ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 12, No. 1, June 2024, pp.223-235

Enhancing machine translation: syntax and semantics-based word type and function extraction through multi-task transfer learning in Indonesian, Tolaki, and English

Muh Yamin[1], Riyanarto Sarno[2], Tambunan Tambunan[3]
[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
[3]Department of English Language Education, Faculty of Teacher Training and Education, Halu Oleo University, Kendari, Indonesia

ABSTRACT

This research aimed at constructing an effective Machine Translation (MT) system for the Indonesian, Tolaki, and English languages by integrating in-depth morphological, syntactic, and semantic analyses. Utilizing both supervised and unsupervised methods such as TF-IDF, Word2vec, BERT, and semantic similarity, this research extracted Indonesian and Tolaki words, categorizing them based on function and type within sentences and documents. The research method involved developing a morph tool to capture morphological elements, followed by rule-based algorithm formulation for syntactic analysis to extract word functions and types influencing translation within sentences. Three MT methods, Rule-Based MT (RBMT), Statistical MT (SMT) and SMT-RBMT (hybrid), were tested for translation accuracy. With an average accuracy of approximately 70%, the evaluation of the hybrid MT method demonstrated its superiority over SMT and RBMT, yielding translation accuracies of 0.71 from English into Indonesian into Tolaki and 0.74 from Indonesian into Tolaki into English.

Keywords: Machine translation, Syntax, Semantics, 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

Within the realm of Machine Translation (MT), while various methodologies have been developed to translate words based on English language rules, similar advancements have not been extensively explored for Indonesian MT. Consequently, there exists a critical need for methodological advancements that not only align with established English MT methodologies but also respect the unique linguistic characteristics of the Indonesian language. However, previous research has attempted to develop Indonesian MT using various approaches, relying on both rules and corpora. The rules approach encompasses literal translation, the transfer method, and the cross-language method, while the corpus-based approach includes statistical and case-based method [1]. In this domain, previous works in Indonesian Machine Translation (MT) have made strides, but significant gaps remain. As in [2] focused on semantic word analysis, developing a tool for capturing nouns and foreign words. However, they did not extend to sentence or document translation. Reference [3] expanded into statistical MT from English to Indonesian but overlooked contextual word cases and morphophonemic analysis. Reference [4] explored statistical MT for Malay-Indonesian translation, addressing limited corpora availability but neglecting morphological and contextual complexities. Reference [5] emphasized affixes and root words in Indonesian to Dayak Taman translation but lacked depth in sentence context and morphology. Ref. [6] highlighted the absence of Sundanese-Indonesian parallel corpora, while ref. [7] identified unused lemmas in online media. Furthermore, Reference [8] underscored the significance of expanding resources for Indonesian and its regional languages, incorporating hybrid neural MT. However, there is no, if any, fully addressed comprehensive translation methodologies, leaving a gap in understanding morphological, syntactic, and

contextual intricacies. This research endeavors to bridge this gap by further developing the existing approach to Indonesian MT.

Therefore, based on the aforementioned gaps, the primary challenge in Indonesian word extraction research stems from the absence of Indonesian corpora with labeling that are directly applicable to analyses of datasets in multiple fields. Indonesian texts frequently present syntax cases, which pose various challenges in word extraction. The extraction process involves considering multiple classification features, comprising lexical assets, analysis of sentiment, surface characteristics, word generalization, language features, and scientific elements [9]. Therefore, it is crucial to develop an extensive method for collecting words from the Indonesian language that can capture semantics as well as syntax using a morphological approach. Besides, English itself has tense forms that may lead to semantic errors when translated into Indonesian, which typically does not have specific tense structures [10]. Words can have entirely different functions and meanings within different sentence contexts. This underscores the need for methodological approaches capable of extracting sentence syntax cases while preserving similar meanings, despite differing word patterns, functions, and types. Syntax issues occur when words in a sentence have identical meaning but distinct in functions and lexical types, whereas semantic issues involve words with the same or different functions and lexical types but convey different meanings. Hence, the goal of this research was to mine the words of Tolaki and Indonesian languages, emphasizing both semantics and syntax for Tolaki and Indonesian machine translation. It expects the existence of detailed classifiers necessary to develop a system that can recognize contextual terms with morphophonemic, phonological, affixing, and semantic elements in Indonesian sentences. Indeed, this research aims to extract sentences relevant to the Tolaki and Indonesian languages containing one or more of these elements, with corpora essentially associated with the Tolaki and Indonesian languages field.

2. Research method

This study involves conducting a literature review of text mining, gathering abstract content from 150 journal articles in PDF format and specific datasets, designing a Tolaki language corpus, developing a system for creating a Tolaki language dictionary using preprocessing and classification methods, implementing the system with Python and Android Studio, and evaluating its accuracy.

2.1. Data collection

Adhering to content analysis principles, data collection involves assigning experts to translate Tolaki texts into pre-selected Indonesian, ensuring meticulous documentation of both explicit and implicit meanings. Subsequently, the Tolaki language data undergoes processing to construct parallel text corpora in English, Indonesian and Tolaki Languages, comprising a total of 4,500 sentence pairs designated for training and testing. The training dataset, consisting of 3,600 sentences, is used to develop the model through pattern formation, while the testing dataset, comprising 900 sentences, evaluates the model's accuracy in classifying word types during the classification process.

2.2. Syntax-based text extraction

A novel approach is developed to extract syntactic patterns, enabling the differentiation of word functions and types within sentences to capture contextual meaning. Interestingly, syntactic text extraction between Indonesian and Tolaki languages shows minimal differences, primarily in morphological elements such as affixes, suffixes, infixes, and compounds. Consequently, the same algorithm is employed in both languages, leveraging separate morphology dictionaries. Initially, POS tagging using the FLAIR tool is conducted, followed by morphological case analysis with the Morphind tool approach. Subsequently, machine learning algorithms are utilized to extract word functions and types within each sentence, with further analysis carried out using TF-IDF and Word2Vec.

Input: Outcome Elements of POS tagging
Output Elements: Outcome of Morph Tool
Start 1. Segment the term into individual units (Tokenization); 2. Identify affixed terms; 3. Analysis the base form of terms; 4. Final Morph Tool; **End**

Figure 1. Algorithm for morphological identification

2.2.1. Morphological extraction

Morphological extraction in Indonesian relies on the StanzaNLP framework, renowned for its precision in natural language processing. In contrast, Tolaki morphological extraction employs a tailored algorithm, complemented by an extended morphology dictionary designed specifically for the Tolaki language.

2.2.2. Extracting the functions of words

The process of identifying a word's function within a sentence is called word function extraction. There are three acknowledged rules for determining a word's function. The first rule defines that a sentence beginning in noun phrase. The second rule, a sentence beginning in verb phrase. The third rule, a sentence beginning in AUX.

Input: Outcome Elements of Morph Tool
Output Elements: Predicate, subject, object, adverb and complement
Start 1. Give every phrase directory a token; 2. Determine the sentence's POS Tag; 3. Analyze the root forms of words, checking for any prefixes or suffixes; 4. If an affix is present, modify the root form accordingly; 5. Predict the next terms based on their sequence; 6. Determine the grammatical role of each term; **End**

Figure 2. Algorithm for extracting the word function

2.2.3. Extracting the category of word

To identify the word categories in a sentence, word category extraction is employed. Fifty-one new rules that regulate parent and child elements have been created. These rules dictate the relationships between word tags, ensuring accurate word classification within sentences. Any misaligned tag relations are swiftly corrected based on the predefined 51 rules.

Input: Outcome of word function
Output Elements: NOUN, PRON, NUM, VERB, AUX, PROPN, ADV, DET, ADJ, PUNCT, SCONJ, SYM, CCONJ, INTJ, ADP, PART, X
Start 1. Encode every list of terms; 2. Determine which sentence's POS Tag is; 3. Base word tag analysis; 4. Base word tag updates; 5. Make word predictions based on word order relations; 6. Outcome of word type; **End**

Figure 3. Algorithm for extracting word category

2.3. Semantic-based text extraction

The previous results of this syntactic extraction are then utilized for semantic extraction in both Indonesian and Tolaki languages. The proposed semantic extraction is carried out using BERT embedding and Cosine Similarity. In this Research, BERT embedding is used to expand document content based on noun types identified through syntactic extraction, enhancing accuracy for subsequent processing. Semantic similarity is then determined using Cosine similarity, measuring tokens generated from BERT embedding against target words in the sentence. Moreover, semantic extraction involves two key processes: single-word meaning extraction and multi-word meaning extraction. Single-word meaning extraction aims to identify words with linear semantic relations to tokens generated from BERT embedding, utilizing a uni-gram approach to capture the true meaning of each word in the sentence. Conversely, multi-word meaning extraction seeks to identify

words with cross-semantic relations to BERT embedding tokens, employing a bi-gram approach to capture the meaning of expressions or phrases within the sentence.

2.4. Machine translation

This research applies statistical-based and rule-based algorithms from the concepts of SMT and RBMT to handle both syntax and semantics. The outlined approach to research method is Hybrid MT, which employs three inputs sourced from the outcomes of SMT, RBMT, and text extraction for comparison to achieve optimal translation outcomes.

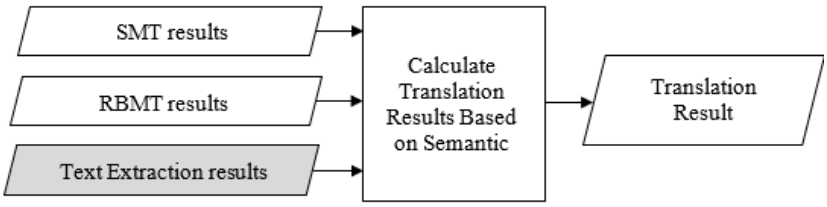

Figure 4. The outlined method of MT hybrid system

The proposed approach of Hybrid MT integrates Neuro Machine Translation as the primary MT method for analysis. This method involves intricate extraction processes, focusing on syntax and semantics, to identify word types and phrases in the input text for translation. However, relying solely on NMT cannot ascertain the accuracy of translation results. In this phase, text extraction is utilized to refine the NMT translations by rectifying errors based on predefined rules from referenced documents. This extraction process encompasses morphology, word function, and word type, with these steps interconnected and their order predetermined. This sequential arrangement is essential as determining the accuracy of word types necessitates prior knowledge of their structure and function within the sentence. Following this, a hybrid technique combining SMT and RBMT is employed to further refine the translation process, ensuring precise adjustments for individual words and phrases. This hybrid approach not only updates existing elements but also aids in the training of a new model to yield more accurate translation results.

3. Results

3.1. Datasets

3.1.1. Assembling sources

The assembled datasets originated from various origins, encompassing resources such as the Indonesian-Tolaki dictionary referenced in [11], Tolaki language function word compilations referenced in [12], texts relating to Tolaki culture cited in [13], translations of the Quran into Tolaki referenced in [14]], and translations into Indonesian as noted in [15]. Manual compilation guaranteed the inclusion of pertinent information, while meticulous cleansing methods eliminated surplus details like superfluous punctuation and unique symbols. The organized datasets were structured into three sections comprising English, Indonesian, and Tolaki, streamlining subsequent stages of training and evaluation during the development of translation models

3.1.2. Tolaki language lemmatization model

The Tolaki language lemmatization model utilized a dataset sourced exclusively from the Indonesian-Tolaki Dictionary, containing 1986 words. The dataset consisted of two columns: one for words with affixes and another for their corresponding base forms, which were manually assigned. Lemmatization was performed using the TF-IDF method to extract word weights and a Random Forest model to map words to their base forms. A Python script was developed using Scikit-learn (Sklearn) tools to train and evaluate the lemmatization model.

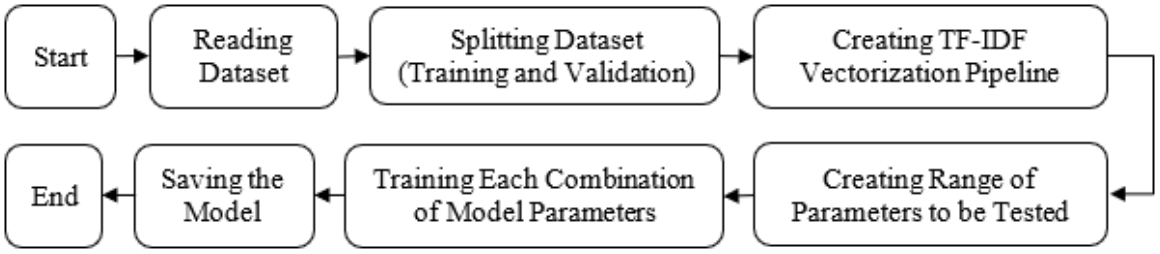

Figure 5. The flowchart is constructing the script for the Tolaki lemmatization model

This script automates the process of building, tuning, and evaluating lemmatization models using Random Forest and text development techniques like TF-IDF. The ultimate outcome is an accurate lemmatization model capable of returning base words from words with affixes.

3.1.3. Dataset cleaning through lemmatization inference

Lemmatization inference is the process of applying the constructed lemmatization model to the dataset. This model is operated using the prediction scoring technique from Scikit-learn. The final results of the inference using the constructed lemmatization model is displayed in the following table:

Table 1. The results of inference using the lemmatization model

Words with Affixes	Lemmatization Results
mepatei	pate
mepoii	poi
mepokomerambi	merambi
mepokondau	pokondau
mepombahora	pombahora
mondotoa	totoa
mondue	tue
monduehi	tue
mondunu	tunu

These results still contain some errors due to inconsistencies in the cultural usage of affixation in the Tolaki language. Therefore, manual correction was performed on the entries with errors.

3.1.4. Translating Indonesian dataset to English

The dataset translation process, spanning from Indonesia to English, utilized the Helsinki NLP Transformers framework, more precisely, leveraging the Opus MT model [16]. Final output samples from the translation inference using the Indonesian-to-English translation model are showcased in the table below.

Table 2. Sample inference of translation using the Indonesian to English translation model

en	id
on the edge of the lake	di pinggir telaga
she has a slim waist	pinggangnya langsing
from inside the basket	dari dalam keranjang
how many times as big as time	berapa kali besar yang bermuara seperti kali
many valleys	lembah-lembah yang banyak
the rat that entered the boat room showed his tail	tikus yang masuk di kamar perahu kelihatan ekornya
he went down the road to the well	dia pergi turun menuju jalan ke sumur
the many wells of parrots bathe	ramai sumurnya burung nuri turun mandi
he saw two three wooden trees by the river	dia melihat dua tiga pohon kayu di tepi sungai
it is called aalahambuti	sehingga dinamakan aalahambuti

3.1.5. Sentence and word dataset tagging

Processing and characterizing datasets in three different languages (Indonesian, English, and Tolaki) using word and sentence tagging methods employing the Transformer model from Stanza NLP [17]. Sample results of tagging for UPOS, XPOS, Ufeats, and Dependency Parser is displayed in the table below:

Table 3. Sample results of tagging for UPOS, XPOS, ufeats, and dependency parser

Lang	Kalimat	Lemma	UPOS	XPOS	Ufeats
en	his sister's almost here	['he'; 'sister'; "'s"; 'almost'; 'here']	['PRON'; 'NOUN'; 'PART'; 'ADV'; 'ADV']	['PRP$'; 'NN'; 'POS'; 'RB'; 'RB']	['Case=Gen|Gender=Masc|Number=Sing|Person=3|Poss=Yes|PronType=Prs'; 'Number=Sing'; None; None; 'PronType=Dem']

Lang	Kalimat	Lemma	UPOS	XPOS	Ufeats
	almost lost the king in bunton	['almost'; 'lose'; 'the'; 'king'; 'in'; 'bunton']	['ADV'; 'VERB'; 'DET'; 'NOUN'; 'ADP'; 'PROPN']	['RB'; 'VBD'; 'DT'; 'NN'; 'IN'; 'NNP']	[None; 'Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin'; 'Definite=Def|PronType=Art'; 'Number=Sing'; None; 'Number=Sing']
Lang	**Kalimat**	**UPOS**	**Lemma**	**XPOS**	**DepPars**
id	adiknya hampir datang	['NOUN'; 'PRON'; 'ADV'; 'VERB']	['adik'; 'dia'; 'hampir'; 'datang']	['NSD'; 'PS3'; 'D--'; 'VSA']	[(4; 'nsubj'); (1; 'det'); (4; 'advmod'); (0; 'root')]
	hampir kalah raja di buton	['ADV'; 'VERB'; 'NOUN'; 'ADP'; 'NOUN']	['hampir'; 'kalah'; 'raja'; 'di'; 'buton']	['D--'; 'ASP'; 'NSD'; 'R--'; 'X--']	[(2; 'advmod'); (0; 'root'); (2; 'obj'); (5; 'case'); (3; 'nmod')]
Lang	**Kalimat**	**Lemma**	**UPOS**	**XPOS**	**UFeats**
tlk	haido aiso leu	-	['INTJ'; 'NOUN'; 'PUNCT']	['INTJ'; 'NOUN'; 'EXCL-POINT']	[None; None; None]
	aisoito kenangia mokolew olio	-	['NOUN'; 'NOUN'; 'NOUN']	['NOUN'; 'NOUN'; 'NOUN']	[None; None; None]

The final results of tagging for Indonesian and English languages are sufficiently accurate. Conversely, Tolaki vocabulary presents some extraction errors, requiring manual correction prior to advancing to the next stage.

3.1.6. Prompt labeling

In this multilingual translation research, three models employed such as UMT5 [18], MT5 [19], and ByT5 [20] are based on Google T5 architecture [21], supporting multitasking capabilities. Prompt labeling aids in directing the models toward specific translation or linguistic analysis tasks. The utilized prompts include translation prompts for Indonesian into English, English into Indonesian, English into Tolaki, Tolaki into English, Indonesian into Tolaki and Tolaki into Indonesian. Additionally, prompts for UPOS tagging, XPOS tagging, Ufeats tagging, lemmatization, and dependency parsing are employed for Indonesian, English, and Tolaki languages. To enhance model understanding, specific lines for translating Indonesian prompts are included, such as transforming "translate english to indonesia:" into "translate" input and "terjemahan" output.

3.1.7. Dataset denoise creation in Tolaki language

The denoising dataset creation for Tolaki language involved employing T5, a model trained using denoising methods. This approach effectively enhances the model's comprehension of languages it hasn't been previously exposed to. The dataset was organized using a specialized CSV format designed specifically for Tolaki, with the assistance of the script provided below.

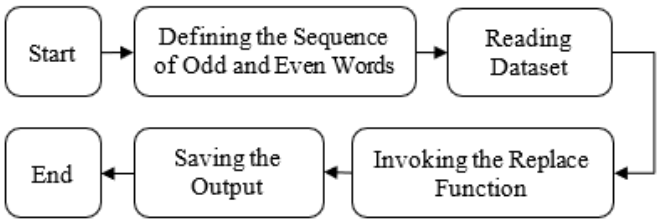

Figure 6. The flowchart of the Tolaki language denoising script

The resulting example from running this script is displayed in the subsequent table:

Table 4. Sample output of the denoising script

Input	Output
<extra_id_0> meo'ana <extra_id_1> pu'u <extra_id_2> ano <extra_id_3> petenano <extra_id_4> sangia <extra_id_5> leu <extra_id_6> te'eni: <extra_id_7> bara	La'iroto <extra_id_0> i <extra_id_1> nohu; <extra_id_2> leu <extra_id_3> baisano <extra_id_4> mbu'u; <extra_id_5> metitiro; <extra_id_6> "po'opo

Input	Output
<extra_id_8> keu <extra_id_9> keu <extra_id_10> horinggi <extra_id_11> nggiro'o <extra_id_12> watu <extra_id_13> Ma; <extra_id_14> hae <extra_id_15> Oheo. <extra_id_16> tekura <extra_id_17> obeke; <extra_id_18> "ohawoto <extra_id_19> Oheo <extra_id_20> " <extra_id_21> i <extra_id_22> kusaruokopo <extra_id_23> tau <extra_id_24> mokowai'ikona. <extra_id_25> hae <extra_id_26> taku <extra_id_27> ona <extra_id_28> sausauru'ikeiioto. <extra_id_29> i <extra_id_30> "iepo <extra_id_31> onggo <extra_id_32> laikano <extra_id_33> sangia <extra_id_34> keku <extra_id_35> humehongge <extra_id_36> batu <extra_id_37> Te'eni'itoka <extra_id_38> nggituo <extra_id_39> buna. <extra_id_40> akimbelako <extra_id_41> mbehehongge. <extra_id_42> mbera <extra_id_43> Asohapoka <extra_id_44> ano <extra_id_45> watu <extra_id_46>	<extra_id_7> Oheo <extra_id_8> pe'eka <extra_id_9> ta <extra_id_10> tumue'i <extra_id_11> mune <extra_id_12> ndumade." <extra_id_13> tekura'ito <extra_id_14> i <extra_id_15> La'ito <extra_id_16> anoleu <extra_id_17> mesuko: <extra_id_18> la <extra_id_19> tinekura'akomu? <extra_id_20> Tumotaha'itoka <extra_id_21> Oheo; <extra_id_22> hae <extra_id_23> onggo <extra_id_24> Te'eni'i <extra_id_25> obeke; <extra_id_26> onggopo <extra_id_27> mokowai'iko'o <extra_id_28> Tumotaha'i <extra_id_29> Oheo; <extra_id_30> aku <extra_id_31> pe'ekai <extra_id_32> baisanggu <extra_id_33> mbu'u <extra_id_34> ari <extra_id_35> nggiro'mune <extra_id_36> ndumade. <extra_id_37> obeke; <extra_id_38> hanu <extra_id_39> Totoa'itoka <extra_id_40> inggami <extra_id_41> Mbendekonorotoka <extra_id_42> obeke. <extra_id_43> pera <extra_id_44> teheho <extra_id_45> ndumade.

Using this approach enables T5-based models to quickly grasp the framework and lexicon of the Tolaki language.

3.1.8. Conversion of the dataset into JSON format

Ultimately, following multiple phases of dataset manipulation, the collective dataset reached a sum of 139,000 entries. To facilitate convenient access to this dataset throughout the translation model development phase, it was transformed into the JSON format.

3.2. Token limit verification

The token limit verification stage plays a crucial role in the development of the translation model. Its purpose is to examine the token length in relation to the memory capacity of the hardware, taking into account the limitations of hardware memory. This stage also assesses the efficiency of text processing and aims to find a compromise between the preferred sentence length and the constraints imposed by the hardware.

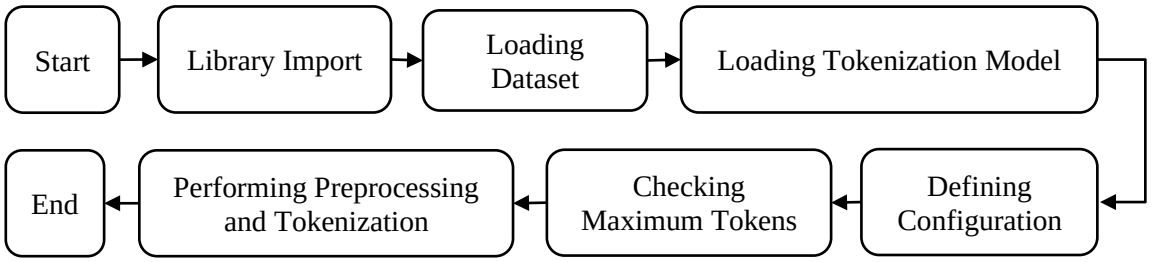

Figure 7. The flowchart of the token limit verification

The script fulfills various natural language processing tasks, including text preprocessing and tokenization. The output generated from this procedure is detailed below:

```
Status:   0%|          | 0/138057 [00:00<?, ? samples/s]
Status:   0%|          | 0/694 [00:00<?, ? samples/s]
Token limit for ByT5: 255
Status:   0%|          | 0/138057 [00:00<?, ? samples/s]
Status:   0%|          | 0/694 [00:00<?, ? samples/s]
Token limit for UMT5: 100
Status:   0%|          | 0/138057 [00:00<?, ? samples/s]
Status:   0%|          | 0/694 [00:00<?, ? samples/s]
Token limit for MT5: 107
```

Figure 8. The token limit result

The findings indicate that ByT5 exhibits the highest level of tokenization, with 255 tokens. This can be attributed to ByT5's unique byte-to-byte tokenization method, which breaks down each letter within a word separately, eliminating the necessity for a predefined vocabulary as observed in MT5 and UMT5. MT5, with a vocabulary of 210,000, results in 107 tokens, while UMT5, featuring a vocabulary of 240,000, records the lowest token count at 100. The decreased token count in UMT5 hints at its potential for more effective training compared to ByT5 and MT5.

3.3. Finetuning

Finetuning is an essential phase in the training of translation models to absorb information from the prepared datasets. This process involves three distinct models: MT5, UMT5, and ByT5, each requiring datasets with unique characteristics. Two scripts are utilized for training these models. The first script, applicable to both MT5 and UMT5, shares similar processing methods, necessitating minimal parameter adjustments for each model. The second script is tailored for ByT5, which employs byte-level tokenization, eschewing word or token vocabularies. Instead, byte-level tokenization processes text at the character or byte level.

Each fine-tuned model underwent evaluation using the BLEU metric, which is a scoring algorithm commonly employed in sequence-to-sequence translation methods for bilingual evaluation. The evaluation yielded scores for each model after 15 epochs of training.

Table 5. The scores obtained by each model

Epoch	MT5 (SMT)	ByT5 (RBMT)	UMT5 (HybridMT)
1	6,812	5,518	7,732
2	7,343	5,948	8,334
3	7,423	6,013	8,425
4	7,545	6,111	8,564
5	7,579	6,139	8,602
6	7,753	6,280	8,800
7	7,843	6,353	8,902
8	7,922	6,417	8,991
9	8,135	6,589	9,233
10	8,157	6,607	9,258
11	8,162	6,611	9,264
12	8,168	6,616	9,271
13	8,215	6,654	9,324
14	8,234	6,670	9,346
15	8,256	6,687	9,371

The table reveals a progressive improvement in scores across each training epoch for all models, albeit with signs of plateauing around the ninth epoch. Judging from the BLEU scores provided, UMT5 demonstrates relatively better performance compared to MT5 and ByT5.

3.4. Comparison of results

Evaluation of machine translation models using deep learning still requires real-time testing to assess usability. Therefore, a script is developed to test the inference results of each model. This script utilizes the text2text-generation pipeline from the Transformers library to translate text from English to Indonesian (id) or Tolaki (tolaki), and vice versa, using several different models.

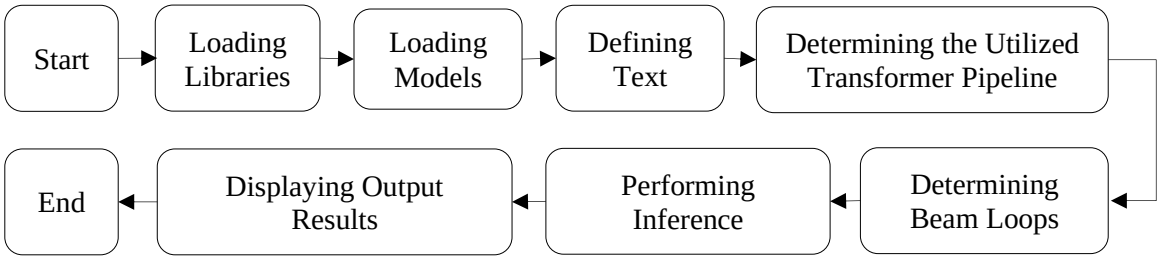

Figure 9. The flowchart of the results comparison script

The inference results of the model with various beam parameter levels are displayed in the following table:

Table 6. Comparison of result

N	Prompt	Target	MT5 (SMT)	UMT5 (HybridMT)	ByT5 (RBMT)
1	terjemahan inggris ke indonesia : I love to swim with my friends.	Saya gemar berenang bersama teman-temanku.	onggo saya pergi ke saya friend	Saya suka untuk pergi bersama-sama temanku.	Kupewiso baru di sana dalam saya.
	terjemahan inggris ke tolaki : I love to swim with my friends.		meena akuto mombaho hende-hendeino	Iee ari inaku ari ari mbera toono laa ari-ari ine ari mombemeeri'ako.	Ku kiio laa meolikee anano nggu
	terjemahan indonesia ke inggris : Aku suka berenang bersama teman-teman.	I love to swim with my friends.	aku always sepa while going to work	I like to go with friends.	I feels to her a fatter with the woman.
2	terjemahan inggris ke indonesia : I love to swim with my friends.	Saya gemar berenang bersama teman-temanku.	i amanggu melihat iwoi	Saya suka pergi bersama-sama temanku.	Kupewiso baru di sana dalam saya.
	terjemahan inggris ke tolaki : I love to swim with my friends.		ku onggo mowawo inaku	Iee ari inaku laa mo'orikee ine ari ari mbera toono laa mombemeeri'ako.	Ku kiio laa meolikee anano nggu
	terjemahan indonesia ke inggris : Aku suka berenang bersama teman-teman.	I love to swim with my friends.	i am looking for a friend	I like to go with friends.	I'm not to regular all the same as hearted.
4	terjemahan inggris ke indonesia : I love to swim with my friends.	Saya gemar berenang bersama teman-temanku.	i amanggu nggo mosusua nggo saya saudara-saudara	Saya inginkan untuk pergi bersama-sama temanku.	Kupewiso di sana menganyam ke bangsanya.
	terjemahan inggris ke tolaki : I love to swim with my friends.		ku laa nggo mbonggaa nggu	Mee-meena'ano ari inggomiu laa mo'orikee.	Ku tarima ito banggonanggu.
	terjemahan indonesia ke inggris : Aku suka berenang bersama teman-teman.	I love to swim with my friends.	i am tired of calling	I like to go with friends.	I do not tell me already for hearts.

N	Prompt	Target	MT5 (SMT)	UMT5 (HybridMT)	ByT5 (RBMT)
6	terjemahan inggris ke indonesia : I love to swim with my friends.	Saya gemar berenang bersama teman-temanku.	i amanggu pergi melihat aku	Saya inginkan untuk pergi bersama-sama temanku.	Kupewiso di sana menganyam ke bangsanya.
	terjemahan inggris ke tolaki : I love to swim with my friends.		ku laa monggaa banggonanggu	Mee-meena'ano inggomiu toono lako ari ine meambo.	Ku tarima ito banggonanggu.
	terjemahan indonesia ke inggris : Aku suka berenang bersama teman-teman.	I love to swim with my friends.	i am looking for a friend	I like to go with friends.	I do not tell me already for hearts.
8	terjemahan inggris ke indonesia : I love to swim with my friends.	Saya gemar berenang bersama teman-temanku.	iamoto nggo mosusua ronga banggonanggu	Saya inginkan untuk pergi bersama-sama temanku.	Kupewiso baru di depan saya bangsanya.
	terjemahan inggris ke tolaki : I love to swim with my friends.		ku onggo monggaa ronga banggonanggu	Mee-meena'ano inggomiu lako ari ine ombumu.	Ku tarima ito banggonanggu.
	terjemahan indonesia ke inggris : Aku suka berenang bersama teman-teman.	I love to swim with my friends.	i am tired of fighting	I like to go with friends.	I have been a cepat from children.
12	terjemahan inggris ke indonesia : I love to swim with my friends.	Saya gemar berenang bersama teman-temanku.	i amanggu nggo mebaho inaku	Saya suka pergi bersama-sama temanku.	Kupewiso kepada saya bangsangkan.
	terjemahan inggris ke tolaki : I love to swim with my friends.		ku onggo monggaa inaku banggonanggu	Mee-meena'ano inggomiu mokondewali'i ari ine Ombu sameena.	Ku tarima ito banggonanggu.
	terjemahan indonesia ke inggris : Aku suka berenang bersama teman-teman.	I love to swim with my friends.	i am looking for friend	I like to go with friends.	I have been a cepat from children.

Based on the sample table above, it's evident that the MT5 model still exhibits numerous errors in translation, where some languages appear mixed and fail to grasp the context of the prompt. Conversely, with UMT5, translations demonstrate a better understanding of the overall prompt context. Although there are translation errors, the sentence structure with UMT5 appears to be more accurate. In contrast, the results for ByT5 indicate that while the model struggles to understand the prompt well, it can manipulate the use of affixes, albeit not always accurately. Additionally, translation outcomes vary for each beam parameter used, with the most optimal beam falling between the eighth and tenth. However, by the twelfth beam, translation results begin to deviate from the context. Furthermore, as the beam count increases, the processing time for output also extends due to backward propagation for reprocessing based on the desired beam parameter count.

3.5. Website inference development

The development of a website for model inference aims to enable users to perform online translations. The website leverages JavaScript and the Huggingface API to execute model inferences. The Huggingface API facilitates the translation model inference process. JavaScript serves as the primary programming language for website development, while HTML implementation utilizes Bootstrap, AJAX, and JQuery frameworks. Users can input tasks according to the provided instructions. The website is hosted at https://tolaki-translator.my.id, and its interface is depicted in the image below.

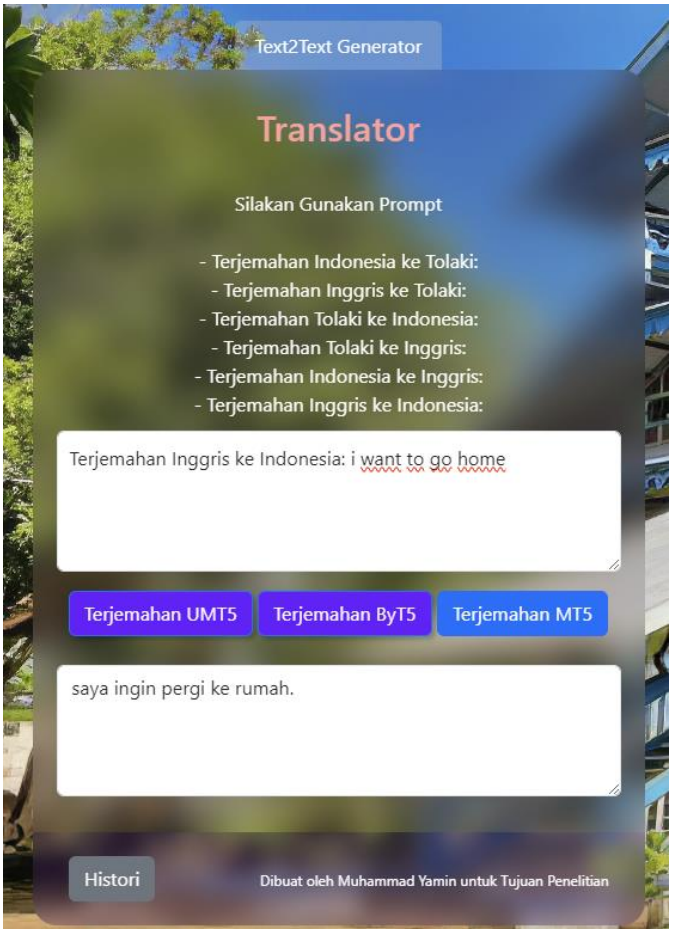

Figure 10. Web interface

4. Discussion

The experimental findings underscore the shortcomings of the MT5-SMT model in grasping the contextual nuances of input prompts, despite its multilingual prowess. Translations across certain languages often appear muddled and inconsistent, necessitating improvements for more reliable and precise outputs, potentially through prolonged training epochs. In contrast, UMT5-HybridMT, an unsupervised multilingual model, surpasses MT5-SMT in comprehending prompt contexts, yielding translations with more accurate sentence structures and overall command of instructions. While UMT5-HybridMT translations aren't flawless, they exhibit a better understanding of context compared to MT5-SMT. ByT5-RBMT demonstrates adeptness in manipulating language affixation, aligning with its linguistic analysis functionality. However, ByT5's-RBMT grasp of prompt context remains suboptimal, resulting in varying translation outcomes depending on the beam parameter used. The optimal outcomes are noticed within the range of the eighth to tenth beams, whereas the twelfth beam tends to deviate excessively from the context.

When assessing translation quality, BLEU scores play a crucial role as a metric. UMT5 achieves the highest BLEU score, indicating superior translation accuracy compared to MT5 and ByT5, suggesting UMT5's capability to generate translations resembling human references.

Despite utilizing a significant dataset from various sources in the Tolaki language, translations into Tolaki still fall short of the desired standard. This issue might arise from limitations within the Tolaki language dataset itself, highlighting the necessity for supplementation with larger and more diverse samples to improve translation accuracy. For instance, while the MC4 corpus used to train T5 comprises 2,000,000 lines of Indonesian words, the Tolaki language dataset utilized in the study consists of only around 50,000 lines. This difference potentially contributes to the higher accuracy observed in English-Indonesian translations compared to English-Tolaki or Indonesian-Tolaki translations.

This study sheds light on potential pathways for advancing machine translation further. Pre-trained T5 models offer versatile applications across various tasks and languages, presenting opportunities for enhanced translation outcomes alongside improvements in managing dataset quality.

5. Conclusions

This study investigates the most recent machine translation (MT) techniques designed specifically for Indonesian, Tolaki, English MT purposes. While conventional methods have concentrated on statistical and rule-based MT, this research explores syntactic and semantic rules. The study translates words considering their functions, which can impact the category of word in a sentence, leading to precise and thorough translations. Statistical MT (SMT) yielded better results for a translation from English into Indonesian into Tolaki that is 65% accurate, compared to the translation from Tolaki into Indonesian into English, which achieved 54.2%. Rule-based MT (RBMT) performed better for the accuracy of the translation from Indonesian into Tolaki into English, achieving 60.8%, while the translation from English into Indonesian into Tolaki was 41.7%. The proposed hybrid MT system rendered English into Indonesian into Tolaki with a higher accuracy of 74.2% compared to the reverse direction, which attained 70.8%. This indicates the superiority of the hybrid SMT-RBMT approach over SMT or RBMT. Manual collection of parallel corpora was also conducted for data training. Further research is needed to explore attention-centric methods aimed at optimizing the effectiveness of the proven techniques across SMT, RBMT, and hybrid SMT-RBMT. Moreover, the study highlights several critical recommendations for future research. Firstly, comprehensive datasets for Indonesian and regional languages spoken in Indonesia should still be further compiled. Future research should focus on implementing and comparing new methods and techniques from existing literature. Developing tailored tools for regional language machine translation (MT) systems is essential. Introducing diverse or new performance metrics for MT research can offer valuable insights. Improving translation system accuracy remains a key objective. Lastly, increasing the number of comparable corpora is crucial for refining evaluation metrics and advancing MT research.

Declaration of competing interest

The authors proclaim that they have no known financial or non-financial conflicts of interest regarding any content discussed in this manuscript.

Funding information

This research received funding from Lembaga Pengelola Dana Pendidikan (LPDP) through the Riset Inovatif-Produktif (RISPRO) Invitation Program, from the Indonesian Ministry of Education and Culture with the support of the Penelitian Terapan Unggulan Perguruan Tinggi (PTUPT) Program, and from Institut Teknologi Sepuluh Nopember (ITS) through the Publication Writing and IPR Incentive Program (PPHKI).

Author contribution

Muh Yamin led the conceptualization and methodology design; developed the software; performed validation, formal analysis, and investigation; managed resources and data curation; prepared the original draft; contributed to reviewing and editing; handled visualization, supervision, project administration, and secured funding. Riyanarto Sarno contributed to the conceptual framework; provided validation and formal analysis; assisted in investigation and resource management; contributed significantly to reviewing and editing the manuscript; supervised the research activities. Tambunan assisted in software development; participated in validation, formal analysis, and investigation; supported resource management.

Acknowledgements

The authors are deeply thankful to Institut Teknologi Sepuluh Nopember (ITS) for providing all the essential facilities. They also extend their heartfelt thanks to the reviewer for their invaluable feedback and suggestions, which have enhanced the quality of this manuscript.

References

[1] P. Li, "A survey of machine translation methods," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 11, no. 12, pp. 7125–7130, 2013, doi: http://dx.doi.org/10.11591/telkomnika.v11i12.2780.

[2] S. D. Larasati, V. Kuboň, and D. Zeman, "Indonesian morphology tool (MorphInd): Towards an Indonesian corpus," in *Systems and Frameworks for Computational Morphology: Second International*

Workshop, SFCM 2011, Zurich, Switzerland, August 26, 2011. Proceedings 2, Springer, 2011, pp. 119–129. doi: https://doi.org/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," in *2013 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M)*, IEEE, 2013, pp. 1–4. doi: 10.1109/ICT4M.2013.6518918.

[4] H. Sujaini, "Mesin Penerjemah Situs Berita Online Bahasa Indonesia ke Bahasa Melayu Pontianak," *ELKHA Journal*, vol. 6, no. 02, 2014, doi: /dx.doi.org/10.26418/elkha.v6i2.9098.

[5] Y. Jarob, H. Sujaini, and N. Safriadi, "Uji Akurasi Penerjemahan Bahasa Indonesia–Dayak Taman Dengan Penandaan Kata Dasar Dan Imbuhan," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 2, no. 2, pp. 78–83, 2016, doi: https://dx.doi.org/10.26418/jp.v2i2.16520.

[6] 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, 2015, pp. 1–6. doi: 10.1109/ICITSI.2015.7437678.

[7] F. Rahutomo, R. A. Asmara, and D. K. P. 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, 2018, pp. 75–80. doi: https://doi.org/10.1109/ICoICT.2018.8528812.

[8] M. Yamin and R. Sarno, "Hybrid neural machine translation with statistical and rule based approach for syntactics and semantics between Tolaki-Indonesian-English languages," *Periodicals of Engineering and Natural Sciences*, vol. 11, no. 5, pp. 117–136, 2023, doi: http://dx.doi.org/10.21533/pen.v11i5.3864.

[9] 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*, 2017, pp. 1–10. doi: https://doi.org/10.18653/v1/W17-1101.

[10] M. Yamin, R. Sarno, and R. Abdullah, "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, 2022, pp. 101–106. doi: https://doi.org/10.1109/ICITRI56423.2022.9970225.

[11] A. Muthalib, J. F. Pattiasina, Alimuddin, C. D., and Husen, *Kamus Tolaki-Indonesia*. Jakarta: Pusat Pembinaan dan Pengembangan Bahasa Departemen Pendidikan dan Kebudayaan., 1985.

[12] A. Muthalib, A. DP., J. F. Pattiasina, Balaka, Haloma, and A. Kadir, *Kata Tugas dalam Bahasa Tolaki*. Jakarta: Pusat Pembinaan dan Pengembangaan Bahasa, 1985.

[13] A. Tarimana, *Kebudayaan Tolaki*. Jakarta: Balai Pustaka, 1993.

[14] H. Insawan, B. Melamba, A. Timbu, Untung, Darmin, and A. Suruambo, *Terjemahan Al Qur'an Bahasa Tolaki*. 2023.

[15] Kemenag, "Qur'an Kemenag." [Online]. Available: https://quran.kemenag.go.id/

[16] J. Tiedemann *et al.*, "Democratizing neural machine translation with OPUS-MT," *Lang Resour Eval*, 2023, doi: https://doi.org/10.1007/s10579-023-09704-w.

[17] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," *arXiv preprint arXiv:2003.07082*, 2020.

[18] H. W. Chung *et al.*, "Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining," *arXiv preprint arXiv:2304.09151*, 2023.

[19] L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020.

[20] L. Xue *et al.*, "Byt5: Towards a token-free future with pre-trained byte-to-byte models," *Trans Assoc Comput Linguist*, vol. 10, pp. 291–306, 2022.

[21] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.