

The Development of Phrase-Based Transfer Rules for ADJ-Based Machine Translation

T.B. Adji¹, B. Baharudin², and N. Zamin²

¹Department of CIS, Universiti Teknologi PETRONAS, Malaysia
EE Department, Gadjah Mada University, Indonesia
Email: adj@mti.ugm.ac.id

²Department of Computer and Information Science
Universiti Teknologi PETRONAS, Malaysia
Email: {baharbh, norshuhani}@petronas.com.my

Abstract— In this paper, annotated disjunct (ADJ) technique is discussed to develop phrase-based transfer rules. We developed an English to Indonesian Machine Translation (MT) system using those transfer rules and compared with three available MT software and our earlier prototype, which uses sentence-based ADJ technique. It was found that the developed system outperforms other systems. In addition, the phrase-based approach generalized the transfer rules, thus reduced the number of transfer rules.

Index Terms— ADJ-Based MT, Phrase-Based Transfer Rules, Machine Translation, Link Grammar

I. INTRODUCTION

Considering the vast amount of available digital information nowadays is in English, there is a need for a mean to translate this information into another language. This work contributes an effective solution which automatically translates information mostly provided digitally in English and provides a method to Natural Language Processing (NLP) society to develop a major – less-resourced language pair MT system, such as an English-Indonesian MT system.

There were several MT research activities for Indonesian language such as the Multilingual Machine Translation System (MMTS) project which uses interlingual approach [1] and two statistical MT systems; namely Google Translate application that uses phrase-based statistical approach [2] and a system developed by BPPT and ANTARA [3]. In the recent years, there are few available English-Indonesian MT software: Rekso Translator¹, Translator XP², and KatakunTM³. However, no details were made available on the algorithm applied in their translation engines. A team of NLP research students from Gadjah Mada University, Indonesia, built English-Indonesian MT system using direct approach [4]. Some modules in this system were embedded into ADJ-based English-Indonesian MT system [5]-[6].

Recently, many statistical MT systems have improved their quality with the use of phrase-based translation such as the statistical phrase-based in [2], the syntax-based phrase translation system in [7], and the joint-probability model for phrase translation [8]. In addition, a comparison between the syntax-based MT model and the

phrase-based statistical MT model based on two models of the most successful MT systems in the NIST⁴ 2006 MT evaluation, namely the statistical phrase-based model [2] and the statistical string-to-tree model [9], has shown that the phrase-based model can consistently gain all the phrase translations based on the computed word alignment, concatenate and reorder those phrases using several cost models. It was also reported that generally rule-based phrase translation contributes to the better results of the sentence translation [10].

Those findings motivate us to incorporate phrase translation method into our previous ADJ-based MT approach [11]. This is done by dividing the source sentence (*SS*) into phrases before the mapping process. Based on the ADJ set of each phrase, each source phrase is mapped into a target phrase. All target phrases are then merged to obtain the target sentence (*TS*). The following briefly explains two examples on how our phrase-based transfer rules can handle standard case and non-standard case translations respectively. A standard case is a projective correspondence e.g. one-to-one word foreseeable mapping. While a non-standard case is not projective correspondence e.g. scrambling, cross serial dependencies, etc. [12].

II. PHRASE-BASED TRANSLATION USING ADJ

The disjunct annotation in the ADJ set represents the uniqueness of word pair alignment. These phenomena were useful in the transfer rules algorithm we developed in [11]. This transfer rules is a sentence-based since it takes into account all disjuncts of all words in an *SS*. This needed tedious work in the development of the transfer rules for all cases. In other words, transfer rules generalization for similar cases in the translation process was never obtained. Hence, we developed phrase-based transfer rules module to resolve the problem. This is done by dividing the source sentence (*SS*) into phrases before the mapping process. Based on the ADJ set of each phrase, each source phrase is mapped into a target phrase. All target phrases are then merged to obtain the target sentence (*TS*). The following sub sections explain two examples on how the phrase-based transfer rules can handle a standard case and a non-standard case translation respectively.

¹ <http://reksotranslator.com>

² <http://translatorxp.com>

³ http://www.toggletext.com/katakun_trial.php

⁴ <http://www.itl.nist.gov/iad/mig/tests/mt/>

A. Handling Standard Case

To handle standard case means that the system need to be able to do the simple one-to-one mapping from English to Indonesian. Example of standard case translation is the mapping of English noun phrase to Indonesian, since this kind of English phrase is one of phrases which are frequently translated in one-to-one mapping. English noun phrase can be in many forms such as:

- determiner + noun e.g. “the car”,
- determiner + superlative adjective + noun e.g. “the best car”,
- determiner + adjective + noun e.g. “the red car”,
- determiner + adjective + adjective + noun e.g. “the big red car”,
- possessive adjective + noun e.g. “his car”,
- noun-modifier + noun e.g. “car seat”.

How the ADJ technique can solve noun phrase translation is explained by examining the most difficult case i.e. “the big red car”, which involves multiple adjectives. Since link parser is a sentence-based [13]-[14], the above phrase has a linkage only in a sentence e.g. “I saw the big red car”. We can obtain from this sentence the ADJ set for the phrase $\{(the, itu, ((D))), (big, besar, ((A))), (red, merah, ((A))), (car, mobil, ((O,D,A)))\}$ as illustrated in Fig. 1. Afterward, our system needs to decompose the input sentence into phrases.

Ref. [15] explained that a phrase is substrings of potentially unlimited size (but not necessarily phrases in any syntactic theory). Our definition of a phrase is an extension to the phrase definition i.e. a phrase is referred to either a single word (in case the word is not part of any phrase in a sentence) or a collection of words with specific connectors. For example, we define a noun phrase made up of single/multiple adjectives (with A right connector) followed by a noun (with A left connector) as an adjective-noun phrase. Meanwhile, a determiner-noun phrase is defined as it consists of a determiner and a noun / noun phrase. Thus, the phrase “the big red car” will be decomposed into two phrases as can be seen in Fig. 2. The 1st decomposition results in an adjective-noun phrase “big red car”. These adjectives (“big” and “red”) precede the noun “car”. It is not valid for the corresponding Indonesian phrase since adjectives (“besar” and “merah”) always follow the noun “mobil” (car). Hence, the mapping is done by swapping the target adjectives and noun. This swapping process is done when a phrase transfer rules identify a source word with A right connector followed by another source word which contains A left connector. This swapping technique is resolved using stack implementation. Afterward, this grammatical target words “mobil merah besar” is grouped into a target adjective-noun phrase. The 2nd decomposition yields an English determiner-noun phrase, which consists of a determiner “the” and a noun phrase “big red car”. This determiner “the” precedes the noun phrase. Meanwhile, the corresponding Indonesian phrase must have its determiner “itu” (the) following the noun “mobil” (car). This problem is resolved by swapping the

target determiner and noun which is done when the phrase transfer rules identify words with D right connector followed by another word or noun phrase which contains D left connector. Finally, a grammatical Indonesian phrase “mobil merah besar itu” is obtained.

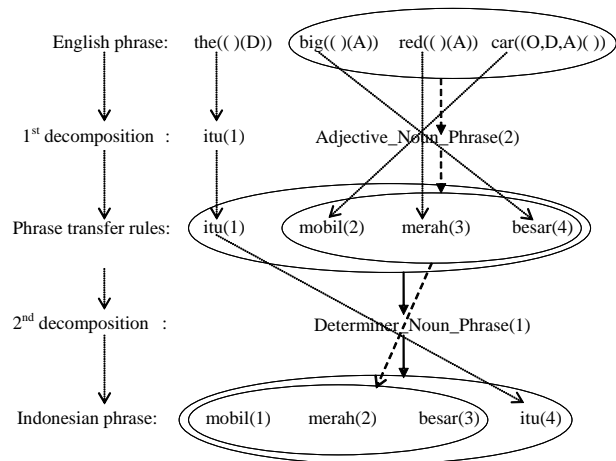


Figure 1. Handling a standard case.

B. Handling Non-Standard Case

There are some non-standard phenomena exist in phrase-based translation between both languages, such as a phrase “might be” in a sentence “She might be useful” which is translated into the Indonesian phrase “mungkin” as illustrated below:

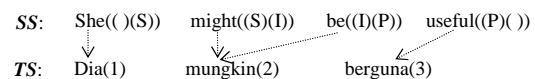


Figure 2. Handling a non standard case.

In this translation, many-to-one word mapping exists where two words “might” and “be” in the SS correspond to “mungkin” in the TS. The ADJ set for the phrase are $\{(might, mungkin, ((S)(I))), (be, “ ”, ((I)(P)))\}$. Thus, the alignment is resolved by mapping “be” into “ ” (blank) if “be” has a disjunct of (I)(P).

III. BUILDING PHRASE-BASED TRANSFER RULES

The previous section explains the way the phrase-based transfer rules handle a standard and a non standard phrase. How does the system translate a sentence with multiple phrases? Firstly, the sentence is pruned and parsed to obtain ADJ set [11]. The ADJ set is then process the transfer rules algorithm which has two layers. The first layer is the Phrase Transfer Rules for mapping each source phrase into its target phrase. This layer has two functions: *Identify_source_phrases()* and *transfer_rules_algorithm()*. The second layer is the Merge Target Phrases for merging all target phrases into a correct target sentence using *Compose_sentence_by_merging()*. Fig. 3 depicts the

entire MT system that consists of two layers transfer rules, drawn inside the dashed box. Note that these two layers transfer rules are the contributions of this work.

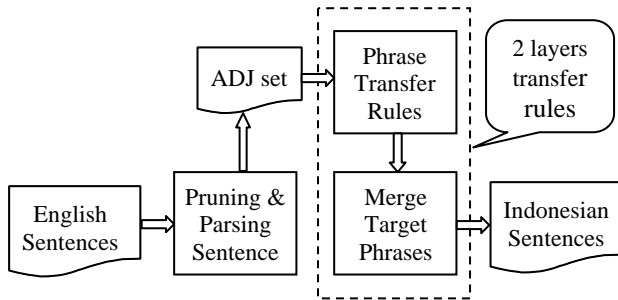


Figure 3. Phrase-based transfer rules diagram.

The algorithm used in the phrase-based transfer rules for applying the above diagram is as below:

1. *Identify_source_phrases(ADJ_SET)*.
2. For each source phrase, do step 2.1.
- 2.1. Run *transfer_rules_algorithm(en_phrase_words, in_phrase_words, en_disjuncts)* to obtain *target_phrases*.
3. *Compose_sentence_by_merging(target_phrases)*.

The variables and functions used in the above algorithm are explained as follows:

- *ADJ_SET* is a set of source words, target words, and the source word disjuncts for an English sentence,
- *transfer_rules_algorithm()* is a function to map a sequence of source words into a correct sequence of target words (explained in detail in [11]),
- *en_phrase_words* is the English phrase words, *in_phrase_words* is the Indonesian phrase words, and *en_disjuncts* is the English phrase word disjuncts,
- *compose_sentence()* is for composing correct target sentence from all target phrases,
- *target_phrases* is all Indonesian phrases as results of line 2.1.

In the above phrase-based transfer rules algorithm, line 1 identifies all phrases in the source sentence. If an input sentence is “That might be the car” (see Fig. 4), the word “that” and two phrases (“might be” and “the car”) are identified. Note in this stage that if a word does not belong to any phrase than it is considered as a phrase with a single word. The *transfer_rules_algorithm* will translate from the most left, starts from “That” translated into Indonesian “Itu”, then the phrases “might be” and “the car” translated into Indonesian phrases “mungkin” and “mobilnya”. The next step is to group these target phrases. As there are no longer available phrase in the source sentence, the *Compose_sentence_by_merging()* function merges the target word “Itu” and all the target phrases (“mungkin” and “mobil itu”) into a complete Indonesian sentence “Itu mungkin mobilnya”.

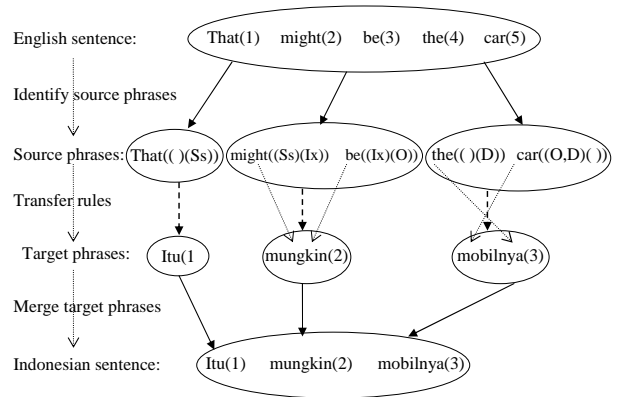


Figure 4. Handling multiple phrases.

IV. EVALUATION OF THE PHRASE-BASED MT SYSTEM USING ADJ

A BLEU metric tool, which we developed in C# for this research, is utilized to evaluate and compare our system performance. Four reference translations for each of 127 example SSs were used during the verification of the BLEU metric [16]. Based on this, we also used four reference translations; but only used 150 SSs, which were selected randomly from 30 English story books. The precision of all tested systems in BLEU metric is shown as follow:

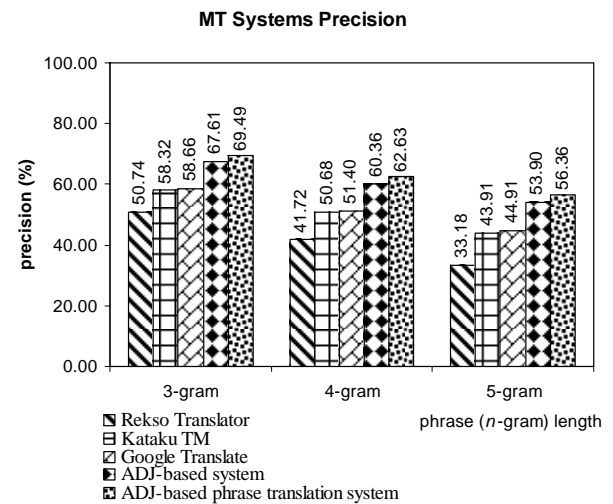


Figure 5. Precision of English-Indonesian MT systems using BLEU metric tool.

Fig. 5 shows consistent results where the phrase-based translation system precision increased slightly about 2% higher than the previous ADJ-based system precision [11] for all *n*-gram length and outperformed other systems with more than 10% different. We classified two cases (among 150 testing data) namely solved cases and unsolved cases into ten categories as shown in Table I. Note that the solved cases category consists of all cases which contribute to the increase of the system accuracy.

TABLE I.
THE NUMBER OF SOLVED AND UNSOLVED CASES FOR EACH PHRASE
CATEGORY

Category	English Phrase Category	Total Solved Cases (%)	Total Unsolved Cases (%)
I	Idiomatic time phrases	1.33	0.00
II	Infinitive phrases	4.67	1.33
III	"ing" form phrases	2.00	2.00
IV	Phrases with pronoun "one"	0.00	2.00
V	Phrases in interrogative sentences	0.60	2.67
VI	Possessive noun phrases	2.67	4.00
VII	Phrases in adjective clauses	0.00	4.00
VIII	Phrases in negative sentences	5.33	4.00
IX	Phrases in passive sentences	0.00	5.33
X	Other phrases	0.00	7.33
Total		16.60	32.67

We define a single case as a sentence which contains at least a problematic phrase that is either solved or unsolved in this work. The solved cases mean cases which were correctly translated by the developed phrase-based transfer rules algorithm.

The number of unsolved cases is about twice of the number of solved cases. This result prompted us to explore the causes. It was found that Category X (other phrases category) contributed to the highest unsolved cases with total cases of 7.33% but 0.00% cases have been solved. Most of this category were phrasal verbs like "dream of" and "go in"; ambiguous phrases such as "there were three bears" and "May I have it?"; and sayings like "What a wonderful day". The disjuncts generated by the link parser for this kind of cases could not be utilized to translate the phrases correctly. Hence, to get the correct translation of the sayings, ambiguous words, and phrasal verbs, a database of all sentences in this category that consists of direct phrase-by-phrase mapping is needed. Another option would be the combination of the developed phrase-based method with the statistical-based method. However, phrases that consist of noun-modifiers such as "the waiting truck" and noun phrases with determiners such as "every other dog", which were also in this category, have connectors that could be used to translate correctly. "waiting" has AN right connector and was translated by the system into a single Indonesian word "menunggu" while the correct translation is the three Indonesian words "yang sedang menunggu". In our transfer rules, AN connector was utilized to solve noun-modifiers in root form of nouns only, not for noun-modifiers in "ing" form of verbs. Thus, adding a rule for "ing" form of verbs will solve the problem.

High percentages of unsolved cases have also been identified in Category V or phrases in interrogative sentence (4.00%) such as "May we go" and "will you fight"; Category VI or possessive noun phrases (4.00%) as in "the fisherman's story" and "the lion's cry"; Category VII or phrases in adjective clauses (4.00%) such as "the money he had" and "Phil, who was half man"; Category VIII or phrases in negative sentences such as "don't stop" and "is not a human"; and Category IX or

phrases in passive sentences such as "will be well paid" and "was grabbed".

Categories V, VII, and IX need to be evaluated extensively since all cases were unsolved except one case (0.60%) in Category V. These categories require morphological analysis and/or construction. For example, a word "May" in "May we go" (Category V) was translated into the Indonesian inflectional interrogative word "Bolehkah". Meanwhile, "may" in a declarative sentence "You may go" is translated into "boleh", without the suffix "kah".

Surprisingly the ADJ algorithm can generate different disjuncts for the word "may" in both forms. The generated ADJ sets are $\{(May, bolehkah, ((Q)(I, SI)))\}$ and $\{(may, boleh, ((S)(I)))\}$ consecutively. Thus, adding a morphological construction to compose interrogative word "bolehkah" can be done simply by adding "kah" to the target word "boleh" when the algorithm identifies Q connector on the left of the word "boleh". Two opportunities arise to improve the system accuracy i.e. adding a word stemmer and a morphological analyzer/ construction for the ADJ approach, which is likely to solve phrases in interrogative sentences, phrases in adjective clauses, and phrases in passive sentences.

Some cases in Categories VI and VIII could be resolved but some cases in the same categories could not be resolved. For instance, in Category VI, a possessive noun phrase of "Aladdin's good fortune" could be solved by the developed transfer rules while another possessive noun phrase of "the fisherman's story" was unable to be solved. It was found that the developed transfer rules only considered a single word of possessor e.g. "Aladdin". Hence, we need to find a mechanism that will not neglect the possessor with two words or more in the possessive noun phrase e.g. "the fisherman" in the phrase "the fisherman's story". However, the mechanism seems to be complicated. Therefore, hierarchical phrase-based mechanism will be one of good options to work out.

The lower percentages of unsolved cases were belong to Category II i.e. infinitive phrases and Category III i.e. "ing" form phrases. Most cases in the infinitive phrases were solved. However, unsolved cases arise when the infinitive must be translated into an Indonesian passive verb. For instance, "to wear" in "for Cinderella to wear to the Ball" must be translated into the Indonesian verb "dipakai". Meanwhile, "to wear" was translated into an active verb "memakai" by the developed system. The other unsolved cases are found when ambiguous infinitives appear e.g. "to", which must be translated into "untuk" (for), was translated into "ke" (to). One way to solve the cases in Category II is by combining phrase-based statistical translation with the developed system.

Cases in the "ing" form phrases can be solved if they are in the present or past continuous forms. The unsolved cases were found in present continuous forms that use apostrophe as in "Nothing's coming" and found in negative forms such as "is not moving"; and also found in present perfect continuous forms such as "has been doing it". These cases can be solved by some modification on the existing phrase-based transfer rules. For example, the

existing transfer rules incorrectly translated “is not moving” into the Indonesian phrase “sedang tidak bergerak” while the correct translation is “tidak sedang bergerak”. Note that there must be a swapping attempt between the word “sedang” and “tidak”. The generated ADJ set for this negative “ing” form phrases is unique e.g. $\{(is, sedang, ((Ss)(Pa, EBm))), (not, tidak, ((EBm)()))\}$ so that the words swapping will be made possible if there is a mechanism to identify all the connectors.

The last categories to be discussed are Category I and IV. The phrases found in Category I were “the next morning” and “next time” and were already solved by the system, showing that the developed phrase-based transfer rules worked well with idiomatic time phrases. Oppositely, all phrases in Category IV or phrases with pronoun “one” were not properly translated. However, it was found that the generated ADJ set for this form of phrases is also unique and thus correct phrase translations can be achieved by using the generated connectors.

Although incorporating phrase-based module in our previous ADJ-based system does not significantly increase the accuracy, but there are other benefits we could get. The total numbers of transfer rules generated in the phrase-based system fewer than that of the previous system. We have developed more than one hundred transfer rules in the previous ADJ-based system and we decided to stop the attempt since the tasks required us to observe the complete disjuncts of all the words in *TSs*. This became difficult in the way that each word can have different disjunct in different sentence or context. We found that generalization of the transfer rules will decrease the total number of transfer rules which in turn will ease the effort of generating the transfer rules. The phrase-based module answered the problems as the generated transfer rules became around 60 only, to be compared with more than one hundred transfer rules of the previous sentence-based module.

However, there is a drawback with the phrase-based module in term of the algorithm complexity, which is of $O(n^5)$, compared with the sentence-based of $O(n^4)$. In this calculation, the link parser algorithm with the complexity of $O(n^3)$ is taken into account since this parser is called by both transfer rules.

V. CONCLUSIONS

In this research, phrase-based transfer rules algorithm is introduced to translate English sentences into Indonesian sentences. Comparison which involves other available systems using the developed BLEU metric tool has shown that the proposed phrase-based MT system using ADJ outperforms other systems. The total number of transfer rules used in the phrase-based MT system has decreased significantly compared with that of the sentence-based MT system. However, this increases the algorithm complexity.

In future works, we need to find the way on how to decompose and compose the phrase-based translation for better accuracy. The use of part of speech as a parameter will also improve the accuracy of the phrase-based transfer rules. Morphological analyzer and phrase-based

statistical module can be incorporated to the developed system to increase the performance.

REFERENCES

- [1] Yusuf, H. “An Analysis of Indonesian Language for Interlingual Machine-Translation System,” In: 15th COLING, Nantes, 23-28 August 1992.
- [2] Och, F. J. and Ney, H. “The alignment template approach to statistical machine translation,” *Computational Linguistics*, Vol. 30, ACL, 2004.
- [3] Riza, H. “Resources Report on Languages of Indonesia,” In: 6th Workshop on Asian Language Resources, Hyderabad, India, 11-12 Jan. 2008.
- [4] Novento, F. “Perangkat Lunak Penerjemah Kalimat Inggris-Indonesia Menggunakan Metode Loading Data Sementara,” Undergraduate final project, Electrical Engineering Department, Gadjah Mada University, 2003.
- [5] Adji, T.B., Baharudin, B., and Zamin, N. “Annotated Disjunct in Link Grammar for Machine Translation,” In: ICIAS (International Conference on Intelligent & Advanced Systems), KL Convention Centre, Kuala Lumpur, 25-28 Nov. 2007.
- [6] Adji, T.B., Baharudin, B., and Zamin, N. “Building Transfer Rules using Annotated Disjunct: An Approach for Machine Translation,” In: 5th SCORED (Student Conference on Research and Development), Malaysia, 11-12 Dec. 2007.
- [7] Yamada, K. and Knight, K. “A syntax-based statistical translation model,” *Proceedings of 39th ACL*, 2001.
- [8] Marcu, D. and Wang, W. “A phrase-based, joint probability model for statistical machine translation,” In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.
- [9] Galley, M., Hopkins, M., Knight, K., and Marcu, D. “What’s in a translation rule?,” *Proceedings of HLT-NAACL*, 2004.
- [10] Bond, F. and Shirai, S. “A Hybrid Rule and Example-based Method for Machine Translation,” Chapter 7 of *Recent Advances in Example-Based Machine Translation*, Springer, 2003.
- [11] Adji, T.B., Baharudin, B., and Zamin, N. “Applying Link Grammar Formalism in the Development of English-Indonesian Machine Translation System”, In: 9th AISC (Artificial Intelligence and Symbolic Computation), Birmingham, UK, 31 July-1 August, 2008.
- [12] Al-Adhaileh, M.H. and Kong, T.E. “Synchronous Structured String-Tree Correspondence (S-SSTC),” In: 20th IASTED02 International Conference, Innsbruck, Austria, February 2002.
- [13] Grinberg, D., Lafferty, J., and Sleator, D. “A Robust Parsing Algorithm for A Link Grammar,” In: 4th International Workshop on Parsing Technologies (IWPT), Prague, 1995.
- [14] Sleator, D.D., Temperley, D. “Parsing English with A Link Grammars,” In: 3rd IWPT, ACL - SIGPARSE conference, University of Tilburg, The Netherlands, 1993.
- [15] Chiang, D. “Hierarchical Phrase-Based Translation,” *Computational Linguistics*, Vol. 33 No. 2, ACL, 2007.
- [16] Papineni, K., Roukos, S., Ward, T., and Zhu, W.J. “BLEU: a method for automatic evaluation of machine translation,” *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, Pennsylvania, 2002, pp. 311–318.