Rule-based Reordering and Post-Processing for Indonesian-Korean Statistical Machine Translation

Candy Olivia Mawalim, Dessi Puji Lestari, Ayu Purwarianti
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
candyoliviamawalim@gmail.com,dessipuji@gmail.com,ayu@informatika.org

Abstract

This paper presents several experiments on constructing Indonesian–Korean Statistical Machine Translation (SMT) system. A parallel corpus containing around 40,000 segments on each side has been developed for training the baseline SMT system that is built based on n-gram language model and the phrase-based translation table model. This system still has several problems, including non-translated phrases, mistranslation, incorrect phrase orders, and remaining Korean particles in the target language. To overcome these problems, some techniques are employed i.e. POS (part-of-speech) tag model, POS-based reordering rules, multiple steps translation, additional post-process, and their combinations. We then test the SMT system by randomly extracting segments from the parallel corpus. In general, the additional techniques lead to better performance in terms of BLEU score compared to the baseline system

1 Introduction

Statistical Machine Translation (SMT) is a corpus-based MT for automatic translation. It has been growing rapidly since this approach gives some advantages, including language-independent and low-cost construction (Koehn, 2010). In the case of Indonesian–Korean translation, there has not been much research done in this field. It is probably because of the difficulty in constructing parallel corpus since both Indonesian and Korean are low-resource languages.

As a valuable resource in developing SMT, we first construct a parallel corpus obtained from Korean learning books, drama and movie subtitles, and Bible text. By using this corpus, we construct the baseline SMT system. Phrase-based translation model is used since the previous studies have shown that phrase-based variant of SMT gives better performance than word-based variant of SMT (Koehn, 2010).

After the baseline system has been built, we analyze the problems found on the translation results. Based on these problems, we investigate several additional techniques which can be used to overcome them. These additional techniques are tested with random segments from the parallel corpus. The quality of each system is determined by using smoothed BLEU metric, known as BLEU+1 (Lin and Och, 2004). BLEU score is calculated by multiplying the geometric mean of the test corpus' modified precision scores with the exponential brevity penalty factor (Papineni, et al. 2002).

2 Related Work

Parallel corpus is a valuable component needed in SMT to train models, optimize the model parameters, and test the translation quality. However, a good parallel corpus of low-resource languages such as Indonesian and Korean is hard to obtain. Therefore, we do not only use books as the source for constructing, but also subtitles and Bible. Automatic parallel corpus extraction from movie subtitles has been introduced in (Caroline et al., 2007). From this study, it was reported that 37,625

aligned pairs with a precision of 92.3% was obtained from 40 movies. Using Bible as the parallel corpus source was also introduced in (Christodouloupoulos and Steedman, 2015). Even though there are missing words and the nature of Bible text problems, Bible corpus can be used as one of parallel corpus source.

The use of pivot language has been a common theme for constructing low resource languages SMT. This approach is also used by well-known available MT, Google Translate. It uses English and Japanese as pivot languages for Indonesian–Korean Translation (Balk et al., 2013). However, it has been reported that direct MT model gives better performance compared to pivot MT model (Costa-jussa et al., 2013). A former study about a speech-to-speech translation for 8 Asian languages in A-STAR project has found that this phenomenon also applies to Indonesian–Korean translation (Sakti et al., 2011).

In (Sakti et al., 2011), the SMT system is designed to translate commonly spoken utterances of travel conversations from a given source language into multiple target languages. Basic travel expression sentences (BTEC) with a comparison of training and testing data of 20:1 is used to construct the system. Each Asian language is treated in a different way. In the case of Korean language, they determine a sequence of morphemes as a word. The quality for this direct Indonesian–Korean SMT system in terms of BLEU score is 30.53 (ID–KR) and 23.62 (KR–ID).

The quality of SMT system for specific languages can be improved by adding models and/or techniques. For Indonesian–Japanese translation, experiments by adding lemma translation, particle elimination, and other processes have been reported to produce a better result (Simbolon and Purwarianti, 2013; Sulaeman and Purwarianti, 2015). Since Japanese and Korean has the most similar characteristics in grammar structures (Kim and Dalrymple, 2013), these additional techniques will also be explored as additional processes.

3 Characteristics of Indonesian and Korean Languages

There are some differences between Indonesian and Korean languages described in Table 1 (Kim et al., 2015).

Characteristics	Indonesian	Korean
Basic pattern	subject-predicate-object-adverb (S-P-O-A)	subject-adverb-object-predicate (S-A-O-P)
Adj. explaining noun	Post-modification	Pre-modification
Preposition	Pre-modification	Post-modification
Aux. verb	Pre-modification	Post-modification
Negation word	Pre-modification	Post-modification
Particle	No	Yes
Time marker	Inflection	Conjugation
Honorific form	No	Yes
Unit	Small to large	Large to small

Table 1: Differences between Indonesia and Korean languages

4 Baseline SMT System

The baseline model was built with the aim to find out the problems that exist in Indonesian–Korean SMT system. The development of this model was carried out using several combinations of the collected corpus. These combinations are conducted to observe which corpus is qualified to be used in constructing a SMT system. There are two main steps that need to be performed in constructing a baseline system.

4.1 Parallel Corpus

The parallel corpus is collected from books, subtitles, and Bible. The segment pairs from each source are taken differently. The book-sourced corpus consists of segments which are already available in two languages and the ones which are available only in one language. The segments which are available only in one language are translated manually.

Unlike (Caroline et al., 2007), corpus from subtitles is built by semi-automatically combining several monolingual drama and movie subtitles. Generally, subtitles for Indonesian are in SRT (Subtitle Resource Tracks) format while for Korean language format are in SAMI (Synchronized Accessible Media Interchange) format. SRT format consists of a number indicating the subtitle's sequence, the start and end time the subtitle is appeared and the caption text. However, SAMI file sets the time to milliseconds and the written style is

similar to HTML and CSS. Due to these differences, the conversion of Korean subtitles from SAMI to SRT is needed. After the subtitles for both languages have the same format, both segments are paired based on the start time and ending time of each subtitle line. In automatic generation of these subtitle pairs, there are some errors that are then fixed manually. The errors are poorly paired subtitles, one subtitle line from one language consists of more/less than one segment from another language, incorrect translation, excessive punctuation, and undefined characters in this study (not alphanumeric or hangul characters).

Using the Bible as a corpus has several advantages. One of them is because it has been translated into numerous languages (Christodouloupoulos and Steedman, 2015). The version used for the Indonesian Bible is the Terjemahan Baru (TB) (published by Indonesian Bible Society) while the Korean Bible is the 현대인의성경 hyeondaein-uiseong-gyeong version (published by International Bible Society). Both of these Bible version are commonly used since they are translated by the official organizations. The unit used for Bible-sourced segments are the Bible verse. Having obtained the verses pairs for both languages, adjustment is needed for the Korean verse translation which has been merged in the previous verse.

After the corpus is collected, corpus cleaning is then performed. Corpus cleaning is employed by removing excessive whitespace characters, converting every word into a lowercase form and separating each punctuation and word with spaces. After that, tokenization is performed in accordance with the language. Tokenization for Indonesian corpus is based on spaces with the addition of tokenization to a word containing prefix ("ku-" and "kau-") and containing suffix ("-ku", "-mu" and "-nya"). This tokenization process is applied because the Korean has different syntax to Indonesian in case of writing proprietary phrases. In Indonesian the writing of proprietary phrases is united like "rumahku" while in Korean the writing is separated into "내 집".

On the other hand, tokenization for Korean corpus is based on Korean morphology by using Mecab class from KoNLPy (Park and Cho, 2014). Table 2 shows the number of paired segments obtained from each source which are used for building baseline system. The comparison between training and testing data follows (Sakti et al. 2011). Besides using only one corpus source, this research also utilizes the combination of the corpus sources, i.e. books and subtitles (bs), books and Bible (bB), Bible and subtitles (Bs) and all.

Source	#paired segments		#vocabulary	
	train	test	ID	KR
books (b)	4,886	243	3,286	3,532
subtitles (s)	5,740	286	3,732	5,600
Bible (B)	28,922	1,446	13,775	13,629

Table 2: Number of paired-segments and vocabulary in corpus

4.2 SMT Model

Phrase-based model is used in constructing baseline system. Generally, it consists of language model, translation model, and decoder. We use the parallel corpus which has been cleaned and tokenized to build the language model and translation model. The n-gram based language model is developed by employing the IRSTLM toolkit (Federico et al. 2008). After that, we create the alignment model of each pair of segments using Giza++ (Och and Ney, 2003). Translation model is built based on the alignment model. We use phrase-based translation table as the translation model. This model was developed from the experiments performed by Dalmia (2014). In the translational model, all punctuation is removed except the hyphen (-) which states the reduplication in Indonesian language. The decoder is built based on stack decoding algorithm (Koehn, 2010).

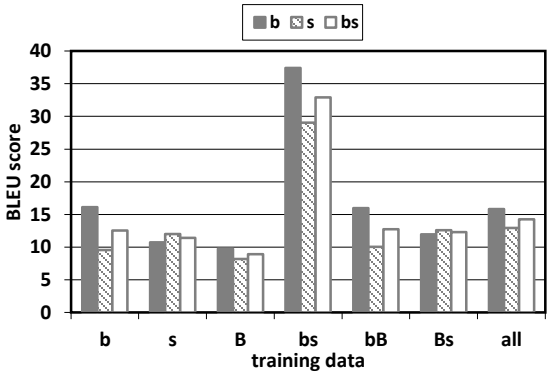

Figure 1: Average BLEU score from baseline system for ID-KR translation

Figure 1 and Figure 2 shows the average BLEU score from the baseline system by using several sources for training and testing data for Indonesian

to Korean (ID–KR) and Korean to Indonesian (KR–ID) respectively. The training data used in this evaluation consists of corpus from each source (shown in Table 2) and their combinations (bs, bB, Bs, and all). The testing data consists of books, subtitles and their combination (bs). Table 3 shows the examples of the translation result from the baseline system.

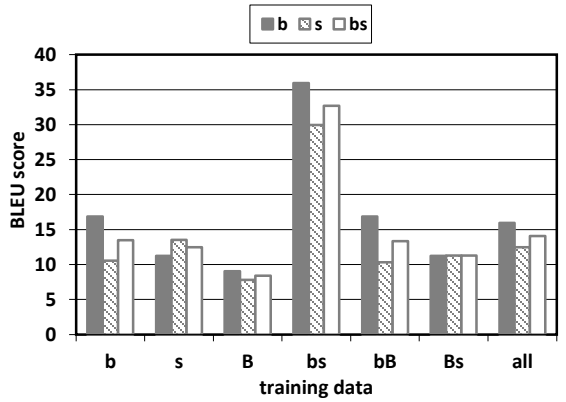

Figure 2: Average BLEU score from baseline system for KR-ID translation

ID–KR	
Source	aku tak sabar bertemu orang-orang
Reference	사람 들 을 정말 만나 보 고 싶 어
Hypothesis	나 tak sabar 만나 서 사람 들
KR–ID	
Source	사람 들 을 정말 만나 보 고 싶 어
Reference	aku tak sabar bertemu orang-orang
Hypothesis	orang-orang di 을 정말 만나 보 kamu ingin 어

Table 3: Example of baseline system translation result

4.3 Analysis

As the baseline, we first observe and determine which parallel corpus to use for training and testing. The quality of translation results are evaluated by using BLEU score. After performing the evaluation by using each source of corpus as testing data, we decide not to use Bible as testing data because the nature of words in Bible is so much different than in books and subtitles. Moreover, machine translation is rarely used for translating Bible because Bible itself has already been translated into numerous languages.

The evaluation of baseline system shows that using Bible corpus as training data obtains worse results than using books or subtitles. However, when we combine the Bible corpus with one of the other corpus, we can obtain slightly better performance for both ID-KR and KR-ID translation. Using books and subtitles as training data increases the BLEU score significantly. It even gives better results than combining all the corpus. Although the nature of Bible words is different than the other corpus, this corpus may increase the BLEU score slightly because adding this corpus reduces out-of-vocabulary (OOV) problem, from 11.8% to 1.08 %. However, because of the number of paired-segments in Bible corpus is approximately 5 times than the other corpus, it contributes much more than the other corpus. Therefore, when translating common phrases, it produces uncommon translation which will make the translation difficult to understand. Table 4 shows the example of this case.

There are several problems which can be found in the baseline system, including non-translated phrases, mistranslation, incorrect phrase orders, and remaining Korean particle(s) in the target language (shown in Table 5). Non-translated phrases can be caused by the phrases are not registered as n-gram model even though the phrase is in the parallel corpus (Sulaeman and Purwarianti, 2015). In addition, the absence of phrases in the parallel corpus (OOV problem) may also lead to the existence of untranslated phrases. Mistranslation problem can be a partial or an entire incorrect phrase translation. This problem can be occurred because there are several possible phrase translation pairs in the translation model.

ID–KR	
Source	kau begitu ingin melawan penjahat
Reference	범죄자 와 싸우 고 싶 어 안달 이 났 나
Hypothesis	그리고 그렇 군요 고 penjahat 싶 은데 하나님 을 대적
KR–ID	
Source	약국 에서 약 을 샀 어요
Reference	saya membeli obat di apotek
Hypothesis	apotek dari hadapan orang israel obat tadi nya kamu membeli apakah kamu

Table 4: Example of SMT result which use bB as training data

The following problem is incorrect phrase orders. The structure of Indonesian and Korean languages which are very different as we explained in section 3 can lead to this problem. Unlike Korean language, Indonesian does not have particle which cause the remaining Korean particle(s) in the KR-ID translation result. In this paper, we conduct some

experiments to overcome these issues. These experiments will be explained in the next section.

Non-Translated phrase	
Source	aku tak sabar bertemu orang-orang
Reference	사람 들 을 정말 만나 보 고 싶 어
Hypothesis	나 tak sabar 만나 서 사람 들
Mistranslation	
Source	saya makan mi instan setiap hari dalam seminggu
Reference	1 주일 동안 매일 라면 을 먹 었 어요
Hypothesis	밥 먹 매일 instan 일주일
Incorrect phrase orders	
Reference	1 주일 동안 매일 라면 을 먹 었 어요 1 2 3
Hypothesis	밥 먹 매일 instan 일 주일 3 2 3 1
Remaining Korean particle(s)	
Source	내일 은 목요일 입니다
Reference	besok hari kamis
Hypothesis	besok adalah 은 목요일

Table 5: Example of translation result with baseline system problems

5 Experiments

There are 5 main techniques that are conducted in this study, i.e. adding POS tag information, POS-based reordering rules, multiple steps translation, additional post-process, and their combinations. The additional POS tag information technique, some additional post-process (lemma translation and particle elimination) are adapted from (Simbolon and Purwarianti, 2013; Sulaeman and Purwarianti, 2015).

5.1 POS Tag Information Addition

Adding POS tag information technique is employed to make the translation phrase more accurate and the POS tag arrangement in the translations more natural. The POS tagger used for Indonesian corpus is the modified Pebahasa library (Wicaksono and Purwarianti, 2010), while for Korean corpus is the Mecab class in KoNLPy (Park and Cho, 2014).

Figure 3 and Figure 4 shows the comparison between baseline system and system with POS tag information addition performance in terms of average BLEU score. From the figure, it can be seen that there is a decrease in BLEU score for both ID-KR and KR-ID translation. This decreasing in the BLEU score indicates that the model with POS tag information does not successfully minimize the phrase translation error. On the other hand, it added the number of non-translated phrases in the translation results (Table 6).

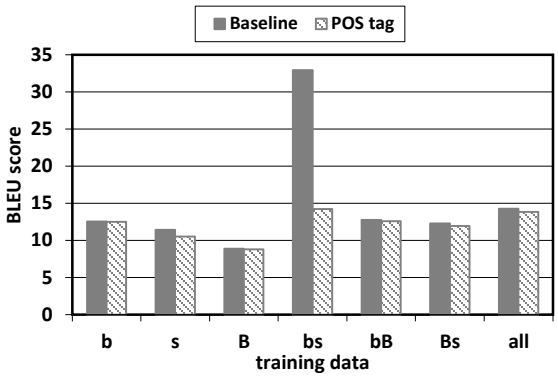

Figure 3: Comparison of baseline system and POS tag information addition system for ID–KR translation

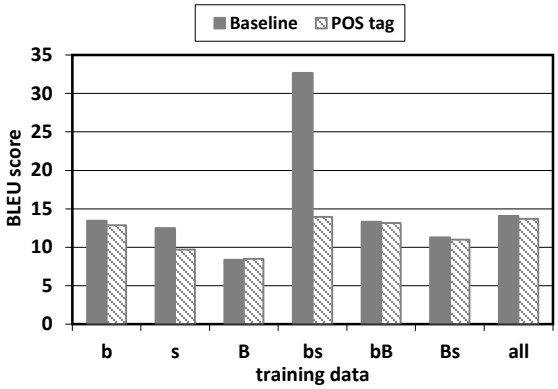

Figure 4: Comparison of baseline system and POS tag information addition system for KR–ID translation

Sumber	할아버지 생신 이 언제 예요
Referensi	kapan ulang tahun kakek
Baseline	kapan ulang tahun kakek
Hipotesis	kakek 생신 yang ini kapan kamu 예 요

Table 6: Example of translation result with POS tag information addition system

5.2 POS-Based Reordering Rules

In this study we do not use the common reordering model, such as syntax-based models (Chiang, 2005) and lexicalized models (Och et al., 2004) because those methods try to solve the common problem which only perform well when the ordering of words does not vary too much (Genzel, 2010). The reordering rule is performed before the source language is translated into the target language. This rule is generated manually based on the POS tag information and the alignment of the segments of

source language and target language. This POS tag information is used to define the part that becomes a unity of subject, predicate, object, and adverb. Table 7 shows the example of the reordering rule.

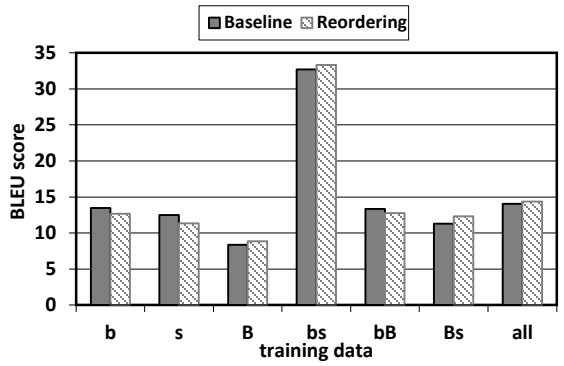

Figure 5: Comparison of baseline system and system with POS-based reordering rule addition for ID–KR translation

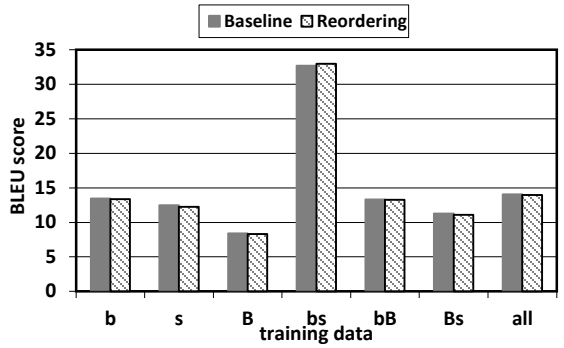

Figure 6: Comparison of baseline system and system with POS-based reordering rule addition for KR–ID translation

lang	segment	POS tag	Rule
ID	bos kami tidak punya banyak uang	NN PRP NEG VBT JJ NN	(1 2) (5 6) (3 4)
KR	약국 에서 약 을 샀 어요	NNG JKB NNG JKO VV+EP EF	(3 4 (5 6)) (1 2)

Table 7: Example of POS-based reordering rule

ID–KR	
Source	jangan berenang
Reference	수영 하 지 마세요
Baseline	지 마세요 수영 을 해요
Hypothesis	수영 을 해요 지 마세요
KR–ID	
Source	약국 에서 약 을 샀 어요
Reference	(saya) membeli obat di apotek
Baseline	obat di apotek saya membeli 어요 obat
Hypothesis	beli 을 obat ini di apotek

Table 8: Example of the translation result by using POS-based reordering rule

The number of rules which are used in this study is 130 for Indonesian language and 50 for Korean language. Figure 5 and Figure 6 shows the comparison of baseline system and system with POS-based reordering rule addition for ID-KR and KR-ID translation respectively. We can see that even though there is only a small number of rules, this technique can improve the quality of ID-KR translation. Table 8 shows the example of the translation result by using this reordering rules.

5.3 Multiple Steps Translation

ID–KR	
Source	saya ingin memberikan sepatu sebagai hadiah kepada pacar saya tapi saya tidak yakin
Reference	남자 친구 에게 신발 을 선물 하 고 싶 은데 고민 이 에요
bs	saya ingin memberikan 신발 을 선물 kepada pacar saya tapi saya tidak yakin
bs-bB	고 싶 은데 여자 친구 memberikan 신발 을 선물 kepada 지만 지 않 아요 yakin
bs-bB-Bs	고 싶 은데 여자 친구 주 었 으며 신발 을 선물 에게 말씀 해 주 셨 지만 지 않 아요 yakin
bs-bB-Bs-all	고 싶 은데 여자 친구 주 었 으며 신발 을 선물 에게 말씀 해 주 셨 지만 지 않 아요 yakin
KR–ID	
Source	오늘 은 저희 학교 졸업식 이 에요
Reference	hari ini adalah hari wisuda sekolah
bs	hari ini 저희 학교 졸업식 이 에요 adalah
bs-bB	hari ini 저희 sekolah wisuda anak manusia juga akan 에요 adalah
bs-bB-Bs	hari ini 저희 sekolah wisuda anak manusia juga akan rupa nya adalah
bs-bB-Bs-all	hari ini 저희 sekolah wisuda anak manusia juga akan rupa nya adalah

Table 9: Example of the translation result by using bs-bB-Bs-all multiple steps translation

Adding corpus does not necessarily improve the quality of the translation but it is able to reduce the OOV problem. This underlies the multiple steps translations both to improve translation quality as well as to reduce OOV. The multiple steps

translation experiments are performed in two ways, i.e. translation with adding b-s-B-all corpus step-by-step and translation with adding bs-bB-Bs-all corpus step by step. Figure 7 shows that multiple steps translation can give a better translation quality, except for bs-bB-Bs-all steps for KR-ID translation. This is caused by Korean morphemes which has no particular meaning, e.g. particles are translated. Table 9 shows the example of multiple steps translation result.

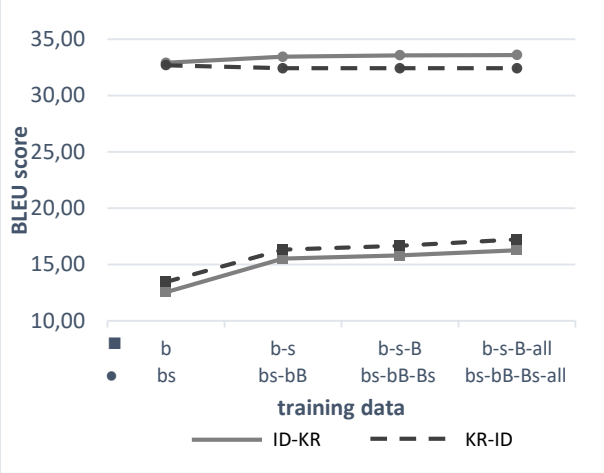

Figure 7: Comparison of baseline system and system with multiple steps translation

5.4 Additional Post-Process

Additional post-processes performed in this study consist of name entity (NE) translation, particle elimination, dictionary translation, lemma translation and basic verb conversion. NE translation process directly translates the word considered as NE from the Indonesian word to the writing of the Korean language and vice versa. The NE is determined by the rules based on its POS tag and lemma. If the NE has high similarity value with vocabulary from training data which is not listed in Kamus Besar Bahasa Indonesia (KBBI) and Son Myun Kwan ID-KR dictionary, the translation of NE is interpreted as that vocabulary.

The following additional process is translating the non-translated phrases by using the ID-KR dictionary help. The contents of this dictionary is not similar to the standard dictionary because it contains examples of sentences and other explanations. Therefore, the translation process is employed by using n-gram matching (from 3-gram to 1-gram). Translation by using dictionary is able to minimize non-translated phrases. However, since there are many possible translations for a single phrase, the translation obtained from the dictionary is only taken from the first phrase found during the search process. This results in the possibility of generated translations is not commonly used in the target language.

Lemma translation is the development of the dictionary translation. For phrases that still can not be translated in dictionary translation, specifically for ID-KR translation which is conducted by using Indonesian lemma. The following additional process is converting Korean verb to its basic form before dictionary translation. This process is conducted because there are many verbs which can not be translated due to the different form. Figure 8 shows the comparison of baseline system and additional post-process system by using bs corpus as training data.

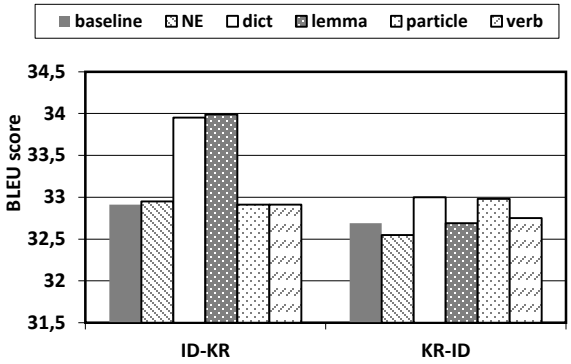

Figure 8: Comparison of baseline system and additional post-process system

5.5 Combination Techniques

Based on the experiment results, we try to combine all the techniques to improve the quality of translations. The combination techniques are performed by not adding the POS tag information since it causes worse result. The experimental combination is divided into two, as follows.

- 1st Combination: Reordering–additional post-processes (particle elimination, dictionary translation, lemma translation, verb conversion, NE translation)–multiple steps translation
- 2nd Combination: Reordering–particle elimination-multiple steps translation-additional post-processes (dictionary translation, lemma translation, verb conversion, NE translation)

Particle elimination is performed first in order to decrease the probability of the Korean particles which usually do not have particular meaning is

being translated as some phrases in Indonesian language. 1st combination and 2nd combination is used to determine whether multiple steps translation or additional post-processes is needed to be performed first. Figure 9 shows the comparison of baseline system and these combination system. 1st combination gives better results in both ID-KR and KR-ID translation.

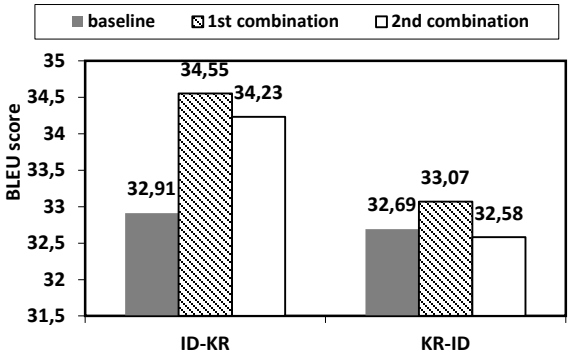

Figure 9: Comparison of baseline system and additional post-process system

ID–KR	
Source	saya ingin memberikan sepatu sebagai hadiah kepada pacar saya tapi saya tidak yakin
Reference	남자 친구 에게 신발 을 선물 하 고 싶 은데 고민 이 에요
Baseline	saya ingin memberikan 신발 을 선물 kepada pacar saya tapi saya tidak yakin
1st Combination	나 바라다 주다 신발 을 선물 에게 애인 나 그러나 나 이 아니다 확신하는
2nd Combination	고 싶 은데 여자 친구 주 었 으며 신발 을 선물 에게 말씀 해 주 셨 지만 지 않 아요 확신하는
KR–ID	
Source	텔레비전 보 기 전 에 숙제 해
Reference	kerjakan prmu sebelum nonton tv
Baseline	dulu sebelum nonton tv 에 숙제 해
1st Combination	dulu sebelum nonton tv pekerjaan rumah Syaka
2nd Combination	dulu sebelum nonton tv pr nya untuk

Table 10: Example of translation by using the combination system

As we can see the result of the 1st combination in Table 10, the untranslated phrases is no longer present in the translation. However, there are more mistranslation phrase, such as "해" which is translated as "Syaka". This word is obtained from dictionary translation and is not related with the reference at all. For ID-KR translation, the dictionary translation help to translate the untranslated verb, such as "tidak yakin" as "아니다 확신하는", this translation is incorrect as a phrase. There are rules to form the Korean verb as explained in section 3. Reordering rules which are provided in this system do not affect this sample translation because of the limitation of the number of the rules. In conclusion, although the problems described in section 4.3 are still found in the translation result, these problems have already been reduced.

On the other hand, the result obtained from the 2nd combination is worse than the 1st combination. The multiple steps translation which performed first causes the unrelated phrase, such as "말씀 해 주 셨" found in the translation result. As shown in Table 9, even though this technique gives the better performance than the baseline system, it causes the appearance of the common Bible phrases, such as "anak manusia".

6 Conclusion and Future Work

In this paper we have presented several experiments on constructing Indonesian–Korean SMT. The combination of books and subtitles corpus is the best corpus which can be used both as training and testing data in this study. Our experiments also show that the corpus collected from Bible is better used as training data after using books and subtitles corpus. Most of the additional techniques can increase the quality of translation in terms of BLEU score, except the adding POS tag information technique. The best technique (1st combination) are able to increase the BLEU score up to 4,97% for ID-KR translation and 1,15% for KR-ID translation.

There are still many things to explore in constructing Indonesian–Korean SMT. Automatic approaches of constructing parallel corpus (Caroline et al., 2007) from subtitles can become alternative in the next study. A source-side reordering model which is introduced in (Genzel, 2010) can also be used to develop the reordering method. Another possibility of improvement is using rules to form the Korean verbs for ID-KR translation. In the future we would like to use these proposed methods in order to improve the Indonesian-Korean SMT.

References

Alfan Farizki Wicaksono and Ayu Purwarianti. 2010. HMM Based POS Tagger for Bahasa Indonesia. Proceedings of 4th International MALINDO (Malay and Indonesian Language) Workshop.

Ayushi Dalmia. 2014. Phrase Based Translation Model. India: International Institute of Information Technology.

Bertoldi, N., Cettolo, M., & Federico, M. (2008). IRSTLM: an open source toolkit for handling large scale language models. INTERSPEECH.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality using Longest Common Subsequence and Skip-Bigram Statistics. 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume. Barcelona, Spain. 605-612.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. Language Resources and Evaluation 375–395.

David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. Proceedings of the ACL'05. Ann Arbor, Michigan. 263-270.

Dmitriy Genzel. 2010. Automatically Learning Source-side Reordering Rules for Large Scale Machine Translation. Proceeding COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China. 376-384.

Ethan M. Balk, Mei Chung, M.L. Chen, T.A. Trikalinos, Kong, Win Chang L. 2013. Assessing the Accuracy of Google Translate To Allow Data Extraction From Trials Published in Non-English Language. Agency for Heatlhcare Research and Quality.

Eunjeong L. Park and Sungzoon Cho. 2014. KoNLPy: Korean natural language processing in Python. 26th Annual Conference on Human & Cognitive Language Technology. Chuncheon, South Korea.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. Association of Computational Linguistics 29: 19-51.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, et al. 2004. A Smorgasbord of Features for Statistical Machine Translation. HLT-NAACL 2004: Main Proceedings. Boston, Massachusetts, USA. 161–168.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia. 311-318.

Lavecchia Caroline, Sma¨ıli Kamel, and Langlois David. 2007. Building Parallel Corpora from Movies. The 4th International Workshop on Natural Language Processing and Cognitive Science - NLPCS. Funchal, Madeira, Portugal.

Marta R. Costa-jussa, Carlos A. Henriquez, and Rafael E. Banchs. 2013. Evaluating Indirect Strategies fo Chinese-Spanish Statistical Machine Translation: Extended Abstract. Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. Beijing, China. 3142-3145.

M.A Sulaeman and Ayu Purwarianti. 2015. Development of Indonesian-Japanese Statistical Machine Translation Using Lemma Translation and Additional Post-Process. The 5th International Conference on Electrical Engineering and Informatics. Bali, Indonesia: IEEE. 54-58.

Philipp Koehn. 2004. Europarl: A Parallel Corpus for Statistical Machine Translation.

Philipp Koehn. 2010. Statistical Machine Translation. New York: Cambridge University Press.

Roger Kim and Mary Dalrymple. 2013. Porting Grammar between Typologically Similar Languages: Japanese to Korean. Pacific Asia Conference on Language, Information and Computation 2013. Taipei. 98-105.

Sakriani Sakti, Michael Paul, Andrew Finch, Shinsuke Sakai, Thang Tat Vu, Noriyuki Kimura, Chiori Hori, Eiichiro Sumita, Satoshi Nakamura, and Jun Park. 2011. A-STAR: Toward Translating Asian Spoken Languages. Computer Speech & Language Vol. 27, Issue 2, Feb 2013 509-527.

Seon Jung Kim, Kyung Mo Min, Sung Tae Park, and Yong Heo. 2015. EPS-TOPIK untuk Orang Indonesia Panduan Belajar Mandiri Bahasa Korea. Ulsan, South Korea: HRD Korea.

Simon Simbolon and Ayu Purwarianti. 2013. Experiment on Indonesian-Japanese Statistical Machine Translation. IEEE Cyberneticscom 2013 80-84.