An Approach to Improving Corpus Quality for Indonesian-English Statistical Machine Translation

Herry Sujaini
Department of Electrical Engineering
University of Tanjungpura
Pontianak, Indonesia

***Abstract*—This paper discusses one of the fundamental issues in statistical machine translation (SMT) : the corpus quality. Corpus is a reference system based on an electronic collection of texts composed in a specific language used for building language model, translation model, or factor model.**

The quality of a SMT system depends heavily on the quality and quantity of the bilingual language resource. However, previous related work mainly focuses on the quantity and tries to collect more bilingual data. Checking the source sentences and their translations manually in a parallel corpus is a very difficult task and requires large resources. In this paper, to optimize the bilingual corpus to improve the quality of the translation system, we propose some approaches to processing the parallel corpus by filtering bad translation sentences from the parallel corpus.

Filter we use is the minimum value of each sentence that is tested by the Bilingual Evaluation Understudy (BLEU) method. The experimental results show that at an optimal point of minimum value, MPS quality can be improved by increasing quality at the expense of quantity of a corpus.

Keywords — Corpus quality; translation quality; statistical machine translation

I. Introduction

Machine translation (MT) is the automatic translation from one source language into target one using computers. In 1949, popular accounts trace its modern origins to a letter written by Warren Weaver, only a few years after ENIAC came online. It has since remained a key application in the field of natural language processing (NLP). A historical overview about MT is given by Hutchins [1], and a comprehensive general survey about MT is given by Dorr, Jordan, and Benoit [2].

Data-driven machine translators or statistical-based machine translators, work by aggregating massive amounts of previously translated bits of information, and uses statistical analysis to determine matches between the source and target language with the previously aggregated corpus. This method is less expensive and requires less development time than transfer-based machine translation method, but the generated translation is often not to the same quality as transfer-based translation [3].

MPS translation quality can be improved by increasing the quantity and quality of corpus. In addition, the translation quality can also be improved by adding linguistic information at the level of words in the corpus. Improvement or development of algorithms that are used both in the pre-process or algorithm in the translation process can also improve the quality of translations. In this paper, we focus on the influence of the corpus quality by means of filtering quality sentences from a parallel corpus. The Task of choosing qualified sentences in parallel corpus should be done manually, but it requires a lot of time and high accuracy, besides it requires large human resources. This paper proposes an alternative strategy, namely by filtering the parallel corpus automatically to improve the quality by using the BLEU method [4] on all existing sentences in the parallel corpus.

II. Corpus Use in Statistical Machine Translation

Machine translation is the translation of text by a computer, with no human involvement. Obviously, computer work faster and cheaper than human. Machine translation can also be referred to as automated translation, automatic or instant translation. Over the last two decades, SMT methods led to considerable improvements. SMT treats translation as a machine learning problem. This means that a learning algorithm is applied to a large body of manually translated text, known as a parallel corpus.

Statistical machine translation is related to alternative data-driven strategies in MT, such as the previous work on example-based machine translation [5]. Statistical methods may be integrated in rule-based systems [6,7]. Parsing and translation decisions may be learned from text data [8]. Multiple scoring procedures may decide between the alternatives generated by the transfer-based system [9]. Conversely, translations from rule-based systems may be used as additional phrase translations in statistical systems [10]. Rule-based systems may be used to generate training data for statistical methods [11], primarily having the statistical method relearn the rule-based system [12]. Often statistical machine translation is used as a fall-back for methods that frequently fail to produce output, but are more accurate when they do [13]. Statistical machine translation models may also be used to automatically post-edit the output of interlingua [14] or rule-based systems [15]. Additional markup from the rule-based system may be exploited for tighter integration [16]. Such post-editing may alleviate the need to customize rule-based systems to a specific domain [17].

In general, the MPS architecture as shown in Fig. 1. The main data source used by the SMT is a parallel corpus and monolingual corpus. The process of training on a parallel corpus produce translation model (TM). The process of training for the target language in the parallel corpus coupled with the target language monolingual corpus generate language model (LM), while the feature model (FM) is generated from the target language in the parallel corpus that every word has been characterized by linguistic features such as the Part of Speech (PoS), lemma, gender, and others. TM, LM and FM results of the above process is used to produce the decoder. Furthermore, the decoder is used as a machine translator to produce the target language of the input sentence in the source language.

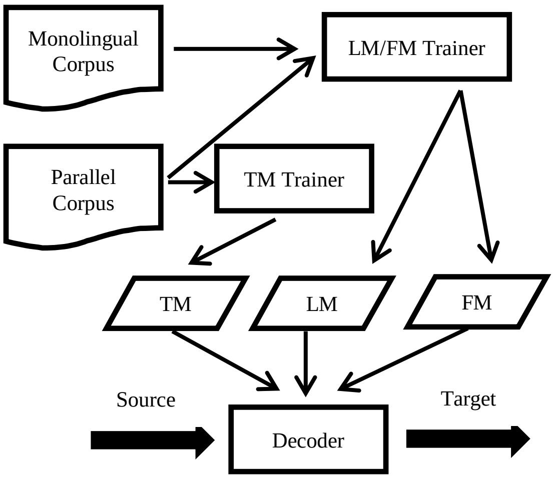

Fig. 1. Statistical Machine Translation Architecture

In Fig. 1, the main data used to generate models of the MPS is parallel corpus. Monolingual corpus can be obtained from the parallel corpus in the target language although usually propagated back from other sources. A study of the influence of the quantity and quality of the corpus has been done on several language pairs, including for the English-Turkish [18], English-Estonian [19] and English-Hindi [20]. In this paper, we use Indonesian as the source language and English as the target language.

III. Experiments

In this paper, we use the BLEU method to conduct elections to all qualified sentences in a parallel corpus. BLEU measure modified n-gram precision scores between the results of automated translation with translation and use a constant reference called brevity penalty.

$$\mathrm{BP_{BLEU}} = \begin{cases} 1 \text{ if } c > r \\ e\left(1-\frac{r}{c}\right)) & \text{if } c \leq r \end{cases} \qquad (1)$$

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} Count(n-gram')} \qquad (2)$$

$$\mathrm{BLEU} = \mathrm{BP}.\exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad (3)$$

wherew_n = 1/N.

BP is a symbol of brevity penalty, c is the number of words from the results of automatic translation, r is the number of words of the referral, and p_n is modified precision score, while the value of w_n is 1 / N. The standard value of N for BLEU is 4, because the value of precision BLEU generally calculated to 4-grams. p_n symbols derived from the number of n-grams in the translation that matches the reference divided by the number of n-grams in the translation.

In this study, we used Moses as a decoder, SRILM to build language models, Giza++ for word alignment process, and BLEU for scoring the translation. Parallel corpus used in this research is a corpus "Identics" [21] at 27 K pairs of sentences in Indonesian-English.

In this study, Indonesian is used as the source language to be translated into English. After decoder MOSES (MPS system) Indonesian-English built, the entire sentence (27 K) on the parallel corpus assessed by the method of BLEU. Sentences that have value BLEU under n (for n = 5%, 10%, 15%, 20%, 30%, 40%, 60% and 80%) are eliminated, while the rest of the sentence is taken to be a new and back parallel corpus conducted the training process to build a new MPS engine.

IV. Results and Discussion

Results of tests performed by the BLEU limit value greater than or equal to n shown in Table I. For example, a sentence which was eliminated from the corpus to the boundary value of n = 10% are shown in Table II. From the examples in the table, the translation sentence pair is not appropriate and feasible eliminated. Sentences that are not qualified will degrade the quality of translation models produced by the corpus and quality of machine translation.

TABLE I. Number of Sentences for BLEU Threshold = n

n (%)	Number of Sentences	Percentage of Sentences (%)
0	27,326	100.00
5	23,335	85.39
10	23,215	84.96
15	22,671	82.96
20	21,740	79.56
30	19,261	70.49
40	16,428	60.12
60	10,753	39.35
80	6,190	22.65

TABLE II. Eliminated Sentences

Indonesian	English	BLEU(%)
bank indonesia akan tetap menjaga rupiah agar tidak terlalu bergejolak seiring dengan terus menguat nya rupiah dan telah mencapai rp 8.750 per dolar as	bank indonesia bi will keep safeguarding the volatility of the country 's currency , the rupiah , which continues to strengthen and now has reached rp 8,750 against the us dollar	0
seperti diketahui partai demokrat menuntut penarikan bertahap atas tentara as di irak	as democrats demand a phased withdrawal of us troops , mccain has argued that more men must be poured in to flush out insurgent strongholds , crush militias and sectarian violence and to train iraqi forces -- a position since taken up by bushi	2.27
sebelum nya , fernandez mengatakan mengerti keputusan klub yang memecat nya	earlier , his predecessor fernandez said he understood the club 's decision to dismiss him given the team 's failure to live up to expectations in the first half of the season	4.84

For every BLEU threshold score n, we built the new MPS system and given 300 test sentences in Indonesian. Thetesting results for each MPS system is shown in Table III and Fig. 2.

TABLE III. SMT BLEU SCORES

n (%)	BLEU Score (%)
0	38.85
5	40.44
10	40.36
15	40.05
20	39.69
30	36.53
40	34,87
60	26,90
80	19,02

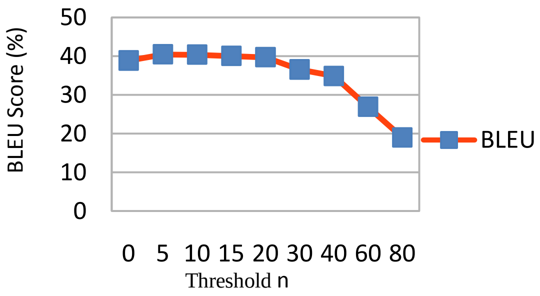

Fig. 2. Graph of SMT BLEU Scores

From the experimental results, the BLEU score of the systems with n = 5% increase from the baseline score, but the BLEU score decrease in the system with a corpus n = 10% and so on. The strategies used to improve corpus quality in this paper have an impact reducing corpus quantity, which also affects the MPS quality.

TABLE IV. TRANSLATION RESULTS

No		Sentences	BLEU (%)
1	Input	hal tersebut lantaran tarif yang dipatok untuk layanan internet kecepatan tinggi di bangladesh terlampau mahal	
	Ref	it is taken because the set tariff for high speed internet service in bangladesh is too expensive	
	K0	it because tariff for high speed internet service in bangladesh is too expensive	56,64
	K10	it is taken because the set tariff for high speed internet service in bangladesh is too expensive	100,00
2	Input	yang perlu diciptakan sekarang ini adalah bagaimana iklim investasi kondusif dan menghilangkan hambatan-hambatan bagi kemajuan sektor riil , bukan pada masalah pembatasan sbi , tambah dia	
	Ref	what is important now is creating a conducive investment atmosphere and eradicate obstacles that hinder real sector development not limiting funds in sbi , he said	
	K0	the need is now is how investment climate conducive and remove constraints for the development of the real sector , not on the issue on sbi , he said	17.53
	K10	the need is now is how the investment climate conducive and eradicate obstacles that hinder real sector development , not on the issue on sbi , he said	44.83
3	Input	dolar as terhadap euro melemah , setelah bank sentral as memutuskan tetapmempertahankan suku bunga overnight 5,25 persen menjadi 1,3412 dari sebelum nya 1,3395 atau melemah 0,15 persen , sedangkan dolar as terhadap yen naik menjadi 117,40 dari 117,15	
	Ref	the us currency weakened against the euro after the fed decided to maintain its overnight interest rate at 5.25 percent to 1.3412 compared to 1.3395 earlier or weakening by 0.15 percent while the us unit went up to 117.40 from 117.15	
	K0	the us dollar against the euro weakening , after the us central bank decided to maintain its overnight interest rate at 5.25 percent to 1.3412 compared from the previous 1,3395 or weakened by 0.15 percent , while the us dollar against unit went up to 117.40 from 117.15	45.76
	K10	the us dollar against the euro weakening , after the us central bank decided to maintain its overnight interest rate at 5.25 percent to 1.3412 compared to 1.3395 earlier or weakened by 0.15 percent , while the us dollar against unit went up to 117.40 from 117.15	55.70

K0 = SMT Output for base corpus
K10 = SMT Output for corpus n = 5%

V. CONCLUSION

MPS translation quality can be improved by finding the balance point between the quality and quantity of a corpus, corpus quantity can be sacrificed at specific points to improve the quality of the corpus by eliminating sentences that have a poor translation in the parallel corpus.

With the strategy used in this study, the quality of the translation system MPS increase by 4.09% after eliminating approximately 14.6% sentence in the corpus. Sentences are eliminated automatically by considering the BLEU score each sentence. Sentences are worth less than 10% is eliminated, while the rest is used to create a new system.

REFERENCES

[1] J. Hutchins, Machine translation: a concise history. In Computer Aided Translation: Theory and Practice, C. S. Wai, Ed. Chinese University of Hong Kong, 2007

[2] B.J. Dorr, P.W. Jordan, and J.W. Benoit, A survey of current paradigms in machine translation, In Advances in Computers, M. Zelkowitz, Ed. Vol. 49. Academic Press, 1999, pp. 1–68.

[3] H. Zhang, "Babel not: machine translation for the technical communicator", ProZ The Translation Workplace, 2008.

[4] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, “BLEU: A method for automatic evaluation of machine translation”, In Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL), Pennsylvania, pp. 311-318, 2002.

[5] H.L. Somers, “Review article: example-based machine translation”, Machine Translation, Vol 14, pp. 113–157, 1999.

[6] K. Knight, I. Chander, M. Haines, V. Hatzivassiloglou, E. Hovy, M. lida, S.K. Luk, A. Okumura, R. Whitney, and K. Yamada, “Integrating knowledge bases and statistics in MT”, In Proceedings of the Conference of the Association for Machine Translation in the Americas, 1994.

[7] N. Habash and B.J. Dorr, “Handling translation divergences: Combining statistical and symbolic techniques in generation-heavy machine translation”, In Richardson, S. D., editor, Machine Translation: From Research to Real Users, 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002 Tiburon, CA, USA, October 6-12, 2002, Proceedings, volume 2499 of Lecture Notes in Computer Science. Springer, 2002

[8] U. Hermjakob and R.J. Mooney, “Learning parse and translation decisions from examples with rich context”, In Proceedings of the 35th

Annual Meeting of the Association for Computational Linguistics (ACL)., 1997.

[9] M. Carl, “METIS-II: The German to English MT system”, In Proceedings of the MT Summit XI. 2007.

[10] Y. Chen, A. Eisele, C. Federmann, E.Hasler, M. Jellinghaus, and S. Theison, “Multi-engine machine translation with an open-source SMT decoder”, In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic. Association for Computational Linguistics, pp. 193–196, 2007.

[11] X. Hu, H. Wang, and H. Wu, “ Using RBMT systems to produce bilingual corpus for SMT”, In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 287–295., 2007.

[12] L. Dugast, J. Senellart, and P. Koehn, “Can we relearn an RBMT system?”, In Proceedings of the Third Workshop on Statistical Machine Translation, pp. 175–178, Columbus, Ohio. Association for Computational Linguistics., 2008.

[13] C. Chai, J. Du, and W. Wei, “NLPR translation system for IWSLT 2006 evaluation campaign”, In Proceedings of the International Workshop on Spoken Language Translation, Kyoto, Japan., 2006.

[14] S. Seneff, C. Wang, and J. Lee, “Combining linguistic and statistical methods for bi-directional English Chinese translation in the flight domain”, In 5th Conference of the Association for Machine Translation in the Americas (AMTA), Boston, Massachusetts., 2006.

[15] M. Simard, C. Goutte, and P. Isabelle, “Statistical phrase-based post-editing”, In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pp. 508–515, Rochester, New York. Association for Computational Linguistics.. 2007

[16] N. Ueffing, J. Stephan, E. Matusov, et al. “Tighter integration of rule-based and statistical MT in serial system combination”, In Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), pp. 913–920, Manchester, UK. COLING 2008 Organizing Committee., 2008.

[17] P. Isabelle, C. Goutte, and M. Simard, “Domain adaptation of MT systems through automatic post-editing”, In Proceedings of the MT Summit XI, 2007.

[18] E. Yıldız, A.C. Tantuğ, and B. Diri, “The effect of parallel corpus quality vs size in English-to-Turkish SMT”, Sixth International Conference on Web services & Semantic Technology (WeST 2014), Chennai, 2014.

[19] H.J. Kaalep and K. Veskis, “Comparing Parallel Corpora and Evaluating their Quality”, Proceedings of MT Summit XI, Copenhagen, pp. 275-279, 1997.

[20] S. Maheshwar and H. Sharma, “improvements in corpus quality for statistical machine translation”, IJSRD - International Journal for Scientific Research & Development, 2(5), pp. 2321-0613, 2014

[21] S.D. Larasati, V. Kuboň, and D. Zeman, “Indonesian morphology tool (morphind): towards an Indonesian corpus”, SFCM 2011. Springer CCIS proceedings of the Workshop on Systems and Frameworks for Computational Morphology, Zurich. pp. 119-129, 2011.