordering sentence, the system will be ordered base on rules to get proper English sentence.

5.1. Input Process

The input sentence is the Myanmar sentence, that is receive from the user input. The two methods can be used for accepting input from user: from keyboard and mouse. In this process, we translated as English raw resulted sentence by using step by step language processing. For better understanding of the given sentence, it is necessary to divide the sentences into unit lexicon which are usually words.

5.2. Translation Process

The translation process of Myanmar to English translation project is developed. In this

phase, the input source sentence must be segmented by POS tag according to the Myanmar grammar rules. The segmentation process is to merge the syllables into word. The input Myanmar sentence can be segmented by the rules of prepositional phrase, for example, simple sentence: "မမသည်/ပန်းကို/ခူးသည်။" the former "သည်" is subject preposition, the word "ကို" is the object preposition and the latter "သည်" is verb preposition. In this system, we defined the noun phrase is "မမသည်" by segmentation subject preposition and "ပန်းကိုခူးသည်" is verb phrase that will be defined verb preposition. In our proposed system, we considered the input sentence to form of noun phrase and verb phrase and then translated this sentence word by word. The basic concept of Myanmar to English translation system are to see the end of source sentence , to adjust the transformation of Myanmar verb and English verb, to know the order of sentence and to take care the agreement process.

6. Example of Reordering

Table 1: Example of Myanmar to English translation

Original Myanmar Sentence(simple)	မမ+**သည်**+ ပန်း+**ကို**+ခူး+**သည်**
Literal English Translation	Ma Ma + flower + pick
Input (Set of English words with POS tag)	Ma Ma(Noun(p))+flower(Noun(CCS)+pick (Verb)
Reordered English Sentence	Ma Ma picks the flower.

The system is based on the rules of POS position. The raw English sentence from translation process will be reordered by using English grammar rules. Myanmar sentence structure is (SOV). For example sentence: "မမသည်ပန်းကိုခူးသည်။". In Myanmar to English translation process, we must be translated and by omitting the prepositional words in this sentence. The preposition words are described as the bold word in Table 1. The translated sentence is entered as input sentence in our system that is not ordering like that "Ma Ma flower pick.". The output of correct English sentence is "Ma Ma picks the flower". The systems are reordered the input set of word with their POS tag to get the proper sentence and add appropriate article word, subject/verb agreement and then produced the correct English sentence by using English grammar rule (SVO) for simple sentence. Firstly, we considered the above example sentence, subject/verb agreement; the verb "pick" must be add "s" because of the subject "Ma Ma" is singular. We must be add the definite article "the" in front of the word "flower" because the definite article is used with countable and uncountable noun by the grammar rules. The word "flower" is common countable noun. Although article is not in Myanmar language, use in English language. The indefinite article (a/an) is used with countable noun. Article checking is can be change upon the usage or sense of English sentence.

7. Reordering

Myanmar sentence feature and English sentence feature are linguistically distinct pairs. Myanmar sentence feature is SOV and English sentence feature is SVO. So, when we translate word by word, their position of POS tag are not the same. We considered to get the proper English sentence must be change the position of the words based on the position of POS tag. For example sentence: "ဘီးတစ်ဘီးသည်ပုံသေတပ်ဆင်ထားသောဝင်ရိုးဌလည်ခြင်းအားဖြင့်ရွေ့လျားသည်။"The ways of Myanmar sentence segmentation are above mentioned. This sentence will be divided by Noun Phrase and Verb Phrase. We defined "ဘီးတစ်ဘီးသည်" is "Noun Phrase" and "ပုံသေတပ်ဆင်ထားသောဝင်ရိုးဌလည်ခြင်းအားဖြင့်ရွေ့လျားသည်။" is "Verb Phrase". In this sentence, Verb Phrase can be subdivided such as place preposition, reasoning preposition and verb preposition. "ပုံသေတပ်ဆင်ထားသောဝင်ရိုးဌ" is place preposition, "လည်ခြင်းအားဖြင့်" is reasoning preposition and "ရွေ့လျားသည်" is verb preposition. The proper

English sentence of the example sentence are “A wheel moves by turning at a fixed axel”. The prepositional segmentation of Myanmar sentence described the following: ဘီးတစ်ဘီးသည်/ပုံသေတပ်ဆင်ထားသောဝင်ရိုး/၌/လည်ခြင်းအားဖြင့်/ရွေ့လျားသည်။

Table2: Literal translation with POS tags

Source sentence	ဘီးတစ်ဘီးသည်/ပုံသေတပ်ဆင်ထားသောဝင်ရိုး၌လည်ခြင်းအားဖြင့် ရွေ့လျားသည်။
Noun Phrase	ဘီး/တစ်/ဘီး/သည်
POS tag	Noun(CCS) +Adj (Numeric) +Particle(Kind) +Sub(prep)
Literal Translation	Wheel + a
Verb Phrase	ပုံသေတပ်ဆင်ထားသော/ဝင်ရိုး/၌ /လည်ခြင်း/အားဖြင့်/ရွေ့လျား/သည်။
POS tag	Adj (descri) + Noun(CCS) + prep(place) + Noun(CCS) + prep(Accusation) +Verb(V1)
Literal Translati -on	Fixed + axle + at + turning + by + move

We will get the resulted raw words with their POS tags by using lexicon described in Table 2. And then, we will get the resulted raw sentence by combining Noun phrase and Verb phrase that is “wheel a fixed axle at by turning move”. This sentence is entered in the system as the input process with their POS tag. In many cases, word by word translation in Myanmar to English can be seen reverse words orders. So, we generate the rules of POS position different between the Myanmar and English sentence feature easily to get the proper English sentence.

8. Rules of POS Position

In this system, we considered to change the position of English POS tag by swapping the words. Although the POS tag of phrase ဘီး/တစ်/ဘီး/သည် are Noun(CCS)/Adj(numeric)/Particle(kind)/Sub(preposition) by POS tag, when we translated can get “wheel a” from translation process because can be omit particle(kind) and sub(preposition) according to Myanmar to English translation. So, we had generated the rules for reordering by changing the position of Myanmar-English POS tag: For above example sentence will be get the following rules.

ဘီး+တစ်ဘီး

Wheel + a → a wheel **[(nounform) = (Adj+ Noun)]** Noun (CCS)+ Adj(numeric) → Adj +Noun(CCS)

(1) Noun (CCS) + Adj→ Adj +Noun (CCS) [noun form]

We generated rule (1), position of adjective in English are always place in front of noun.

ပုံသေတပ်ဆင်ထားသော+ဝင်ရိုး

Fixed+axle(noun form) Adj(dscri) + Noun (CCS) = not change

လည်ခြင်း+အားဖြင့်

Turning +by→ by turning

Noun(CCS)+Pre (Accusation) → Pre (Accu) + NCCS

(2) ing(NCCS)+Pre(Accusation)→Pre(Accusation)+ing(NCCS)

We generated rule (2), position of preposition “by” in English always take in front of noun.

ပုံသေတပ်ဆင်ထားသော+ဝင်ရိုး+၌

Fixed+axle+at→ at fixed axle

Noun form+Pre(place)→ Pre(place) + Noun form

(3) Noun form + Pre(place) → Pre (place) + Noun form

We generated rule (3), position of place preposition in English will always have in front of noun form.

In the resulted raw sentence with POS tag is “wheel a fixed axle at turning by move”. We must be defined the “wheel a” is Noun Phrase and “fixed axle at turning by move” is Verb Phrase. Therefore, we can be order these words in this two phrases by using the reordering rules. From the above example resulted raw sentences will be swap the word “wheel” and “a” by the rule (1). We will get the phrase “a wheel”, if the words are order form not consider to order , the word “fixed” and “axle” are match rule (1) but the phrase “fixed axle” and “at” must be swap because of rule (3). We can be get the phrase “at fixed axle” which is refer to place preposition in the sentence. And then, swap the words

“turning” and “by” according to the rule (2), the phrase “by turning” is represented the manner in the sentence. After swapping the words, we will get the proper ordering phrase from the above example sentence by changing the position of POS tag. According to the English grammar rules these sentence can be order (S + V + M + P) pattern. In this system, we generate proper English sentence “A wheel moves by turning at fixed axle”. Here, we consider the subject/ verb agreement the word (verb) “move” must be add suffix “s” because the subject “a wheel” is singular noun. Another example sentence is “ဘီးတစ်ဘီး၏ပုံသဏ္ဌာန်သည်လုံးဝိုင်းသည်။” We can be defined ဘီးတစ်ဘီး၏ပုံသဏ္ဌာန်သည် is Noun Phrase and လုံးဝိုင်းသည် is Verb Phrase. We can get the resulted raw sentence from the translation is “wheel a of shape” and “is round”. In this sentence we must be order the Noun phrase. The word “wheel” and “a” can be swap by rule (1).

၏+ပုံသဏ္ဌာန်

Of + shape→ shape + of

Pre (possessive) + Noun (CCS) → NCCS +Pre (posses)

(4) Pre (possessive) + Noun (CCS) → Pre (posses) +NCCS

We generated rule (4), position of possessive preposition in English is always take place in front of noun.

So, the word “of” and “shape” must be swap by rule (4). We will get the two phrases “a wheel” and “shape of”, the phrase “a wheel” is noun form and the phrase “shape of” is a possessive form, which can be swap by using rule (5).

(5) Noun form+ Possesive form→Possesive form+Noun form

ပုံစံ+အမျိုးမျိုး+ဖြင့်

Form+various+in→in various form

[Noun(CCS)+Adj(numeric)]=noun form Pre(Accusation)

(6) Noun form + Pre (Accusation) → Pre (Accusation) + Noun form

စက်သီး+၏ +ရည်ရွယ်ချက်

Pully+ of + purpose→ purpose of pully

Noun(CCS)+pre Noun(Bj)→ Noun(Bj) + pre(possessive) + Noun(CCS)

(7) Noun (CCS) + Pre(possessive) +Noun (Bj) → Noun (Bj) + Pre (possessive) + Noun (CCS)

ပြောင်း+ရန်

Change + to→ to change

Verb + pre(intention) → pre(intention) + verb

(8) verb + Pre (intention) → Pre(intention) + verb

So, we can get the sentence “shape of a wheel” and then we consider the article checking process. Because of the proper English sentence should be “The shape of a wheel is round”. The article word “the” must be add in front of the word “shape” which is definite noun and then the prepositional phrase “of a wheel” is follow the “shape”, so we must be add the article in this sentence by the English grammar rules. This sentence can be ordered (S + V) pattern. Above all of these rules will cover almost 6 sentences and match the English patterns which are: (S + V + O), (S + V + O + M), (S + VI + P)

9. Kinds of Sentence in Myanmar and English

In English, a sentence refers to a set of words expressing a statement, a question or an order, usually containing a subject and a verb. A combination of words which are systematically arranged in order to convey intended meanings forms a sentence in Myanmar. A sentence consists of at least a subject and a verb. In English, sentences can be divided into four kinds: Declarative Sentence, Interrogative Sentence, Imperative Sentence and Exclamatory Sentence. Besides, it has another classification of three types: Simple Sentence, Compound Sentence and Complex Sentence. In Myanmar, based on their meanings, sentences can be divided into five kinds: Declarative Sentence, Interrogative Sentence, Negative Sentence, Imperative Sentence and Desiderative Sentence. Moreover, according to their formation, sentences are of two kinds: Simple Sentence and Compound Sentence position of pos tag in English. In English, owing to the fewness of the inflexions, the order or arrangement of the words in a sentence is of the first importance.

8.1. Proposed System

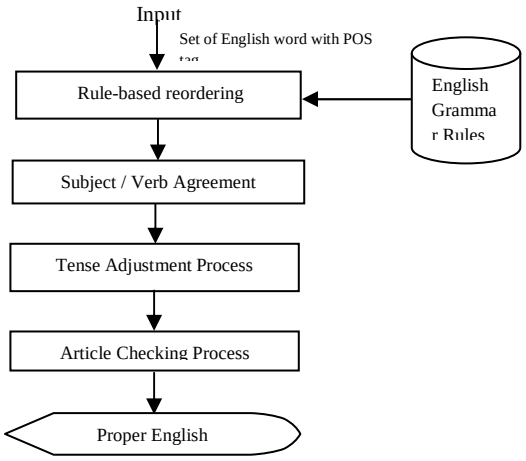

Figure 2: Proposed system of Architecture

The input sentence is set of English word; these words are reordered by English grammar rules. We will get the set of English words from translation process, these words are not arrangement word. So, the system must be reordered by using above mention rules. In the subject/verb agreement process, firstly we will see the subject if the subject is singular which must be followed singular verb(be) if the subject is plural which must be followed plural verb(be) for example: "မောင်ဘသည်ကျောင်းသားဖြစ်သည်။" Myanmar sentence feature (SOV) and English sentence feature (SVO) are distinct language pairs. We will get set of English words from translation process is "Mg Ba student is". In our system will be reordered these words like that "Mg Ba is student." In article checking process if necessary we will add article (a, an, the) in the sentence. We must be add article "a" above example sentence because "Mg Ba" is singular subject and "student" is subject complement (S.C). So, we will add article "a" in front of the word "student" by the proper English sentence. And then, we will get the proper English sentence is "Mg Ba is a student". According to English grammar rules we must be ordered above example sentence by S+V grammar pattern.

10. Conclusion

In this paper we presented words to phrase reordering and generate the rules which can be incorporated to Machine Translation System. Consider this system is reordered the raw English sentence from translation process in Myanmar to English NLP system. The system can also predefined source sentence structure to encourage clear writing which can improve the quality of the source text and the translation output. In this paper, we presented the example of reordering rules for simple sentence and complex sentence. The ongoing research will be described many algorithm for reordering that must be cover any type of sentence and mapping many English grammar patterns.

References

[1] A.Z. Min, N.N. Hlaing, "A Comparative Study of the Two Grammatical Systems of Written English & Myanmar and its Significance to Learning English as a Foreign Language", Department of English, University of Mandalay, Myanmar, 2009, May.

[2] D.Gupta, M.Cettolo, and M. Federro, "POS-Based Reordering Models for Statistical Machine Translation", FBK-irst, Centro per la Ricerca Scientifica e Technologica.

[3] E.E. Han, N.L Thein, "A Novel Approach for Myanmar Language Synthesization", University of Computer Studies, Yangon, 2007, February.

[4] K. Rottmann, S.Vogel, " Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model.

[5] M.T.Tun, N.L.Thein, " Automatic Phrase Reordering Approach for Machine Translstion", Fourth International Conference ICCA 2006, University of Computer Studies, Yangon, Myanmar.

[6] T.P Nguyen, " A Tree-to-String Phrase-Based Model for Statistical Machine Translation System, College of Technology Vietnam National University, Hanoi, CoNLL 2008: