2025 Volume: 01 Issue: 07 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication

Page: 01-21

DEVELOPMENT OF STATISTICAL AND RULE-BASED MACHINE TRANSLATION MODELS FOR MULTILINGUAL TRANSLATION BETWEEN INDONESIAN, TOLAKI, AND ENGLISH

***Muh Yamin**

Department of Informatics, Faculty of Enggineering, Halu Oleo University, Kendari, Indonesia.

Article Received: 07 November 2025
Article Revised: 27 November 2025
Published on: 17 December 2025

***Corresponding Author: Muh Yamin**
Department of Informatics, Faculty of Enggineering, Halu Oleo University, Kendari, Indonesia.

DOI: https://doi-doi.org/101555/ijrpa.8256

***ABSTRACT**:*

Accurate Indonesian Machine Translation (MT) requires more than syntactic processing; it must also incorporate morphological and semantic context to ensure meaningful translation output. Lexical resources such as dictionaries are essential for interpreting Indonesian root words and generating contextually appropriate translations. This research utilizes Indonesian and Tolaki lexical datasets to develop a more reliable MT system through enhanced morphological and syntactic analysis. A dedicated morphotool was developed to analyze word morphology, while syntactic rule modeling was applied to determine grammatical roles and categories influencing translation performance. The study integrates supervised and unsupervised learning techniques, including TF-IDF, Word2Vec, BERT, and semantic similarity, to perform classification at word, sentence, and document levels based on Indonesian–Tolaki morphonemic and syntactic patterns. A hybrid translation model combining Statistical Machine Translation (SMT) and Rule-Based Machine Translation (RBMT) was applied for sentence translation. The experimental results show accuracy scores of 0.74 for Indonesian–Tolaki to English translation and 0.71 for English to Indonesian–Tolaki translation. These results confirm that the hybrid MT approach provides better performance than standalone SMT and RBMT models, with an overall average accuracy of about 70%.

***KEYWORDS**: Machine translation, SMT, RBMT, hybrid MT.*

INTRODUCTION

Various methods have been introduced for developing English Machine Translation (MT) systems, which typically translate text by applying English grammatical structures. However, equivalent advancements have not been fully realized in Indonesian MT research. Consequently, there is a need to design MT methods that parallel the progress of English MT while remaining consistent with the linguistic properties of the Indonesian language. In general, MT systems are categorized into two main paradigms: rule-based and corpus-based approaches [1]. Rule-based MT encompasses direct, transfer-based, and interlingua-based translation, while corpus-based MT includes statistical and case-based techniques. This section reviews prior studies to examine the development of Indonesian MT methods and technologies.

Research on Indonesian MT at the semantic word level was conducted by Larasati [2], who developed morphological tools for identifying nouns and foreign lexical terms. However, the study was limited to word-level analysis and did not address sentence or document translation. Mantoro [3] later extended this work by developing a statistical MT system for English-to-Indonesian translation using four weighting components: translation models, language models, distortion (reordering), and word penalties, evaluated using BLEU and NIST scores. Nevertheless, this work did not comprehensively analyze contextual word usage in Indonesian sentences or deeply investigate morphonemic features.

Sujaini [4] explored Indonesian–Pontianak Malay translation using statistical MT, although limited corpus availability constrained translation quality. Another study on Indonesian–Japanese lemma-based translation utilized lemma and POS tagging to address issues such as limited corpus coverage, unknown words, and sentence reordering [5]. However, morphological and contextual linguistic aspects were not thoroughly addressed. Jarob [6] improved previous translation methods by focusing on Indonesian–Dayak Taman translation, particularly affixation and root words, using statistical MT, but sentence-level contextual and morphological analysis remained limited. Suryani [7] examined Sundanese-to-Indonesian translation and highlighted the lack of ready-to-use parallel corpora, leading to translation errors caused by typos and inconsistent corpus writing. Rahutomo [8] further analyzed Indonesian vocabulary usage from a computational perspective and found that 26,887 dictionary lemmas are rarely used in daily online media communication.

Word extraction can be performed using several feature types, including surface features, word generalization, sentiment features, lexical resources, linguistic features, and knowledge-based features [9]. Therefore, an advanced Indonesian word extraction mechanism is required that integrates syntactic extraction based on morphological analysis [2] with semantic extraction capabilities [10], [11]. A key challenge in Indonesian word extraction research is the limited availability of labeled Indonesian corpora that can be directly applied across multiple datasets and domains.

Syntactic ambiguity frequently appears in Indonesian writing and presents challenges for word extraction. For example, the word “naik” (Tolaki: pe’eka; English: go up) may function differently depending on sentence context. In one sentence, it may function as a verb predicate; in another, it may act as a noun subject; and in another, it may function as an adjective or adverbial modifier. These cases demonstrate syntactic variation, where words maintain similar meanings but serve different grammatical roles. This differs from semantic ambiguity, where identical words may have different meanings regardless of grammatical function. Therefore, an approach is required to identify syntactic patterns and differentiate word functions within sentence structures.

This research focuses on extracting Indonesian–Tolaki lexical units by considering both syntactic and semantic features, given the significant development potential of Indonesian–Tolaki MT. The dataset was manually compiled from various Indonesian linguistic sources with emphasis on Tolaki regional language data. The study assumes that fine-grained classification is necessary to detect contextual linguistic features such as morphonemics, pronouns, affixation, and semantic relations. The objective is to extract Indonesian–Tolaki sentences containing one or more of these linguistic characteristics while ensuring topical relevance to the Indonesian–Tolaki domain.

A rule-based framework is applied to analyze Indonesian–Tolaki lexical data. The process begins with preprocessing, followed by text extraction based on sentence structure. The FLAIR framework is used to generate word-level tags, particularly for Noun Phrases (NP) and Adjective Phrases (AP), to capture syntactic patterns in Indonesian text [12]. Based on tagging results, classification is performed using machine learning methods integrated with the Term Frequency–Inverse Document Frequency (TF-IDF) technique [13].

The proposed extraction system is evaluated using a manually annotated corpus consisting of 800 training samples and 300 testing samples. Feature representation for syntax-based extraction utilizes TF-IDF, Word2Vec, and BERT embeddings. Furthermore, a hybrid translation framework combining Statistical-Based Machine Translation (SBMT) and Rule-Based Machine Translation (RBMT) is implemented for semantic translation. System performance for classification and translation tasks is evaluated using precision, recall, F1-score, and accuracy metrics.

Related theory

Tolaki language

Tolaki language is a language that comes from the Tolaki Grammar Book [14]. The following are instructions for writing and spelling Tolaki.

Table 1. Tolaki Writing and Spelling Instructions.

	Explanation	**Example**
Letter **y**	Writing only on certain foreign words or at the beginning of person name.	*oyuta, i Yondi.*
Quotation marks (Apostrof) **'**	writing quotation marks according to the original form of the base word or affix (not deleted).	*ki'oki, me'ambo, sumosa'a'i, indi'o'i, mokonda'u'i.*
Twin Vowels	Complete writing of two vowels according to the root form (not deleted).	*wee'ikee, saa nokii'i.*
All **Affixes**	Writing combined with basic words.	*mombeka'o'olu'ako, meosandoono, iko'aso, iamo'oha, oruoikaa.*
Element **i** and **to**	The writing is separated from names that start with a capital letter. The writing of the element **i** is combined with words that start with a lowercase letter.	*i Ali, i Kandari, to Wuna (band. Tolaki, Toraa)* *i'ama, ikota*
Element **o**, **ke**, **kei**, **a**, **ha**, **ma**, **ko**, and **no**	Writing is combined with the next word, including words starting with a capital letter.	*o'Ombu, odahu, noinaku, noi Ali, keinaku, kei'iee, aku, keku, maku, noku.*
Kata ganti bebas	Written separately.	*inaku, inggo'o, iee, inggito, inggami, inggomiu, ihiro.*

Dataset

The dataset is the main basic material that is very important because it is the initial input for the whole process in this research. This study focuses on Indonesian datasets that have been worked on before to obtain updated contributions from problems that have not been worked on before. In addition, the Tolaki language dataset which was compiled manually was also used. The following table shows the representations used in this study. The process for creating a dataset consists of two main steps: collecting and annotating the dataset.

Table 2. Dataset representation.

Domain	Train	Test	Total
Indonesian	800	300	1100
Tolaki	800	300	1100

Pre-processing

In natural language processing, the pre-processing stage is carried out to process raw data so that it is ready to be processed based on data requirements that will be used as input for further analysis processes. Generally, the following pre-processing stages: 1) case folding, 2) filtering, 3) normalization, 4) stopword removal, 5) stemming, 6) tokenizing.

Text Extraction

The Indonesian text extraction stage requires a very detailed process because Indonesian reviews have very complex word types that can affect the word extraction process in a sentence. Common problems that occur when extracting Indonesian text include: unstructured syntax [15] [16], morphemes [17], word functions, and word types. In general, 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.

The function of Tolaki language words consists of Subject, Predicate, Object, and Description. The following table shows the relationship between words and function words in sentences, where the position of a word can affect the taking of the function of the word itself.

Table 3. Word function sentence.

No	Sentence	Word function
	Indonesian-Tolaki	
1	Saya **naik** kelas-(Inaku **pe'eka** kalasi)	Subject **Predicate** Object
2	**Naik** *terasa* melelahkan-(**Pe'eka** *kupenasa'i* mokongango)	**Subject** *Predicate* Complement Adverbial
3	Harga minyak **naik-**(Oli luwi **pe'eka**)	Subject **Complement Adjunct**

Word type

Types of words in the Tolaki language consist of 17 tags which are generally taken from the universal POS tag. The following table shows the relationship between words and types of words in sentences, where the position of a word can affect the choice of the type of word itself.

Table 4. Word type sentence.

No	Sentence	Word type
	Indonesian-Tolaki	
1	Saya **naik** kelas - (Inaku **pe'eka** kalasi)	Noun **Verb** Noun
2	**Naik** *terasa* melelahkan **-** (**Pe'eka** *kupenasa'i* mokongango)	**Noun** *Verb* Adverb
3	Harga minyak **naik -** (Oli luwi **pe'eka**)	Noun **Adjective**

Machine translation (MT)

In previous research, Machine Translation can be divided into two, namely rule-based and corpus-based as shown in Figure 1 and 2. The rule-based method consists of three methods, namely: Literal translation, Transfer-based, and Interlingua-based. While the corpus-based method consists of two methods, namely: statistical-based and case-based.

- **Corpus Based**

The corpus based approach or better known as statistical machine translation (SMT) works based on statistical models taken from parallel-parallel corpora of bilingual texts. The SMT approach assumes that every word in the target language is a translation of the source language words with several possibilities. The words that have the highest probability of giving the best translation are taken as the result of the translation. The consistent pattern of divergence between languages when translating from one language to another is one of the fundamental problems in MT when dealing with reordering divergence. The main steps in SMT are: Corpus preparation, Training, Decoding and Testing.

Corpus preparation, alignment and cleaning are carried out at the Pre-Processing stage. Training is the process by which a supervised or unsupervised statistical machine learning algorithm is used to build statistical tables from parallel corpora. In SMT, alignment based on words and phrases plays a major role during parallel corpus training. The translation model, Language Model, Distortion Table, Phrase Table and so on were carried out at this stage of the training. Whereas Decoding is the most complex task in MT where the trained model will be decoded. These processes are the main processes of SMT for translation into the target language using the phrase table, translation model and previously generated language model.

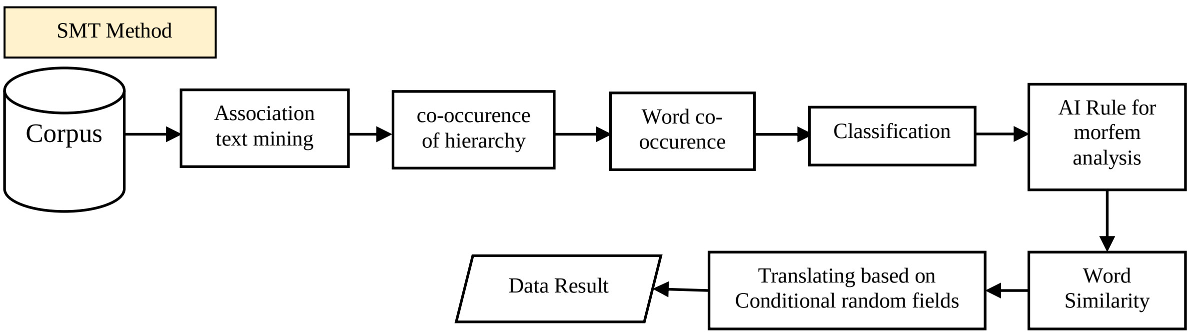

Figure 1. Word Extraction Methods and Techniques for SMT.

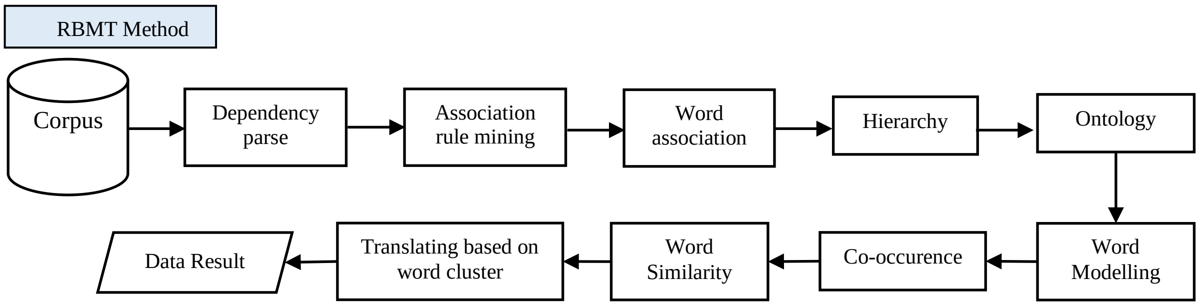

Figure 2. Word Extraction Methods and Techniques for RBMT.

- **Rule Based**

The rule based approach or better known as rule based machine translation (RBMT) works based on the specification of rules for morphology, syntax, lexical selection, and also transfer and generation. The set of rules and the bilingual or multilingual lexicon are the main materials used in the RBMT. The transfer model involves three stages: analysis, transfer, and generation. RBMT workflows in general are morph analyzer, Part of Speech (POS) tagger and chunker, name entity recognition (NER), word sense disambiguation (WSD), lexical transfer, word generator, translation result. The RBMT workflow is grouped into three process phases, namely:

1. Analysis phase, a linguistic analysis process is carried out on the input source sentences to extract information in terms of morphology, speech parts, phrases, named entities, and word meaning disambiguation.
2. Lexical transfer phase, there are two steps, namely word translation and grammatical translation.

a. In word translation, the root word of the source language is replaced with the root word of the target language with the help of a bilingual dictionary

b. In grammatical translation, the suffix is translated.

3. Generation phase, the chunking process is carried out to translate words such as gender and ownership that are adapted to the form of the verb or object of a subject.

METHODOLOGY

In this section, analysis and discussion of previous articles on machine translation is carried out. The analysis and discussion is based on several factors such as the text extraction, classification, and machine translation methods used. Furthermore, the strengths and weakness, data collection and its language, and the performance result from the other researched is used to develop this proposed method.

Dataset

The dataset used in this research describes the source of the data obtained. Table 5 shows the representation of the Indonesian and Tolaki sentences data collection. The following is a representation of the dataset that was built manually in this study:

Table 5. 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
4.	Saya berjalan naik	Inaku *lumako* pe'eka
5.	Saya merasakan naik melelahkan	Ku *penasa'i* pe'eka mokongango

Text extraction

In the text extraction stage, this research uses a pre-processing process to remove symbols that are not used and prepare the data to be ready for processing. In this stage, an approach method is developed to extract syntactic cases that can distinguish each sentence pattern that has different functions and types of words based on the context and meaning of words. First, the POS tagging process is carried out using FLAIR.

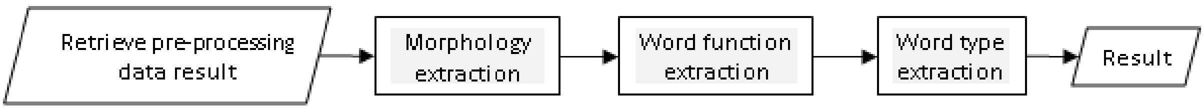

Figure 3. Text extraction flow chart.

The results of POS tagging are shown in Table 6. Furthermore, the stage of analysis of morphological cases is carried out using the concept of the Morphind approach. Then the extraction of functions and types of words is carried out using a machine learning algorithm.

Table 6. 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>
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>

Morphology extraction

The Indonesian morphology extraction process was carried out using the MorphInd concept from previous studies. Meanwhile, the morphology extraction process in the Tolaki language is carried out using an algorithm. The morphology extraction algorithm that used in this study. First, the pre-processing results are used as input for POS tagging. Then take the token 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.

Word function extraction

Word function extraction is used to get word function in the sentence. The results of the morphology extraction process are used as input in this process. The flow of the word function extraction process. Determination of the function of the word subject, predicate, object, compl adverb, compl 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.

Word type extraction is used to get word type in the sentence. The results of the morphology extraction process are used as input in this process. The flow of the word function extraction process. In this stage, we compiled 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 have compiled these 51 rules as the basis for word tag relation rules. correct in the sentence. So if there is an incorrect tag relation, the system will automatically update the correct word based on the 51 rules.

Classification

In the previous explanation in Section 1, it has been known that the extraction of word functions on word types is needed to work on syntactic cases. Furthermore, from these results, word analysis related to semantic cases was also carried out in sentences. In this classification process, four methods are used to extract a word based on the syntax and semantics of the word. In the classification stage, Word2vec, TF-IDF, BERT embedding, and Cosine Similarity are used. The following is a detailed explanation of the 4 methods.

- TF-IDF

TF-IDF untuk menghitung daftar kata dalam dokumen yang telah diberi label agar didapatkan hasil akurasi yang lebih baik

- Word2Vec

This process converts the results of the text feature into a vector value. We used gensim library in python to implement word2vec to train and test data. Table x shows an example of word2vec representing a vector.

- BERT embedding

Document expansion in terms of aspect words from Word2vec is then followed by a similarity matching process using BERT embedding. BERT embedding is used to improve the accuracy of word retrieval aspects that will be processed in the next process. Following are the stages of BERT embedding in this research.

- Cosine similarity

Semantic similarity in this study uses the Cosine method with equations which have been described.

Machine Translation (MT)

In the MT process in this study, Rule-based is used because it requires a set of rules that can work on syntactic cases, namely identifying and extracting word functions against the types of words in sentences so that they can work on semantic cases of words in sentences.

Analysis and Result

Based on the analysis of the previous sections, it can be concluded that the problem is that the need for Indonesian machine translation is getting higher. This is because the MT, which has been widely developed for the proposed Indonesian translation system, still does not cover all the rules that exist in Indonesian, such as the morphonemic case and the case of a word that has different types of words based on the function of the word itself. Therefore, in Section 4, we present the results of the analysis of the research work that has been carried out, including: text extraction results, classification results, and machine translation results.

Text extraction result

Table 7 shows the sample results of syntactic cases extraction that consist 9 sentences of Indonesian-Tolaki. Sentence 1, can be seen that the word “naik” as a predicate with a verb type. Sentence 2, the word “naik” as a subject with a noun type. Sentence 3, the word “naik” as a complement with adjective type. Sentence 4, the word “naik” as a complement with adverb type. Sentence 5, the word “naik” as an objects with a noun type. While sentences 6 until 9, the words “naik” are the morphonemic example cases that has affixes and suffixes. Their function are predicates with verb type. Based on these 9 example sentences, this proposed approach method can work properly to extract the cases of the syntactic sentences based on functions and types of words. Then, Table 7 shows an example of the identification of the word "naik" for morphonemic cases extraction.

Table 7. Analysis of function and type of words.

No	Sentences	Extraction Results	
		Word function	**Word type**
	Indonesian-Tolaki		
1	Saya **naik** kelas (Inaku **pe’eka** kalasi)	Subject **Predicate** Object	Noun **Verb** Noun
2	**Naik** *terasa* melelahkan (**Pe’eka** *kupenasa’i* mokongango)	**Subject** *Predicate* Complement Adverbial	**Noun** *Verb* Adverb
3	Harga minyak **naik** (Oli luwi **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 meena*)	Subject **Predicate** Object *Complement Adjunct*	Noun **Verb** Noun *Adjective*
7	Saya **menaiki** tangga *susah sekali* (Inaku **pe’ekari’i** la’usa *masusa meena*)	Subject **Predicate** Object *Complement Adjunct*	Noun **Verb** Noun *Adjective*
8	**Kenaikan** harga minyak *disiarkan di televisi* (***Nope’eka*** *oli luwi bawo i televisi*)	**Subject** Predicate *Complement Adverbial*	**Noun** Verb Adverb
9	**Kenaikan** harga minyak akan **menaikkan** harga sembako (***Nope’eka*** *oli luwi nggo* ***pe’eka itoono*** *oli sombako*)	**Subject** Predicate **Object** *Complement Adjunct*	**Noun Verb** Noun

Classification result

The comparison results of one-way and back-way translations can be seen in Table 8. Table 8 shows the words that marked as mistranslations because they do not have the similar meaning with the actual input sentences. Therefore, an analysis was carried out based on the word probabilities of the documents that used to get better accuracy results of words meaning. Accurate translation results are influenced by the word class based on the function, type, and meaning of the word in the sentence

Testing the proposed method for the word classification process using TF-IDF, Word2vec, and BERT embedding are showed good results. TF-IDF can be able to get terms from each target word. Next, Word2Vec works by calculating the vector value of each trm that has been taken. Table 9 shows the results of the TF-IDF and Word2vec processes for examples of the word “naik” target in this study. Finally, BERT embedding calculates the similarity of the target term with the entire word form in the document. The term with the highest similarity value is taken as the result of the actual term for the analysis of word types and functions. Table 9 shows the results of the probability calculation between terms extracted using Word2vec and the word term pairs that labeled as wrong translation. BERT and cosine similarity are used for this calculation method.

Table 8. TF-IDF and Word2vec for SMT analysis.

Sent[1]	**I'm going to class**
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: [('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)]

Table 9. BERT + cosine for SMT analysis.

Sent[1]	**I'm going to class**
Terms similarity	[(‘going: class’, 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)]

Table 10 shows proposed rules implementation for RBMT analysis. Indonesian POS tagging results are used to measure the completeness of word structure in sentences. Then, a good word structure used in a sentence should at least consist of a subject (NOUN) and a predicate (VERB). The result of determining the translation is used to compare the probability of word similarity between the results of rule-based analysis and statistical-based analysis, with the highest value will be taken as the translation result. This 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.
- if there is an incomplete word structure without VERB after subject NP or object NOUN in the sentence, then the VERB tobe is added automatically after the subject NP or object NOUN.

First case, sentence “saya naik kelas”, it can be seen that the word “naik” as a predicate VERB has difference result while using Indonesian-Tolaki to English translation and English to Indonesian-Tolaki. The word “naik” this sentence context, 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 result of identification based on the proposed rules in this study, can be obtained the translation of the VERB “naik” when paired with the NOUN “kelas” is “promoted to next grade”. So, updating process result of the sentence is “I am promoted to next grade”.

Second case, sentence “naik terasa melelahkan”, it can be seen that the word “naik” as a subject PROPN has difference result while using Indonesian-Tolaki to English translation and English to Indonesian-Tolaki. The word “naik” this sentence context, 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 result of identification based on the proposed rules in this study, 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 of the sentence is “hike feels tiring”.

Third case, sentence “harga minyak naik”, it can be seen that 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 result of identification based on the proposed rules in this study, 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 of the sentence is “oil prices are going up”.

Fourth case, sentence “saya berjalan naik”, it can be seen that the word “naik” as a complement adverb ADV has difference result while using Indonesian-Tolaki to English translation and English to Indonesian-Tolaki. The word “naik” this sentence context, if it is translated from Indonesian-Tolaki to English then the result is “up” with predicate VERB “walked”. Actually, the pairing word “walked up” while is translated to Indonesian has the meaning “berjalan ke atas” that express past tense of the sentence. While the input sentence used does not state a form of past tense at all. Therefore, the pairing word “walked up” is marked as a word translation error. The result of identification based on the proposed rules in this study, can be obtained the translation of the VERB “berjalan” when paired with the NOUN “naik” is “walk up”. So, updating process result of the sentence is “I walk up”.

Fifth case, sentence “saya merasakan naik melelahkan”, it can be seen that the word “naik” as a object NOUN has difference result while using Indonesian-Tolaki to English translation and English to Indonesian-Tolaki. The word “naik” this sentence context, if it is translated from Indonesian-Tolaki to English then the result is “ride” with predicate VERB “feel” and complement anjunct ADJ “tiring”. Actually, the word “ride” while is translated to Indonesian has the meaning “perjalanan”. Therefore, the word “ride” is marked as a word translation error. The result of identification based on the proposed rules in this study, can be obtained the object form expansion of word "naik" with the NOUN type from the corpus existing is "kenaikan". It is also added automatic VERB tobe "adalah" after object NOUN expansion "kenaikan" to complete the sentence structure. The actually sentence in Indonesian is changed to be "saya merasakan kenaikan adalah melelahkan". Then, based on hidden word translation using NOUN-VERB-Complement can be obtained hidden word translation of "kenaikan adalah melelahkan" is "hike is tiring". So, updating process result of the sentence is "I feel the hike is tiring".

Table 3. Proposed rule implementation for RBMT analysis.

Indonesian to English				**English to Indonesian**					
saya	**naik**	kelas		i	am	going	to	class	
PRON	VERB	NOUN							
i	*am going to*	*class*		*Saya*	*pergi*		*ke*	*kelas*	
S:NP	Hidden topic: 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*	*berikutnya*	*kelas*

Machine translation result

In this section, first, the implementation of morphological extraction is carried out to get the types of words in sentences. As an illustration, Indonesian-Tolaki sentences were used to be translated into English. Table 12 shows the process of translating Indonesian-Tolaki into English. Then, the English translation result is used for back translation as the input to be translated into Indonesian-Tolaki.

Tables 11 show the differences between the results of one-way translation and reverse translation. The text classification process that has been carried out is able to increase the accuracy of text translation but has not been 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, Indonesian-Tolaki translation into English, Indonesian-Tolaki has affixes and word endings, while English does not have them, causing hybrid machine translation 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 that the back way translation process, 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 taht obtained is "I'm promoted to next grade."

Table 4. Comparison result of SMT, RBMT, and Hybrid SMT-RBMT.

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

Evaluation process to compare the MT approach using SMT, RBMT, and Hybrid MT have also been carried out in this study. 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 that obtained from the proposed Hybrid MT method are still better when compared to SMT and RBMT. Finally, the results of the MT evaluation process are shown in Table 12.

Based on our experiments with the sample results in Tables 12, the proposed hybrid MT method can work well with an average accuracy of 74.17% for one-way translation of Indonesian-Tolaki to English. Meanwhile, back way translation, English to Indonesian-Tolaki achieves an average accuracy of 70.83%.

Table 5. Evaluation process of MT.

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

CONCLUSIONS

The findings of this study present an investigation into Machine Translation (MT) development by emphasizing the adoption of recent global MT approaches for Indonesian MT applications. Previous research on Indonesian MT has largely focused on statistical and rule-based approaches, with limited exploration of detailed syntactic rule integration. This study performs word translation by considering grammatical functions that influence word categories within sentence structures, which is shown to improve translation precision and detail.

SMT-based translation achieved higher performance in English to Indonesian–Tolaki translation, with an accuracy of 65.00%, compared to Indonesian–Tolaki to English translation, which achieved 54.17%. In contrast, RBMT produced better performance for Indonesian–Tolaki to English translation, achieving 60.83% accuracy, compared to 41.67% for English to Indonesian–Tolaki translation. The proposed hybrid MT method demonstrated superior performance in both translation directions. Specifically, it achieved 74.17% accuracy for English to Indonesian–Tolaki translation and 70.83% for Indonesian–Tolaki to English translation.

These results indicate that the proposed hybrid SMT–RBMT framework outperforms individual SMT and RBMT methods. Additionally, the parallel corpus used in this research was manually constructed to support the training process.

Acknowledgments

This research was funded by the Research and Community Service Institute (LPPM-UHO) at Halu Oleo University.

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. 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.
3. 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, ICT4M 2013*, 2013, doi: 10.1109/ICT4M.2013.6518918.
4. H. Sujaini, "Mesin Penerjemah Situs Berita Online Bahasa Indonesia ke Bahasa Melayu Pontianak," vol. 6, no. 2, pp. 38–44, 2014.
5. 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, ICEEI 2015*, pp. 54–58, Dec. 2015, doi: 10.1109/ICEEI.2015.7352469.
6. 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.
7. 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.
8. F. Rahutomo, R. A. Asmara, and D. K. P. Aji, "Computational Analysis on Rise and Fall of Indonesian Vocabulary During a Period of Time," *undefined*, pp. 75–80, Nov. 2018, doi: 10.1109/ICOICT.2018.8528812.

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. I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," *2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017*, vol. 2018-Janua, no. October, pp. 233–237, 2018, doi: 10.1109/ICACSIS.2017.8355039.
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. Z. Sailan and Pusat Pembinaan dan Pengembangan Bahasa., "Tata bahasa Tolaki," p. 180, 1995.
15. A. F. Hidayatullah and M. R. Ma'arif, "Pre-processing Tasks in Indonesian Twitter Messages," *J. Phys. Conf. Ser.*, 2017, doi: 10.1088/1742-6596/801/1/012072.
16. 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.*, 2017, doi: 10.1088/1742-6596/801/1/012066.
17. 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 (LREC-2014)*, 2014, doi: 10.1306/C1EA47CA-16C9-11D7-8645000102C1865D.
18. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.
19. M. Kulmanov, F. Z. Smaili, X. Gao, and R. Hoehndorf, "Semantic similarity and machine learning with ontologies," *Brief. Bioinform.*, vol. 22, no. 4, pp. 1–18, 2021, doi: 10.1093/bib/bbaa199.
20. A. Mahmoud and M. Zrigui, "Semantic similarity analysis for paraphrase identification in Arabic texts," *PACLIC 2017 - Proc. 31st Pacific Asia Conf. Lang. Inf. Comput.*, no.

January, pp. 274–281, 2019.

21. 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.
22. M. Kamayani and A. Purwarianti, "Dependency parsing for Indonesian," *Proc. 2011 Int. Conf. Electr. Eng. Informatics, ICEEI 2011*, no. July, 2011, doi: 10.1109/ICEEI.2011.6021552.
23. W. Suwarningsih and I. Supriana, "PreBI : Indonesian Predictive Parser," *Proc. IIAI Int. Conf. Adv. Inf. Technol.*, no. November, pp. 36–40, 2013.
24. C. Latiri, K. Sma\"\ili, C. Lavecchia, D. Langlois, and C. Nasri, "Phrase-based machine translation based on text mining and statistical language modeling techniques," *Proc. 12^{th} Int. Conf. Intell. Text Process. Comput. Linguist. CICLING'2011, Int. J. Comput. Linguist. Appl.*, vol. 2, no. 1–2, pp. 193–208, 2011.
25. S. K. D. S. Niranjan, "Design and Implementation of Association Rules Based System for Evaluating WSD," *Int. J. Sci. Res.*, vol. 3, no. 6, pp. 2791–2796, 2014.
26. Z. Zhang and S. Zhu, "A new approach to word sense disambiguation in MT system," *2009 WRI World Congr. Comput. Sci. Inf. Eng. CSIE 2009*, vol. 7, pp. 407–411, 2009, doi: 10.1109/CSIE.2009.1105.
27. N. Rahman and B. Borah, "An unsupervised method for word sense disambiguation," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2021, doi: 10.1016/j.jksuci.2021.07.022.
28. J. Su, J. Zeng, D. Xiong, Y. Liu, M. Wang, and J. Xie, "A Hierarchy-to-Sequence Attentional Neural Machine Translation Model," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 3, pp. 623–632, Mar. 2018, doi: 10.1109/TASLP.2018.2789721.
29. M. Moradshahi, G. Campagna, S. Semnani, S. Xu, and M. Lam, "Localizing Open-Ontology QA Semantic Parsers in a Day Using Machine Translation," pp. 5970–5983, 2020, doi: 10.18653/v1/2020.emnlp-main.481.
30. D. Han, J. Li, Y. Li, M. Zhang, and G. Zhou, "Explicitly Modeling Word Translations in Neural Machine Translation," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 19, no. 1, Jul. 2019, doi: 10.1145/3342353.
31. C. Mi, Y. Yang, L. Wang, and X. Li, "Co-occurrence degree based word alignment in statistical machine translation," *Open Autom. Control Syst. J.*, vol. 6, no. 1, pp. 561–565, 2014, doi: 10.2174/1874444301406010561.
32. C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. F. Zaidan, "Findings of the 2010 Workshop on Statistical Machine Translation and Metrics for

Machine Translation," *Wmt-2010*, no. July, pp. 17–53, 2010.

33. X. Tan, J. Chen, D. He, Y. Xia, T. Qin, and T. Y. Liu, "Multilingual neural machine translation with language clustering," *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 963–973, 2020, doi: 10.18653/v1/d19-1089.