Homonym and Polysemy Approaches in Term Weighting for Indonesian-English Machine Translation

Rachmad Abdullah [a, b], Riyanarto Sarno [a], Diana Purwitasari [a], Alifa Izzan Akhsani [a], Suhariyanto [a, c]
[a] *Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[b] *Sekolah Tinggi Ilmu Ekonomi Indonesia, Surabaya, Indonesia*
[c] *Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia*
Email: rabdullah1506@gmail.com, izzan@mhs.if.its.ac.id, riyanarto@if.its.ac.id, diana@if.its.ac.id, suhariyanto@dsn.dinus.ac.id

2023 14th International Conference on Information & Communication Technology and System (ICTS) | 979-8-3503-1216-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICTS58770.2023.10330875

***Abstract*— Many previous studies have discussed extracting ambiguous sentences in the Indonesian language to increase text extraction performance, especially for improving the accuracy of Indonesian-English Machine Translation (MT). One of the factors that can influence the presence of ambiguous sentences is the word features of homonym and polysemy; however, not many extraction methods have been developed to solve these issues. This research proposes word feature extraction of homonyms and polysemy in Indonesian to improve Indonesian-English MT accuracy. First, POS tagging is used to obtain the word types in the sentences. To measure the word similarity for extracting homonyms and polysemy within uni-gram and bi-gram features, Word2vec and Synonym-based term expansion are used. The terms indicated as homonym and polysemy are compiled in dictionaries 1 and 2. Then, Semantic similarity is used to calculate word similarity for extracting the highest similarity value as the updated terms between existing terms in the sentence and extracted synonym terms. Neural Machine Translation (NMT) is used for the translation process, which has two translation steps as the proposed MT: NMT and NMT with proposed word feature extraction. The proposed MT modifies the translation result to obtain a more accurate translation result according to the two translation results. Finally, the proposed MT performance is evaluated. The evaluation results based on precision, recall, f-1 measure, and accuracy, respectively, are 0.7791, 0.8428, 0.8097, and 0.7975.**

Keywords*— Machine Translation, Word features extraction, Word2vec, Synonym-based term expansion, Semantic similarity.

I. INTRODUCTION

Many Indonesian Machine Translation (MT) studies work on disambiguous sentence features but not on ambiguous sentence features [1]. The ambiguous sentences [2] occur because the adjectives cannot be identified or paired directly with the target noun using the dependency parse. One of the backgrounds of the ambiguous sentences in Indonesian is affected by homonym and polysemy word features. These factors can affect the existence of errors in many implementations using the Indonesian language, especially in Indonesian-English MT. For example, *malang* can mean a city name *malang* (malang) as a noun and an opinion *malang* (unlucky) as an adjective.

Many MT developments [1], [3], [4] can solve translation tasks in sentences with a single meaning. However, they do not solve the word feature translation of homonyms and polysemy that influence a term with multiple meanings. For example, *Bapak rekreasi ke Kota Malang* (Father recreation to Malang City), where previous studies still extract *Malang* (Poor) as an adjective and have not been able to extract *Malang* (Malang) as a noun because of inaccurate POS tagging results. This problem is due to an error in capturing the context of the word in the sentence, impacting POS tagging errors. Another example is *bis itu jatuh terpeleset ke jurang* (The bus slipped into a ravine); where previous studies still extract *jatuh terpeleset* (slipped) as two verbs *jatuh* (fell) and *terpeleset* (slipped). In fact, *jatuh terpeleset* (slipped) is the verb phrase and has not been discussed. Therefore, the developed MT is needed to capture homonym and polysemy features which can be a word or a phrase.

Polysemy research works based on syllable and phonemic lengths [5], relies on lexicon and corpus data [6], and deals with large-scale corpus [7]. The results achieved higher precision and F1 in sentence classification by considering polysemy word elements. They have not explored the structural differences between common and rare words. They should apply the methodology to other quantities, such as contextual diversity, that may be more relevant than just word frequency in some lexical assignments. However, they are limited to high correlation or hierarchy.

Indonesian MT [8]–[12] worked on translation problems based on Indonesian syntaxis and semantics but has not focused on a word that can have different meanings based on the type and function of the word itself. Phrase translation [8] has not yet focused on translating noun and verb phrases specifically, which have a translation output different from the input. Translation of speech fluency [9] is still limited to using a lexicon. Translation of spoken language [12] is still limited to detecting synonyms. Then,

parallel corpus [10], [11] to detect broader words and phrases are still limited to detecting word synonyms.

Indonesian MT [3] can translate the case of a word based on the type and function of different words in several sentences. However, this study has not discussed cases of translations containing homonyms and polysemy elements. Indonesian word extraction for polysemy case by transforming the words into uni-gram and bi-gram [13] has been worked. The word similarity tasks still depend on the result of POS tagging. Furthermore, the tasks impact errors in the word classification task. The impact is that they cannot take accurate words in ambiguous sentences with homonym and polysemy cases. In addition, these studies have not been able to distinguish the context of words based on the meaning of each word in the sentence.

This research proposes homonym and polysemy approaches in term weighting for Indonesian-English MT. First, we use POS tagging to obtain the word type in the dataset sentences. We measure the word similarity for extracting homonyms and polysemy within uni-gram and bi-gram features using Word2vec and Synonym-based term expansion. The terms indicated as homonym and polysemy are compiled in dictionaries 1 and 2. Then, we use Semantic similarity to calculate word similarity for extracting the highest similarity value as the updated terms between existing terms in the sentence and extracted synonym terms in the sentence. Then, the translation process uses Neural Machine Translation (NMT). We use two translation steps: NMT and NMT with proposed word feature extraction. We modify the result to obtain a more accurate translation result according to the two translation results. Finally, we evaluate the proposed MT performance.

II. RELATED WORKS

This research explains several works as follows.

A. POS Tagging

The POS tagging process uses the FLAIR method [14] to get the initial tag of each term in the sentences. FLAIR can tag all Indonesian words. However, FLAIR has not been able to label homonym and polysemy words properly. Thus, we need a method that can capture homonym and polysemy features.

B. Word Features of Homonym and Polysemy

Indonesian word features [15] have word variables with the same writing characteristics that can affect word meaning and root structure differences. Table I shows the definition and example of homonym and polysemy word features.

C. Word Feature Extraction of Homonym and Polysemy

The vector representation using the skip-gram model [16] is an approach to predict the target and surrounding terms. The basic skip-gram formulation is defined using the softmax function, as shown in Eq. (1). The probability of output and input words is denoted by $p(w_O|w_I)$. The input v_w and the output v'_w are vector representations of the word *w*. At the same time, the number of words in the vocabulary is denoted by W. The synonym-based expansion [17] is used to catch the synonym of every word and phrase based on uni-gram and bi-gram. The synonym-based expansion functions to identify each word containing homonym and polysemy elements based on the type of synonym word with the highest similarity value to the word vectors in the sentence. Then, Cosine Similarity [18] is used to predict term similarity between the two input terms. Cosine similarity calculates the similarity of two words based on the meaning using Eq. (2), 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. Specifically, Eq. (2) functions to get synonyms based on words and phrases obtained from the results of Eq. (1).

TABLE I. DEFINITION OF HOMONYM AND POLYSEMY

Word Features	Definition	Example
Homonym	two or more words that have the same spelling or pronunciation but have different meanings.	*malang* (malang) and *malang* (unlucky)
Polysemy	two or more words that have the same spelling or pronunciation but have different meanings but are related to each other.	*jatuh terpeleset* (slipped) and *jatuh cinta* (fall in love)

$$p(w_O|w_I) = \frac{exp\left(v'^{\top}_{w_O} v_{w_I}\right)}{\sum_{w=1}^{W} exp\left(v'^{\top}_{w} v_{w_I}\right)} \quad (1)$$

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

D. Neural Machine Translation

Neural Machine Translation (NMT) [4] trains data on a large scale through an artificial neural network. Then, the training process results are modeled for the translation process to get accurate results. The NMT approach consists of embedding, encoding, attending, and decoding. The NMT approach has been widely used in various language translations, including the Indonesian language [12]. However, NMT [4], [12] needs to be developed in the embedding task to capture the target translation word that contains certain features in the Indonesian language, such as homonyms and polysemy, so the translation result is accurate. The NMT softmax layer for calculating the score to generate the target word as output P is shown in Eq (3), 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.

TABLE II. COMPARISON OF MACHINE TRANSLATION TASKS

No.	Semantics Features in the Sentence	Researchers	
		Dwiastuti [12]	Yamin [3]
	Lexicals		
1	Basic	✓	✓
2	Prefix	✓	✓
3	Suffix	✓	✓
	Phrases		
4	Neutral	✓	✓
5	Homonym	X	X
6	Polysemy	X	X

$$P(y_j|y_{<j}, x) = softmax\left(f(s_j, y_{j-1}, c_j)\right) \quad (3)$$

The main job of translating sentences is translating words based on lexical and phrase semantics. Of course, an approach is needed to measure the semantic distance between the term source and the translation term target so that errors do not occur in capturing the context of the lexical or phrase translation. A comparison of Indonesian MT development tasks for semantic cases of Indonesian sentences that have been done and that have not been done is presented in Table II.

Based on Table II, it can be seen that Dwiastuti [12] developed NMT [4] to translate Indonesian sentences based on lexical semantics (basic, prefix, and suffix) and phrase semantics (neutral) but has not been able to translate sentences containing homonym and polysemy. Apart from that, the translation based on lexical semantics has not focused on a word with different word functions and types. It can lead to errors in capturing the context of the meaning of the target word that is precise and accurate. Accordingly, the development of hybrid MT [3] has been carried out to translate lexical features based on word functions and types. However, this research has not worked on the case of sentences containing homonyms and polysemy features. For further MT development, it needs to solve semantic issues of lexical and phrases in Indonesian sentences, especially those containing homonym and polysemy features.

III. METHODOLOGY

This study works on developing word feature extraction of homonyms and polysemy to improve Indonesian-English Machine Translation accuracy.

A. Dataset

The Indonesian language dataset [3] is a dataset that contains universal sentences based on types of Indonesian sentence structure where the sentences contain words based on functions and types of Indonesian terms. This dataset was collected manually by Indonesian language experts by entering types of sentences that contain functions and types of words in Indonesian. This dataset consists of several types of sentences: simple, compound, complex, and compound-complex. The sentences contain several types of terms: neutral, homonym, and polysemy. The collection of this dataset aims to identify the extent to which Machine Translation (MT) performance can accurately translate Indonesian. The dataset consists of 1000 train data and 200 test data. The dataset had been pre-processed and properly annotated manually by Indonesian language experts. Indonesian language experts have classified this dataset into each sentence type based on its structure. Then, Indonesian language experts have also classified each word in a sentence based on its type and function. We use this dataset as a goal to test the performance of the proposed MT method to solve the cases of sentences containing homonyms and polysemy features.

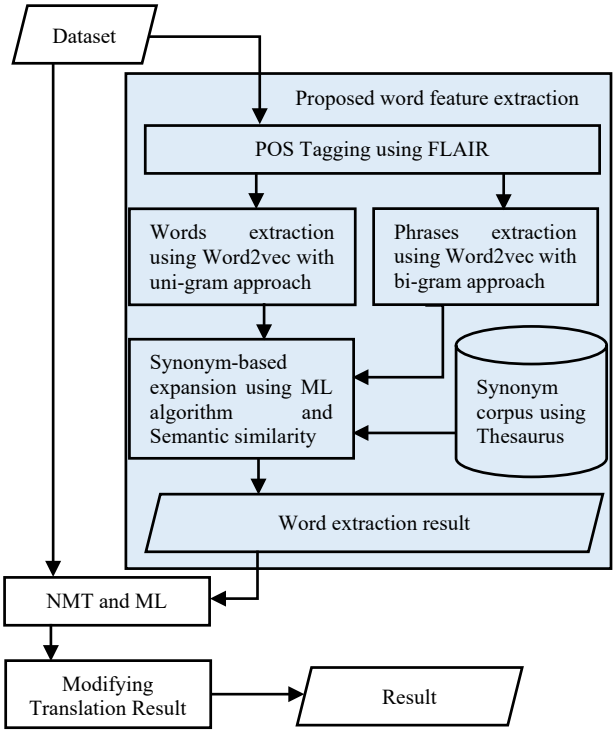

Fig. I. Proposed machine translation method

B. POS Tagging

FLAIR method [14] using embeddings of Token, Word, Stacked, and BERT has been used in many research for the POS tagging process through the 17 labels of universal POS. Table III shows that POS tagging illustrations of the four sentences describe the homonym and polysemy issue in Indonesian sentences. For example, *malang* has multiple meanings: *sial* (unfortunate) and *malang* (malang). Therefore, this POS tagging should be processed and developed to extract uni-gram features to capture the words and to extract bi-gram features to capture the phrases.

C. Proposed Word Feature Extraction

This research proposes word feature extraction using Word2vec, Synonym-based expansion, and Semantic similarity. The purpose is to extract the word features of homonym and polysemy. Figure I shows the flowchart of the proposed word feature extraction. We set the label tag result of PROPN and PRON as NOUN. Then, we extract the terms with uni-gram and bi-gram features from each sentence

Input: Indonesian Sentence S
Output: English Sentence S'
Start
1. Generate POS tag of S using FLAIR embedding;
2. Extract the word W using Word2vec with a uni-gram approach;
3. Extract the phrase P using Word2vec with a bi-gram approach;
4. When S has W and P that exist in synonym corpus, then:

a. If a W has the synonym W' with different word type x, then:
   i. Predict similarity between W' and W
   ii. Extract the highest similarity value as the synonym W'
   iii. Labelling W as homonym
   iv. Obtains W into dictionary 1

b. If a P consists of w_1 and w_2 is exist as an entity, then:
   i. If w_2 is exist as a synonym of P, then extract w_2 as the synonym P'
   ii. If w_2 is not exist as a synonym of P, then extract the highest similarity value as the synonym P'
   iii. If w_1 is exist in other P then labeling w_1 as the polysemy
   iv. Obtains w_1 into dictionary 2

5. When both W' and P' are extracted, then obtain P' as the synonym result;
6. Translate input S using NMT;
7. Translate input S from output proposed word feature extraction step using NMT;

a. If the term W' or P' is an existing token in synonym entities of people, animal, and city, then extracts W' or P' as the translation result
b. If the term W' or P' is not an existing token in synonym entities of people, animal, and city, then translates W' or P' using NMT

8. Detects the aggregate translation result of S' from similar S using NMT and NMT with proposed word feature extraction;
9. Modify S' by replacing the NMT result with NMT and word extraction result;
10. Obtain S' as the translation result.

End

Fig. II. Proposed machine translation algorithm

using Word2vec. Here, the uni-gram feature represents the form of a word, and the bi-gram feature represents the form of a phrase. In addition, we expand the terms with uni-gram and bi-gram features using synonym-based expansion to identify the other word types of terms. We develop rules based on two conditions for generating and updating terms and label tags of terms.

The rules based on two conditions are as follows.

1. *If the synonyms of the term bi-gram are exist*
   We measure the similarity between each synonym result and the sentence using semantic similarity to obtain the highest similarity value. The extracted synonym term with the highest similarity value to the sentence has represented the updating term. The synonym-based expansion result is used to substitute the term uni-gram with the updating term; substitute the existing label tag of terms with the label tag of the updating term.
2. *If the synonyms of the term uni-gram exist and the label tag is different from the label tag of the existing term*
   We extract the synonym with the highest similarity value as the updating term. Then, we mark the term uni-gram with the updating term and substitute the label tag of the existing uni-gram with the label tag of the updating term.

We extract the homonym features based on a term in which the number of label tag types is higher than one consisting of noun <NOUN>, verb <VERB>, adjective <ADJ>, adverb <ADV>, and auxiliary <AUX>. Then, we take the polysemy word feature based on a term with one label tag, extracted as the bi-gram, where the number of term context types is higher than one. Finally, the word features of homonym and polysemy are taken based on the number of existing label tags and the number of the term context type.

TABLE III. DATASET REPRESENTATION

ID	Sentences	Sentence types	Word types
S1	*Saya tinggal di Kota Surabaya*	Simple	Neutral
S2	*Bapak rekreasi ke Kota Malang*	Simple	Homonym
S3	*Bis itu jatuh terpeleset ke jurang*	Simple	Polysemy
S4	*Saya tinggal di Surabaya dan ibu tinggal di Jakarta*	Compound	Neutral
S5	*Malang dan Surabaya adalah kota destinasi liburan yang sangat diminati*	Compound	Homonym
S6	*Jatuh cinta dapat membuat bahagia dan sedih*	Compound	Polysemy
S7	*Saya bermain di taman ketika ibu memasak di dapur*	Complex	Neutral
S8	*Anak itu bernasib malang karena orang tuanya telah meninggal dunia*	Complex	Homonym
S9	*Bis itu tidak kuat menanjak sehingga jatuh terpeleset ke jurang*	Complex	Polysemy
S10	*Saya dan ibu tinggal di Surabaya ketika bapak tinggal di Jakarta*	Compound -complex	Neutral
S11	*Saya pergi sekolah dan ibu memasak ketika bapak pergi ke Malang*	Compound -complex	Homonym
S12	*Truk yang bermuatan pasir itu tidak kuat menanjak dan jatuh terpeleset ke jurang*	Compound -complex	Polysemy

D. Proposed Machine Translation

The proposed MT method system diagram is shown in Figure I. The dataset as the input is processed by two methods, using NMT and NMT with proposed word feature extraction. Here, NMT works in two stages. First, NMT uses Eq. (3) functions to translate the results of synonym extraction using Eq. (2). Second, NMT uses Eq. (3) to translate sentence input from the dataset. Then, the two results are modified using a Machine Learning (ML) algorithm to obtain more accurate translation results. Suppose there are terms in the proposed word feature extraction results. In that case, we modify the translation results by substituting the NMT results with the proposed word feature extraction and NMT as the translation results. Meanwhile, if there are no term results in the proposed word feature extraction, we take the NMT results as a translation result. The proposed MT algorithm is shown in Figure II.

IV. RESULT AND DISCUSSION

This section shows the result and discussion of the proposed method.

A. Dataset Representation

The dataset representation is shown in Table III. Table III shows examples of sentences based on the structure containing words with different functions and types. Then, Table IV shows details of the types of sentences and words from the training data.

TABLE IV. DETAILS OF THE TYPES OF SENTENCE AND WORD

No	Sentence types	Word types			Total
		Neutral	Homonym	Polysemy	
1	Simple	100	75	75	250
2	Compound	100	75	75	250
3	Complex	100	75	75	250
4	Compound-complex	100	75	75	250

B. Proposed Word Feature Extraction Result

After the POS tagging process, we extract the terms with uni-gram and bi-gram features. We carry out a synonym-based expansion to get synonyms and other types of words from the extracted terms and substitute existing terms in uni-gram using bi-gram. For instance, the synonym-based expansion results are shown in Table V. The results of the uni-gram expansion can be seen in the synonym result of the uni-gram column, where existing terms marked in bold can be expanded based on synonyms word to find out the existing types of the term. Then, the results of the bi-gram expansion can be seen in the synonym result of the bi-gram column.

Then, semantic similarity is carried out to measure the similarity of words between the results of the terms expansion and the sentences. For instance, the term *malang* <NOUN> in S2 was obtained as the *kota malang* <NOUN> context. Malang <NOUN> is extracted as a homonym, and *kota* <NOUN> is extracted as a polysemy. In S3, the term *terpeleset* <VERB> was obtained as the *jatuh terpeleset* <VERB> context. J*atuh* <VERB> is extracted as polysemy.

C. Proposed Machine Translation Result

The translation results of the proposed MT method are shown in Table V. Table V shows that NMT [4], [12] can not translate S2 properly because there was an error in translating the term *malang* marked in red. NMT has not distinguished the context of the term *malang* in S2. NMT can still capture the context of the term *malang* in S2 as an adjective.

TABLE V. DATASET REPRESENTATION AND RESULTS OF POS TAGGING, WORD FEATURE EXTRACTION, AND PROPOSED MT

ID	Sentences	POS Tagging Result	Synonym Result		Word Feature Result		Translation Result		
			Uni-gram	Bi-gram	Homonym	Polysemy	by expert	by NMT [4], [12]	by proposed method
S2	*Bapak rekreasi ke Kota Malang* (Father recreation to Malang City)	*Bapak* <NOUN> *rekreasi* <NOUN> *ke* <ADP> *Kota* <NOUN> *Malang* <PROPN>	-	***Kota Malang*** <NOUN>*: ["malang" <NOUN>: 0.6325]*	***Malang*** <NOUN>	***Kota*** <NOUN>	Father recreation to malang city	Father recreation to poor city	Father recreation to malang city
S3	*Bis itu jatuh terpeleset ke jurang* (The bus slipped into a ravine)	*Bis* <NOUN> *itu* <DET> *jatuh* <VERB> *terpeleset* <VERB> *ke* <ADP> *jurang* <NOUN>	-	***Jatuh terpeleset*** <VERB>*: [" terpeleset" <VERB>: 0.5774]*	-	***Jatuh*** <VERB>	The bus slipped into a ravine	The bus fell and slipped into a ravine	The bus slipped into a ravine

NMT also has not been able to translate S3 properly because there is an error in translating the phrase *jatuh terpeleset* marked in red. NMT has not grasped the meaning of *jatuh* in S3, which expresses *jatuh terpeleset*. NMT can still capture the context of the meaning of *jatuh terpeleset* as two different words by adding the conjunction *and* in the translation result, *fell and slipped*.

The proposed MT method can successfully translate Indonesian-English sentences containing semantics features: lexical (basic, prefix, and suffix) and phrase features (neutral, homonym, and polysemy). The proposed MT method can properly distinguish the meaning of the word contexts of *malang* and *jatuh*. The proposed MT method properly translates homonyms and polysemy words in Indonesian sentences into English. In S2, the word *malang* is captured as a phrase *kota malang* and can be translated correctly as *malang city*. In S3, the word *jatuh* is captured as a phrase *jatuh terpeleset* and can be correctly translated as *slipped*.

In the data testing performance result, the proposed MT method can properly translate sentences based on their structure, which contains homonym and polysemy features in words and phrases. Simple sentence translation can work well for cases where sentences consist of an adverb with complement (e.g., *malang sekali*), a pair of a proper noun and verb (e.g., *saya jatuh cinta*), a pair of a proper noun, verb, and adjective (e.g., *Michael bernasib malang*), and a pair of a proper noun, verb, adjective, and complement (e.g., *Bulan bersinar terang di malam ini*).

However, the proposed MT method can not properly translate several sentences consisting of compound, complex, and compound-complex that contain homonym features in the form of prefixes and suffixes. For example, the sentence *Pada umumnya, manusia akan hidup tenang dan bahagia ketika beruang banyak*, where *beruang banyak* is still translated as *a lot of bears*. *Beruang banyak,* in some contexts, is a prefix that means *has a lot of money*. The proposed MT method can still not translate sentence cases containing pairs of prefixes and phrases that indicate homonym features.

TABLE VI. COMPARISON OF MT EVALUATION RESULT

Method	Language Translation	P	R	F	A
NMT [12]	Indonesian to English	0.7292	0.6539	0.6895	0.7075
M Yamin [3]		0.7231	0.7000	0.7114	0.7417
Proposed method		**0.7791**	**0.8428**	**0.8097**	**0.7975**

D. Evaluation Result

The evaluation of the proposed MT using word feature extraction determines the best performance of the term list against homonym and polysemy word features. The evaluation result of the proposed MT for Precision (P), Recall (R), F-1 measure (F), and Accuracy (A), respectively, are 0.7791, 0.8428, 0.8097, and 0.7975. We also compared the evaluation test with two other studies, as shown in Table VI. The evaluation results showed that the proposed MT method performed better than the other two. The proposed MT method is proven to improve the accuracy of NMT performance.

V. CONCLUSION

We propose homonym and polysemy approaches in term weighting for Indonesian-English MT. The proposed word feature extraction using Word2vec, Synonym-based expansion, and Semantic similarity can extract words and phrases to identify the different types of words in each sentence. Then, semantic similarity can properly measure the similarity value between existing words and phrases with the synonyms to get the context terms based on the highest similarity value. We grouped words labeled homonyms in Dictionary 1 and polysemy in Dictionary 2.

Finally, we perform the proposed machine translation using NMT with the proposed word feature extraction. The proposed machine translation can properly translate sentences that contain homonym and polysemy word features. The evaluation results obtained better results than the other two methods with values of precision, recall, f-1 measure, and accuracy, respectively, of 0.7791, 0.8428, 0.8097, and 0.7975.

In future work, this proposed method needs to be developed to solve more complex homonym and polysemy problems in various other types of sentences to increase the evaluation value of the proposed method. We also hope that the results of this research can be used to support text extraction in other fields, such as sentiment analysis, hate speech detection, and others.

ACKNOWLEDGMENT

This research was funded by Institut Teknologi Sepuluh Nopember (ITS) under the Penelitian Dana Departemen Program, the Indonesian Ministry of Education and Culture under the Penelitian Terapan Unggulan Perguruan Tinggi (PTUPT) Program; and Institut Teknologi Sepuluh Nopember (ITS) under project scheme of the Publication Writing and IPR Incentive Program (PPHKI).

REFERENCES

[1] 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] B. S. Rintyarna, R. Sarno, and C. Fatichah, "Semantic features for optimizing supervised approach of sentiment analysis on product reviews," *Computers*, vol. 8, no. 3, 2019, doi: 10.3390/computers8030055.

[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," *Mach. Transl.*, vol. 35, no. 2, pp. 289–299, 2021, doi: 10.1007/s10590-021-09277-x.

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

[6] S. Skoufaki and B. Petrić, "Exploring polysemy in the Academic Vocabulary List: A lexicographic approach," *J. English Acad. Purp.*, vol. 54, 2021, doi: 10.1016/j.jeap.2021.101038.

[7] 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 Syst.*, vol. 218, 2021, doi: 10.1016/j.knosys.2021.106827.

[8] A. A. Suryani, D. H. Widyantoro, A. Purwarianti, and Y. Sudaryat, "Experiment on a phrase-based statistical machine translation using PoS Tag information for Sundanese into Indonesian," *2015 Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2015 - Proc.*, pp. 1–6, 2016, doi: 10.1109/ICITSI.2015.7437678.

[9] K. M. Shahih and A. Purwarianti, "Utterance disfluency handling in Indonesian-English machine translation," *4th IGNITE Conf. 2016 Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA 2016*, pp. 1–5, 2016, doi: 10.1109/ICAICTA.2016.7803104.

[10] A. Eka, P. Lestari, A. Ardiyanti, and I. Asror, "Phrase Based Statistical Machine Translation Javanese-Indonesian," *J. Media Inform. 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] M. Dwiastuti, "English-Indonesian Neural Machine Translation for Spoken Language Domains," in *Proceedings ofthe 57th Annual Meeting ofthe Association for Computational Linguistics: Student Research Workshop*, 2019, pp. 309–314, doi: 10.18653/v1/P19-2043.

[13] F. Husein Wattiheluw and R. Sarno, "Developing Word Sense Disambiguation Corpuses Using Word2vec and Wu Palmer for Disambiguation," *Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018*, pp. 244–248, Nov. 2018, doi: 10.1109/ISEMANTIC.2018.8549843.

[14] A. Akbik, S. Schweter, D. Blythe, and R. Vollgraf, "F LAIR : An Easy-to-Use Framework for State-of-the-Art NLP," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 54–59, doi: 10.18653/v1/N19-4010.

[15] A. B. Retnomurti, "English Homonym and Polysemy Words Through Semantic Approach: Novels Woy & The Dancer," *Deiksis*, vol. 13, no. 1, p. 21, 2021, doi: 10.30998/deiksis.v13i1.6608.

[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations ofwords and phrases and their compositionality," *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2013, doi: 10.48550/arXiv.1310.4546.

[17] I. Ghozali, K. R. Sungkono, R. Sarno, and R. Abdullah, "Synonym based feature expansion for Indonesian hate speech detection," vol. 13, no. 1, 2023, doi: 10.11591/ijece.v13i1.pp1-1x.

[18] P. Xia, L. Zhang, and F. Li, "Learning similarity with cosine similarity ensemble," *Inf. Sci. (Ny).*, vol. 307, pp. 39–52, 2015, doi: 10.1016/j.ins.2015.02.024.