Homonym and polysemy approaches with morphology extraction in weighting terms for Indonesian to English machine translation

Budi Harjo[1], Muljono[1], Rachmad Abdullah[2,3]
[1]Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia
[2]Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[3]Deparment of Management, Sekolah Tinggi Ilmu Ekonomi Indonesia, Surabaya, Indonesia

Article Info

Article history:

Received May 6, 2024
Revised Jul 12, 2024
Accepted Aug 6, 2024

Keywords:

Bidirectional encoder representations from transformers embedding
Morphology extraction
Neural machine translation
Semantic similarity
Synonym-based term expansion

ABSTRACT

Homonym and polysemy features can influence some errors in translation from a source language to another target language, for example, from Indonesian to English. A lemma or a morphology factor can cause the configuration of Indonesian homonym features. For example, the word *beruang* can mean an animal *beruang* (bear) and can mean a verb alternation *ber+uang* (has/have money). The Indonesian polysemy feature can also impact an error in the translation process because it can have a literal meaning and a symbolic meaning. For example, the terms *bunga melati* (jasmine flower) and *bunga hati* (lover), where *bunga* does not only mean flower. Therefore, the development machine translation (MT) method needs to capture homonym and polysemy features in the form of a word or a phrase. This research proposes homonym and polysemy approaches with morphology extraction in weighting terms for Indonesian to English MT. First, this research uses morphology extraction to detect sentences that contain prefixes, lemma, and suffixes. Then, the word similarity measurement functions to extract homonym and polysemy in the form of uni-gram and bi-gram using bidirectional encoder representations from transformers (BERT) embedding, named entity recognition (NER), synonym-based term expansion, and semantic similarity. This research uses neural machine translation for the translation process.

Corresponding Author:

Muljono
Faculty of Computer Science, Dian Nuswantoro University
207 Imam Bonjol Road, Semarang, Central Java 50131, Indonesia
Email: muljono@dsn.dinus.ac.id

1. INTRODUCTION

The machine translation (MT) developments [1], [2] intend to solve sentence translation problems from source to target language. Homonyms and polysemy are word features that can influence some errors in translation from a source language to another target language, for example, from Indonesian to English. A lemma with multiple meanings can cause the configuration of the Indonesian homonym feature. For instance, *palu* can mean a city name *Palu* (Palu) and a tool *palu* (hammer).

The hybrid MT method development using statistical MT and rule-based MT [3] discusses how to solve translation tasks from Indonesian to English sentences but not how to translate the homonym feature. Neural machine translation (NMT) [4], [5] has achieved a significant breakthrough in translation performance. But how to solve the sentence translation that contains the Indonesian homonym feature is still not discussed.

A web-based translation engine, *e.g.*, [6]–[8], can find vocabulary items and enhance the completion of Indonesian-English translating but is not accurate enough for translating homonym features. For example, Google Translator still translates the sentence *Ayah pergi ke Palu* (Father goes to Palu) into *Father went to the hammer*, where *Palu* still translates as the hammer and has been unable to translate *Palu* as a city name of *Palu*. This translation error occurs because the MT method cannot capture the word's context in the sentence.

Morphological factors can also cause the homonym feature in Indonesian. For example, *beruang* has multiple meanings: an animal name *beruang* (bear) and a verb with *ber+uang* (has/have money). Apart from the homonym factor, the Indonesian polysemy factor can also cause errors in translating sentences from Indonesian to English. Indonesian polysemy phrases may have a literal meaning but can also have a symbolic meaning. For example, *bunga* has the phrase *bunga melati* (jasmine flower) and *bunga hati* (lover).

Many MT method development for the Indonesian language still focus on spoken language translation [9] and build some parallel corpus to translate broader words and phrases [10], [11], but still do not work on how to translate homonym and polysemy terms in the sentence. Indonesian MT [12] works on translations for cases of homonym and polysemy. However, the developed method cannot detect homonym terms in the form of prefixes. Besides that, the implementation can still solve polysemy terms that indicate a denotation and not connotation. Lexicon and corpus data [13], [14] and syllable and phonemic lengths [15] for polysemy feature extraction achieve higher precision and F1-measure but have not explored contextual diversity. The MT development should be able to capture homonym and polysemy features in the form of a word or phrase.

Based on the problems in previous Indonesian MT studies, this research proposes homonym and polysemy approaches with morphology extraction in weighting terms for Indonesian to English MT. This research aims to translate Indonesian sentences with homonyms and polysemy features into English sentences, whether they contain morphological features or not. First, this research uses morphology extraction to detect sentences that contain prefixes and suffixes. Then, the word similarity measurement extracts homonym and polysemy features using bidirectional encoder representations from transformers (BERT) embedding, name entity recognition (NER), synonym-based term expansion, and semantic similarity. This research uses NMT for the translation process. Finally, this research evaluates the proposed MT performance.

2. RELATED WORK

2.1. Dataset

Dataset [12] is an Indonesian and English dataset for machine translation testing. This dataset of 1,200 sentences consists of four types of sentences: simple, compound, complex, and compound-complex. These sentences are general sentences that can contain neutral, homonym, and polysemy words. This dataset has been pre-processed and well-annotated. The Indonesian syntaxis and semantic problems contained in this dataset are quite complex, and some have not yet been discussed. For example, how can we translate a word into a sentence where the word can be an affix word and can also be a basic word, where each one has a different meaning based on the context of the sentence. Therefore, this dataset can be used as input to test the performance of the Indonesian machine translation method.

2.2. Morphology extraction

Indonesian morphology extraction [16] identifies a word's prefixes and suffixes based on the alternation scheme of nouns, adjectives, verbs, and numbers. This identification is used to analyze the form of words accurately and in accordance with the context of a sentence, whether the word is a complete word or a word with a prefix or suffix. The alternation scheme for Indonesian nouns, adjectives, verbs, and numerals is shown in Table 1.

2.3. Metaphor corpus

The metaphor corpus, which consists of 2508 Indonesian phrases, is extracted from the Indonesian thesaurus dictionary [17]. Table 2 shows the metaphor corpus representation. Table 2 shows that the three terms listed can have synonyms, which can be denotations or connotations.

2.4. Word feature extraction

Generally, word feature extraction aims to capture the meaningful characteristics of words, phrases, or documents in a way that algorithms can interpret and process. The deep contextualized word representations are introduced to effectively model complex characteristics of word use (*e.g.*, syntax and semantics) and the variations of these words in the context (*i.e.*, to model polysemy), thereby generating context information [18]. In this research, the word feature extraction functions to identify contextual words in sentences and extract terms more accurately based on synonyms. For this reason, this research uses a combination of BERT embedding methods, name entity recognition, synonym-based term expansion, and semantic similarity.

Table 1. Alternation scheme for morphology extraction

Alternation		Pre-prefix	Prefix	Lemma	Suffix
Noun	ε+		ε+		+ε
			pen+		
			ke+		
			per+	Lemma	
			ke+		
			tidak+		
Adjective	ε+		ε+		+ε
	non+		*ter+*		*+an*
			ke+	Lemma	*+nya*
			se+		
Verb		ε+	ε+		+ε
		men+	*per+*		*+kan*
		di+		Lemma	*+i*
		ber+			
Numeral			ε+		+ε
			ke+		*+nya*
			ber+	Lemma	*+belas*

Table 2. Metaphor corpus representation

Terms	Synonyms
Bunga	*Kembang* (flower), *ornament* (ornament), *perempuan cantic* (beautiful women), *embel-embel* (frill)
Bunga api	*Cetusan* (spark), *kilatan api* (fire flash), *lelatu* (flash)
Bunga hati	*belahan jiwa* (soul-mate), *kekasih* (lover), *kesayangan* (beloved)

2.4.1. BERT embedding

Bidirectional encoder representations from transformers (BERT) [19], [20] functions to train deep two-way representation of unlabeled text by co-conditioning the left and right contexts across all layers. BERT uses a corpus of 3,300 million words for the pre-training process. Then, the fine-tuning process functions to model single text extraction tasks or text pairs by swapping the appropriate input and output. For applications involving text pairs, BERT uses the self-attention mechanism to encode the combined text pairs by effectively implementing the self-attention mechanism that simultaneously includes two-way cross-attention between two sentences.

2.4.2. Named entity recognition

The named entity recognition (NER) approach for the Indonesian language [21] is based on a set of rules capturing the contextual, morphological, and part-of-speech knowledge necessary to recognize named entities in Indonesian texts. The Singgalang dataset [22] is a NER dataset comprising 48,957 sentences. The dataset is divided into four named entity classes: person names, place names, organization names, and others.

2.4.3. Synonym-based term expansion

The synonym-based term expansion [23] captures the existing term synonym to get the best accuration of the term according to the sentence context. The synonym-based term expansion identifies each term that can indicate homonym and polysemy features based on the extracted synonym term with the highest similarity value to the term vectors. For example, synonym-based term expansion [12] expands the term *malang*, to include synonyms like *Malang* (city name), and *sial* (unlucky).

2.4.4. Semantic similarity

Semantic similarity using Cosine [24] predicts word similarity between the two input words. Cosine calculates the similarity of two words based on the meaning using (1), where w_a denotes the word 1, w_b denotes the word 2, w_{ai} denotes the vector member from w_a, and w_{bi} denotes the vector member from w_b.

$$cosine(w_a, w_b) = \frac{\sum_{i=1}^{n} w_{ai}\, w_{bi}}{\sqrt{\sum_{i=1}^{n}(w_{ai})^2}\sqrt{\sum_{i=1}^{n}(w_{bi})^2}} \quad (1)$$

2.5. Neural machine translation

The NMT [4] trains data on a large scale through an artificial neural network to build a model for the translation process. The NMT stages consist of embedding, encoding, attending, and decoding. The NMT softmax layer for calculating the score to generate the target word as output P is shown in (2), where $x = [x_1, \ldots, x_n]$ denotes the input source sentence, y_j denotes the current generated word, $y_{<j}$ denotes the previously generated words, s_j denotes the decoder hidden state for the current word, y_{j-1} denotes the embedding of the previous word, c_j denotes the context vector and f denotes the feedforward layer.

$$\mathrm{P}\left(y_j \middle| y_{<j}, x\right) = SoftMax\left(f\left(s_j, y_{j-1}, c_j\right)\right) \quad (2)$$

The NMT approach has been widely used in various language translations, including the Indonesian language [9]. However, NMT [4], [9] needs to be developed in the embedding task to capture the target word translation containing certain Indonesian language features, such as morphology, homonyms, and polysemy. The leading translation task is translating words based on lexical and phrase semantic meaning. The semantic distance measurement between the term source and translated term target is needed to capture the context of the word or phrase translation more accurately.

3. METHOD

This research develops a word feature approach of homonym and polysemy with morphology extraction in term weighting. The purpose is to enhance the performance of Indonesian to English machine translation. This section explains the proposed dataset, proposed morphology extraction, proposed word feature extraction, proposed machine translation, and evaluation.

3.1. Proposed dataset

As the proposed dataset, the Indonesian dataset [3] consists of 1,000 train data and 200 test data. The dataset had been pre-processed and properly annotated manually by Indonesian language experts. This research uses this dataset to test the proposed MT performance. The dataset representation is shown in Table 3.

Table 3. Proposed dataset representation

ID	Sentences
0	*Ayah pergi ke Palu* (Father goes to Palu)
1	*Ayah membeli palu* (Father buys a hammer)
2	*Anda perlu beruang banyak untuk membeli mobil ini* (You need to have a lot of money to buy this car)
3	*Kami menikmati pertunjukan beruang yang luar biasa* (we enjoy a great bear show)
4	*Aroma teh bunga melati ini sungguh nikmat* (The aroma of this jasmine flower tea is really delicious)
5	*Saya menikmati suasana outdoor yang indah dengan bunga hatiku* (I enjoy the beautiful outdoor atmosphere with my lover)

3.2. Proposed word feature extraction with morphology extraction

First, the proposed morphology extraction uses the concept of morphology with four alternation schemes: noun alternation, adjective alternation, verb alternation, and numeral alternation [16]. Figure 1 shows the proposed morphology extraction flow chart to extract lemma features and derivatives based on prefixes, lemmas, and suffixes. The tokens from the input sentences are generated using the tokenization process. The token results are processed through the alternation scheme: suffixes extraction and prefixes extraction. The results of suffix and prefix extraction are processed in the lemma update process to obtain morphological term identification results. Then, lemmatization is carried out to extract modified tokens. Finally, we save the modified token results. Table 4 illustrates sentences 3 and 5 morphology extraction.

Second, the proposed word feature extraction uses BERT embedding [19], synonym-based term expansion [23], and Semantic similarity [25] to obtain homonym and polysemy terms. The flowchart of the proposed word feature extraction is shown in Figure 2. We use sentences, identified morphology terms, and modified token results as the input. We extract the synonym terms using Semantic similarity and BERT based on the term with the highest similarity value through the Singgalang entity corpus and the thesaurus synonym corpus. Moreover, in the synonym extraction process, if there is an extracted prefix *ber-*, the term *ber-* is modified to be the term *mempunyai* (has/have). Then, if there are bi-gram term extraction results and the bi-gram term exists in the metaphor corpus, take the synonym term from the metaphor corpus with the highest similarity value to be the extracted synonym term. Furthermore, BERT and Semantic similarity measure the similarity between existing terms, modified tokens, and extracted synonym terms. Finally, the extracted term result with the highest similarity is used to substitute the existing terms in the input sentence. Table 5 illustrates sentences 3 and 5 word feature extraction.

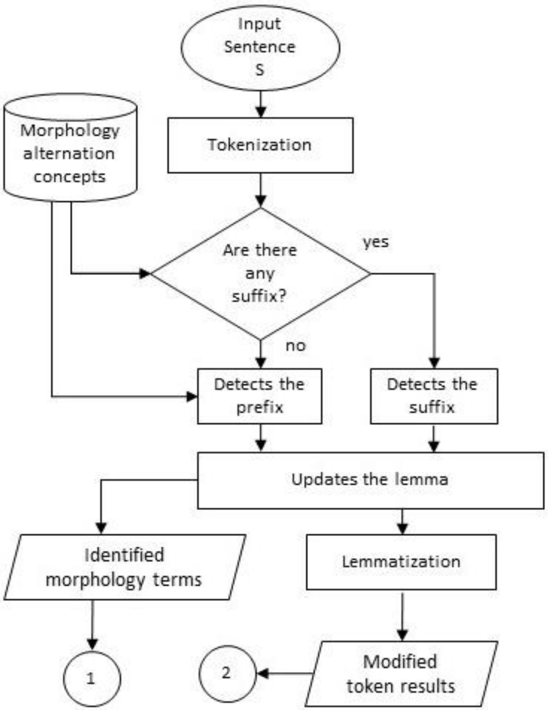

Figure 1. Proposed morphology extraction

Table 4. Illustration of morphology extraction

ID	Sentences	Tokens	Identified morphology terms	Modified tokens
3	*Kami menikmati pertunjukan beruang yang luar biasa*	*menikmati, pertunjukan, beruang*	*me+nikmat+i, per+tunjuk+an, ber+uang*	*menikmati, pertunjukan, beruang*
5	*Saya menikmati suasana outdoor yang indah dengan bunga hatiku*	*menikmati, suasana, outdoor, bunga, hatiku*	*me+nikmat+i, hati+ku*	*menikmati, suasana, outdoor, bunga, hati*

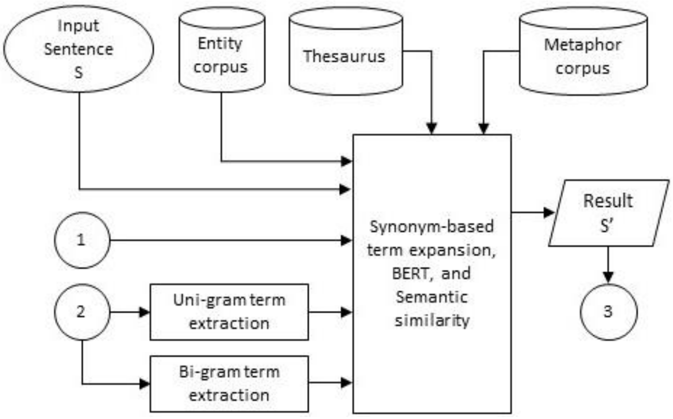

Figure 2. Proposed word feature extraction

Table 5. Illustration of word feature extraction

ID	Sentences	Identified morphology terms	Modified tokens	Identified terms of homonym and polysemy	Modified terms	Similarity score between identified terms and modified tokens	Similarity score between modified terms and modified tokens	Results
3	*Kami menikmati pertunjukan beruang yang luar biasa*	*me+nikmat+i, per+ tunjuk+an, ber+uang*	*menikmati, pertunjukan, beruang*	*beruang*	*mempunyai, uang*	**0.7363**	0.7259	*Kami menikmati pertunjukan beruang yang luar biasa*
5	*Saya menikmati suasana outdoor yang indah dengan bunga hatiku*	*me+nikmat+i, hati+ku*	*menikmati, suasana, outdoor, bunga, hati*	*bunga hati*	*kekasih*	0.6571	**0.7557**	*Saya menikmati suasana outdoor yang indah dengan kekasihku*

3.3. Proposed machine translation

The algorithm of the proposed MT method using NMT [4] is shown in Table 6. First, the word extraction result is used as the input. In the encoding steps, the input sequence will be converted into a fixed-length representation or a series of vectors that capture the meaning of the input. Then, the attention mechanism computes a context vector as a weighted sum of all encoder hidden states. The weights are determined by alignment scores, which measure the relevance of each encoder's hidden state to the current decoding step. The alignment scores are normalized using a SoftMax function to produce a probability distribution over the encoder's hidden states. The decoding generates the output sequence one token at a time, using the context vectors from the attention mechanism and the previous decoder’s hidden states. Finally, the translation result is saved.

Table 6. The algorithm of the proposed machine translation

Input: Word feature extraction result
1. Taking the word feature extraction results as the input.
2. Encoding
3. Attending.
4. Decoding.
5. Save the translation result.

3.4. Evaluation

The evaluation process aims to measure the performance of the proposed method. The evaluation uses a confusion matrix to calculate scores of precision (P), recall (R), F1-measure (F), and accuracy (A) [25]. The precision, recall, F1-measure, and accuracy are calculated using (3), (4), (5), and (6), respectively.

$$P = \frac{TP}{TP+FP} \tag{3}$$

$$R = \frac{TP}{TP+FN} \tag{4}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{5}$$

$$A = \frac{TP+TN}{TP+FP+TN+FN} \tag{6}$$

4. RESULTS AND DISCUSSION

In this section, the results of the proposed morphology extraction and proposed word feature extraction are explained first. Then, the proposed MT and Google Translator results are also compared. Finally, the performance results are compared with other performance methods.

4.1. Word feature extraction result

The homonym and polysemy features can be adequately extracted. Table 7 shows the illustration result of the word feature extraction for six sentences. The highest similarity score of sentence 0 was 0.7446 for *palu*, and *palu* is used to update the result of the modified sentence. The highest similarity score of

Sentence 1 was 0.7283 for *martil*, and *martil* is used to update the result of the modified sentence. The highest similarity score of sentence 2 was 0.7711 for *mempunyai uang*, and *mempunyai uang* is used to update the result of the modified sentence. The highest similarity score of sentence 3 was 0.7363 for *beruang*, and *beruang* is used to update the result of the modified sentence. The highest similarity score of sentence 4 was null because there was no identified term, and the result of the modified sentence was the same as the input sentence. The highest similarity score of Sentence 5 was 0.7557 for *kekasih*, and *kekasih* is used to update the result of the modified sentence.

Table 7. Illustration results of word feature extraction

ID	Tokens	Identified terms	Comparison similarity between terms and tokens				Modified sentence results
			Identified		Modified		
			Terms	Score	Terms	Score	
0	*pergi, palu*	*palu*	*palu*	**0.7446**	*martil*	0.7283	*Ayah pergi ke Palu*
1	*membeli, palu*	*palu*	*palu*	0.6694	*martil*	**0.7283**	*Ayah membeli martil*
2	*perlu, beruang, membeli, mobil*	*beruang*	*beruang*	0.7548	*mempunyai, uang*	**0.7711**	*Anda perlu mempunyai uang banyak untuk membeli mobil ini*
3	*menikmati, pertunjukan, beruang*	*beruang*	*beruang*	**0.7363**	*mempunyai, uang*	0.7259	*Kami menikmati pertunjukan beruang yang luar biasa*
4	*aroma, teh, bunga, melati, nikmat*	null	null	null	null	null	*Aroma teh bunga melati ini sungguh nikmat*
5	*menikmati, suasana, outdoor, bunga, hati*	*bunga hati*	*bunga hati*	0.6571	*kekasih*	**0.7557**	*Saya menikmati suasana outdoor yang indah dengan kekasihku*

The use of the morphology alternation concept in word feature extraction can well capture the morphological features of homonym terms (e.g., *beruang* to be *ber+uang*) and polysemy terms (e.g., *bunga hatiku* to be *bunga hati+ku*). Then, BERT embedding and Semantic similarity expand terms through synonyms based on terms in the entity corpus (e.g., *palu* has synonym *martil*) and thesaurus (e.g., *bunga hati* has synonym *kekasih*). The retrieval of modified term results for modified sentences is done by measuring the similarity value between the identified term and the extracted token. The resulting term with the highest similarity value is used as a modifier for the input sentence. This modified sentence is then used as input for the proposed machine translation.

The model performance measurement results are shown in Table 8. Table 8 compares the average accuracy results of the three models: semantic similarity, BERT embedding+semantic similarity, and morphology extraction+synonym-based term expansion+BERT embedding+semantic similarity. Performance testing of 200 test data was carried out using 50 epochs. The best model results obtained were the proposed method model using morphology extraction, synonym-based term expansion, BERT embedding, and Semantic similarity with an average accuracy value of 0.81.

Table 8. Model average accuracy of word extraction

Method	Average accuracy
Semantic similarity	0.73
BERT embedding+Semantic similarity	0.78
Morphology extraction+Synonym-based term expansion+BERT embedding+Semantic similarity	0.81

4.2. Comparison results of machine translations

The previous method using NMT, Google Translator, can adequately translate the sentences 1, 3, and 4 but cannot correctly translate the sentences 0, 2, and 5. In Sentence 0, Google Translator still translates *palu* as the *hammer*, which means a tool, and it cannot translate the correct term *palu* as the city name *Palu*. In Sentence 2, Google Translator still translates *beruang* as the *bear*, which means an animal, and cannot translate the correct term *beruang* as the alternation verb *ber+uang*, which means *have money*. In Sentence 5, Google Translator still translates *bunga hatiku* as the *flowers of my heart*, which means a denotation term, and cannot translate the correct term *bunga hatiku* as *my lover*, which means a connotation term.

We compare our proposed MT with Google Translator. The word feature extraction with morphology extraction can significantly enhance NMT performance. The proposed MT method can translate correctly for six sentences 0, 1, 2, 3, 4, and 5. The problems of translation tasks from Indonesian to English for input sentences containing homonym and polysemy features with morphological features can be solved by the proposed method. However, the proposed word feature extraction method still measures the similarity of an identified term to the token of a sentence. We have not tested this approach for multiple homonym and polysemy terms in a sentence. Then, the MT results in determining the verb forms of the target sentence are still different from the actual result that the expert determines. For example, in the input sentence 1 *Ayah membeli palu*, the expert determines the result is *Father buys a hammer*, while the proposed method determines the result is *Father bought a hammer*. The verb *membeli* as the input is still translated to *bought* by the proposed method and is not the same as the translated result by the expert, namely *buys*.

The differentiation occurs because the expert defines the verb form based on the existing adverb of time in the sentence. If the sentence does not contain any adverb of time, then the expert labels the sentence as the present tense, which has the verb in the first form. Meanwhile, the proposed method and Google Translator define the sentence without an adverb of time, which might be present tense or past tense because of the default assumptions made by the model. In this research, the differentiation of translation results between doing by expert and doing by the proposed method for determining the verb form in the target language based on the input sentence without the adverb of time is assessed as reasonable and has true value.

4.3. Evaluation result

The comparison of evaluation results is shown in Table 9. Table 9 shows the proposed MT method can perform better than the previous methods. The proposed word feature extraction using homonym and polysemy approach with morphology extraction determines good performance to increase NMT performance. The evaluation results are 0.8357 for precision, 0.7862 for recall, 0.8102 for the F-1 measure, and 0.8098 for accuracy.

Table 9. Comparison results of machine translation performances

Method	Language translation	P	R	F	A
NMT [9]	Indonesian to English	0.7292	0.6539	0.6895	0.7075
Hybrid RBMT and SMT [3]		0.7231	0.7000	0.7114	0.7417
NMT with homonym and polysemy extraction [12]		0.7791	0.8428	0.8097	0.7975
Proposed method		0.8357	0.7862	0.8102	0.8098

The NMT method [9], with an accuracy score of 0.7075, has been developed using the NMT method with homonym and polysemy extraction [12]. The accuracy score of the NMT method with homonym and polysemy extraction [12] is 0.7975 and is better than the accuracy score result of the NMT method [9]. Meanwhile, the results of other methods using hybrid RBMT and SMT [3] could still get an accuracy score of 0.7417. The accuracy score result of the proposed method of 0.8098 shows that this proposed method is better than the previous methods [3], [9], [12] in translating Indonesian sentences that contain morphological features in the homonym and polysemy features. However, the proposed MT method still works on one-way translation from Indonesian to English and has not worked on the reverse translation from English to Indonesian for a similar case. Apart from that, there are still false positives for translation results from input sentences containing the same terms but with different contexts. For example, in the sentence *Dia membeli palu di palu*, where the result of the proposed method is wrong, *He went to buy a hammer on a hammer*.

Overall, the performance of the proposed method has shown significant results in translating Indonesian sentences into English. Identifying homonym and polysemy features in Indonesian, whether they contain morphological features or not, is important in the word embedding stage in machine translation so that it can increase translation accuracy. Development of the MT method is still needed to translate sentences that contain the same two terms but have different meanings. Moreover, the sentences still labeled as false positive and false negative need to be analyzed. So, further MT method development can work better.

5. CONCLUSION

We propose homonym and polysemy approaches with morphology extraction in weighting terms for Indonesian to English MT. The proposed morphology extraction uses morphology alternation schemes that can adequately detect terms of prefixes and suffixes. The morphology extraction result is used as input for word feature extraction using metaphor corpus, entity corpus, BERT embedding, synonym-based expansion, and Semantic similarity, which can extract proper terms and different synonyms to get more accurate terms based on the context of each sentence. The word feature extraction result can detect the homonym and polysemy terms in the form of words and phrases with morphology features, which can have multiple meanings, denotation, and connotation meanings. The word feature extraction result is then used as input for machine translation. Finally, the proposed machine translation using NMT can correctly translate the sentences that contain homonym and polysemy features using updated sentences from the proposed word extraction results. The evaluation results were better than the previous methods, with values of 0.8357 for precision, 0.7862 for recall, 0.8102 for F1-measure, and 0.8098 for accuracy. For future research, this proposed method needs to be developed to solve the other problems in Indonesian to English translation, such as how to determine the precise translation of the English target verb based on the adverb of time that is contained in the sentence, how to translate two similar terms in a sentence that has different meanings, and the other Indonesian sentence features. We also hope these research results can support machine translation development from English to Indonesian translation or the Indonesian language with another language.

ACKNOWLEDGEMENTS

This research is supported by Universitas Dian Nuswantoro and the Directorate of Research and Community Service, Directorate General of Research and Development Strengthening of the Ministry of Research, Technology and Higher Education of the Republic of Indonesia.

REFERENCES

[1] F. Rahutomo, A. A. Septarina, M. Sarosa, A. Setiawan, and M. M. Huda, “A review on Indonesian machine translation,” in *4th Annual Applied Science and Engineering Conference*, 2019, pp. 1–6, doi: 10.1088/1742-6596/1402/7/077040.

[2] I. Rivera-trigueros, “Machine translation systems and quality assessment: a systematic review,” *Language Resources and Evaluation*, vol. 56, no. 2, pp. 593–619, 2022, doi: 10.1007/s10579-021-09537-5.

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

[4] W. Jooste, R. Haque, and A. Way, “Philipp Koehn: neural machine translation,” *Machine Translation*, vol. 35, no. 2, pp. 289–299, 2021, doi: 10.1007/s10590-021-09277-x.

[5] F. Stahlberg, “Neural machine translation: A review,” *Journal of Artificial Intelligence Research*, vol. 69, pp. 343–418, 2020, doi: 10.1613/JAIR.1.12007.

[6] M. Groves and K. Mundt, “Friend or foe? Google translate in language for academic purposes,” *English for Specific Purposes*, vol. 37, pp. 112–121, 2015, doi: 10.1016/j.esp.2014.09.001.

[7] S. C. Tsai, “Using google translate in EFL drafts: a preliminary investigation,” *Computer Assisted Language Learning*, vol. 32, no. 5–6, pp. 510–526, Jul. 2019, doi: 10.1080/09588221.2018.1527361.

[8] S. Mall and U. C. Jaiswal, “Shallow parsing and word sense disambiguation used for machine translation from Hindi to English languages,” *International Journal of Intelligent Engineering and Systems*, vol. 10, no. 3, pp. 381–390, 2017, doi: 10.22266/ijies2017.0630.43.

[9] M. Dwiastuti, “English-Indonesian neural machine translation for spoken language domains,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 2019, pp. 309–314, doi: 10.18653/v1/P19-2043.

[10] A. Eka, P. Lestari, A. Ardiyanti, and I. Asror, “Phrase based statistical machine translation Javanese-Indonesian,” *Jurnal Media Informatika Budidarma*, vol. 5, no. 2, pp. 378–386, 2021, doi: 10.30865/mib.v5i2.2812.

[11] Z. Abidin, Permata, and F. Ariyani, “Translation of the Lampung language text dialect of Nyo into the Indonesian language with DMT and SMT approach,” *INTENSIF*, vol. 5, no. 1, pp. 58–71, 2021, doi: 10.29407/intensif.v5i1.14670.

[12] R. Abdullah, R. Sarno, D. Purwitasari, A. I. Akhsani, and Suhariyanto, “Homonym and polysemy approaches in term weighting for Indonesian-English machine translation,” in *2023 14th International Conference on Information and Communication Technology and System, ICTS 2023*, 2023, pp. 232–237, doi: 10.1109/ICTS58770.2023.10330875.

[13] S. Skoufaki and B. Petrić, “Exploring polysemy in the academic vocabulary list: a lexicographic approach,” *Journal of English for Academic Purposes*, vol. 54, 2021, doi: 10.1016/j.jeap.2021.101038.

[14] S. Li, R. Pan, H. Luo, X. Liu, and G. Zhao, “Adaptive cross-contextual word embedding for word polysemy with unsupervised topic modeling,” *Knowledge-Based Systems*, vol. 218, 2021, doi: 10.1016/j.knosys.2021.106827.

[15] B. Casas, A. Hernández-Fernández, N. Català, R. Ferrer-i-Cancho, and J. Baixeries, “Polysemy and brevity versus frequency in language,” *Computer Speech and Language*, vol. 58, pp. 19–50, 2019, doi: 10.1016/j.csl.2019.03.007.

[16] S. D. Larasati, V. Kuboň, and D. Zeman, “Indonesian morphology tool (MorphInd): towards an Indonesian corpus,” *Communications in Computer and Information Science*, 2011, doi: 10.1007/978-3-642-23138-4_8.

[17] 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*, Nov. 2018, pp. 75–80, doi: 10.1109/ICOICT.2018.8528812.

[18] C. Sun *et al.*, “A deep learning approach with deep contextualized word representations for chemical-protein interaction extraction from biomedical literature,” *IEEE Access*, vol. 7, pp. 151034–151046, 2019, doi: 10.1109/ACCESS.2019.2948155.

[19] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2019.

[20] N. Reddy, P. Singh, and M. M. Srivastava, “Does BERT understand sentiment? leveraging comparisons between contextual and non-contextual embeddings to improve aspect-based sentiment models,” *arXiv preprint arXiv:2011.11673*, 2020.

[21] M. O. Ibrohim and I. Budi, “Multi-label hate speech and abusive language detection in Indonesian Twitter,” in *Proceedings of the third workshop on abusive language online*, 2019, pp. 46–57, doi: 10.18653/v1/w19-3506.

[22] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, “Hate speech detection in the Indonesian language: A dataset and preliminary study,” *2017 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017*, pp. 233–238,

2017, doi: 10.1109/ICACSIS.2017.8355039.

[23] I. Ghozali, K. R. Sungkono, R. Sarno, and R. Abdullah, “Synonym based feature expansion for Indonesian hate speech detection,” *International Journal of Electrical and Computer Engineering*, vol. 13, no. 1, pp. 1105–1112, Feb. 2023, doi: 10.11591/ijece.v13i1.pp1105-1112.

[24] P. Xia, L. Zhang, and F. Li, “Learning similarity with cosine similarity ensemble,” *Information Sciences*, vol. 307, pp. 39–52, 2015, doi: 10.1016/j.ins.2015.02.024.

[25] B. Harjo, Muljono, and R. Abdullah, “Attention-based sentence extraction for aspect-based sentiment analysis with implicit aspect cases in hotel review using machine learning algorithm, semantic similarity, and BERT,” *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 3, pp. 189–200, 2023, doi: 10.22266/ijies2023.0630.15.

BIOGRAPHIES OF AUTHORS

Budi Harjo received the Doctor (Ph.D.) degree in informatics engineering from the Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember (ITS). He is a lecturer at the Faculty of Computer Science, Dian Nuswantoro University, Semarang. His research includes artificial intelligence, machine learning, data mining, data science, computational linguistics, and natural language processing. He can be contacted at email: budi.harjo@dsn.dinus.ac.id.

Muljono holds a Doctor of Electrical Engineering degree from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia in 2016. He joined an Internship Program at the School of Media Science, Tokyo University of Technology Japan in 2014. He received his Magister of Computer (Informatics) from STTIBI Jakarta, Indonesia in 2001, and he received his B.Sc (Mathematics) from Universitas Diponegoro (UNDIP) in 1996. He is currently an associate professor at the Faculty of Computer Science at Dian Nuswantoro University, Semarang, Indonesia. His research includes artificial intelligence, machine learning, data mining, data science, computational linguistics, and natural language processing. He can be contacted at email: muljono@dsn.dinus.ac.id.

Rachmad Abdullah received the M.MT. degree in information technology management from Institut Teknologi Sepuluh Nopember (ITS). He is currently pursuing a Ph.D. degree in informatics engineering from the Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember (ITS). His research interests include data science, artificial intelligence, natural language processing, focusing text mining, sentiment analysis, machine learning, and the internet of things (IoT). He can be contacted at email: rabdullah1506@gmail.com.