HIERARCHICAL PHRASE-BASED ENGLISH – INDONESIAN MACHINE TRANSLATION USING ADJ TECHNIQUE

Teguh Bharata Adji
Electrical Engineering & Information Technology Department
Engineering Faculty, Gadjah Mada University
Yogyakarta, Indonesia
E-mail : adji@mti.ugm.ac.id

Abstract
In this paper, annotated disjunct (ADJ) technique is employed to develop hierarchical phrase-based transfer rules. We also developed an English to Indonesian Machine Translation (MT) system using those transfer rules to be compared with the phrase-based system, which is our previous version that outperforms other available English to Indonesian MT systems.

Keyword: hyrarchical, phrase-based, machine translation, ADJ technique.

I. INTRODUCTION
There is a need for a mean to translate information in many kinds of digital information into Indonesian language since most Indonesian people is not an English active speaker. A notable MT activity for Indonesian language is the Multilingual Machine Translation System (MMTS) project [1]. This MMTS includes BIAS (Bahasa Indonesia Analyzer System), an analysis component for Indonesian language part [2]. BIAS uses interlingual approach which relies on the syntactic and semantic rules.

Recently, there are some available English-Indonesian MT software: Rekso Translator1, Translator XP2, and KatakuTM3. No details were made available on the algorithm applied in their translation engines. Another English-Indonesian MT system is found in Google Translate application that uses phrase-based statistical approach [3]. However, [4] still found that the accuracy of this work depends on the language pairs and also varies with the training corpus size.

Indonesian language has the same root and hence shares many aspects with the Malay language. An intensive work in the field of MT was conducted in University of Science Malaysia (USM) by using example-based methods. A technique to construct the Structured String-Tree Correspondence (SSTC) for the Malay sentence by means of a synchronous parsing technique was introduced [5]. The advantage is this technique can solve non-projective cases. The limitations include extra work required to annotate all constituent levels and the need to formalize the English and Malay grammars.

Meanwhile, a team of NLP students from Gadjah Mada University, Indonesia, has built English-Indonesian MT application using Visual Basic [6]. They proposed a direct approach. Pure direct approach is no longer used, but the technique is still employed in more modern MT system at present. Instead of merely using this approach, some modules in [6] was embedded into the rule-based MT system as explained in [7]-[8]. This hybrid approach – namely ADJ technique – is able to solve one-to-one, one-to-many, and many-to-many word(s) mapping which are very important requirements in an MT system. The limitation of the system is its inability to solve non-projective cases (the problems arise in the dependent or constituent levels). The advantages of the technique are: 1) ADJ uses Link Grammar (LG) formalism [9]-[10], which is more closely related to human intuition than dependency or constituency grammar formalism [11], 2) ADJ only needs parsing for the source language (English), 3) ADJ is suitable to MT where the target language does not have corpus and parser, such as Indonesian language, 4) ADJ is a syntact-based MT, in which its components (the grammars, parsers, and generators) are useful for other specific syntactic research purposes [12].

On the other hand, many statistical MT systems have improved their quality with the use of phrase-based translation [4], such as phrase-based model of the ADJ-based MT system that outperforms other available English to Indonesian MT systems [13]. Moreover, [14] stated the possibility to gain insights from the strengths of the phrase-based extraction model to increase both the phrasal coverage and translation accuracy of the syntax-based model.
However, problems arise during the decomposition of hierarchical phrases. This problem is solved by the implementation of hierarchical phrase-based transfer rules into the ADJ-based MT system as discussed in following lines.

It was presented by Chiang [15] that a hierarchical phrase-based MT system performed significantly better than the Alignment Template System (a state-of-the-art phrase-based system proposed by Och and Ney [16]) in a comparison using BLEU as a metric of translation precision. Additionally, Lopez [17] stated the following statement for the implementation of hierarchical phrase-based MT system:

“Given an input sentence, efficiently find all hierarchical phrase-based translation rules for that sentence in the training corpus.” [17]

Those achievements encourage us to incorporate hierarchical phrase translation method into the ADJ-based method [7] as explained in the following section.

2. METHODOLOGY

300 training sentences are used in this research. After looking through these training data then they can be assumed to have three layers of phrases maximally. This finding let us to define a hierarchical phrase as a phrase that consists of sub phrases where each sub phrase may or may not consist of sub sub phrases, which has more explanation but is not opposed to the definition by Chiang [15]:

“… hierarchical phrases—phrases that contain subphrases.” [15]

We call the first top phrase layer (or a phrase that consists of two phrase layers) as the third group phrase, the second top phrase layer (or a phrase that consists another phrase layer) as the second group phrase, and the last layer from the top (or a phrase that no longer consists any phrase layer) as the first group phrase. Each of these groups is solved by generating what we call first group transfer rules, second group transfer rules, and third group transfer rules respectively. The diagram of the transfer rules are thus divided into three as seen in Figure 1.

![Diagram](http://example.com/diagram.png)

Figure 1. The word pergi and its translations

This diagram explained that the transfer rules receive an ADJ set as its input. ADJ set is a set of source word (English), target word (Indonesian), and the associated disjunct as explained in [18]. This disjunct represents the uniqueness of word pair alignment that were useful in the transfer rules algorithm in English to Indonesian translation mapping. Based on the source words and annotated disjunct, which are the elements of the ADJ set, the transfer rules then classify which of the sequence of source words can be categorized as the first group phrase. When a first group phrase is found, first group transfer rules process and translate the phrase into a correct first group target phrase. The transfer rules then identify other sequence which can be categorized as the second group phrase, and translates the phrase into second group target phrase while repositioning the first group target phrase inside the second group target phrase. Finally, the transfer rules categorize other sequence of the third group phrase, translate the phrase into the third group target phrase, and reposition the second group target phrase inside the third group target phrase.

3. HIERARCHICAL PHRASE-BASED TRANSFER RULES

32 transfer rules have been developed for the ADJ-based MT system and these rules are categorized into three groups. Those three groups and the examples of the transfer rules belonging to each of the groups are described in the following sub sections.
A. First Group Transfer Rules

This group consists of simple rules, which handle phrases consisting of words (e.g. prenominal adjectives, superlative adjectives, and noun-modifiers) that modify nouns, handle phrases of adverbs modifying adjectives, and handle phrases of determiners followed by nouns in idiomatic time expressions. The following lines explain how the first group transfer rules handle phrases consisting of prenominal adjectives that modify nouns. In English grammar, prenominal adjectives always precede nouns. Oppositely, in Indonesian grammar, adjectives always come after nouns. Therefore, the translation of an English phrase consisting of those kinds of words into Indonesian is done with the swapping between the Indonesian prenominal adjective and noun. Let us consider an English phrase “red car” with its link as seen in Figure 2.

![Figure 2. A phrase with A link connecting a prenominal adjective and a noun](image)

The linkage in Figure 2 has an A link which shows the prenominal adjective “red” which precedes the noun “car”. The English-to-Indonesian words mapping of the phrase is illustrated in Figure 3. The mapping is divided into two processes. Process A translates each word in the English phrase into the Indonesian word. The English phrase “red car” consists of two words, “red” and “car”.

![Figure 3. Translation mapping of prenominal adjective-noun phrase](image)

The word “red” has an empty left connector and a A right connector. The word “car” has a A left connector and an empty right connector. Based on these source words and their disjuncts, the English prenominal adjective “red” is translated into the Indonesian adjective “merah” and the English noun “car” is translated into the Indonesian noun “mobil”. Thus, process A results in the Indonesian target words “merah mobil”. Subsequently, process B swaps the position of the target words “merah” and “mobil” to get a grammatical Indonesian phrase “mobil merah”.

B. Second Group Transfer Rules

This group consists of rules with more complexity than those of the first group, such as rules for phrases that consist of demonstrative pronouns or determiner “the”, possessive adjectives, and possessive nouns. How the second group handles phrases consisting of demonstrative pronouns. Demonstrative pronouns or determiner “the” always precede nouns or noun phrases in English grammar. These contradict with Indonesian grammar where Indonesian demonstrative pronouns or determiner “the” are located after nouns or noun phrases. Let us assume an English phrase “this red car” and its link as seen in Figure 4.
D link in Figure 4 shows the demonstrative pronoun “this” which precedes the noun phrase “red car”. The translation of the phrase in Figure 4 is depicted in Figure 5.

Figure 5. Translation of a demonstrative pronoun preceding a noun phrase

C. Third Group Transfer Rules

4. ANALYSIS OF THE DEVELOPED SYSTEM

A hierarchical phrase-based MT is developed and the precision is compared with that of the previous version of our system using a tool so called BLEU tool. This tool is based on BLEU metric and was developed in C# by Adji as also stated in [13]. The previous version is phrase-based system. The hierarchical phrase-based system precision increased with a slight difference higher than the previous phrase-based system precision for 3-gram, 4-gram, and 5-gram by 0.38%, 0.34%, and 0.28% gaps, respectively. Most of the solved and unsolved cases that were found in the previous phrase-based system were also found in the hierarchical phrase-based system. However, there were cases that could be translated by the hierarchical phrase-based system with higher precision which were found in interrogative sentences like “What magic do you use?”, in adjective clauses such as “the money he had”, “the moment I dreamed of”, “Robin Hood, who slipped away”, and “The old woman, who was really the fairy”; in passive sentences for examples “will be well paid”, “had been promised”. In addition, the use of hierarchical phrase-based module, as suggested in the discussion of phrase-based MT system [18], could correctly translate cases in possessive noun phrases. Possessors with two words or more were no longer ignored by the Hierarchical Phrase-Based Transfer Rules Algorithm for instances in possessive noun phrases “the fisherman’s story”, “the lion’s cry”, “Baby Bear’s little bed”, “the handsome prince’s future”, “the Incredible family’s problems”, and “a girl’s game”.

Phrases in interrogative sentences, possessive noun phrases, phrases in adjective clauses, and phrases in passive sentences contributed significantly to the unsolved cases in the previous phrase-based MT system using ADJ technique. Thus, incorporating the hierarchical phrase-base module is valuable. The numbers of transfer rules generated in the hierarchical
phrase-based translation system are 32 only. These are fewer than those generated in the phrase-based system, of which 60 rules. Moreover, the algorithm hierarchy makes ease the effort of generating and managing the transfer rules as the rules can be classified into simple (or the first group), more complex (or the second group), and most complex (or the third group) rules.

5. CONCLUSION AND FUTURE RESEARCH

The hierarchical phrase-based transfer rules can generalize the transfer rules for similar translation cases which in turn reduce the number of transfer rules thus will ease the effort of transfer rules generation. The accuracy depends on the number of transfer rules. Higher accuracy is expected to be achieved with more example data fed to the MT system for generating transfer rules, instead of using only 300 bilingual sentences. The developed system outperforms the phase-based system, however its accuracy is still low (69.87%). Generally, the developed hierarchical phrase-based MT system translated simple, compound, and complex English sentences in present, present continuous, present perfect, past, past perfect, and future tenses with better precision than other systems.

The limitation of the hierarchical phrase-based MT system still cannot solve several phrases which are found mostly in interrogative sentences, adjective clauses, negative sentences, and passive sentences.

There are more issues that can be discussed in this research which were not found in the example data, such as sayings and phrasal verbs. Hence, a database of all sentences in both categories which consists of phrase-by-phrase mapping, that translates the whole source phrase into the whole target phrase as it is, is needed. Another way is by the combination of the developed hierarchical phrase-based method with the hierarchical phrase-based SMT method. The ambiguous Indonesian meaning of the English words for some cases could not be solved with the disjunct annotation. Thus, ontology and the use of other parameters, such as the word class, are possible approaches to address this problem.

REFERENCES


