6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)
29-31 August 2018, Gurugram, India

Corpus Construction and Semantic Analysis of Indonesian Image Description

Khumaisa Nur'aini[1,3], *Johanes Effendi*[1], *Sakriani Sakti*[1,2],
Mirna Adriani[3], *and Satoshi Nakamura*[1,2]

[1]Nara Institute of Science and Technology, Japan
[2]RIKEN, Center for Advanced Intelligence Project AIP, Japan
[3]Faculty of Computer Science, Universitas Indonesia, Indonesia

khumaisa.nuraini@ui.ac.id, {johanes.effendi.ix4, ssakti, s-nakamura}@is.naist.jp,
mirna@cs.ui.ac.id

Abstract

Understanding language grounded in visual content is a challenging problem that has raised interest in both the computer vision and natural language processing communities. Flickr30k, which is one of the corpora that have become a standard benchmark to study sentence-based image description, was initially limited to English descriptions, but it has been extended to German, French, and Czech. This paper describes our construction of an image description dataset in the Indonesian language. We translated English descriptions from the Flickr30K dataset into Indonesian with automatic machine translation and performed human validation for the portion of the result. We then constructed Indonesian image descriptions of 10k images by crowdsourcing without English descriptions or translations, and found semantic differences between translations and descriptions. We conclude that the cultural differences between the native speakers of English and Indonesian create different perceptions for constructing natural language expressions that describe an image.

Index Terms: Indonesian image description, corpus construction, semantic analysis

1. Introduction

Sentence-based image description has become an active research topic for both computer vision and natural language processing. Some applications with sentence-based image description datasets include automatic image description [1, 2], image retrieval based on textual data [3], and visual question answering [4]. To satisfy these studies, some available datasets contain images alongside human-generated English text description, including Flickr8K [5], Flickr30K [6], and MSCOCO [7]. In recent years, the English image descriptions in Flickr30K have been manually translated into German, French, and Czech. The resulting corpora can be utilized in multimodal machine translation [8, 9]. Some English datasets have also been extended to other languages, such as Japanese descriptions of MSCOCO [10], Chinese descriptions from Flickr8K [11], etc.

Sentence-based image description in a new language is manually constructed by human annotators, generally by either looking at the image and creating sentence descriptions that correspond to the pictures or translating the source languages into the target languages. However, manually collecting image descriptions is expensive and time-consuming. Image descriptions in target languages can also be extended using automatic text machine translation [12], where target image descriptions are automatically created, given the source language. Multimodal machine translation can also generate image descriptions in a target language using both image and its descriptions [8, 9].

Using translation methods, whether multimodal or other types, will create a new dataset of image descriptions in target languages that have identical meaning as the image description in the source languages. In other words, although using different languages, the semantic meaning of these two datasets is assumed to be identical, regardless of the differences in cultural background. However, neuroscience studies have found a difference in visual perceptions based on different cultural backgrounds [13, 14]. For example, European Americans generally pay more significant attention to foreground objects than East Asians who often focus more substantial attention on background objects [13]. Further study of the differences in visual perception is needed in the context of natural language expressions for describing an image.

This paper describes our attempt to construct an image description dataset in the Indonesian language. We translated the English descriptions from the Flickr30K dataset into Indonesian with automatic machine translation (denoted by "Eng2Ind_Translation") and performed human validation on a portion of the result (denoted by "Eng2Ind_PostEdit"). We then made Indonesian image descriptions of 10k images by crowdsourcing without giving English descriptions or translations to the worker (denoted by "Ind_Caption"). We investigated whether substantial differences exist between the translated sentences that were originally based on natural language expressions by native English speakers and Indonesian descriptions that were expressed by native Indonesian speakers. We calculated their semantic distances using Word2Vec and FastText embeddings.

This paper is structured as follows. Section 1 explains the reason and purpose of the semantic-based image description dataset, and Section 2 describes the existing research of multilingual image description datasets and their corpus construction method and analysis. Sections 3 and 4 describe our approach for constructing an Indonesian image description dataset and method to calculate the semantic distances. Section 5 analyzes the automatic text translation and direct image construction result. Finally, Section 6 concludes our paper.

2. Related works

Multilingual image description datasets can be constructed in many different ways, such as manual annotation as well as human, automatic, and multimodal translations. Some datasets are created by combining two different methods to improve their image description results.

The IAPR TC-12 datasets [15] contain 20,000 image descriptions that were collected for the CLEF cross-language image retrieval track. Most of their image descriptions were written in German. The German sentences were then validated and

10.21437/SLTU.2018-9

translated into English and Spanish by professional translators. Another image description dataset, created by a professional translator, is the Pascal dataset [16] that contains 1000 pairs of English-Japan image descriptions. Here, the translated sentences closely resemble the source sentences.

Multi30K [17] consists of multilingual sentence-based image descriptions that were created from Flickr30K, which uses two different methods: (1) the translation and (2) independent captions. For the translation case, the dataset was collected from professional English-German translators using Flickr30K English descriptions as the source language. The translators were given the English captions of the Flickr30K images without the images themselves. The independent captions of the pictures were provided by crowdsourcing without the English descriptions. This study also analyzed the difference between the translations and the independent image descriptions by calculating the sentence length and the vocabulary size. The results reveal that the English image descriptions are generally longer than the German descriptions, both in the number of words and characters.

Even though there are already available image description datasets in multilingual settings, none of the datasets use the Indonesian language. We constructed an Indonesian image description dataset in two ways: (1) a translation method and (2) direct image description. In contrast with the above previous studies, we investigated the differences between the translation and the independent image descriptions (Eng2Ind_Translation versus Ind_Caption), not only by the sentence length and the vocabulary size but also in the semantic distance using Word2Vec and FastText embeddings.

3. Corpus construction

The Flickr30K dataset contains 31,783 images with five corresponding English sentences for each picture that were used for many kind of tasks. One of the tasks with the Flickr30K dataset is a WMT multimodal machine translation that includes training, development, and test sets. The development set contains 1015 images with five descriptions per image, and the 2017 and 2018 test sets include 1000 and 1071 images with one description per image.

With the details described below, we constructed Indonesian image descriptions in two different ways: automatic translation from English descriptions and direct image descriptions.

3.1. English-to-Indonesian translation without image data

First, we translated the English image description of the Flickr30k dataset to Indonesian (Eng2Ind_Translation) without the image itself by utilizing automatic translation by the Google Translate API[1]. In addition to the Flickr30k dataset, we also automatically translated the WMT2017 and WMT2018 test data[2]. Thus, the resulting dataset includes training, development, and test sets, like the one used in the WMT Multimodal Machine Translation Task.

To ensure the translation quality, we manually validated the translation result (Eng2Ind_PostEdit) of the WMT development and test datasets by crowdsourcing. We asked the registered workers to perform the task on several sentences, and selected nine native Indonesian crowdworkers (four males, five females, from 20-30 years old) that demonstrated to have a good understanding of English to participate in the task. We provided 250 sentences per session, and each crowdworker performed more than one session.

The crowdworkers performed post-editing to correct any errors. To ensure that they only fixed errors based on the translation results, a validation process was also performed without any images that correspond to the descriptions. The mistakes in the translation results included word selection, misplacement, and grammatical errors. Our crowdworkers validated 7146 sentences (5075 sentences of a development set, 1000 sentences of the WMT2017 test set, and 1071 sentences of the WMT2018 test set).

3.2. Direct Indonesian image description without English captions

Next, we directly constructed Indonesian descriptions from the images without their English captions (Ind_Caption). However, due to limited budget and time, we only used 10k images of the Flickr30k datasets, including the WMT development and test datasets by crowdsourcing. We successfully gathered 22 such workers (7 males, 15 females) whose ages ranged from around 20 to 30.

Given an image, the crowdworker wrote a sentence that describes it. Since no English descriptions or Indonesian translations were provided, they wrote their own natural language expressions based on their perception of the image. We suggested a range of about 5 to 25 words per sentence, like in the length of the English descriptions, without limiting their sentences to that suggested range. For one session, each crowdworker described 200 images (one caption per image) and be able to take another session if they want.

4. Semantic embeddings

The word-embedding method has successfully identified the semantic distances between two sentences better than the traditional approach for text similarity (e.g., the distance of the tf-idf vector) [18]. In this research, we used two word-embedding methods to calculate the semantic distance between Eng2Ind_Translation and Ind_Caption:

- **Word2Vec**
  Word2Vec [19] is an embedding method where the target words are represented using surroundings words with neural network whose hidden layer encodes the word representation. For example, in the sentence, "I ate a slice of pizza," the word vector representation of "ate" is affected by "I," "a," "slice," "of," and "pizza." The main idea is that the words that share common contexts in the corpus are located near one another in the space.
- **FastText**
  FastText [20] is an extension of Word2Vec, where instead of using individual words as neural network input, it uses sub-words (n-grams). For example, the tri-grams for *pizza* are *piz*, *izz*, and *zza*, and the word embedding vector for *pizza* is the sum of those n-grams.

Here we utilized the pre-trained Indonesian model of Word2Vec and FastText [20]. Assuming that $W(n)$ is the word embedding of word n in a sentence where the number of words is N, we calculated the sentence embedding for the image descriptions in three different ways:

- Average of word vector embeddings in a sentence:

$$mean = \frac{1}{N}\sum_{n}^{N} W(n). \quad (1)$$

[1] Google Translation API – https://translate.google.com/

[2] WMT – http://www.statmt.org/wmt18/multimodal-task.html

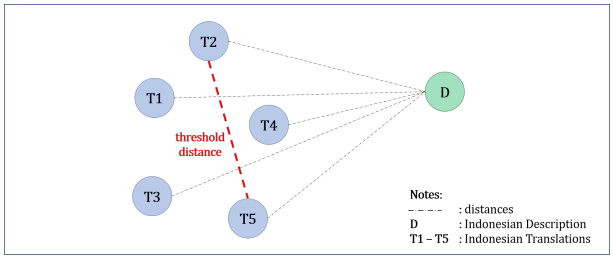

Figure 1: *Illustration of semantic distances calculation for Ind_Caption and Eng2Ind_Translation.*

- Sum of word vector embeddings in a sentence:

$$sum = \sum_{n}^{N} W(n). \quad (2)$$

- Maximum of word vector embeddings in a sentence:

$$max = \text{element_maximum}(W(n) \text{ for } n \text{ in } N). \quad (3)$$

- Minimum of word vector embeddings in a sentence:

$$min = \text{element_minimum}(W(n) \text{ for } n \text{ in } N). \quad (4)$$

Even though calculating the sentence embeddings with the mean, sum, maximum, and minimum of the word vectors is a simple approach, they provide good results for semantic analysis [18, 21].

As mentioned above, since Flickr30K [6] used five English descriptions per image, we have five corresponding Eng2Ind_Translation sentences. But for the Indonesia image description, we only created one Ind_Caption per image. In this study, we first calculated the sentence-embedding distances among the five sentences of the Eng2Ind_Translation. After that, we calculated the sentence-embedding distances between the Ind_Caption and all of the Eng2Ind_Translations. To decide whether the Ind_Caption remains on an acceptable semantic embedding range, we set the maximum of the embedding distances among the five Eng2Ind_Translations as a threshold (Fig. 1). $D(i, j)$ is the distance between translations i and j of an image, and k is the total number of the translations for each image. The threshold T is calculated as follows:

$$T = max(D(i,j) for\ i\ in\ [1..k],\ j\ in\ [1..k], i \neq j) \quad (5)$$

If the embedding distance between the Ind_Caption and the Eng2Ind_Translation exceeds the threshold, a substantial semantic difference exists between the Ind_Caption and the Eng2Ind_translation.

5. Analysis

5.1. Quality of automatic translation

To investigate the quality of the automatic translation, we compared the performance of the English-to-Indonesian automatic translation with human post-editing which is treated as a reference (Eng2Ind_Translation versus Eng2Ind_PostEdit). First, for the number of words in the sentences, we found no significant differences between the Eng2Ind_Translation and the Eng2Ind_PostEdit, where both types of data averaged about 12 words per sentence. Second, in terms of quality, we calculated the translation error rate (TER) [22], which is defined as the minimum number of edits in the translation so that it exactly matches the corresponding reference. The number of TER edits is calculated from the number of insertions, deletions, substitutions, and shifts. The average of the TER scores between all of the Eng2Ind_Translation and the Eng2Ind_PostEdit was 5%, which means that there was little difference in the structure of the words in the automatic translation and the manual post-edits. The resulting automatic translations are satisfactory as Indonesian image descriptions for Flickr30k.

Table 1: *Frequencies of several tags in Ind_Caption and Eng2Ind_Translation.*

Tag	Tag description	Percentage	
		Translations	Indonesian descriptions
NN	Noun	43.94%	38.47%
VB	Verb	14.61%	15.38%
IN	Preposition	12.83%	7.81%
JJ	Adjective	7.10%	5.13%
NND	Classifier, partitive and measurement noun	6.19%	5.04%
SC	Subordinate conjunction	3.56%	5.16%
CD	Cardinal number	2.98%	2.57%
CC	Cordinating conjunction	2.66%	1.46%
FW	Foreign word	1.49%	2.49%
NNP	Proper noun	0.52%	7.63%
Others	Ordinal number, determiner, modal auxiliary, negation, etc.	4.13%	8.86%

5.2. Translation vs description

5.2.1. Syntax analysis

The Eng2Ind_Translation sentences are 7.5% longer than the Ind_Caption. Unlike the Eng2Ind_Translation and Eng2Ind_PostEdit that have almost the same amount of words, the Ind_Caption might have very different words than those in Eng2Ind_Translations. However, since Indonesian attaches many suffixes and affixes to words, the same two words with different affixes may be viewed as two very different words. To reduce the differences that are just caused by different affixes, we removed affixes with the Indonesian stemmer [23] and used Indonesian POStag [24] in both the Eng2Ind_Translation and Ind_Caption.

Table 1 shows the statistics of the POS tags in the Ind_Caption and Eng2Ind_Translation. The Ind_Caption used more proper nouns which specified name of a person, thing, or place, than the Eng2Ind_Translation. On the other hand, Eng2Ind_Translation used mode adjective. Furthermore, many loanwords in both Eng2Ind_Translation and Ind_Caption that were adopted from English word. This mean either some English words cannot be translated into standard Indonesian words or some images cannot be expressed in standard Indonesian words.

5.2.2. Semantic analysis

Different words do not necessarily have different semantic meanings. By calculating the cosine distance between the Ind_Caption and the Eng2Ind_Translation embeddings, we analyzed the semantic differences between these two datasets. Table 2 shows that the embedding distances of the Ind_Caption to the Eng2Ind_Translation are always farther away than the distance among the Eng2Ind_Translation themselves.

To measure the semantic distance between sentences, we used two different embeddings methods, Word2Vec and FastText. Word2Vec calculated embeddings based on word granularity, while FastText works on subword granularity. This might have some affect on the distance measurements since Indonesian language has suffixes and affixes, which makes the Word2Vec regards the same word with different suffix or affix

Table 2: *Semantic distances between Ind_Caption and Eng2Ind_Translation.*

Word embeddings	Sentence embeddings	Distances between Eng2Ind_Translation			Distances between Ind_Caption and Eng2Ind_Translation			Percentage of Ind_Caption lies outside the threshold
		min	mean	max	min	mean	max	
Word2Vec	min	0.055	0.109	0.164	0.099	0.130	0.171	45.52%
	mean	0.147	0.294	0.446	0.258	0.349	0.454	48.03%
	max	0.055	0.110	0.166	0.098	0.130	0.172	44.15%
	sum	0.147	0.294	0.446	0.258	0.349	0.454	48.03%
FastText	min	0.056	0.103	0.148	0.093	0.118	0.149	42.06%
	mean	0.089	0.176	0.264	0.155	0.207	0.266	46.22%
	max	0.060	0.112	0.160	0.101	0.129	0.162	44.30%
	sum	0.089	0.176	0.264	0.155	0.207	0.266	46.22%

Table 3: *Example of sentences between images with shortest and longest semantic distances from Fig. 2.*

	Distance between Ind_Caption and Eng2Ind_Translation	
	Shortest distance (Fig.2, image a3)	**Longest distance (Fig. 2, image b2)**
(1) Eng_Caption	*A black dog is running along the beach.*	*Green Bay Packer player cooling off.*
(2) Eng2Ind_Translation	*Seekor anjing hitam berlari di sepanjang pantai.*	*Pemain Green Bay Packer sedang mendinginkan diri.*
(3) Ind_Caption	*Seekor anjing hitam sedang berlari-lari di pantai.*	*Pemain dengan nomor punggung 4.*
(4) Ind2Eng_Translation	*A black dog is running around on the beach.*	*Player whose number is 4.*

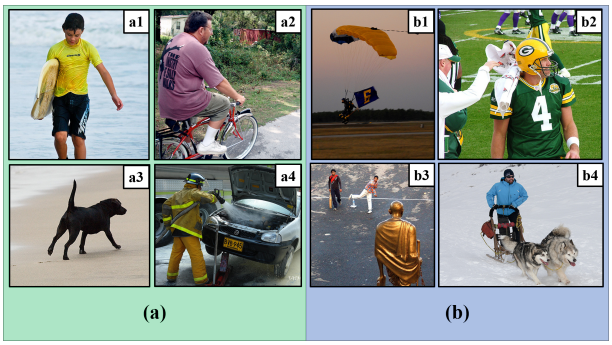

Figure 2: *Image examples that have (a) shortest and (b) longest distance between Ind_Caption and Eng2Ind_Translation.*

as different words. On the other hand, FastText functions at the word's sub-word unit level instead of the word itself and allows it to extract the sub-word matching within a word that results in a smaller embedding distance.

Next, comparing the distance among the Eng2Ind_Translation and between the Ind_Caption and Eng2Ind_Translation, an average of almost 50% of the Ind_Caption lie outside the maximum range of the distance among the Eng2Ind_Translation. Differences in the embedding vectors may, of course, occur due to the syntax problems discussed above. However, they might also be caused by differences in visual interpretation.

Next we analyzed several examples of images with the shortest and longest semantic distances between the Ind_Caption and Eng2Ind_Translation (Fig. 2). To simplify the problem, we analyzed the mean sentence embedding with FastText.

Figure 2(a) shows examples of images with the shortest distance between the Ind_Caption and Eng2Ind_Translation. Most describe such simple everyday activities as *"jumping," "playing,"* or *"running"*, all of which are also commonly experienced in Indonesia. The background of the images more or less consists of one solid image, such as *"grass hill"* or *"beach."* The image focus on a simple, main object: *"dogs," "a boy,"*, or *"a man"*, all of which again are commonly seen in Indonesia.

On the other hand, Fig. 2(b) shows examples of the images with the longest distance between the Ind_Caption and Eng2Ind_Translation. In Fig. 2 (b4), the words, *"the huskies"*, were misinterpreted as *"the wolves"*. Since Indonesia has no winter, such large dogs are uncommon in the country. Other images illustrate a complex or an unusual activity that is seldom done in Indonesia, such us *"parasailing"* or *"Green Bay Packer."*

Table 3 lists the created sentences of Figs. 2(a3) and (b2). The sentences in the Ind_Caption and Eng2Ind_Translation for Fig. 2(a3) are similar. On the other hand for Fig. 2(b2), since the Indonesian annotators failed to identify *"Green Bay Packer,"* they could only describe the *"player"* instead of *"Green Bay Packer player,"* resulting in a wide distance in the embedding space. Different cultural backgrounds may indeed affect visual perceptions.

6. Conclusions

This paper presents a corpus construction of Indonesian image descriptions based on the Flickr30k dataset. Our dataset consists of the following: (1) Eng2Ind_Translation: the English-to-Indonesian automatic translations of the full set of Flickr30k plus WMT2017 and WMT2018 benchmark test sets; (2) Eng2Ind_PostEdit: the manual post-edits on translation sentences on the development and test sets of WMT2017 and WMT2018; and (3) Ind_Caption: the 10k Indonesia image descriptions. Our dataset was developed by crowdsourcing by native Indonesians. We performed syntactic and semantic analysis of the differences in the Indonesian descriptions and translations. An average of almost 50% of the Indonesian captions fall outside of the maximum range of the distance among the translations, suggesting that many substantial differences are found in the visual perception of images between native Indonesian and English language users.

We often assume that an image represents a universal concept, but languages do not. However, visual perception also greatly depends on cultural backgrounds. Currently, we only constructed different captions given the same image. Future work will investigate whether people from different cultural backgrounds can produce similar images given identical captions or translated versions in their own native languages.

7. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP 17K00237.

8. References

[1] X. He, B. Shi, X. Bai, G.-S. Xia, Z. Zhang, and W. Dong, "Image caption generation with part of speech guidance," *Pattern Recognition Letters*, pp. 1–9, 2017.

[2] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," vol. 39, December 2014.

[3] Y. Feng and M. Lapata, "Topic models for image annotaion and text illustration," *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Assosiation for Computational Linguistics*, pp. 831–839, 2010.

[4] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention network for image question answering," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21–29, June 2016.

[5] C. Rashtchiana, P. Young, M. Hudosh, and J. Hockenmaier, "Collecting image annotations using Amazon's Mechanical Turk," *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, June 2010.

[6] P. Young, A. lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metric for semantic inference over event descriptions," *Transaction of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[8] D. Elliot, S. Frank, L. Barrault, F. Bougares, and L. Specia, "Finding of the secong shared task on multimodal machine translation and multilingual image description," *Proceedings of the Conference on Machine Translation (WMT)*, vol. 2, pp. 215–233, September 2017.

[9] L. Specia, S. Frank, K. Simaán, and D. Elliott, "A shared task on multimodal machine translation and crosslingual image description," *Proceedings of the Conference on Machine Translation (WMT)*, vol. 2, pp. 543–553, August 2016.

[10] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "STAIR captions: Constructing a large-scale Japanese image caption dataset," *CoRR*, vol. abs/1705.00823, 2017. [Online]. Available: http://arxiv.org/abs/1705.00823

[11] X. Zeng and X. Wang, "Add English to image Chinese captioning," *2017 IEEE 2nd International Conference on cloud computing and big data analysis (ICCCBDA)*, April 2017.

[12] M. Kay, M. King, J. Lehrberger, A. Melby, and J. Slocum, "Machine translation," *American Journal of Computational Linguistics*, vol. 8, pp. 74–78, April-June 1982.

[13] S. G. Goto, Y. Ando, C. Huang, A. Yee, and R. S. Lewis, "Cultural differences in the visual processing of meaning: Detection incongruities between background and foreground object using the N400," *Social Cognitive and Affective Neuroscience*, vol. 5, pp. 242–253, June 2010.

[14] J. Ćenék and Ś. Ćenék, "Cross-cultural differences in visual perception," *Journal of Education Culture and Society*, vol. 1, 2015.

[15] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The IAPR TC12 benchmark: A new evaluation resource for visual information systems," 10 2006.

[16] R. Funaki and H. Nakayama, "Image-mediated learning for zero-shot cross-lingual document retrieval," *Proceddings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 585–590, September 2015.

[17] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30K: Multilingual English-German image descriptions," *CoRR*, vol. abs/1605.00459, 2016.

[18] C. D. Boom, S. V. Canneyt, S. Bohez, T. Demeester, and B. Dhoedt, "Learning semantic similarity for very short texts," *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[20] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *TACL*, vol. 5, pp. 135–146, 2017.

[21] Y. Adi, E. Kermany, Y. Belinkov, O. Lavi, and Y. Goldberg, "Fine-grained analysis of sentence embeddings using auxilary prediction tasks," *5th International Conference on Representations*, 2017.

[22] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *In Proceedings of Association for Machine Translation in the Americas*, 2006, pp. 223–231.

[23] M. Adriani, J. Asian, B. Nazief, S. M. Tahaghoghi, and H. E. Williams, "Stemming Indonesian: A confic-stripping approach," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 6, December 2007.

[24] A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung, "Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus," *2014 International Conference on Asian Language Processing (IALP)*, 2014.