Syntaxis-based extraction method with type and function of word detection approach for machine translation of Indonesian-Tolaki and English sentences

Muh Yamin
Department of Informatics
Faculty of Intelligent Electrical and Informatics Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
muh_yamin@uho.ac.id

Riyanarto Sarno
Department of Informatics
Faculty of Intelligent Electrical and Informatics Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Rachmad Abdullah
Department of Informatics
Faculty of Intelligent Electrical and Informatics Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
rabdullah1506@gmail.com

Untung
Department of Sastra
Faculty of Sastra Bahasa
Halu Oleo University
Kendari, Indonesia
unesa200852@yahoo.com

2022 International Conference on Information Technology Research and Innovation (ICITRI) | 978-1-6654-6184-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICITRI56423.2022.9970225

***Abstract*—To build an Indonesian Machine Translation (MT), it is not only needed a related syntactic analysis to the correct spelling of words but also needed related contextual analysis, consist type and function of word, morphology, and semantic. The dictionaries usage is needed to translates Indonesian basic words and to captures good word translations through the semantic and context of words in a sentence or document. This study purposes to extracts Indonesian and Tolaki words for building a good MT by comparing the development of Indonesian MT which focuses on deep cases of morphology and syntactic. We developed morphtool to captures the morphological elements of Indonesian and Tolaki words. For working in deep syntactic case, we build a rule to captures the function and type of word that can affect the word itself translation in the sentence. We combine supervised and unsupervised techniques to work on the text extraction in the words, sentences, and documents through the morphonemic rules of Indonesian-Tolaki syntaxis manner. Then, we use hybrid MT, combining Statistical MT (SMT) and Rule Based MT (RBMT), for sentence translation process. The hybrid MT evaluation process from the Indonesian-Tolaki to English translation performance test shows the accuracy result is 0.74. Meanwhile, the performance test of the English to Indonesian-Tolaki translation shows the accuracy result is 0.71. These results indicate that the proposed MT method can work better than the SMT and RBMT methods with an average accuracy of around 70%.**

Keywords—machine translation, SMT, RBMT, hybrid MT

I. INTRODUCTION

Many methods are resulted in the development of English Machine Translation (MT) for translating words through the English rules but it does not happen for the Indonesian MT. So it is needed a method development that is in line with the development of the English MT method and in accordance with the Indonesian rules. MT can be classified into two main models, namely rule-based and corpus-based [1]. Here, we explain more detail the developments regarding the work of Indonesian MT through the techniques and methods described in the previous research.

The Indonesian MT [2] through the word semantic using morphological tools to capture nouns and foreign words has not worked on the translation of Indonesian sentences or documents. It has not worked on the word contextual cases in sentences and morphonemics in depth. The other Indonesian MT [3] can solve the problems in the translation process, such as low coverage corpus data, unknown words, and sentence rearrangement problems. While the case of morphology and contextual words has not been done. Jarob [4] corrected the problems of the previous translation process by working on the translation that related to affixes and basic words. However, it has not worked on translations based on sentence context and morphology in depth. The related issue of Indonesian MT [5] shows that there is not yet a ready-to-use parallel corpus because it is still very dependent on the corpus used. There are translation errors due to typos and inconsistencies in the writing of words in the corpus. Rahutomo [6] examines the phenomenon of using Indonesian vocabulary listed in dictionaries more deeply from the point of computer science view. There are 26,887 lemmas that never used from daily analysis in online media.

Based on these problems and reviews, we propose a method to analyze and provide comprehensive word translations from various studies conducted on the translation of single words and phrases in Indonesian sentences. All related research to Indonesian MT are classified and summarized to demonstrate the novelty that exists in this area and to assist other researchers to further discover findings, allow development of new algorithms, and improvements to available work. Word extraction has various kinds classification features, namely: simple surface features, word generalization, sentiment analysis [7], [8], lexical resources, linguistic features, and knowledge-based features [9].

Therefore, we need a detailed Indonesian word extraction process that is not only perform syntaxis extraction based on morphological analysis [10] but also to extract Indonesian semantics [11]. However, the main problem of related research to Indonesian word extraction is the lack of a corpus labeled Indonesian that can be applied directly to dataset reviews and also to different domains.

Syntaxis cases are often found in Indonesian word extraction. For example, the 3 sentences that contain syntaxis cases in Indonesian-Tolaki and English which are caused by a word "naik (in Tolaki: pe'eka; in English: go up)" that can have different types and functions of word. Sentence S1 contains the word "naik" as the predicate with verb type. Sentence S2 contains the word "naik" as the subject with noun type. Sentence S3 contains the word "naik" as the complement with adjective type. So, a method is needed to extract the syntaxis cases which can distinguish each word pattern that has different functions and types of word but the word meaning context is similar. Syntaxis cases are problems that have the same word with the same meaning but have different word function and type in a sentence. Certainly, syntaxis cases are different with semantic cases that have the same word, both have the same or different word function and type and have different word meanings.

S1 : Saya naik kelas (Inaku pe'eka kalasi) → I go to class

S2 : Naik terasa melelahkan → Riding is tiring

(Pe'eka kupenasa'i mokongango)

S3 : Harga minyak naik (Oli luwi pe'eka) → Oil prices go up

We work on Indonesian-Tolaki word extraction for the Indonesian-Tolaki MT, which is not only based on syntaxis but also semantic, because there are still many novelties that can be done regarding Indonesian-Tolaki MT. We compile dataset manually from several Indonesian datasets, especially the Tolaki Regional Language. We assume that a detailed classifier is needed to develop a system that can assist in determining whether the sentence is related to contextual words that contain elements of morphonemic, pronoun, affixation, and semantic. We aim to extract related Indonesian-Tolaki sentences that contain these elements or nothing. We use rule-based approach to analysis the Indonesian-Tolaki words. First, the pre-processing stage is carried out. Then, text extraction process to detect words based on sentence types is carried out using FLAIR tools to generate tags for each word that contains elements of Noun Phrase (NP) and Adjective Phrase (AP). This tool is used to detect the syntaxis element of Indonesian words [12]. After getting the word tag result, the classification process is carried out. To develop the classification process, a combination of machine learning and Term Frequency Inverse Document Frequency (TF-IDF) [13] methods was used. We test the proposed extraction system using an annotated corpus which compiled manually, consists of 1200 training and 300 testing data to obtain the best classification performance test. We use TF-IDF, Word2Vec, and BERT embedding for syntaxis-based word extraction. Then, we combine Statistical Machine Translation (SMT) and Rule-Based Machine Translation (RBMT) for proposed MT semantically. Finally, evaluation process based on Precision (P), Recall (R), F1-measure (F), and Accuracy (A) for the classification and translation processes is obtained.

II. Related Work

This research explains several works as follows.

A. Dataset

This study uses Indonesian and Tolaki datasets which is compiled manually. Table I shows the proposed dataset representation.

TABLE I. Dataset Representation

Domain	Train	Test	Total
Indonesian	1200	300	1500
Tolaki	1200	300	1500

B. Pre-processing

Generally, the pre-processing stage consists of: 1) case folding, 2) filtering, 3) normalization, 4) stopword removal, 5) stemming, 6) tokenizing.

C. Text Extracting

Generally, the text extraction method consists dependency parsing, association rule mining, association, hierarchies, ontology, word modeling, co-occurrence, rule-based, and clustering. Common problems that occur when extracting Indonesian text include: unstructured syntax [14], morphemes [15], word functions, and word types. The text extraction stage aims to analyze the relationship between word functions and word types with the assumption that there is a word that has a different word type.

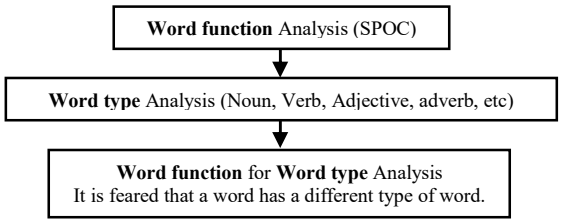

Fig. 1. Function and type of words analysis

1) Morphonemics

Indonesian and Tolaki has morphonemics that can affect the function and type of words from the basic words themselves. Table II shows the morphonemics representation in Indonesian and Tolaki.

TABLE II. Morphonemics Representation of Words

No	Indonesian		Tolaki	
	Words	Morphonemics	Words	Morphonemics
1	Jemur	-	Puai	-
2	Menjemur	Men	**Mom**buai	**Mom**
3	Jemuran	an	**Pipi**nuai	**Pipi**

TABLE III. STRENGTH AND WEAKNESSES OF SMT AND RBMT

Description	SMT	RBMT
Translation quality.	Relatively high.	Relatively low, because the effort required to build an RBMT system is so large that it affects the performance of the translator machine.
Morphological translation.	Performs better for morphologically low to rich languages.	Performs better for morphologically rich languages to low languages.
Morphological translation process.	Manually, cannot separate the suffix by itself.	Automatically, can use a morphology analyzer.
Morphological type translation process.	Unable to handle translation of inflected suffix words in some cases.	Can easily separate suffixes from inflected words and generate translations.

2) *Word Function*

The Tolaki word functions consist of subject, predicate, object, and complement where the word position in the sentence can affect the word function as shown in Section IV.

3) *Word Type*

The Tolaki word types consist of 17 tags which are taken from the universal POS tag. The word position in the sentence can affect the word type as shown in Section IV.

D. *Machine Translation*

Machine Translation (MT) can be divided into two, namely corpus-based and rule-based. The corpus based approach or Statistical Machine Translation (SMT) works based on statistical models that taken from parallel corpora of bilingual texts. The main steps in SMT are: Corpus preparation, Training, Decoding and Testing. Meanwhile, the rule based approach or Rule Based Machine Translation (RBMT) works based on the specification of rules for morphology, syntax, lexical selection, and also transfer and generation. The RBMT workflow is grouped into three process phases, namely: Analysis phase, Lexical transfer phase, and Generation phase.

Table III shows the strength and weakness comparisons of SMT and RBMT from several previous studies. The process of translating sentences to or from Indonesian, also known as two-way language translation, often results in poor translation because the MT that has been made has not paid attention to the Indonesian rules. These rules include: morphological, syntactic, and semantic rules.

III. METHODOLOGY

In this section, analysis and discussion of previous articles on MT is carried out. The analysis and discussion is based on several factors such as the text extraction, classification, and MT methods used. Figure 2 shows the proposed method flowchart using hybrid SMT-RBMT.

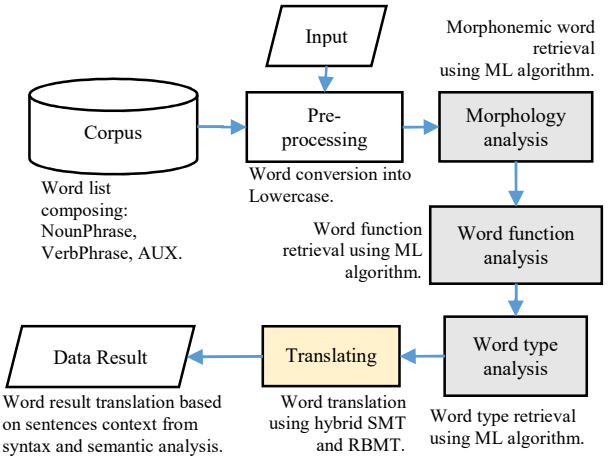

Fig. 2. Proposed method flowchart using hybrid SMT-RBMT

A. *Dataset*

The proposed dataset is shown in Table IV. Table IV shows the dataset representation of Indonesian and Tolaki sentences that is compiled manually.

TABLE IV. DATA COLLECTION

No.	Indonesian sentences	Tolaki Sentences
1.	Saya naik kelas	Inaku pe'eka kalasi
2.	Naik terasa melelahkan	Pe'eka *kupenasa'i* mokongango
3.	Harga minyak naik	Oli luwi pe'eka

B. *Text Extraction*

The pre-processing is used to remove symbols that are not used and prepare the data to be ready for text extraction process. An approach is developed to extract syntaxis cases that can distinguish each sentence pattern that has different word function and type based on the word context and meaning. First, the POS tagging process is carried out using FLAIR. The POS tagging result is shown in Table V. Then, the morphological analysis is carried out using the MorphInd concept. Finally, the extraction of word function and type is carried out using a machine learning algorithm.

TABLE V. POS TAGGING RESULT

ID	Indonesian sentences	Tolaki Sentences	POS tagging result
1	Saya **naik** kelas	Inaku **pe'eka** kalasi	Saya <PRON> naik <VERB> kelas <NOUN>
2	**Naik** terasa melelahkan	**Pe'eka** *kupenasa'i* mokongango	Naik <PROPN> terasa <VERB> melelahkan <ADJ>
3	Harga minyak **naik**	Oli luwi **pe'eka**	Harga <NOUN> minyak <NOUN> naik <ADJ>

1) *Morphology Extraction*

The morphology extraction process was carried out using the MorphInd concept. Figure 3 shows the proposed morphology extraction algorithm. First, the pre-processing results are used as input for POS tagging. Then, take the token using TF-IDF. After the tokenization results are obtained, the vector calculation of each token is carried out using

Word2vec. The result of the highest vector value is used as the BERT embedding input to get the actual target token based on the number of word forms in the document.

***Input**: POS tagging result, **Output**: MorphTool result*
Start
1. *Tokenize each term list;*
2. *Extraction of affixed words;*
3. *Basic word analysis;*
4. *MorphTool Result;*

End

Fig. 1. Morphology extraction algorithm

2) Word Function Extraction

The morphology extraction result is used as input in word function extraction. Fig. 4 shows the word function extraction algorithm. The word function determination of subject, predicate, object, complement of adverb and adjunct is done based on the word sequence in the sentence. We compiled 3 rules to identify the function of a word. Rule 1, if a sentence begins with NP. Rule 2 if a sentence starts with VP. While rule 3, if a sentence begins with AUX.

***Input**: MorphTool result*
***Output**: Subject, predicate, object, compl adverb, compl adjunct*
Start
1. *Tokenize each term list;*
2. *Identify the POS Tag in the sentence;*
3. *Analysis of basic words if there are affixed words;*
4. *Update the basic word if there is an affixed word;*
5. *Predict word based on word sequence;*
6. *Word function Result;*

End

Fig. 2. Word function extraction algorithm

3) Word Type Extraction

The word function extraction result is used as input in word type extraction. Figure 5 shows the word function extraction algorithm to get word type. We compile 51 rules of parent and child nodes. The parent node consists of NP, VP, and AUX. While the child nodes consist of ADJ, ADP, ADV, AUX, CCONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROPN, PUNCT, SCONJ, SYM, VERB, X. We compile 51 rules as the basis to correct word tag relation in the sentence. So if there is an incorrect tag relation, the system will automatically update the correct word.

***Input**: Word function result*
***Output**: ADJ, ADP, ADV, AUX, CCONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROPN, PUNCT, SCONJ, SYM, VERB, X*
Start
1. *Tokenize each term list;*
2. *Identify the POS Tag in the sentence;*
3. *Analysis of basic word tags;*
4. *Update basic word Tags;*
5. *Predict word based on word sequence relation;*
6. *Word function Result;*

End

Fig. 3. Word function extraction algorithm

C. Machine Translation

The proposed Hybrid SMT and RBMT is used because it requires a rule set that can work on syntaxis cases, namely identifying and extracting word functions against the word type in sentence so they can work on semantic cases.

IV. Results and Analysis

This section shows the analysis and result of proposed method of text extraction and machine translation.

A. Text Extraction Analysis

Table 6 shows proposed rules implementation for RBMT analysis. Indonesian POS tagging results are used to measure the word structure completeness in sentences. Then, a good word structure used in a sentence should at least consist of a subject (NP) and a predicate (VP). The result of determining the translation is used to compare the word similarity probability between the results of rule-based analysis and statistical-based analysis, with the highest value will be taken as the translation result. These rules work in the following two cases:

- if there is a difference in the results between the Indonesian-Tolaki to English translation and the English to Indonesian-Tolaki translation, then an analysis is carried out based on the position of the word error, including the following:
  - if there is an error in Subject Noun Phrase translation, then word updating is carried out based on similarity (Noun Phrase, Verb Phrase)
  - if there is an error in Predicate VERB translation, then word updating is carried out based on result of hidden word translation between Predicate VERB – Object NOUN, Predicate VERB – Complement, or Predicate VERB – Object NOUN – Complement from existing sentence structure
  - if there is an error in Object NOUN translation, then word updating is carried out based on word similarity in the object NOUN form that obtained from all document existing
  - if there is an error in Complement of adverb ADV and adjunct ADJ, then word updating is carried out based on word translation between predicate VERB – object NOUN – complement of adverb ADV or adjective ADJ.
- if there is an incomplete word structure without VERB after subject NP or object NOUN, then the VERB tobe is added automatically after the subject NP or object NOUN.

First case, sentence “saya naik kelas”, the word “naik” as a predicate VERB has difference result while using Indonesian-Tolaki to English translation and English to Indonesian-Tolaki. The word context “naik”, if it is translated from Indonesian-Tolaki to English then the result is “going” using automatic tobe “am” from subject “I”. Actually, the pairing word “am going” while is translated to Indonesian has the meaning “sedang pergi”. Therefore, the word “going” is marked as a word translation error. The identification result based on the proposed rules can be obtained the translation of the VERB “naik” when paired with the NOUN “kelas” is “promoted to next grade”. So, updating process result is “I am promoted to next grade”.

Second case, sentence "naik terasa melelahkan", the word "naik" as a subject PROPN has difference translation result between Indonesian-Tolaki to English and English to Indonesian-Tolaki. The word context "naik", if it is translated from Indonesian-Tolaki to English then the result is "riding" with predicate VERB "feels" and complement adjunct "tiring". Actually, the word "riding" while is translated to Indonesian has the meaning "berkendara". Therefore, the word "riding" is marked as a word translation error. The identification result based on the proposed rules can be obtained the subject form expansion of word "naik" with the NOUN type from the corpus existing is "kenaikan". The actually sentence in Indonesian is changed to be "kenaikan terasa melelahkan". So, updating process result is "hike feels tiring".

Third case, sentence "harga minyak naik", the result of Indonesian-Tolaki to English translation and English to Indonesian-Tolaki translation can be obtained the similar result. However, the sentence is not complete because there is not has predicate VERB. Therefore, the sentence is marked as a wrong sentence, as well as the resulting word translation is wrong. The identification result based on the proposed rules, the VERB tobe "adalah" is added after the subject NOUN "harga minyak". The actually sentence in Indonesian is changed to be "harga minyak adalah naik". So, updating process result is "oil prices are going up".

TABLE VI. THE RULE IMPLEMENTATION FOR RBMT ANALYSIS

Case 1	
Input	Saya (PRON) naik (VERB) kelas (NOUN)
Translation	I am going to class
Back translation	Saya pergi ke kelas
Hidden topic analysis for different result of two way translation	Subject: NP Saya (PRON) → naik (VERB) → kelas (NOUN) Naik kelas → promoted to next grade promoted to next grade → naik kelas
Result	I am promoted to next grade (Saya dipromosikan ke kelas berikutnya)
Case 2	
Input	Naik (PROPN) terasa (VERB) melelahkan (ADJ)
Translation	Riding feels tiring
Back translation	berkendara terasa melelahkan
Hidden topic analysis for different result of two way translation	Subject: NP naik (VERB) → kenaikan (NOUN) kenaikan (NOUN) → terasa (VERB) → melelahkan (ADJ) kenaikan terasa melelahkan → hike is tiring hike is tiring → mendaki itu melelahkan
Result	hike feels tiring (mendaki terasa melelahkan)
Case 3	
Input	Harga (NOUN) minyak (NOUN) naik (ADJ)
Translation	Oil prices rise
Back translation	harga minyak naik
Hidden topic analysis for different result of two way translation	Subject: NP harga (NOUN) → minyak (NOUN) → adalah (VERB) → naik (ADJ) harga minyak adalah naik → oil prices are going up oil prices are going up → harga minyak naik
Result	Oil prices are going up (harga minyak naik)

B. *Machine Translation Analysis*

The morphological extraction implementation is carried out to get the word types in sentences. As an illustration, 3 Indonesian-Tolaki sentences are used to be translated into English. Table VII shows the translating process Indonesian-Tolaki to English. The English translation result is used for back translation as the input to be translated into Indonesian-Tolaki. Then, Table VIII shows the back translation result.

TABLE VII. MACHINE TRANSLATION ANALYSIS RESULT FOR INDONESIAN-TOLAKI TO ENGLISH

ID	Example of Sentence	Results	
		Word function	**Word type**
	Indonesian-Tolaki-English		
1	Saya **naik** kelas (Inaku **pe'eka** kalasi) (I'm **promoted** to next grade)	Subject **Predicate** Object	Noun **Verb** Noun
2	**Naik** *terasa* melelahkan (**Pe'eka** *kupenasa'i* mokongango) (**Hike** feels tiring)	**Subject** *Predicate* Complement adverbial	**Noun** *Verb* Adverb
3	Harga minyak **naik** (Oli luwi **pe'eka**) (Oil prices **are going up**)	Subject **Complement adjunct**	Noun **Adjective**

TABLE VIII. MACHINE TRANSLATION ANALYSIS RESULT FOR ENGLISH TO INDONESIAN-TOLAKI

ID	Input (English)	Output (Indonesian – Tolaki)
1	I'm **promoted** to next grade	Saya dipromosikan ke kelas berikutnya *(Inaku nggo pine'eka'ako ine kalase lakotu'uno)*
2	**Hike** feels tiring	Mendaki terasa melelahkan *(Monduka'ako kupenasa'i mokongango)*
3	Oil prices **are going up**	Harga minyak naik *(Oli luwi pe'eka)*

TABLE IX. EVALUATION PROCESS OF MACHINE TRANSLATION

Method	Language translation	P	R	F	A
SMT	Indonesian-Tolaki to English	0.5397	0.5167	0.5279	0.5417
RBMT		0.6102	0.6167	0.6134	0.6083
Proposed method		0.7231	0.7000	0.7114	0.7417
SMT	English to Indonesian-Tolaki	0.6406	0.6167	0.6284	0.6500
RBMT		0.4219	0.3833	0.4017	0.4167
Proposed method		0.7119	0.7167	0.7143	0.7083

Tables VII and VIII show the differences between the results of one-way translation and back-way translation. The text classification process that has been carried out is able to increase the accuracy of text translation but has not be able to produce an accuracy close to 100%. This is due to the difference in word structure between Indonesian-Tolaki and English. In one-way translation from Indonesian-Tolaki to English, Indonesian-Tolaki has affixes and suffixes of word, while English does not have them, causing hybrid MT errors to understand to pick up very precise translation words. The word translation analysis proposed on hidden topics is proven to be able to capture the context of the word more accurately.

So, the back-way translation process from English to Indonesian-Tolaki can work better and more accurately according to the actual meaning of the sentence. For instance, the word "naik" when translated into English has two classes, namely adverb and verb. Whereas in English corpus, the adverb form of the word "naik" has membership [go on, go up] and the verb form of the word "naik" has membership [going, ride, rise, increase, raised, increased, …]. Furthermore, English has tenses-based word forms which impacted in error translation, even though the actual word input was used did not use the adverb of time. This case occurs based on the word probability factor in the document, which is also one of the methods to get the target word translation in the proposed hybrid MT. For instance, "Saya naik kelas" which has the translation "I'm going to class". While using proposed hybrid MT, the more accurate result that obtained is "I'm promoted to next grade."

V. Conclusion

This study results describe the research conducted on MT which focuses on the latest MT methods in the world to be applied to Indonesian-Tolaki MT. Most Indonesian researchers have worked on SMT and RBMT, but have not covered syntaxis rules in depth. We translate words based on the word function that can affect the word type in a sentence. This is proven to affect the result of a detailed and accurate translation. The SMT obtain better result for the English to Indonesian-Tolaki translation with an accuracy of 65.00% than the Indonesian-Tolaki to English translation with an accuracy of 54.17%. The RBMT obtain better result for the Indonesian-Tolaki to English translation with an accuracy of 60.83% than the English to Indonesian-Tolaki translation with an accuracy of 41.67%. While the proposed method, hybrid SMT-RBMT, obtain better result for the English to Indonesian-Tolaki translation with an accuracy of 74.17% than the Indonesian-Tolaki to English translation with an accuracy of 70.83%. These results indicate that the proposed method can work better than SMT or RBMT. Parallel corpus collection was also done manually for the data training process. Attention-based approaches also need to be developed in depth to improve the performance of this proposed method, starting from SMT, RBMT, and hybrid SMT-RBMT. Here, the conclusion of the new novelties that is needed for further research: The collection of new data related to the Indonesian language and the rules as a whole; The collection of new data related to Regional Languages in Indonesia and the rules as a whole; Experiment with new methods and techniques from existing work and compare the performance results; Development of new tools for Indonesian MT; Experiment with different performance metrics or new performance metrics in the MT research area; Increasing the accuracy of the translation system which is influenced by many factors; Increasing the number of parallel corpus in order to increase the evaluation value.

Acknowledgment

This research was funded by Lembaga Pengelola Dana Pendidikan (LPDP) under Riset Inovatif-Produktif (RISPRO) Program, the Indonesian Ministry of Education and Culture under Penelitian Terapan Unggulan Perguruan Tinggi (PTUPT) Program, and Institut Teknologi Sepuluh Nopember (ITS) under project scheme of the Publication Writing and IPR Incentive Program (PPHKI).

References

[1] P. Li, "A Survey of Machine Translation Methods," *TELKOMNIKA Indones. J. Electr. Eng.*, vol. 11, no. 12, pp. 7125–7130, 2013, doi: 10.11591/telkomnika.v11i12.2780.

[2] T. Mantoro, J. Asian, R. Octavian, and M. A. Ayu, "Optimal translation of English to Bahasa Indonesia using statistical machine translation system," *2013 5th Int. Conf. Inf. Commun. Technol. Muslim World*, 2013, doi: 10.1109/ICT4M.2013.6518918.

[3] M. A. Sulaeman and A. Purwarianti, "Development of Indonesian-Japanese statistical machine translation using lemma translation and additional post-process," *Proc. - 5th Int. Conf. Electr. Eng. Informatics Bridg. Knowl. between Acad. Ind. Community*, pp. 54–58, Dec. 2015, doi: 10.1109/ICEEI.2015.7352469.

[4] Y. Jarob, H. Sujaini, and N. Safriadi, "Uji Akurasi Penerjemahan Bahasa Indonesia – Dayak Taman Dengan Penandaan Kata Dasar Dan Imbuhan," *J. Edukasi dan Penelit. Inform.*, vol. 2, no. 2, pp. 78–83, 2016, doi: 10.26418/jp.v2i2.16520.

[5] A. A. Suryani, D. H. Widyantoro, A. Purwarianti, and Y. Sudaryat, "Experiment on a phrase-based statistical machine translation using PoS Tag information for Sundanese into Indonesian," *2015 Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2015 - Proc.*, Mar. 2016, doi: 10.1109/ICITSI.2015.7437678.

[6] F. Rahutomo, R. A. Asmara, and D. K. P. Aji, "Computational Analysis on Rise and Fall of Indonesian Vocabulary During a Period of Time," *2018 6th Int. Conf. Inf. Commun. Technol.*, pp. 75–80, Nov. 2018, doi: 10.1109/ICOICT.2018.8528812.

[7] R. Priyantina and R. Sarno, "Sentiment Analysis of Hotel Reviews Using Latent Dirichlet Allocation, Semantic Similarity and LSTM," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 4, pp. 142–155, Aug. 2019, doi: 10.22266/ijies2019.0831.14.

[8] R. Abdullah, S. Suhariyanto, and R. Sarno, "Aspect Based Sentiment Analysis for Explicit and Implicit Aspects in Restaurant Review using Grammatical Rules, Hybrid Approach, and SentiCircle," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 5, pp. 294–305, 2021, doi: 10.22266/ijies2021.1031.27.

[9] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," no. 2012, pp. 1–10, 2017, doi: 10.18653/v1/w17-1101.

[10] S. D. Larasati, V. Kuboň, and D. Zeman, "Indonesian morphology tool (MorphInd): Towards an Indonesian corpus," *Commun. Comput. Inf. Sci.*, vol. 100 CCIS, pp. 119–129, 2011, doi: 10.1007/978-3-642-23138-4_8.

[11] N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri, and A. Jamal, "Colloquial Indonesian Lexicon," *Proc. 2018 Int. Conf. Asian Lang. Process. IALP 2018*, no. August 2020, pp. 226–229, 2019, doi: 10.1109/IALP.2018.8629151.

[12] A. Akbik, S. Schweter, D. Blythe, and R. Vollgraf, "F LAIR : An Easy-to-Use Framework for State-of-the-Art NLP," pp. 54–59, 2019.

[13] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, 2018, doi: 10.5120/ijca2018917395.

[14] D. S. Maylawati and G. A. P. Saptawati, "Set of Frequent Word Item sets as Feature Representation for Text with Indonesian Slang," *J. Phys. Conf. Ser.*, vol. 801, p. 012066, 2017, doi: 10.1088/1742-6596/801/1/012066.

[15] M.-C. de Marneffe *et al.*, "Universal Stanford Dependencies: A cross-linguistic typology," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 2014, pp. 4585–4592, doi: 10.1306/C1EA47CA-16C9-11D7-8645000102C1865D.