ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 11, No. 5, October 2023, pp.117-136

Hybrid neural machine translation with statistical and rule-based approach for syntactics and semantics between Tolaki-Indonesian-English languages

Muh Yamin[1], Riyanarto Sarno[2]
[1,2]Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[1]Department of Informatics Technology, Faculty of Engineering, Halu Oleo University, Kendari, Indonesia

ABSTRACT

Machine Translation (MT) incorporates syntax lexical extraction and semantics to predict accurate results. Indonesian have many factors compared to English that related with syntax, especially morphophonemic factors in the language study. These factors are influenced by Lexical type and function while effected MT to frequently mistranslate sentences containing these factors. Meanwhile, semantic extraction is heavily reliant on syntaxis extraction results to predict accurate Lexical translations. In this study, we propose a hybrid statistical and rule-based for MT method that can solve syntaxis and semantic Indonesian problems that conducted the Local Languages in it, particularly Tolaki. First, we developed lexical extraction techniques in Statistical and Rule Based Approach to compile into hybrid MT. This lexical extraction technique is divided into three major tasks: morphophonemic extraction, Lexical Function, and Lexical type extraction. Then we forecast each output of forwards and backwards translations. We compare the predicted output to find accurate translations. Following that, we update the Lexical type based on the actual Lexical function for the translation updating process, which we mark as incorrect translation. Finally, we evaluated MT in both directions. As a result, the proposed method received significant evaluation results, with a percentage success of Indonesian-Tolaki to English translation achieved Precision 0.7231; Recall 0.7; F1-measure: 0.7114; Accuracy: 0.7417 and percentage of success English to Indonesian-Tolaki translation Precision: 0.7119; Recall: 0.7167; F1-measure: 0.7143; Accuracy: 0.7083.

Keywords: Machine translation, Semantic similarity, SMT, RBMT, Hybrid MT.

Corresponding Author:

Riyanarto Sarno
Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology
Institut Teknologi Sepuluh Nopember
Jl. Teknik Kimia, Keputih, Kec. Sukolilo, Surabaya, Jawa Timur, 60111, Indonesia
E-mail: riyanarto@if.its.ac.id

1. Introduction

Machine Translation (MT) for English language has been develop with multiple approaches by determining English rules which is not applicable to Indonesia Machine Translation (IMT). Consequently, it’s required to develop an approached that dependable with progression of the English MT approach and compliant by Indonesian language regulations. The translation phase of the dataset from Indonesian to English was carried out using the Transformers model from Helsinki NLP was called Opus MT [1], which is a language model that can be used to translate from Indonesian to English. This model is known for its ability in natural language translation tasks. Basically, MT can be divided into rule or corpus-based [2]. Literal translation methods, transfer-based methods, and interlingua-based methods are all part of the rule-based approach. Meanwhile, the corpus-based approach combines statistical and case-based methods. Several studies have been conducted in IMT development. Reference [3] was conducted using Indonesian morphological tools to identify nouns and foreign words within the semantic content of Indonesian sentences or documents without analyzing their translation. Reference [4] is working on a statistical-based MT from English to Indonesian that considers four weighting variables, namely the translation model, language model, distortion (rearrangement), and word penalties, using the BLEU and NIST methods. However, this study did not go into detail about the contextual case of words in Indonesian sentences and morphonemics. Reference [5] was a study on the translation of Indonesian to Pontianak Malay using a statistical-based MT. The limited corpus, on the other hand, becomes an impediment to the translation quality that worked on the Indonesian-Japanese lemma translation using the terms lemma and POSTAG in the translation process [6]. This study, on the other hand, can resolve Indonesian Japanese translation issues such as sentence rearrangement problems, insufficient corpus data analysis and anonymous words. Even though word structure and contextual words need further studied. Reference [7] was proposed the translation of Indonesian-Dayak Taman affixes and basic words by utilizing statistical MT to correct problems in the previous translation process. However, they did not explore context and morphology into translations based on sentences. According to reference [8] there is no ready-to-use parallel corpus of Sundanese to Indonesian, as demonstrated by the difficulty of translating Sundanese text into Indonesian. This study continues to rely heavily on the corpus employed. As a result of typo error and writing inconsistencies in word, the Sundanese still contains translation errors. Reference [9] stated that the standpoint of computer science, which examines in greater detail the experience of applying Indonesian vocabulary listed in thesauruses. From daily analysis in online media, there are 26,887 lemmas that are never used. Furthermore, to understand Indonesian MT this study was identified several studies that conducting method, Local language of Indonesia and analysis that has been developed as follows:

Table 1. Related Study

Author	Method, Language	Analysis		
		Semantic	Contextual	Others
[3]	Indonesian Morphology Tool	-	-	Morphology
[4]	Rule based, English - Indonesian	Lex	-	-
[5]	Statistical based, Indonesia – Melayu Pontianak	Lex	-	-
[6]	Rule based, Indonesian, Japanese	Lex	-	-
[7]	Corpus based, Indonesian - English	Lex		Morphology
[10]	Rule Based, Indonesia –Minang Dan Minang – Indonesia	Lex	-	-
[11]	Rule Based, Indonesia - Gorontalo	Lex	-	-
[12]	Rule Based, Inggris - Bali	Lex	-	-
[13]	Statistical based, Indonesia to Local language (karo)	Lex	-	-
[14]	Statistical and memory based, Indonesian - Javanese	Lex	-	Pragmatic: Krama, Krama Alus
[15]	Phrase-based statistical MT, Sunda - Indonesia	Lex	-	Phrase
[16]	Indonesia - Tolaki	Lex	-	-
[17]	Indonesia - Tolaki	Lex	-	-
[18]	Indonesia – Sulawesi Selatan	Lex	-	-
[19]	NMT using RNN, Lampung - Indonesia	Lex: single and compound sentences	-	-
[20]	NMT attention based, Lampung - Indonesia	Lex: single and compound sentences	-	-
[21]	Rule-based, Indonesia - Tolaki	Lex	-	-
[22]	Rule based, Indonesia - Aceh	Lex	-	-
[23]	Rule based, Inggris – Jawa Krama	Lex		Morphology
[24]	Rule Web-based, Tolaki - Indonesia	Lex		Synonym
[25]	Rule based, Melayu Riau - Indonesia	Lex	-	-
[26]	Direct and Statistical based, Lampung - Indonesia	Lex	-	-
[27]	Direct and Statistical based, Lampung - Indonesia	Lex	-	-
[28]	Direct and Statistical based, Lampung - Indonesia	Lex	-	-

Notes: Lex: Lexical.

The classification features used in lexical extraction include basic surface features, word generalization, sentiment analysis, lexical resources, linguistic characteristics, and knowledge-based features [29]. This study proposes a method for analyzing and providing complete word translations since numerous studies on the translation individual words and phrases in Indonesian sentences still becoming hot topic. Therefore, we require a comprehensive Indonesian lexical extraction procedure that can perform non syntaxis extraction on structural analysis nevertheless syntaxis extraction based on semantic analysis [3]. Nevertheless, additionally be capable of extracting Indonesian semantics [30], [31]. Conversely, the main obstacle in researching Indonesian lexical extraction is the lack of an annotated Indonesian corpus that can be employed as a dataset across diverse domains. Therefore, this research aims to extract Indonesian-Tolaki words not only based on word syntax but also word semantics. This is because there are still many possibilities that can be explored with Indonesian-Tolaki machine translation. This paper extended our study that has been done earlier and explain more detail related with the proposed method Hybrid Neural Machine Translation with Statistical and Rule Based Approach for syntactics and semantics between Tolaki-Indonesian-English languages [32]. The Tolaki Regional Language dataset, which was manually compiled from several Indonesian datasets, was utilized. The hypothesis of this study is that a robust classifier is required to create a system capable of identifying Indonesian sentences that contain Morphophonemic, Pronoun, Affixation, and Semantic contextual words. The aim of this study is to detect sentences in Indonesian-Tolaki language that have either one or none of these elements. This study determined that existing documents must be pertinent to the Indonesian-Tolaki language.

2. Material and methods

In this part, we consider the proposed method that was performed in this study. The present study examines two Machine Translation (MT) methods: corpus and rule based. The corpus-based technique, also called Statistical Machine Translation (SMT), employs statistical models obtained from multilingual corpora. On the other hand, the rule-based approach, also referred to as Rule-Based Machine Translation (RBMT), operates on rules designed for translation. In Figure 1 show the Proposed Hybrid Machine Translation Statistical and Rule Based Approach as follows:

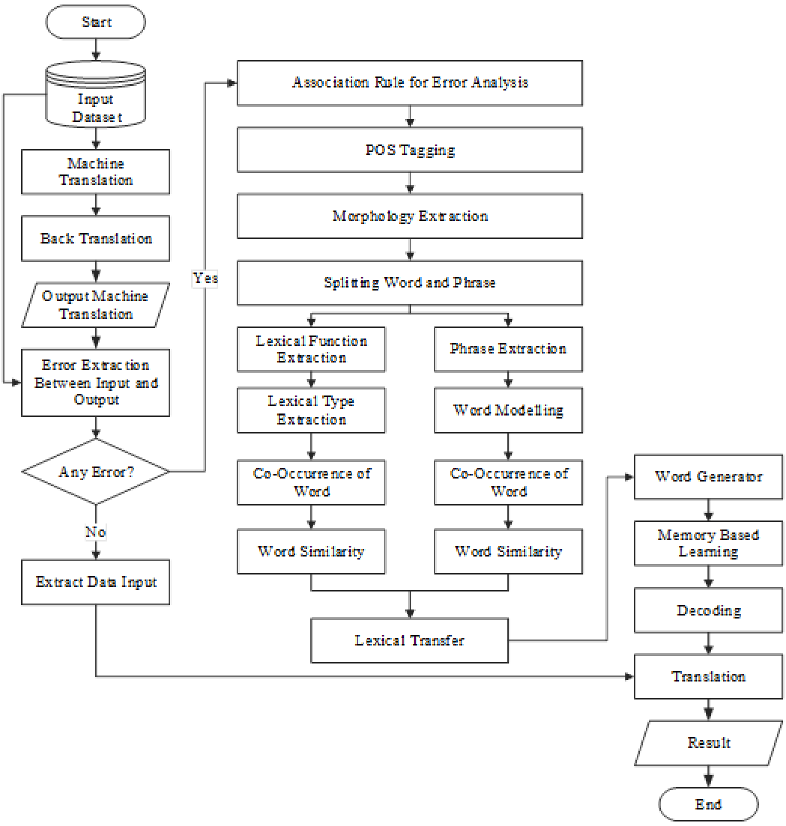

Figure 1. Proposed Hybrid Machine Translation Statistical and Rule Based Approach

The dataset used in this research is crucial because it serves as the starting point for the entire research procedure. This study focuses on previously worked-on Indonesian datasets to provide updated contributions for unaddressed issues. Additionally, a manually compiled Tolaki language dataset was also used. The representations used in this study are shown in Table 2. The dataset creation process involves collecting and annotating the data. The Tolaki research data will be processed into a corpus of parallel texts in Tolaki, Indonesian, and English. A total of 4500 pairs of parallel corpus sentences were utilized, consisting of training and test data. The training data has undergone a process to create a pattern model of a sentence with a known translation and the correct Lexical type, with 3,600 Tolaki, Indonesian, and English sentences used for this purpose. The test data, consisting of 900 sentences in Tolaki, Indonesian, and English with unknown translations and Lexical types, were inputted, stored, and then used for system testing/prediction to determine the accuracy of the pattern model at the classification stage.

Table 2. Dataset Example

No.	Tolaki Languanges	Indonesian Languanges
1.	Ibio pe’eka kalasi limo	Ibio naik kelas lima
2.	Pe’eka *kupenasa’i* mokongango	Naik terasa melelahkan
3.	Oli gola pe’eka	Harga gula naik
4.	Inaku *lumako* pe’eka	Saya berjalan naik
5.	Ku *penasa’i* pe’eka mokongango	Saya merasakan naik melelahkan

In Figure 1 depicts the initial phase of the lexical extraction process for Hybrid MT translation. In most other MT methods, text extraction methods are used to determine the type of word or phrase from the input to be translated, based on the steps of each MT. The first MT method used as the basis for analysis in this proposed method is Neuro Machine Translation (NMT), where general, detailed, and specific extraction stages have been arranged based on syntax and semantics to determine the type and phrase of words from an input to be translated. However, this NMT method cannot determine whether the translation results obtained are extremely accurate. Then, the objective of the text extraction phase is to obtain retranslation data from NMT as a correction of translation errors based on rules made against existing cases of documents that have been used. We have identified an association rule to determine the function of a word, which consists of three rules. Rule 1 applies if the first word of a sentence is a noun phrase (NP), while rule 2 applies if it begins with an adjective phrase (AP). Rule 3 is used when the first word of a sentence is an auxiliary (AUX). The next step in the text extraction process involves POS tagging, which is a crucial step in natural language processing (NLP). We used FLAIR [33], which is one of the state-of-the-art NLP libraries in language processing, for the POS tagging process. The FLAIR tools generated tags for each word that contained elements of Noun Phrase (NP) and Adjective Phrase (AP), and the results are shown in table 3. In addition, we used the MorphInd concept for the extraction of Indonesian morphology. The morphology extraction algorithm used in this study can be described as follows:

```
Algorithm:
Input: Result of POS tagging
Output: Result of MorphTool
    Separate each term in the list.
    Extract the affixed lexical.
    Conduct a standard analysis of the words.
    Obtain the outcome of MorphTool.
End.
```

This study utilizes a morphology extraction algorithm that takes POS tagging as its input. To improve text extraction accuracy, the algorithm counts Lexical lists, Lexical functions, and Lexical types in labeled documents using TF-IDF. The tokenization process produces results, and Word2vec performs vector calculation on each token to convert text feature results into vector values. The training and testing data are handled by the gensim Python library. By considering the number of Lexical forms in the document, the vector with the highest value is used to obtain the BERT embedding input for the actual target token. The expansion of the document in terms of noun type generated by Word2vec in syntactic extraction is followed by the matching of Lexical similarities to words in the sentence. To improve the accuracy of extracting target words, BERT embedding is employed in the subsequent process.

Table 3. POS tagging result

ID	Indonesian Languages	Tolaki Languages	POS tagging result
1	Ibio naik kelas lima	Ibio pe’eka kalasi limo	Ibio <PRON> naik <VERB> kelas lima<NOUN>
2	Naik terasa melelahkan	Pe’eka *kupenasa’i* mokongango	Naik <PROPN> terasa <VERB> melelahkan <ADJ>
3	Harga gula naik	Oli gola pe’eka	Harga <NOUN> gula <NOUN> naik <ADJ>
4	Saya berjalan naik	Inaku *lumako* pe’eka	Saya <PRON> berjalan <VERB> naik <ADV>
5	Saya merasakan naik melelahkan	Ku *penasa’i* pe’eka mokongango	Saya <PRON> merasakan <VERB> naik <NOUN> melelahkan <ADJ>

These steps are interrelated and cannot be changed in the order of the process. This is because, to determine a type of word worth true or false to the word itself, it is necessary to know in advance the form of the word, whether it is a root word or not. In the next step is Split Word and Phrase. Firstly, Split Word starts with the extraction of Lexical Function to determine the Lexical Function in a sentence. The function of the subject, predicate, object, complement adverb, and complement adjunct is determined by the word order in a sentence, which serves as the input for this process. In Tolaki Lexical Functions, the position of the word in the sentence influences its function, which includes subject, predicate, object, and complement. The next step is the Extraction of Lexical type, which determines the type of words in each sentence. This process takes the output of the morphology extraction procedure as input and uses 51 parent and child node rules. The three main nodes of the sentence structure are NP, VP, and AUX, while the 18 types of child nodes include ADJ, ADP, ADV, AUX, CCONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROPN, PUNCT, SCONJ, SYM, VERB, and X. A set of 51 rules are established based on these nodes and child nodes, which determine the acceptable word tag relations within a sentence. If the word tag relation is incorrect, the system can identify and correct it based on these rules. The function of this word must be analyzed to anticipate a word with multiple Lexical Functions, thereby preventing an error in determining the type of word the word itself is. Phrase extraction also extends word acquisition to phrases. Throughout this extension, phrase recognition is a top priority. Certain hypotheses are designed to select candidate phrases, or all word sequences are candidates. Consider phrase extraction a task that necessitates supervised learning. In this step, the words selected in the previous phase are combined into multiword keywords if they occur in the text together. The score of newly created keywords is equal to the sum of the scores of the individual words that are used to create word models. Then, we primarily quantify semantic relationships between words based on their co-occurrence. The distributional hypothesis suggests that words with similar meanings will appear in similar contexts and co-occur with the same other words [34]. As an alternative to assessing semantic similarity based on the immediate co-occurrence of two terms, we propose comparing their co-occurrences with all other terms. To achieve this, we define the co-occurrence distribution of each word as the weighted average of the word distributions of all documents containing the word. We use similarity measures for the co-occurrence distributions of two terms to quantify their "semantic similarity" [35]. One can also compare the cooccurrence distribution of a word to the distribution of words in a text. This provides a metric for determining the frequency of a given word in a text. Afterward, lexical transfer, A forward transfer is, logically a transfer from L1 (Tolaki) to L2 (Indonesia) or L2 (Indonesia) to L3 (English), and the reverse transfer is a transfer from L3 to L2 or L2 to L1. Pre-processing aims to reduce the complexity of a text to translate the text into actual syntactic analysis. However, it cannot produce identification relating to the problem significantly. It is likely that grammar and spelling are incorrect, caused because the texts are derived or edited from humans with varying language skills whereas the existing solution methods will only work on perfect text i.e. sentence text with completely correct grammar and spelling. For syntactic and semantic extraction of sentence translation processes with cases: simple, complex, compound, and complex compound sentences.

Table 4. Comparation of word translation result

No	Word analysis			
	Input	Output	Input	Output
	Indonesian-Tolaki	*English*	*English*	*Indonesian-Tolaki*
1	Ibio naik kelas lima (Ibio pe'eka kalasi limo)	Ibio going to class five	Ibio going to class five	Ibio pergi ke kelas lima
2	Naik *terasa* melelahkan (Pe'eka *kupenasa'i* mokongango)	Riding feels tiring	Riding feels tiring	Berkendara terasa melelahkan
3	Harga gula naik (Oli gola pe'eka)	Sugar prices rise	Sugar prices rise	Harga gula naik
4	Saya *berjalan* naik (Inaku *lumako* pe'eka)	I walked up	I walked up	aku berjalan ke atas
5	Saya *merasakan* naik melelahkan (Ku *penasa'i* pe'eka mokongango)	I feel the ride is tiring	I feel the ride is tiring	Saya merasa perjalanan ini melelahkan
6	Saya menaikkan bendera *tinggi sekali* (Inaku pe'ekatingge bandera *me'ita dahu*)	I raised the flag very high	I raised the flag very high	Saya mengibarkan bendera sangat tinggi
7	Saya menaiki tangga *susah sekali* (Inaku pe'ekari'i la'usa *masusa dahu*)	I climbed the stairs very hard	I climbed the stairs very hard	Saya menaiki tangga dengan sangat keras
8	Kenaikan harga gula *disiarkan di televisi* (*Nope'eka oli gola bawo I televisi*)	Rising sugar prices broadcast on television	Rising sugar prices broadcast on television	Kenaikan harga gula disiarkan di televisi
9	Kenaikan harga gula akan menaikkan harga sembako (*Nope'eka oli gola nggo pe'eka itoono oli sombako*)	An increase in sugar prices will increase the price of basic necessities	An increase in sugar prices will increase the price of basic necessities	Kenaikan harga gula akan menaikkan harga kebutuhan pokok

Word Generator module generates text in the target language based on its structure. It gets into the transfer of lexical verbs, auxiliary verbs for tense, aspect, and mood, and information about gender, number, and person. In terms of resolving syntactic and lexical ambiguities, this method is superior to direct translation. Moreover, Memory Based Learning has been successfully applied to the related problem of word sense disambiguation [36-46]. In this study, we trained classifiers using memory-based learning. Memory-based classifiers prevent overgeneralization by storing all training examples as feature vectors in memory without removing exceptional instances. At runtime, a new instance is compared to the saved instances and classified based on the closest match (nearest neighbors). To decipher the sentences is assumed $x = \{x_1, ...x_n\}$ NMT translates the source sentence into the corresponding target sentence $y = \{y_1, ...y_m\}$ utilizing a trained NMT model. In practice, this transform decoding into a searching problem, for which a beam searcher is used to find the target sentence with the highest generation probability. A typical NMT model generates in an auto-regressive manner. Therefore, the generation of each token depends on the source sentence and the prefix of the target sentence that has been generated. The entire generation of the target sentence can be expressed as a conditional $P(y \mid x)$ as described in Equation 1 below:

$$P(y \mid x) = \prod_{i=0}^{m} P(y_i \mid x, y < i) \tag{1}$$

Where $y < i = \{y_1, y_2, ..., y_i - 1\}$ represents the prefix tokens generated for target sentences at time-step i. Based on the results of text extraction, in the translation engine task for the process of generating text translation error correction results, the Hybrid SMT-RBMT method is used. The purpose of this proposed MT method is so that the data obtained for correction of translation errors, for single words and phrases, can be stored to update existing entities. Moreover, this proposed method is also used to build a new model based on these results as a stage of MT training to obtain more accurate translation results.

3. Experimental results

In this study, the outcomes of the experimental work conducted are discussed, text extraction outcomes, Classification result and machine translation.

3.1. Result of text extractions

The sample results of syntactic case extraction from nine Indonesian-Tolaki sentences are shown in Table 5 The word "naik" serves as a verb-type predicate in the first clause, while in the second clause, it functions as a noun-type subject. The word "naik" is used as an adjective complement in the third clause. In the fourth clause, the word "naik" functions as a complement to an adverb. The word "naik" is used as an object of the noun type in clause 5. While sentences 6 through 9 contain affixes and suffixes, "naik" is an example of morphonemics. They serve as verb-typed predicates. The proposed method can correctly extract cases of syntactic sentences based on functions and Lexical types using these nine example sentences.

Table 5. Analysis of function and type of words

No	Sentences	Extraction Results	
		Word function	Word type
	Indonesian-Tolaki		
1	Ibio naik kelas lima (Ibio pe’eka kalasi limo)	Subject Predicate Object	Noun Verb Noun
2	Naik *terasa* melelahkan (Pe’eka *kupenasa’i* mokongango)	Subject *Predicate* Complement Adverbial	Noun *Verb* Adverb
3	Harga gula naik (Oli gola pe’eka)	Subject Complement Adjunct	Noun Adjective
4	Saya *berjalan* naik (Inaku *lumako* pe’eka)	Subject *Predicate* Complement Adverbial	Noun *Verb* Adverb
5	Saya *merasakan* naik melelahkan (Ku *penasa’i* pe’eka mokongango)	Subject *Predicate* Object Complement Adjunct	Noun *Verb* Noun Adjective
6	Saya menaikkan bendera *tinggi sekali* (Inaku pe’ekatingge bandera *me’ita dahu*)	Subject Predicate Object *Complement Adjunct*	Noun Verb Noun *Adjective*
7	Saya menaiki tangga *susah sekali* (Inaku pe’ekari’i la’usa *masusa dahu*)	Subject Predicate Object *Complement Adjunct*	Noun Verb Noun *Adjective*
8	Kenaikan harga gula *disiarkan di televisi* (*Nope’eka oli gola bawo i televisi*)	Subject Predicate *Complement Adverbial*	Noun Verb Adverb

9	Kenaikan harga gula akan menaikkan harga sembako (*Nope'eka oli gola nggo pe'eka itoono oli sombako*)	Subject Predicate Object *Complement Adjunct*	Noun Verb Noun

The following example identifies the word "naik" for morphonemic case extraction, as shown in Table 6 below:

Table 6 Morphonemic cases extraction analysis result

No	Indonesian	Morphonemic	Tolaki	Morphonemic	Label
1	naik	-	pe'eka	-	*Verb*
2	menaikkan	me##, ##naik##, ##kan	pe'ekanggee	nggee	*Verb*
3	menaiki	me##, ##naik##, ##kan	pe'ekari'i	ri'i	*Verb*
4	kenaikan	ke##, ##naik##, ##kan	pe'ekano	no	*Noun*

3.2. Classification result

Table 9 compares the results of one-way and reverse translations displays the words that have been marked as incorrect translations because their meaning differs from the original input sentences. As a result, an analysis based on the word probabilities of the documents used to obtain more precise results for word meaning was conducted. Based on the function, type, and meaning of the word in the sentence, the word class influences accurate translation results. Table 7 and table 8 the proposed method for classifying words was evaluated using TF-IDF, Word2vec, and BERT embeddings, and the results were positive. TF-IDF is able to extract terms from each word target. Word2Vec then computes the vector value of each term that has been extracted. Finally, BERT embedding calculates the target term's similarity to the document's entire word form. As the actual term for the analysis of Lexical types and functions, the term with the highest similarity value is used.

Table 7. TF-IDF and Word2vec for SMT analysis

Sentence[1]	Ibio **going** to class five
Terms	Going: [('goes', 0.663), ('coming', 0.657), ('went', 0.635), ('gone', 0.632), ('heading', 0.630), ('trying', 0.617), ('moving', 0.594), ('go', 0.582), ('wanting', 0.567), ('slipping', 0.567)] Class five: [('classes', 0.603), ('grade', 0.581), ('batch', 0.510), ('kaichu', 0.494), ('subclass', 0.485), ('classman', 0.471), ('moudge', 0.467), ('grades', 0.453), ('viiis', 0.444), ('quartile', 0.444)]
Sentence[2]	**Riding** feels very tiring
Terms	riding: [('sidesaddle', 0.583), ('broomhaugh', 0.565), ('mameah', 0.555), ('pillion', 0.538), ('bareback', 0.532), ('galloping', 0.526), ('unicycles', 0.524), ('equitation', 0.519), ('rode', 0.507), ('prancing', 0.506)] feels: [('thinks', 0.765), ('feeling', 0.744), ('isn', 0.722), ('feel', 0.713), ('looks', 0.713), ('understands', 0.704), ('felt', 0.694), ('knows', 0.677), ('realizes', 0.669), ('admits', 0.665)] very tiring: [('fatigued', 0.686), ('tedious', 0.666), ('frustrating', 0.656), ('exhausting', 0.651), ('tiresome', 0.644), ('wearying', 0.625), ('strenuous', 0.621), ('hectic', 0.615), ('monotonous', 0.609), ('grueling', 0.598)]
Sentence[3]	Sugar prices **rise**
Terms	sugar: [('petroleum', 0.774), ('gas', 0.688), ('colza', 0.642), ('sugarfield', 0.637), ('refinery', 0.634), ('coal', 0.630), ('hydrocarbon', 0.610), ('canvasboard', 0.604), ('arpechim', 0.599), ('neatsfoot', 0.595)] prices: [('price', 0.775), ('inflation', 0.722), ('rates', 0.702), ('demand', 0.699), ('tariffs', 0.699), ('costs', 0.669), ('stocks', 0.652), ('premiums', 0.651), ('wages', 0.643), ('pricing', 0.631)] rise: [('rising', 0.584), ('decline', 0.553), ('surge', 0.519), ('emergence', 0.501), ('collapse', 0.493), ('resurgence', 0.487), ('rises', 0.482), ('growth', 0.478), ('fall', 0.477), ('flourish', 0.474)]
…	**…**

Table 8. BERT and cosine for SMT analysis

Sent[1]	I'm **going** to class five
Terms similarity	[(‘going: class five’, 0.9046)] [('goes: class', 0.8986), ('coming: class', 0.9070), ('went: class', 0.9011), ('gone: class', 0.8952), ('heading: class', 0.9008), ('trying: class', 0.9108), ('moving: class', 9115), ('go: class', 0.8821), ('wanting: class', 0.8904), ('slipping: class', 0.8997)] [('goes: classes', 0.9265), ('coming: classes', 0.9422), ('went: classes', 0.9338), ('gone: classes', 0.9386), ('heading: classes', 0.9296), ('trying: classes', 0.9451), ('moving: classes', 0.9435), ('go: classes', 0.8983), ('wanting: classes', 0.9224), ('slipping: classes', 0.9256)] [('goes: grade', 0.9136), ('coming: grade', 0.9115), ('went: grade', 0.9064), ('gone: grade', 0.8986), ('heading: grade', 0.8986), ('trying: grade', 0.9148), ('moving: grade', 0.9150), ('go: grade', 0.8989), ('wanting: grade', 0.9032), ('slipping: grade', 0.9097)] [('goes: batch', 0.9008), ('coming: batch', 0.8987), ('went: batch', 0.8939), ('gone: batch', 0.8798), ('heading: batch', 0.8999), ('trying: batch', 0.9062), ('moving: batch', 0.9006), ('go: batch', 0.8952), ('wanting: batch', 0.8947), ('slipping: batch', 0.9133)] [('goes: kaichu', 0.4176), ('coming: kaichu', 0.3681), ('went: kaichu', 0.3723), ('gone: kaichu', 0.3293), ('heading: kaichu', 0.3949), ('trying: kaichu', 0.3877), ('moving: kaichu', 0.3898), ('go: kaichu', 0.4799), ('wanting: kaichu', 0.4160), ('slipping: kaichu', 0.4696)] [('goes: subclass', 0.4410), ('coming: subclass', 0.3652), ('went: subclass', 0.3774), ('gone: subclass', 0.3310), ('heading: subclass', 0.4411), ('trying: subclass', 0.3826), ('moving: subclass', 0.3978), ('go: subclass', 0.4838), ('wanting: subclass', 0.4130), ('slipping: subclass', 0.4745)] [('goes: classman', 0.8952), ('coming: classman', 0.9359), ('went: classman', 0.9182), ('gone: classman', 0.9396), ('heading: classman', 0.8996), ('trying: classman', 0.9288), ('moving: classman', 0.9384), ('go: classman', 0.8555), ('wanting: classman', 0.9104), ('slipping: classman', 0.8831)] [('goes: moudge', 0.5360), ('coming: moudge', 0.4577), ('went: moudge', 0.4792), ('gone: moudge', 0.4288), ('heading: moudge', 0.5161), ('trying: moudge', 0.4866), ('moving: moudge', 0.4871), ('go: moudge', 0.6022), ('wanting: moudge', 0.5078), ('slipping: moudge', 0.5724)] [('goes: grades', 0.8782), ('coming: grades', 0.8644), ('went: grades', 0.8630), ('gone: grades', 0.8499), ('heading: grades', 0.8883), ('trying: grades', 0.8831), ('moving: grades', 0.8754), ('go: grades', 0.8706), ('wanting: grades', 0.8741), ('slipping: grades', 0.9096)] [('goes: viiis', 0.4623), ('coming: viiis', 0.3799), ('went: viiis', 0.4091), ('gone: viiis', 0.3494), ('heading: viiis', 0.4426), ('trying: viiis', 0.4099), ('moving: viiis', 0.4071), ('go: viiis', 0.4910), ('wanting: viiis', 0.4286), ('slipping: viiis', 0.4638)] [('goes: quartile', 0.4365), ('coming: quartile', 0.3703), ('went: quartile', 0.3880), ('gone: quartile', 0.3517), ('heading: quartile', 0.4292), ('trying: quartile', 0.4128), ('moving: quartile', 0.3973), ('go: quartile', 0.4794), ('wanting: quartile', 0.4429), ('slipping: quartile', 0.4668)]
Sent[2]	**Riding** feels very tiring
Terms similarity	[(‘riding: feels very tiring’, 0.7490)] [('sidesaddle: feels tiring', 0.5350), ('broomhaugh: feels tiring', 0.6452), ('mameah: feels tiring', 0.5886), ('pillion: feels tiring', 0.6211), ('bareback: feels tiring', 0.6700), ('galloping: feels tiring', 0.6048), ('unicycles: feels tiring', 0.6491), ('equitation: feels tiring', 0.6309), ('rode: feels tiring', 0.7557), ('prancing: feels tiring', 0.6466)] [('sidesaddle: thinks tiring', 0.4783), ('broomhaugh: thinks tiring', 0.6006), ('mameah: thinks tiring', 0.5547), ('pillion: thinks tiring', 0.6205), ('bareback: thinks tiring', 0.6712), ('galloping: thinks tiring', 0.5788), ('unicycles: thinks tiring', 0.6043), ('equitation: thinks tiring', 0.5919), ('rode: thinks tiring', 0.7430), ('prancing: thinks tiring', 0.6196)] [('sidesaddle: feeling tiring', 0.4949), ('broomhaugh: feeling tiring', 0.6339), ('mameah: feeling tiring', 0.5892), ('pillion: feeling tiring', 0.6970), ('bareback: feeling tiring', 0.7223),

Sent[1]	I'm **going** to class five
	('galloping: feeling tiring', 0.6115), ('unicycles: feeling tiring', 0.6644), ('equitation: feeling tiring', 0.6292), ('rode: feeling tiring', 0.7512), ('prancing: feeling tiring', 0.6518)] [('sidesaddle: isn tiring', 0.3837), ('broomhaugh: isn tiring', 0.5295), ('mameah: isn tiring', 0.4962), ('pillion: isn tiring', 0.7566), ('bareback: isn tiring', 0.7019), ('galloping: isn tiring', 0.4694), ('unicycles: isn tiring', 0.5678), ('equitation: isn tiring', 0.5253), ('rode: isn tiring', 0.6768), ('prancing: isn tiring', 0.5270)] [('sidesaddle: feel tiring', 0.5110), ('broomhaugh: feel tiring', 0.6392), ('mameah: feel tiring', 0.5929), ('pillion: feel tiring', 6314), ('bareback: feel tiring', 0.6762), ('galloping: feel tiring', 0.5959), ('unicycles: feel tiring', 0.6524), ('equitation: feels tiring', 0.6341), ('rode: feel tiring', 0.7534), ('prancing: feel tiring', 0.6529)] [('sidesaddle: looks tiring', 0.5210), ('broomhaugh: looks tiring', 0.6451), ('mameah: looks tiring', 0.5917), ('pillion: looks tiring', 0.6093), ('bareback: looks tiring', 0.6706), ('galloping: looks tiring', 0.6130), ('unicycles: looks tiring', 0.6464), ('equitation: looks tiring', 0.6199), ('rode: looks tiring', 0.7442), ('prancing: looks tiring', 0.6447)] [('sidesaddle: understands tiring', 0.5138), ('broomhaugh: understands tiring', 0.6383), ('mameah: understands tiring', 0.6052), ('pillion: understands tiring', 0.6326), ('bareback: understands tiring', 0.7006), ('galloping: understands tiring', 0.6124), ('unicycles: understands tiring', 0.6461), ('equitation: understands tiring', 0.6262), ('rode: understands tiring', 0.7642), ('prancing: understands tiring', 0.6585)] [('sidesaddle: felt tiring', 0.5436), ('broomhaugh: felt tiring', 0.6655), ('mameah: felt tiring', 0.6078), ('pillion: felt tiring', 0.5151), ('bareback: felt tiring', 0.6078), ('galloping: felt tiring', 0.6199), ('unicycles: felt tiring', 0.6534), ('equitation: felt tiring', 0.6437), ('rode: felt tiring', 0.7345), ('prancing: felt tiring', 0.6597)] [('sidesaddle: knows tiring', 0.5143), ('broomhaugh: knows tiring', 0.6683), ('mameah: knows tiring', 0.6361), ('pillion: knows tiring', 0.6445), ('bareback: knows tiring', 0.7124), ('galloping: knows tiring', 0.6289), ('unicycles: knows tiring', 0.6715), ('equitation: knows tiring', 0.6508), ('rode: knows tiring', 0.7839), ('prancing: knows tiring', 0.6784)] [('sidesaddle: realizes tiring', 0.5277), ('broomhaugh: realizes tiring', 0.6558), ('mameah: realizes tiring', 0.6027), ('pillion: realizes tiring', 0.6056), ('bareback: realizes tiring', 0.6766), ('galloping: realizes tiring', 0.6326), ('unicycles: realizes tiring', 6504), ('equitation: realizes tiring', 0.6432), ('rode: realizes tiring', 0.7886), ('prancing: realizes tiring', 0.6693)] [('sidesaddle: admits tiring', 0.4893), ('broomhaugh: admits tiring', 0.6248), ('mameah: admits tiring', 0.5777), ('pillion: admits tiring', 0.7077), ('bareback: admits tiring', 0.7386), ('galloping: admits tiring', 0.6243), ('unicycles: admits tiring', 0.6540), ('equitation: admits tiring', 0.6209), ('rode: admits tiring', 0.7719), ('prancing: admits tiring', 0.6516)]
Sent[3]	Sugar prices **rise**
Terms similarity	[('Sugar prices: rise'), 0.7580]
…	**…**

Table 9 Comparation of word translation result

Word analysis				
No	Input	Output	Input	Output
	Indonesian-Tolaki			*English*
1	Ibio naik kelas lima (Ibio pe'eka kalasi limo)	Ibio going to class five	Ibio going to class five	Ibio pergi ke kelas lima
2	Naik *terasa* melelahkan (Pe'eka *kupenasa'i* mokongango)	Riding feels tiring	Riding feels tiring	Berkendara terasa melelahkan
3	Harga gula naik	Sugar prices rise	Sugar prices rise	Harga gula naik

Word analysis				
No	Input	Output	Input	Output
	(Oli gola pe’eka)			
4	Saya *berjalan* naik (Inaku *lumako* pe’eka)	I walked up	I walked up	aku berjalan ke atas
5	Saya *merasakan* naik melelahkan (Ku *penasa’i* pe’eka mokongango)	I feel the ride is tiring	I feel the ride is tiring	Saya merasa perjalanan ini melelahkan
6	Saya menaikkan bendera *tinggi sekali* (Inaku pe’ekatingge bandera *me’ita dahu*)	I raised the flag very high	I raised the flag very high	Saya mengibarkan bendera sangat tinggi
7	Saya menaiki tangga *susah sekali* (Inaku pe’ekari’i la’usa *masusa dahu*)	I climbed the stairs very hard	I climbed the stairs very hard	Saya menaiki tangga dengan sangat keras
8	Kenaikan harga gula *disiarkan di televisi* (*Nope’eka oli gola bawo I televisi*)	Rising sugar prices broadcast on television	Rising sugar prices broadcast on television	Kenaikan harga gula disiarkan di televisi
9	Kenaikan harga gula akan menaikkan harga sembako (*Nope’eka oli gola nggo pe’eka itoono oli sombako*)	An increase in sugar prices will increase the price of basic necessities	An increase in sugar prices will increase the price of basic necessities	Kenaikan harga gula akan menaikkan harga kebutuhan pokok

Table 10 displays the rules suggested in this research that employ POS tagging results of Indonesian language to assess the completeness of word structure in sentences. A sentence is considered grammatically correct if it has, at a minimum, a subject (NOUN) and a predicate (VERB). The translation output is subsequently compared to determine the likelihood of word similarity between the outcomes obtained from statistical analysis and rule-based methods, with the translation outcome having the highest probability being chosen. These rules are suitable for use in the following two situations:

Table 10. Proposed rule implementation

Indonesian to English				English to Indonesian					
ibio	naik	Kelas lima		ibio		going	to	Class five	
PRON	VERB	NOUN							
ibio	*going to*	*Class five*		*Ibio*	*pergi*		*ke*	*Kelas lima*	
S:NP	Hidden topic: Saya (PRON) → naik (VERB) → kelas lima (NOUN)								
	go to fifth grade			naik kelas lima					
	naik kelas lima			go to fifth grade					
Result				ibio		go	to	fifth	grade
				ibio		*naik*	*ke*	*lima*	*kelas*
naik	terasa	Melelah kan		riding	feels	tiring			

PROPN	VERB	ADJ							
riding	*feels*	*tiring*		*berkendara*	*terasa*	*Melelahkan*			
S:NP	Hidden topic: naik (VERB) → kenaikan (NOUN) kenaikan (NOUN) → terasa (VERB) → melelahkan (ADJ)								
	hike is tiring			mendaki itu melelahkan					
	kenaikan terasa melelahkan			hike is tiring					
Result				*mendaki*	*terasa*	*Melelahkan*			
				hike	feels	tiring			
harga	gula	naik		sugar	prices	rise			
NOUN	NOUN	ADJ							
sugar prices		*rise*		*harga gula*		*naik*			
S:NP	Hidden topic: - (VERB) → adalah (VERB) harga (NOUN) → gula (NOUN) → adalah (VERB) → naik (ADJ)								
	sugar prices are going up			harga gula naik					
	harga gula adalah naik			sugar prices are going up					
Result				sugar	prices	are	going	up	
				harga gula		*naik*			
saya	berjalan	naik		i	walked	up			
PRON	VERB	ADV							
i	*walked*	*up*		*saya*	*berjalan*	*ke atas*			
S:NP	Hidden topic: saya (PRON) → berjalan (VERB) → naik (ADV)								
	berjalan naik			walk up					
	walk up			berjalan ke atas					
Result				i	walk	up			
				saya	berjalan	ke atas			
saya	merasakan	naik	melelahkan	i	feel	the	ride	is	tiring
PRON	VERB	NOUN	ADJ						
i	*feel*	*the ride*	*is tiring*	*saya*	*merasa*	*perjalanan ini*		*melelahkan*	
S:NP	Hidden topic: Naik (VERB) → kenaikan (NOUN) - (VERB) → adalah (VERB) Kenaikan (NOUN) → adalah (VERB) → melelahkan (ADJ)								
	hike is tiring			mendaki itu melelahkan					
	kenaikan adalah melelahkan			hike is tiring					
Result				i	feel	the	hike	is	tiring
				saya	*merasa*	*pendakian ini*		*melelahkan*	

The analysis must be done when there is a significant difference in translation on both directions based on word error position. We provided a set of rules or guidelines for word updating in case of translation errors in a sentence that we determined as follows:

i. In case of an error in translating the subject noun phrase, update the word by finding similarity with the noun phrase and verb phrase.
ii. In case of an error in translating the predicate verb, update the word based on the result of hidden word translation between the predicate verb and the object noun, complement, or both, from the existing sentence structure.
iii. In case of an error in translating the object noun, update the word by finding word similarity in the object noun form obtained from all available documents.
iv. In case of an error in translating the complement of an adverb or adjunct, update the word based on the word translation between the predicate verb, object noun, and the complement of the adverb or adjective.

If a sentence has an incomplete word structure where a verb is missing after the subject noun phrase or object noun, the verb 'to be' will be automatically added after the subject noun phrase or object noun.

Based on the given example, the sentence "Ibio naik kelas lima" in Indonesian-Tolaki translates to "I am going to fifth grade" in English, which has been identified as a translation error due to the incorrect usage of the verb "going" in this context. The proposed rule for identifying and updating translation errors has been applied to suggest a more accurate translation of the predicate verb "naik" when paired with the noun "kelas" to "promoted to next grade". Therefore, the updated translation of the sentence is "Ibio is promoted to the next grade". Other given example, the sentence "Naik terasa melelahkan" in Indonesian-Tolaki translates to "Riding feels tiring" in English, which has been identified as a translation error due to the incorrect usage of the word "riding" in this context. The proposed rule for identifying and updating translation errors has been applied to suggest a more accurate expansion of the subject form of the word "naik" with the noun type from the existing corpus to "kenaikan". Therefore, the updated translation of the sentence is "Hiking feels tiring". The actual sentence in Indonesian is changed to "Kenaikan terasa melelahkan". Based on the third given example, the sentence "Harga gula naik" in Indonesian-Tolaki translates to "Sugar prices go up" in English, which has been identified as a wrong sentence due to the absence of a predicate verb. The proposed rule for adding the verb "to be" after the subject noun phrase has been applied to suggest a more complete sentence structure. Therefore, the updated translation of the sentence is "Harga gula adalah naik" which translates to "Sugar prices are going up" in English. Based on the fourth given example, the sentence "saya berjalan naik" in Indonesian-Tolaki translates to "I walked up" in English, which has been identified as a word translation error as the original sentence does not state a form of past tense. The proposed rule for updating the word based on the complement of the adverb or adjective has been applied to suggest a more accurate translation of the sentence. Therefore, the updated translation of the sentence is "I walk up" in English, which translates to "saya berjalan ke atas" in Indonesian-Tolaki. Based on the firth given example, the result of identification based on the proposed rules in this study is that the word "naik" is an incomplete object NOUN. Hence, the term "naik" with the NOUN type is expanded to "kenaikan" for object form based on the available corpus. Furthermore, the sentence structure is deemed incomplete as it lacks a predicate VERB. Thus, the addition of the VERB "adalah" after the object NOUN expansion "kenaikan" is necessary to complete the sentence structure. The original Indonesian sentence is modified to "saya merasakan kenaikan adalah melelahkan". Next, by utilizing the NOUN-VERB-Complement hidden word translation method, the hidden word translation of "kenaikan adalah melelahkan" is determined to be "the increase is tiring". Consequently, the result of the updating process for the sentence is "I feel that the increase is tiring".

3.3. Machine translation result

The Evaluation process to compare the MT approach using SMT, RBMT, and Hybrid MT has also been carried out in this study. In table 11 shows the comparison result of sentences translation with the case: simple sentences, complex, compound, complex compound. As the input, we use Indonesian-Tolaki and English as the output. The results obtained from the proposed Hybrid MT method are still better when compared to SMT and RBMT. To conclude, the results of the MT evaluation process are shown in Figure 2 and 3 below.

Table 11. Comparison results of SMT, RBMT, and Proposed Method

Input (Indonesian/Tolaki)	**Output** (English) **SMT**	**RBMT**	**Hybrid MT**
harga gula mengalami kenaikan tinggi sekali. *oli gola no pe'eka me'ita dahu.*	sugar prices have increased very high.	sugar prices increased very high.	sugar prices have very high increment
harga gula mengalami kenaikan tinggi sekali dan membuat harga sembako juga ikut naik. *oli gola no pe'eka me'ita dahu ronga mowai oli sombako itoono etai pe'eka.*	sugar prices experienced a very high increase and made the prices of basic necessities also increase.	sugar prices increased very high and made price of groceries also went up.	sugar prices have very high increment and make the prices of basic necessities also increase.
harga gula mengalami kenaikan tinggi sekali, jika tidak ada regulasi pemerintah terhadap harga jual gula di pasar. *oli gola no pe'eka me'ita dahu, keno taanionggi atorano odisi ine oli gola pine'oliako idaoa.*	the price of sugar will rise very high, if there is no government regulation on the selling price of sugar in the market.	sugar prices increased very high, if there is no government regulation on the selling price of sugar in the market.	sugar prices have very high increment, if there is no government regulation on the selling price of sugar in the market.
jika tidak ada regulasi pemerintah terhadap harga jual gula di pasar, harga gula akan mengalami kenaikan tinggi sekali dan membuat harga sembako juga ikut naik. *keno taanionggi atorano odisi ine oli gola pine'oliako idaoa, oli gola no pe'eka me'ita dahu ronga mowai gola sombako itoono etai pe'eka.*	if there is no government regulation on the selling price of sugar in the market, the price of sugar will rise very high and make the price of basic necessities also rise.	if there is no government regulation on the selling price of sugar in the market, sugar prices will increased very high and make the price of groceries also go up.	if there is no government regulation on the selling price of sugar in the market, sugar prices will have very high increment and make the prices of basic necessities also increase.

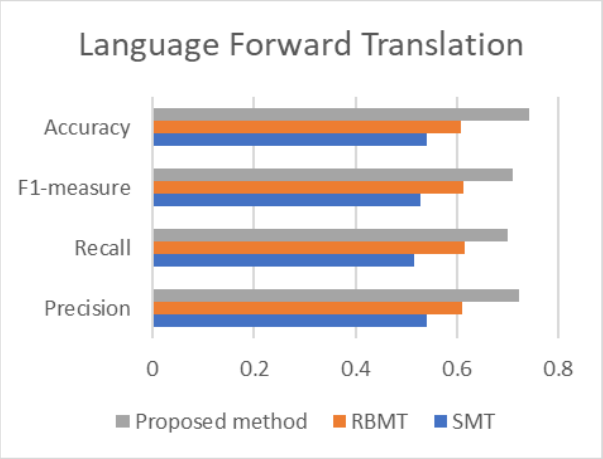

Figure 2. Comparison Result Languange Translation Indonesian Tolaki to English

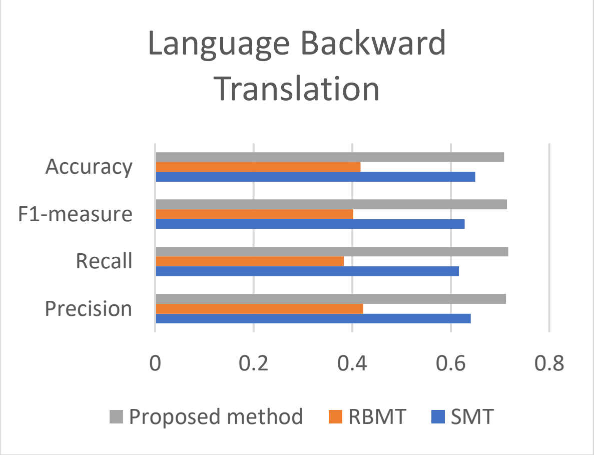

Figure 3. Comparison Result English to Indonesian Tolaki

As an example, the translation of nine Indonesian-Tolaki sentences into English was used. The process of translating Indonesian-Tolaki to English is outlined in Table 12. The result of the English translation is then used as the input for the Indonesian-Tolaki translation. Table 13 displays the outcome of the backward translation procedure.

Table 12. Result for Indonesian Tolaki to English Machine Translation

No	Instance of Sentences	Results		
		Lexical Function	Lexical type	Translation
	Indonesian-Tolaki			*English*
1	Harga gula naik (Oli gola pe'eka)	Subject Complement Adjunct	Noun Adjective	Sugar prices are going up
2	Saya *berjalan* naik (Inaku *lumako* pe'eka)	Subject *Predicate* Complement Adverbial	Noun *Verb* Adverb	I walk up
3	Saya *merasakan* naik melelahkan (Ku *penasa'i* pe'eka mokongango)	Subject *Predicate* Object Complement Adjunct	Noun *Verb* Noun Adjective	I feel the hike is tiring
4	Saya menaikkan bendera *tinggi sekali* (Inaku pe'ekatingge bandera *me'ita dahu*)	Subject Predicate Object *Complement* *Adjunct*	Noun Verb Noun *Adjective*	I raise the flag very high
5	Saya menaiki tangga *susah sekali* (Inaku pe'ekari'i la'usa *masusa dahu*)	Subject Predicate Object *Complement* *Adjunct*	Noun Verb Noun *Adjective*	I climb the stairs very hard
6	Kenaikan harga gula *disiarkan di televisi* (*Nope'eka oli luwi bawo I televisi*)	Subject Predicate *Complement* *Adverbial*	Noun Verb Adverb	Sugar prices increment reported in television
7	Kenaikan harga gula akan menaikkan harga sembako (*Nope'eka oli luwi nggo pe'eka itoono oli sombako*)	Subject Predicate Object *Complement* *Adjunct*	Noun Verb Noun	Sugar prices increment will increase the price of basic necessities

Table 13. Result for English to Indonesian Tolaki Machine Translation

ID Sentence	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	Sugar prices are going up	Harga gula naik *(Oli luwi pe'eka)*
4	I walk up	Aku berjalan ke atas *(inaku lumako ine wawo)*

ID Sentence	Input (English)	Output (Indonesian – Tolaki)
5	I feel the hike is tiring	Saya merasa pendakian ini melelahkan *(Kupenasa'i ponduka'ako'a ni'ino mokongango)*
6	I raise the flag very high	Saya mengibarkan bendera sangat tinggi *(Inaku mondangako bandera me'ita mbu'upu'u)*
7	I climb the stairs very hard	Saya menaiki tangga dengan sangat keras *(Inaku pe'ekari'i la'usa mokora mbu'upu'u)*
8	Sugar prices increment reported in television	Kenaikan harga gula disiarkan di televisi *(Pe'ekano oli luwi nibuangako ine televisi)*
9	Sugar prices increment will increase the price of basic necessities	Kenaikan harga gula akan menaikkan harga kebutuhan pokok *(Pe'ekano oli luwi nggo pe'ekanggee oli pipinaralungi kondu'uma)*

The average accuracy of the proposed method in translating Indonesian-Tolaki to English is 74.17 percent, while the average accuracy of the reverse translation from English to Indonesian-Tolaki is 70.83 percent, as shown in Figure 2 and Figure 3. Despite implementing a text classification process to enhance translation accuracy, the differences in grammatical structures between Indonesian-Tolaki and English have prevented near-perfect accuracy. One-way translation is particularly challenging due to the presence of affixes and word endings in Indonesian-Tolaki, which results in hybrid machine translation errors when translating to English. The proposed word translation analysis has the potential to capture the context of the word more precisely, but the English-to-Indonesian-Tolaki back-translation process must be more effective and accurate in conveying the sentence's actual meaning. For example, the English word "naik" can serve as an adverb or verb, and English has various word forms based on tenses, leading to translation errors despite the absence of a time adverb in the input word. The document's word probability factor, one of the proposed hybrid MT methods for obtaining the target word's translation, contributes to this issue.

4. Conclusions

This study investigates the application of the most recent MT methods in the field of IMT. While previous work on Indonesian MT has focused on statistical and rule-based MT, little attention has been paid to syntactic rules. This study proposes a method that considers the function of words in a sentence, as this can affect the accuracy of translation. The proposed hybrid MT method outperformed both SMT and RBMT in terms of accuracy for English to Indonesian-Tolaki translation (74.17%) and Indonesian-Tolaki to English translation (70.83%). The RBMT method achieved higher accuracy for Indonesian-Tolaki to English translation (60.83%) than for English to Indonesian-Tolaki translation (41.67%). The study's results indicate that the proposed hybrid SMT-RBMT approach can outperform both individual SMT and RBMT methods. However, further research is needed to investigate parallel corpus collection methods and the development of attention-based approaches to enhance the performance of the proposed method. The research-oriented workspace concludes with the following recommendations:

1. Conduct further research on the Indonesian language and its rules to gather new information.
2. Gather new information on regional languages in Indonesia and their governing principles.
3. A new methods and techniques and compare outcomes based on previous work.
4. Extended new tools for Indonesian MT.

5. Explore new or alternative performance metrics for MT research.
6. Improve the translation system's accuracy by addressing various factors.
7. Increase the number of parallel corpora to improve evaluation quality.

Declaration of competing interest

The authors declare that they have no known financial or non-financial competing interests in any material discussed in this paper.

Funding information

This research was funded by Lembaga Pengelola Dana Pendidikan (LPDP) under Riset Inovatif-Produktif (RISPRO) Invitation Program managed, 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).

Acknowledgements

The authors are grateful to the Institut Teknologi Sepuluh Nopember (ITS) for providing all facilities and the authors are very thankful to the reviewer for their valuable feedback and comments to improve the contents of this manuscript.

References

[1] J. Tiedemann, M. Aulamo, S. Hardwick, and T. Nieminen, “Open Translation Models, Tools and Services,” G. Rehm, Ed., in Cognitive Technologies. Cham: Springer International Publishing, 2023, pp. 325–330. doi: 10.1007/978-3-031-17258-8_24.

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

[3] S. D. Larasati, V. Kuboň, and D. Zeman, “Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus,” in *Communications in Computer and Information Science*, 2011, pp. 119–129. doi: 10.1007/978-3-642-23138-4_8.

[4] T. Mantoro, J. Asian, R. Octavian, and M. A. Ayu, “Optimal translation of English to Bahasa Indonesia using statistical machine translation system,” in *2013 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M)*, IEEE, Mar. 2013, pp. 1–4. doi: 10.1109/ICT4M.2013.6518918.

[5] H. Sujaini, “Mesin Penerjemah Situs Berita Online Bahasa Indonesia ke Bahasa Melayu Pontianak,” *J. ELKHA*, vol. 6, no. 2, pp. 38–44, 2014.

[6] M. A. Sulaeman and A. Purwarianti, “Development of Indonesian-Japanese statistical machine translation using lemma translation and additional post-process,” in *2015 International Conference on Electrical Engineering and Informatics (ICEEI)*, IEEE, Aug. 2015, pp. 54–58. doi: 10.1109/ICEEI.2015.7352469.

[7] 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, Sep. 2016, doi: 10.26418/jp.v2i2.16520.

[8] 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,” in *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*, IEEE, Nov. 2015, pp. 1–6. doi: 10.1109/ICITSI.2015.7437678.

[9] F. Rahutomo, R. A. Asmara, and D. K. Purwoko Aji, “Computational Analysis on Rise and Fall of Indonesian Vocabulary During a Period of Time,” in *2018 6th International Conference on Information and Communication Technology (ICoICT)*, IEEE, May 2018, pp. 75–80. doi: 10.1109/ICoICT.2018.8528812.

[10] D. Soyusiawaty, “E-Translator With Rule Based Indonesia-Minang Dan Minang-Indonesia,” *J. Inform.*, vol. 2, no. 2, pp. 234–247, 2008, doi: 10.26555/jifo.v2i2.a5255.

[11] W. Ridwan and R. D. R. Dako, “Bidirectional Indonesian-Gorontalo Text Translator: Rule-Based Approach,” *Int. J. Appl. Eng. Res.*, vol. 10, no. 13, pp. 33847–33852, 2015.

[12] I. P. D. Pratama and A. Muliani, “Perancangan dan Implementasi sistem penerjemahan teks Bahasa Inggris ke Bahasa Bali dengan menggunakan pendekan berbasis aturan (rules based),” *J. Ilmu Komput.*, vol. 5, no. 1, pp. 47–54, 2012, [Online]. Available: https://ojs.unud.ac.id/index.php/jik/article/view/2711

[13] A. Ginting and N. AZ, “Penerjemah Dua Arah Bahasa Indonesia Ke Bahasa Daerah (Karo) Menggunakan Teknik Statistical Machine Translation (Smt) Sebagai Fitur Pada Situs Web Untuk Meningkatkan Web Traffic,” *Telemat. MKOM*, vol. 4, no. 1, pp. 60–72, 2016, [Online]. Available: https://journal.budiluhur.ac.id/index.php/telematika/article/view/158

[14] A. P. Wibawa, A. Nafalski, J. Tweedale, N. Murray, and A. E. Kadarisman, “Hybrid machine translation for Javanese speech levels,” in *2013 5th International Conference on Knowledge and Smart Technology (KST)*, 2013, pp. 64–69. doi: 10.1109/KST.2013.6512789.

[15] 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,” in *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*, 2015, pp. 1–6. doi: 10.1109/ICITSI.2015.7437678.

[16] A. N. Arrasyid and M. S. Said, “Aplikasi Kamus Bahasa Daerah Tolaki Berbasis Android,” *Simtek J. Sist. Inf. dan Tek. Komput.*, vol. 1, no. 1, pp. 62–68, 2016, doi: 10.51876/simtek.v1i1.9.

[17] M. A. Aryano, “Aplikasi Kamus Bahasa Indonesia-Tolaki Pada Smartphone Berbasis Android,” 2016. [Online]. Available: http://repository.unissula.ac.id/id/eprint/5728

[18] M.- Maslan, Y. Setiono, and F. Alfazri, “Pengembangan Smart Application Translation Aneka Bahasa Sulawesi Berbasis Android,” *J. Nas. Teknol. dan Sist. Inf.*, vol. 2, no. 1, pp. 55–64, Apr. 2016, doi: 10.25077/TEKNOSI.v2i1.2016.55-64.

[19] Z. Abidin, “Penerapan Neural Machine Translation untuk Eksperimen Penerjemahan secara Otomatis pada Bahasa Lampung – Indonesia,” *Pros. Semin. Nas. Metod. Kuantitatif*, no. 978, pp. 53–68, 2017.

[20] Z. Abidin, “Translation of Sentence Lampung-Indonesian Languages with Neural Machine Translation Attention Based Approach,” *Inov. Pembang. J. Kelitbangan*, vol. 6, no. 02, pp. 191–206, 2018, doi: 10.35450/jip.v6i02.97.

[21] L. O. Kasema, S. R. Sentinuwo, and A. M. Sambul, “Aplikasi Kamus Bahasa Daerah Pasan Berbasis Android,” *J. Tek. Inform.*, vol. 13, no. 2, pp. 1–6, 2018, doi: 10.35793/jti.13.2.2018.22489.

[22] R. Pebrijayanti and Z. Ardian, “Rancang Bangun Aplikasi Kamusbahasaindonesia - Bahasa Aceh Menggunakan Metode Rule Based Berbasis Android,” *J. Informatics Comput. Sci.*, vol. 4, no. 2, p. 91, 2019, doi: 10.33143/jics.vol4.iss1.534.

[23] D. Savira and Y. Widiastiwi, “Sistem Penerjemah Teks Bahasa Inggris Ke Dalam Bahasa Jawa Krama Dengan Pendekatan Berbasis Aturan (Rule Based),” Seinasi-Kesi, pp. 24–25, 2019.

[24] N. Hikmah, B. T. Cahyo, H. Rianto, and S. Dewi, “Rancang Bangun Pembuatan Program Kamus Plesetan Berbasis Pwa (Progressive Web Application),” J. Sist. Inf., vol. 4, no. 4, pp. 1–8, 2020.

[25] R. Prasetia, “Alih Bahasa Teks Bahasa Melayu Riau Ke Teks Bahasa Indonesia Dengan Pendekatan Berbasis Aturan (Rule Based),” 2021.

[26] Z. Abidin and P. Permata, “Pengaruh Penambahan Korpus Paralel Pada Mesin Penerjemah Statistik Bahasa Indonesia Ke Bahasa Lampung Dialek Nyo,” J. Teknoinfo, vol. 15, no. 1, p. 13, 2021, doi: 10.33365/jti.v15i1.889.

[27] Z. Abidin, Permata, I. Ahmad, and Rusliyawati, “Effect of mono corpus quantity on statistical machine translation Indonesian – Lampung dialect of nyo,” J. Phys. Conf. Ser., vol. 1751, no. 1, p. 012036, Jan. 2021, doi: 10.1088/1742-6596/1751/1/012036.

[28] Z. Abidin, P. Permata, and F. Ariyani, “Translation of the Lampung Language Text Dialect of Nyo into the Indonesian Language with DMT and SMT Approach,” INTENSIF J. Ilm. Penelit. dan Penerapan Teknol. Sist. Inf., vol. 5, no. 1, pp. 58–71, 2021, doi: 10.29407/intensif.v5i1.14670.

[29] A. Schmidt and M. Wiegand, “A Survey on Hate Speech Detection using Natural Language Processing,” in Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 1–10. doi: 10.18653/v1/W17-1101.

[30] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, “Hate speech detection in the Indonesian language: A dataset and preliminary study,” in 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, Oct. 2017, pp. 233–238. doi: 10.1109/ICACSIS.2017.8355039.

[31] N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri, and A. Jamal, “Colloquial Indonesian Lexicon,” in 2018 International Conference on Asian Language Processing (IALP), IEEE, Nov. 2018, pp. 226–229. doi: 10.1109/IALP.2018.8629151.

[32] M. Yamin, R. Sarno, R. Abdullah, and Untung, “Syntaxis-based extraction method with type and function of word detection approach for machine translation of Indonesian-Tolaki and English sentences,” in 2022 International Conference on Information Technology Research and Innovation (ICITRI), IEEE, Nov. 2022, pp. 101–106. doi: 10.1109/ICITRI56423.2022.9970225.

[33] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, “FLAIR: An easy-to-use framework for state-of-the-art NLP,” NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Demonstr. Sess., pp. 54–59, 2019.

[34] C. Wartena, R. Brussee, and W. Slakhorst, “Keyword Extraction Using Word Co-occurrence,” in 2010 Workshops on Database and Expert Systems Applications, IEEE, Aug. 2010, pp. 54–58. doi: 10.1109/DEXA.2010.32.

[35] D. Khotimah and R. Sarno, “Sentiment Analysis of Hotel Aspect Using Probabilistic Latent Semantic Analysis, Word Embedding and LSTM,” Int. J. Intell. Eng. Syst., vol. 12, no. 4, pp. 275–290, Aug. 2019, doi: 10.22266/ijies2019.0831.26.

[36] B. H. Majeed, "Impact of a Proposed Strategy According to Luria's Model in Realistic Thinking and Achievement in Mathematics," International Journal of Emerging Technologies in Learning, vol. 17, no. 24, 2022.

[37] M. S. K. Wahib, Z. A. A. Alamiry, and B. H. Majeed "Digital citizenship for faculty of Iraqi universities," Periodicals of Engineering and Natural Sciences, vol. 11, no. 2, pp. 262-274, 2023.

[38] A. Ghazi et al., "Performance Analysis of ZCC-Optical-CDMA over SMF for Fiber-To-The-Home Access Network," in Journal of Physics: Conference Series, 2020, vol. 1529, no. 2: IOP Publishing, p. 022013.

[39] A. H. M. Alaidi, and F. T. Abed, "Attendance System Design And Implementation Based On Radio Frequency Identification (RFID) And Arduino," Journal of Advanced Research in Dynamical Control Systems, vol. 10, no. SI4, pp. 1342-1347, 2018.

[40] A. Fareed et al., "Comparison of Laguerre-Gaussian, Hermite–Gaussian and linearly polarized modes in SDM over FMF with electrical nonlinear equalizer," in AIP Conference Proceedings, 2020, vol. 2203, no. 1: AIP Publishing LLC, p. 020045.

[41] S. M. Najeeb, and S. M. Ali, "Finding the discriminative frequencies of motor electroencephalography signal using genetic algorithm," Telkomnika vol. 19, no. 1, 2020.

[42] A. Ghazi et al., "Donut Modes in Space Wavelength Division Multiplexing: Multimode Optical Fiber Transmission based on Electrical Feedback Equalizer," in Journal of Physics: Conference Series, 2021, vol. 1755, no. 1: IOP Publishing, p. 012046.

[43] B. K. Mohammed, M. B. Mortatha, A. S. Abdalrada, , "A comprehensive system for detection of flammable and toxic gases using IoT," Periodicals of Engineering and Natural Science, vol. 9, no. 2, pp. 702-711, 2021.

[44] B. H. Majeed, "Computational Thinking (CT) Among University Students," International Journal of Interactive Mobile Technologies, vol. 16, no. 10, 2022.

[45] A. Ghazi et al., "Hybrid WDM and Optical-CDMA over Multi-Mode Fiber Transmission System based on Optical Vortex," in Journal of Physics: Conference Series, 2021, vol. 1755, no. 1: IOP Publishing, p. 012001.

[46] M. Holmqvist, “Memory-based learning of word translation,” in Proceedings of the 16th Nordic Conference of Computational Linguistics, NODALIDA 2007, 2007, pp. 231–234.