TELKOMNIKA, Vol.16, No.2, April 2018, pp. 718~727
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v16i2.8687

Quality Translation Enhancement Using Sequence Knowledge and Pruning in Statistical Machine Translation

Media A. Ayu*[1], Teddy Mantoro[2], Jelita Asian[3]
[1,2] Faculty Science and Technology, Sampoerna University, Jakarta, Indonesia
[3]UBITEC Lab, Surya University, Tangerang, Indonesia
*Corresponding Author, e-mail: media.ayu@sampoernauniversity.ac.id

Abstract

Machine translation has two important parts, a learning process which followed by a translation process. Unfortunately, most of the translation process requires complex operations and in-depth knowledge of the languages in order to give a good quality translation. This study proposes a better approach, which does not require in-depth knowledge of the linguistic properties of the languages, but it produces a good quality translation. This study evaluated 28 different parameters in IRSTLM language modeling, which resulting 270 millions experiments, and proposes a sequence evaluation mechanism based on a maximum evaluation of each parameter in producing a good quality translation based on NIST and BLEU. The parallel corpus and statistical machine learning for English and Bahasa Indonesia were used in this study. The pruning process, user interface, and the personalization of translation have a very important role in implementing of this machine translation. The result is quite promising. It shows that pruning process increases of the translation process time. The particular sequence knowledge/value parameter in translation process has a better performance than the other method using in-depth linguistic knowledge approaches. All these processes, including the process of parsing from a stand-alone mode to an online mode, are also discussed in detail.

Keywords*: statistical machine translation, parallel corpus, pruning, personalization translation, hierarchical ontology, ontology matching, IRSTLM*

1. Introduction

A translation process, as a critical part of a machine translation, can be put simply as a process of decoding the meaning of the source text and re-encoding the meaning into the target language. Unfortunately, this translation process requires complex operations in order to give a good quality translation. To decode the meaning of a sample piece of text completely means that the translation process must be able to analyze all the features of the text. This approach requires in-depth knowledge of the grammatical structure, semantics, idioms, syntax, etc. including the culture in which the source language is used. The same in-depth knowledge is also required for the machine translation to re-encode the meaning into the target language to minimize the mis-interpretation of the source text. This complex process is used in linguistic approaches to machine translation.

This study proposes a better approach in producing a good quality translation, which does not require in-depth knowledge of the linguistic properties of the languages. This approach uses statistic machine learning to increase the accuracy of the translation which uses corpora and statistical based translation in machine translation process.

As a case study, the translation from Bahasa Indonesia to English and vice versa is chosen. As stated by Lewis, Indonesia is a big country with very diverse languages, about 722 living languages [1], however Bahasa Indonesia is considered as a low resource language, meaning the large amounts of bilingual texts are not readily available for the research and development of statistical based translation. For this, we contribute to the development of our own parallel corpus with 3 million words for each corpus.

Producing an optimum evaluation is not an easy task, as many parameters need to be considered, including the translation model, n-gram, smoothing algorithm, alignment, distortion/reordering and word penalty. This study evaluates 28 different parameters in IRSTLM

Received November 15, 2017; Revised January 24, 2018; Accepted February 8, 2018

language modeling and proposes a sequence evaluation mechanism based on the maximum evaluation of each parameter. We explore more than 228 (about 270 millions) experiments which range from 1.5 to 9 hours for each loop of learning and translation process in finding the best quality translation. In this study, the optimum quality translation are measured based on NIST and BLEU. Interested reader may read our previous work in [2-4].

Our current result is quite promising. It shows that by using this particular sequence knowledge/value parameter in translation process, it has a better performance than the other linguistic approaches with in-depth knowledge. In addition to this, in this keynote, the role of pruning, the user interface, the personalization of translation, and the implementation of this machine translation approach in a stand-alone and an online mode is also discussed.

In our study, translation tool is a critical part, but pre-processing, distribution of CPU process to speed up the translation process, user interface, personalization of translation by using user defined dictionary, and user registration are among parts that have important roles and they are not part of the translation tools. A complete set of this machine translation application, we call it Surelator, which currently available in two versions: stand alone and web version.

A translation process can also be speed up by implemented pruning process. The pruning process increases the translation process time by reducing the branch of low probability of the learning translation process.

The remainder of this paper is organized as follows: Section 2 discusses related work on Machine Translation. User Translation Process is described in Section 3. Section 4 presents the process on web translation including the result and discussion. Section 5 discusses the pruning process including the algorithm and the result. Section 6 concludes the paper and outlines areas for future work.

2. Related Work

A complete translation system with all the components required from preprocessing corpus data, training the language models and the translation model is rarely available in the machine translation literatures. In 2006, an open source toolkit was introduced, by Koehn et al. [5]. and it called Moses. Unfortunately Moses has several weaknesses, including have no translation on capital letter and special characters properly. Surelator development fixed this approach in pre and post translation process.

As mentioned earlier, our development concerns on the performance of the translation process in producing a good quality and speed of translation. Daelemans et. al (1999) [6] suspect that the decrease in performance can be linked to the degree of abstraction from exceptions such as pruning or eagerness. Lots of researcher didn’t use a dictionary based and an English lemmatizer approach to translate, but some still do such as in [7].

Long sentences with complex syntax and long-distance dependencies pose difficulties for machine translation systems. Short sentences, on the other hand, are usually easier to translate. Skeleton-based translation can be used to translate Long sentences by decomposing the input sentences into syntactically meaningful chunks. The central part of the sentence can be identified and remains unaltered while other parts of the sentence are simplified. This process produces a set of partial, potentially overlapping translations which are recombined to form the final translation [8]. Word similarity approach can also be used to improve the long sentences translation in statistical based machine translation, by performing word clustering on corpus. Part of speech (PoS) can be developed as Word similarity approach to improve the quality of automatic machine translation [9]. Some researchers also studied to translate for short message such as frm twitter [10].

Statistical Machine Translation (SMT) must be evaluated in terms of many aspects of translation, including fidelity, adequacy, grammaticality and fluency of the output. When the range of variation concerning answers to yes-no questions is taken into consideration, SMTs fail in most of these criteria. Binomial logistic regression can be considered to get a quality translation by carried out with the same corpus data and evaluated in terms of missing arguments, incorrect subject type, correct verbal (item and agreement), particle presence (yes or no), tag answer utilization, full sentence output and non-sentential “other” elements [11].

In term of quality translation, BLUE and NIST has been used by many researchers for quite some time. In 2002, Pepineni et. al (2002) proposed an automatic evaluation of Machine

Translation using BLUE with the range of 0 to 1 and only an identical translation that can get score 1 [12]. NIST has been used as an evaluation metric which uses arithmetic average of N-gram and it gives weight for n-grams that occurs less frequently as it is more likely to carry more information [13]. Automatic metrics to evaluate quality translation can sometimes be misleading and believe a future human evaluation with bilingual judge would required to gain a more complete understanding of the relative merits of the machine translation approaches [14] when it applied to the low-resources languages (Thai, Lao, Bahasa Indonesia). Many researchers also implement the tool for refining the result of quality translation using example based machine translation [3], inclusing fuzzy logic [15].

For web translation, detecting a type of language has been developed also such as in [16] by detecting linguistic pattern. They used Arabic and English, while we used Bahasa Indonesia and English. A bidirectional attention-based encoder-decoder model for the task of machine transliteration has been presented in [17] which described a method that allows automatic construction of a transliteration corpus from a raw English-to-Arabic parallel corpus.

For web translation, the Semantic web which aims to use semantics in the retrieval process, can be considered as the candidate for the web translation model. The semantics which can be captured in ontologies or at the very least in concept hierarchies. After finding the pairs of concepts from different meta-data schemes of ontologies including hierarchical ontologies, then we have an equivalent meaning; and this approach is known as ontology matching [18-20]. However, this study didn’t use any phenotype ontology, such as in [21]. In this study, the ontology match is used for web translation from Bahasa Indonesia to English and vice versa.

3. Translation Process and User Interface Design

A translation begins when a user open document or type text to source window as shown in Figure 1 and click the translation button. The first step of a translation process is the detection of the type of language. Surelator will check the input text language and match it with the chosen translation direction. If it does not match, Surelator will show warning and ask the user to change the translation.

The second step of a translation process is sentence tokenization. User text may contain meaningless characters such as punctuation marks, etc. In this step, every meaningful sentences clause will be extracted from the text. Sentences and meaningless characters will be stored in separate containers. Surelator aggregates these sentences into one clean text for translation process, while the meaningless characters will be recorded and it will be restored after the translation process done.

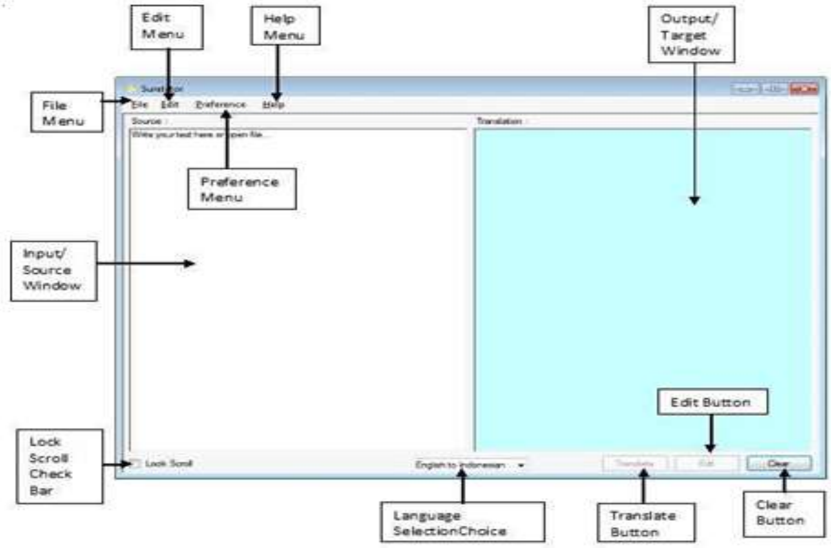


Figure 1. Generic user interface translation menu

Before going into the translation process, there are several step that need to be done. As Surelator has feature to have a personalized translation, the user needs to define a set of user defined dictionary. The user can enable and disable this feature. Each word that matches in this user dictionary will be replaced using string matching function. The words replacement also padded with certain flag so that the translation tool can treat them as a unique word (such as proper nouns) and it will not be translated. The whole translation process in the user interface can be depicted in Figure 2.

When the clean text is ready, it is then sent to translation tools for translation process based on the highest probability. The language direction and pruning parameters will also be given along with the text so that the translation tool will load the correct language model based on user translation preference, i.e. a quick translation for speed process or accurate translation for a better translation accuracy. A progress bar indicating a translation progress will be shown to user during the translation process.

When the translation process finishes translating the text, Surelator will retrieve the text result and split it again sentence by sentence. These sentences will be restored to get the attribute based on the original text format. This is done by combining it with previously stripped meaningless characters and aggregated them in a single sentence with the original attributes. The final translation text then presented to user in a target window.

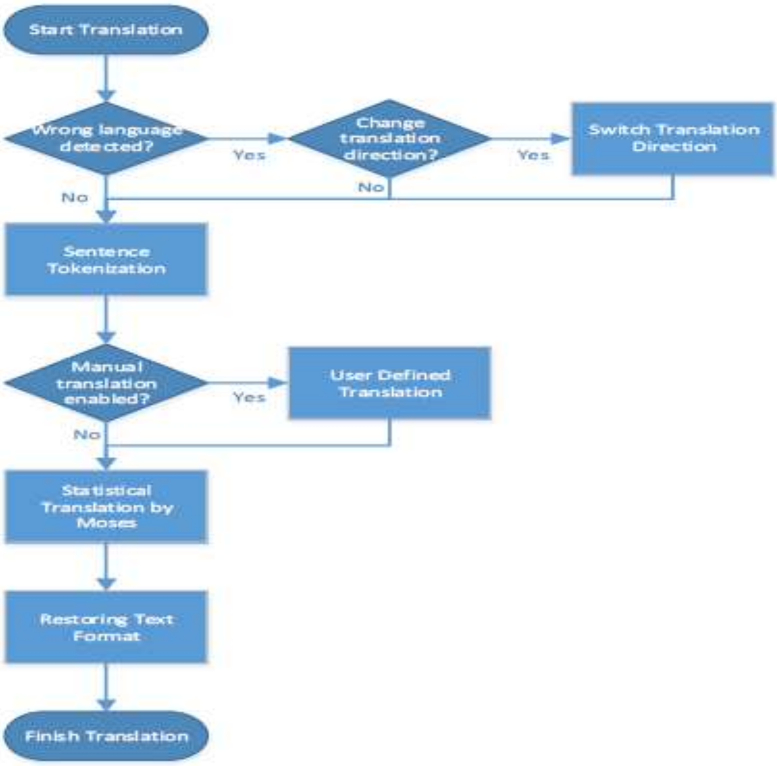


Figure 2. Translation algorithm for the user interface

4. Online (web) Translation

Surelator web translator is a web-based system which supports text translation and web page translation. With Surelator translation tool running in the backend, Surelator web translator supports translation both from English to Indonesia and vice versa.

Surelator web translator consists of a Java based daemon server that runs continuously in the background which is connected to Surelator translation tool and PHP scripts which interfaces the system to clients. For every query from clients, the PHP scripts forward the original text to the Java server through socket. The Java server which consists of several active threads put the accepted text in a queue. A special thread runs continuously and waits until there is a translation thread that is not translating any text. When a translation thread is ready to accept text, the special thread takes the message from the front of the queue, parses it, and sends it to the translation thread. The translation thread translates the text by forwarding it to Surelator translation tool and taking the result by redirecting its I/O streams. Then, it stores the result in a database, from which the PHP scripts query the translated text and show it to clients.

The difference between text translation and web page translation lies on the way the system parses the text before feeding to Surelator translation tool. To parse the web page, Surelator Web Translator utilizes jsoup, a java HTML parser library. Texts are extracted using jsoup, and then parsed as normal text to be translated by the system. Figure 3 depicts the relationship in the system.

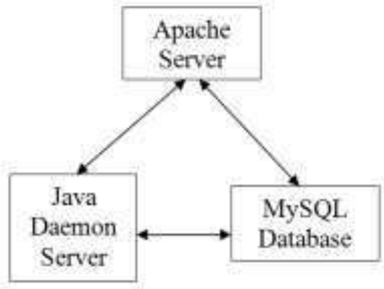


Figure 3. Web translation relationship

The flowchart shows the process of translating a text or web page through Surelator web translator.

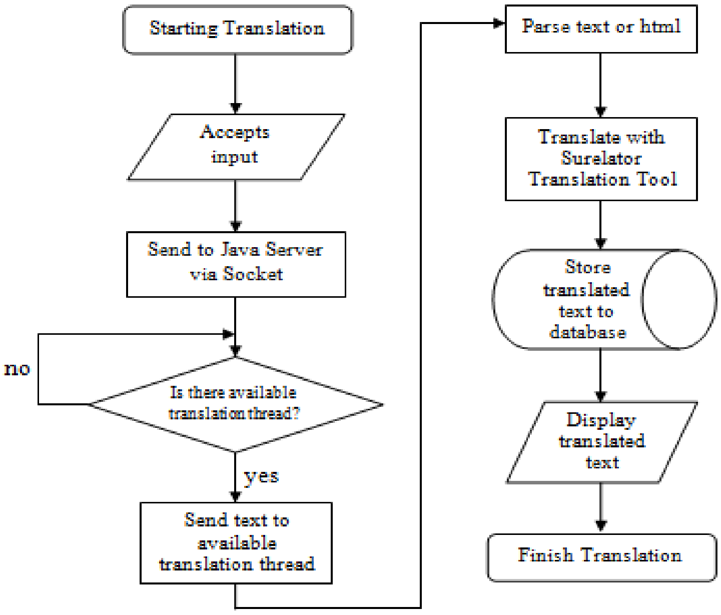
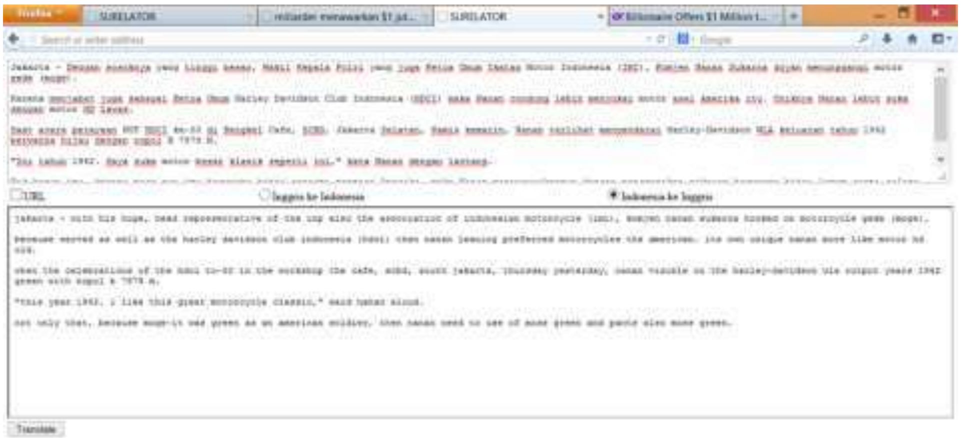

Starting Translation
Accepts input
Send to Java Server via Socket
no
Is there available translation thread?
yes
Send text to available translation thread
Parse text or html
Translate with Surelator Translation Tool
Store translated text to database
Display translated text
Finish Translation

Figure 4. The flowchart for web translation process

As proof of concept, Figure 5 presents a text translation using web that user enter manually from bahasa Indonesia to English, whereas Figure 6 presents a link of web translation from English to Bahasa Indonesia, without changing the location of the text. This prototype also capable to translate a link inside of a web as well.



Jakarta - Dengan posturnya yang tinggi besar, Wakil Kepala Polri yang juga Ketua Umum Ikatan Motor Indonesia (IMI), [illegible] motor gede (moge).

Karena [illegible] Harley Davidson Club Indonesia (HDCI) [illegible]

[illegible] Harley-Davidson WLA keluaran tahun 1942 [illegible] B 7878 H.

"Ini tahun 1942. Saya suka motor [illegible] klasik seperti ini," kata Nanan [illegible].

□ URL ○ Inggris ke Indonesia ● Indonesia ke Inggris

jakarta - with his huge, head representative of the [illegible] also the association of indonesian motorcycle (imi), [illegible] hooked on motorcycle gede (moge).

because served as well as the harley davidson club indonesia (hdci) then nanan leaning preferred motorcycles the american. [illegible] unique nanan more like motor hd old.

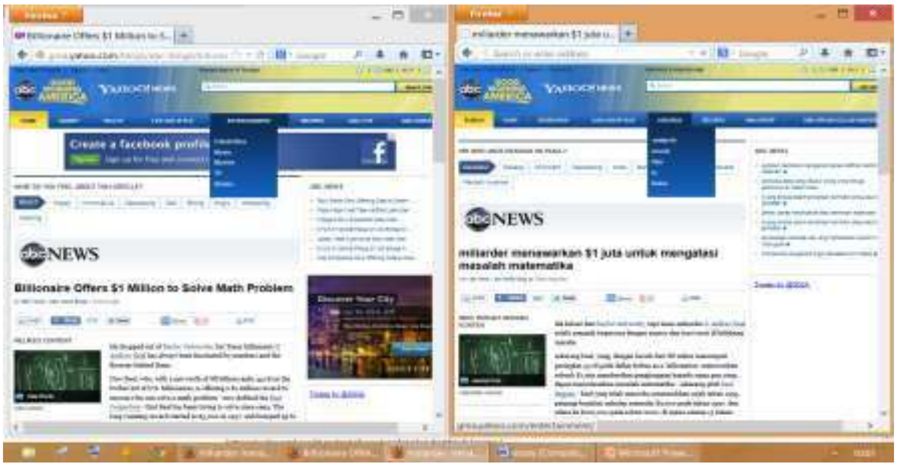
when the celebrations of the hdci [illegible] in the workshop the cafe, scbd, south jakarta, thursday yesterday, nanan visible on the harley-davidson wla output years 1942 green with nopol b 7878 h.

"this year 1942, i like this great motorcycle classic," said nanan aloud.

not only that, because moge-it was green as an american soldier, then nanan used to see of more green and pants also more green.

SURELATOR
miliarder menawarkan $1 jut...
SURELATOR
Billionaire Offers $1 Million t...
Translate

Figure 5. Web translation for text from Indonesia to English



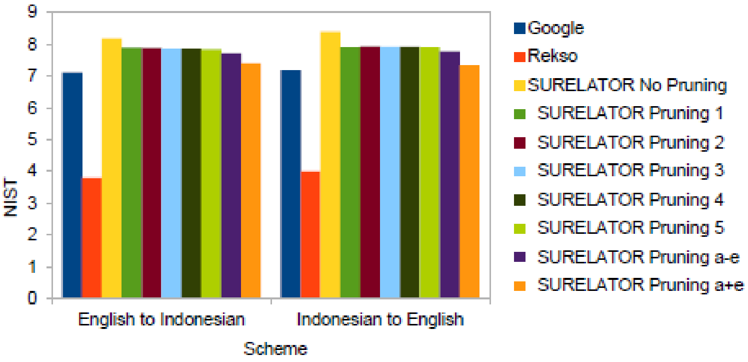
Yahoo! News
Create a facebook profile
abcNEWS
Billionaire Offers $1 Million to Solve Math Problem
Discover Your City
abcNEWS
miliarder menawarkan $1 juta untuk mengatasi masalah matematika

Figure 6. Web translation for a link from English to Indonesia

5. Pruning Process

A translation process can be speed up by implemented pruning process which reduces the branch of low probability of the learning translation process. The more pruning is done, the faster the translation process will be and the quality of the translation will decrease which can be recognized by exploring the values of NIST and BLEU. Theoretically, the pruning process should be stopped when the quality of the translation is starting to become un-acceptable by the user.

In this experiment, we use 1000 sentences from generic domain as the input text for both English to Indonesian and Indonesian to English. These sentences are not inside of the corpus. We compared the pruning process with Google Translate and Rekso Translator (the commercial local translator). As a result, Surelator performs better than Google Translate and Rekso. Even with pruning factor of a from 1 to 5, a-e, and a+e, the NIST and BLEU results are higher than those translation tools as shown in Figure 7 and 8.



Figuure 7. NIST values for google translate, rekso, and SURELATOR with and without pruning

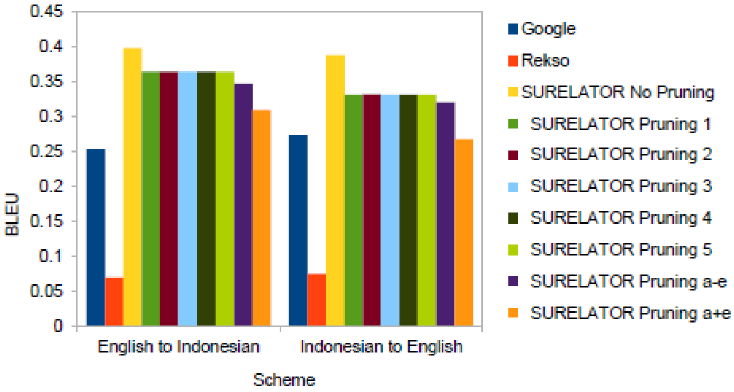


Figure 8. BLEU values for google translate, rekso, and SURELATOR with and without pruning

In another experiment, Surelator also provides better translations for both English to Indonesian and Indonesian to English as compared to Google Translate, Rekso Translor, and Sederet. We measure them using standard translation evaluation tools, NIST and BLEU with the test data from generic domain of 25 and 1000 sentences and political domain of 50 sentences. We perform Sederet translations manually, by feeding one by one sentence of 25 sentences as Sederet web translation application only provide sentence by sentence translation.

The result of this experiment is quite promising, the NIST and BLEU for Surelator reach to 8.5383 and 0.4284 respectively, for English to Indonesian translation while the second best, Google Translate reach the value of 7.3235 and 0.2723. Other tools such as the popular dictionary based tool Rekso translator and the website Sederet with their three different variations do not perform as well as Surelator nor Google Translate. Similar cases also happen for Indonesian to English translation, the highest NIST and BLEU are achieved by Surelator with the values of 8.7216 and 0.4189 respectively. Google Translate also perform second best with NIST and BLEU of 7.2634 and 0.2783.

When we bring to experiment by using the political domain data, Google Translate performs better than Surelator, the NIST and BLEU reach to 5.9017 and 0.3138 for Indonesian to English translation while the values of Surelator are 5.2216 and 0.2232 (Figure 9-11). However, Surelator still performs better than Rekso translator. Similar cases also apply to the English to Indonesian translation with Google Translate performs the best, the second one is Surelator and the last one is Rekso translator. We believe that Google performs better in the political domain because the Google's corpus has higher training data of political domain than Surelator's corpus. If we increase our training data for each domain and train them separately, we may achieve similar results to Google Translate in political domain.

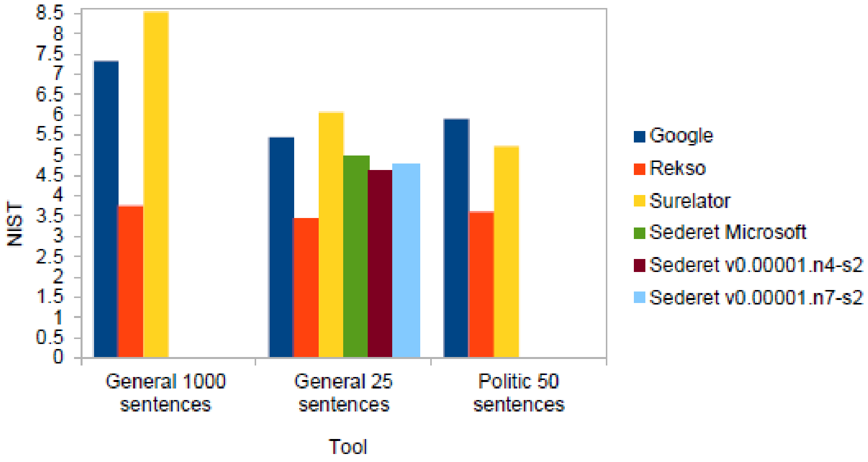


Figure 9. NIST values for English to Indonesian translation in general domain and politic domain

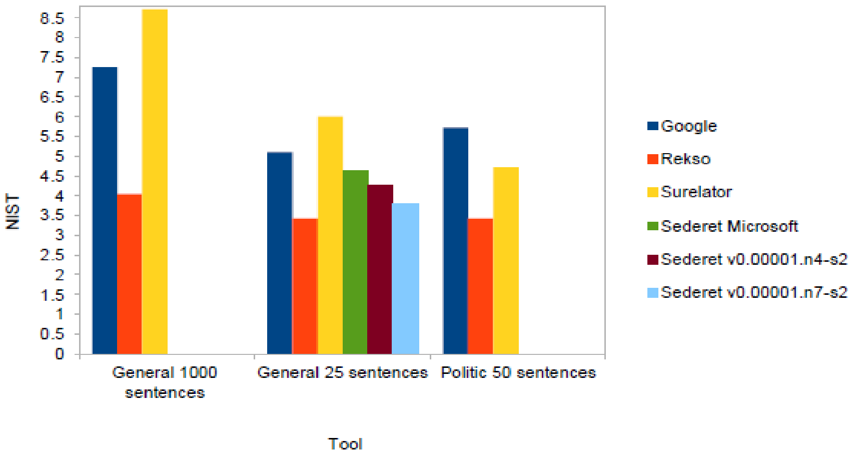


Figure 10. NIST values for Indonesian to English translation in general domain and politic domain

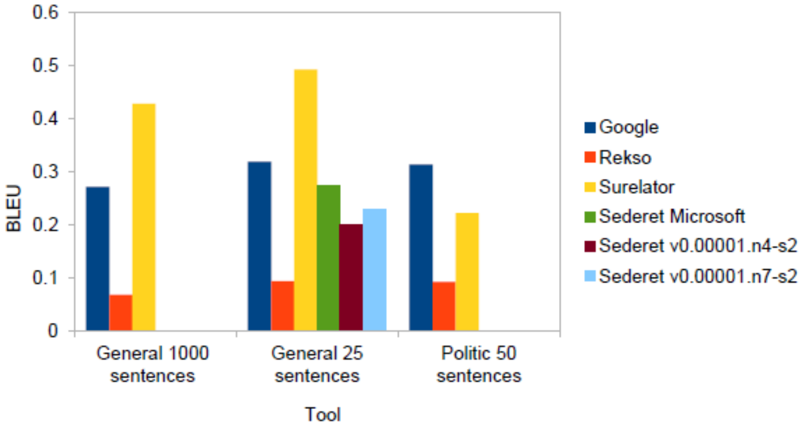


Figure 11. BLEU values for English to Indonesian translation in general domain and politic domain

4. Conclusion

This paper study a complete translation system with all the components required from preprocessing corpus data, training the language models and the translation model. This paper focus on the roles of pruning, the user interface design including pre-processing, the personalization of translation by defining user dictionary, and the development of the prototype in a stand-alone and an online mode. The aim of this study is to find a possible optimum translation based on two variables: quality and speed of translation. During the study, the learning process and translation process of the translation tools have a significant role for both variables, to find the optimum translation. For this purpose, 28 different parameters in IRSTLM language modeling was evaluated. As a result, a sequence evaluated optimum variable formed after the experiments, and it produces a promising result. During the test, two domains were explored, generic domain and political domain. Surelator was performing very well in generic domain but for political domain Google translate produces better quality translation based on NIST and BLEU value.

Pruning plays an important role for speeding up a translation process but it reduces the quality of the translation. In this paper, 7 types of prunings were tested and the best pruning approach is pruning type 2 for BLUE and pruning type 1 for NIST for both translations from Bahasa Indonesia to English and vice versa.

As for the future work, the next target would be 1. The development of a multi-domain Surelator which includes Math, Biology, Chemistry, Physic, Politic, Economy, Sport, etc. and 2. To develop a machine translation for local languages in Indonesia with the highest number of speakers or other consideration, such as Bahasa Java, Sunda and Papua.

References

[1] MP Lewis (ed). Ethnologue: Languages of the World, Sixteenth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com. 2009.

[2] J Asian, T Mantoro*. Optimum Mechanism for Good Quality Translations Evaluation in IRSTLM Language Modelling*. The 6th International Workshop on Malay and Indonesian Language Engineering (MALINDO) 2012, Sarawak, Malaysia.

[3] MA Ayu, T Mantoro. *An Example-Based Machine Translation Approach for Bahasa Indonesia to English*, The IEEE Symposium on Industrial Electronics and Applications (ISIEA), Langkawi, Malaysia. 2011.

[4] T Mantoro, J Asian, R Octavian, MA Ayu. *Optimal translation of English to Bahasa Indonesia using statistical machine translation system*, 2013 5th International Conference on Information and Communication Technology for the Muslim World. Rabat, Morocco. 2013.

[5] P Koehn, H Hoang, A Birch, C Callison-Burch, M Federico, N Bertoldi, B Cowan, W Shen, C Moran, R Zens, C Dyer, O Bojar. A Constantin, E Herbst. *Moses: Open Source Toolkit for Statistical Machine Translation*. In Proc. of the 45th Anual Meeting of the ACL (ACL'07), Stroudsburg, PA, USA. 2007: 177-180.

[6] W Daelemans, A van den Bosch, J Zavrel. Forgetting Exceptions is Harmful in Language Learning (1999), *Machine Learning Special Issue on Natural Language Learning.* 1999; 34 (1): 11-41

[7] Y Yeong, T Tan, SK Mohammad. *Using Dictionary and Lemmatizer to Improve Low Resource English-Malay Statistical Machine Translation System*, Procedia Computer Science. 2016; 81: 243-249.

[8] E Hasler, A Gispert, F Stahlberg, A Waite, B Byrne. Source sentence simplification for statistical machine translation. *Computer Speech & Language*. 2017; 45: 221-235.

[9] H Sujaini, Kuspriyanto, AA Arman, A Purwarianti. A Novel Part-of-Speech Set Developing Method for Statistical Machine Translation. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*. 2014; 12(3): 581-588.

[10] F Mallek, B Belainine, F Sadat. *Arabic Social Media Analysis and Translation,* Procedia Computer Science. 2017; 117: 298-303.

[11] EC Soares, *Yes-No Answers. Partial Pro-drop Languages and Machine Translation.* Procedia-Social and Behavioral Sciences. 2016; 231: 135-142.

[12] K Papineni, S Roukos, T Ward, WJ Zhu. *BLEU: A Method for Automatic Evaluation of Machine Translation.* In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02). Stroudsburg, PA, USA. 2002: 311-318.

[13] G Doddington. *Automatic Evaluation of Machine Translation Quality Using N-Gram Co- Occurrence Statistics.* In proc. of the 2nd Int'l Conf. on Human Language Technology Research (HLT'02). Morgan Kaufmann Pub. Inc., CA, USA, pp. 138-145. 2002.

[14] WPPa, YK Thu, A Finch, E Sumita. *A Study of Statistical Machine Translation Methods for Under Resourced Languages.* Procedia Computer Science. 2016; 81: 250-257.

[15] M Rana, M Atique. *Use of Fuzzy Tool for Example Based Machine Translation.* Procedia Computer Science. 2016; 79: 199-206.

[16] S Ebrahim, D Hegazy, MGM Mostafa, SR El-Beltagy. *Detecting and Integrating Multiword Expression into English-Arabic Statistical Machine Translation.* Procedia Computer Science. 2017; 117: 111-118.

[17] MSH Ameur, F Meziane, A.Guessoum. *Arabic Machine Transliteration Using an Attention-Based Encoder-Decoder Model.* Procedia Computer Science. 2017; 117: 287-297.

[18] Risto Gligorov, Zharko Aleksovski, Warner ten Kate, Frank van Harmelen. *Using Google Distance to Weight, Approximate Ontology Matches*. International World Wide Web Conference Committee, (IW3C2) WWW 2007/Track: Semantic Web. 2007;.767-776.

[19] S Khan, M Safyan. Semantic matching in hierarchical ontologies, *Journal of King Saud University-Computer and Information Sciences.* 2014; 26(3): 247-257.

[20] AlNazer, S Albukhitan, T Helmy. C*ross-Domain Semantic Web Model for Understanding Multilingual Natural Language Queries: English/Arabic Health/Food Domain Use Case*. Procedia Computer Science. 2016; 83: 607-614.

[21] K Wołk, A Wołk. *Machine enhanced translation of the Human Phenotype Ontology project*, Procedia Computer Science. 2017; 121: 11-18.