Adapting Google Translate using Dictionary and Word Embedding for Arabic-Indonesian Cross-lingual Information Retrieval

1st Maryamah Maryamah
Departement of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
maryamah.18051@mhs.its.ac.id

2nd Agus Zainal Arifin
Departement of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
agusza@if.its.ac.id

3rd Riyanarto Sarno
Departement of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

4th Ahmad Makki Hasan
Department of Arabic Language Education
Universitas Islam Negeri Maulana Malik Ibrahim
Malang, Indonesia
ahmadmakkih@pba.uin-malang.ac.id

2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS) | 978-1-7281-9448-6/20/$31.00 ©2021 IEEE | DOI: 10.1109/IoTaIS50849.2021.9359708

***Abstract*— The translation has an essential role in Cross-lingual Information Retrieval. Translation using a dictionary is reliable even though it has a limited vocabulary. Translation using google translate, in some cases, using different words used in document target words. The translation process causes word translation to be less accurate to get relevant documents. In this paper, we proposed a new translation approach by adapting google translate using a dictionary and word embedding in Arabic-Indonesian Cross-lingual Information Retrieval. The dictionary is the primary resource used for translation improved by Levenshtein distance and FastText for finding the correct word translation. Google translate is used to complete translation when the word does not exist in the dictionary resource. The proposed method archive a BLEU score of 0.47. This score is higher than the other comparison resource score. The proposed method successfully improves the translated query to retrieve more relevant documents in cross-lingual information retrieval based on this implementation.**

Keywords—Cross-lingual information retrieval, Dictionary, Google Translate, Levenshtein distance, FastText

I. INTRODUCTION

Cross-lingual Information Retrieval (CLIR) is a searching document using a query in one language and retrieving documents in diverse languages [1]. CLIR can help users find information from documents which have a different language than the user's source language [2]. CLIR is needed because multilingual content has been growing significantly with rapid internet access growth all over the world [3]. CLIR is a challenging task in the absence of aligned parallel corpora by translating the query into the same word as the target document [4].

An essential role in CLIR is Translation process [5]. Machine translation is translating a text from source language into a target language with computers [6]. Translation results significantly affect the results of retrieving documents. Incorrect translation results can cause the resulting document to be irrelevant. The development of machine translation can use deep learning methods such as Recurrent Neural Network [7], Long Short-Term Memory [8], and Transformer [9]. Deep learning has a limitation. It requires extensive data to get the best model in training data. The more data used, the better the results model will be obtained, and the better to predicted testing data. Creating sufficient datasets is labor-intensive and time-consuming [10].

The translation often uses a dictionary as a resource. The dictionary is also a reliable resource for use as a translation. The dictionary is more trusted because it is made by experts who can be trusted with the results. Translation using a dictionary is very dependent on the number of words in the dictionary. However, many words do not exist in the dictionary, and using the dictionary alone is not sufficient. The dictionary also uses standard language, and it needs the append of additional methods to search for words that are similar to the words in the dictionary because habitual users use everyday words for the query when searching the document.

The development of a translation system that can accept input from everyday language or spoken language must check the similarity of words between word input and words in the dictionary. Common errors often arise when making queries because users use spoken language in searching for documents, not written language. This causes the translation of the query not to be detected in the dictionary to translation results less accurate. A word check using spoken language is done to confirm its presence in the dictionary. An example in the Indonesian dictionary is "*shalat*" (pray), but in everyday life, users often use the word "*salat*" or "*sholat*". Manual checking of a spell is difficult and time-consuming [11].

Automatic checking spell or string correction can be done using Levenshtein distance [12]. Levenshtein distance is finding the shortest length between words [13], [14]. Translation using the resource dictionary alone is not enough because of the limited words. This problem can be overcome by append other resources such as google translate.

Google translate provides an API that can be used for translation. In some cases, the google translate API results are less precise than the google translate website version. Lower precision is because the Translation API web page uses the new Neural Machine Translation (NMT) model, while the translation API uses the default Phrase-Based Machine Translation (PBMT) model. In some cases, according to experts knowledge, the translation results using google translate API is not correct. An example, the Arabic word of zakat fitrah is "عشر". The correct word, according to experts, is "زكاة الفطرة". In CLIR, this problem can cause the returned document is irrelevant.

In this paper, we proposed a new translation approach by adapting google translate using dictionary and word embedding in Arabic-Indonesian Cross-lingual Information Retrieval. We expect the combination of dictionary and google translate can complement reliable but limited dictionary resources. The translation is done using the dictionary as the primary resource and google translate as the second resource. Translation of words that are not in the dictionary does not directly use google translation. This study ensures that the query word is not in the dictionary by checking similarity using Levenshtein distance and FastText. Levenshtein distance is used to find the word that most closely matches the query when the query does not exist. The choice of words from the minimum of Levenshtein distance result is selected using the similarity of FastText to ensure that the words are similar or not. FastText results for less than the threshold will be searched using google translate and the words will be combined after all the words in the query have been translated.

II. METHODOLOGY

The methodology in this study is shown in Fig. 1. The first process in the methodology is to enter the input query in the form of a question sentence used to search for documents. Search for documents using a language different from the document language or CLIR problem is need translation process. The translation is needed to match the language between the query and the document. System translation is made using a combination of dictionary and google translate resources to get the translation output. The dictionary is the primary resource and google translate will be used when the word does not exist in the dictionary. When the word in the dictionary already contains all the words in the query sentences, google translate resources do not need.

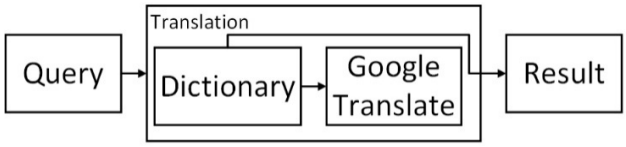

Fig. 1. Methodology paper

Fig.2 explains that the first process that is execute after the input query is to separate the query sentence into several words. The translation process will be process by each word in the sentence to ensure that each word is the correct translation based on the dictionary resource. The word in the dictionary ensures that the translation result is the word used in information retrieval matters. In this study, the dictionary is considered the primary translation reference. Words that are not in the dictionary will be translated using google translate to complete the translation result.

The first word in the query sentence will be searched according to the target language in the dictionary. When the word is exist in the dictionary, the word will be output result and stored in the output temp to be combined with other words in the query. Words that are not in the dictionary will be done by several steps. First, calculating the similarity of query words with all of words in the dictionary using Levenshtein distance. It is used to find other words that are similar to the word query. The word in the query can be mistyping or using non-standard words such as in a dictionary. An example in the Indonesian dictionary is *"shalat"* (pray), but in everyday life, users often use the word *"salat"* or *"sholat"*. Another similar word search is required to ensure that the word in the query exists, and the result is translation relevant.

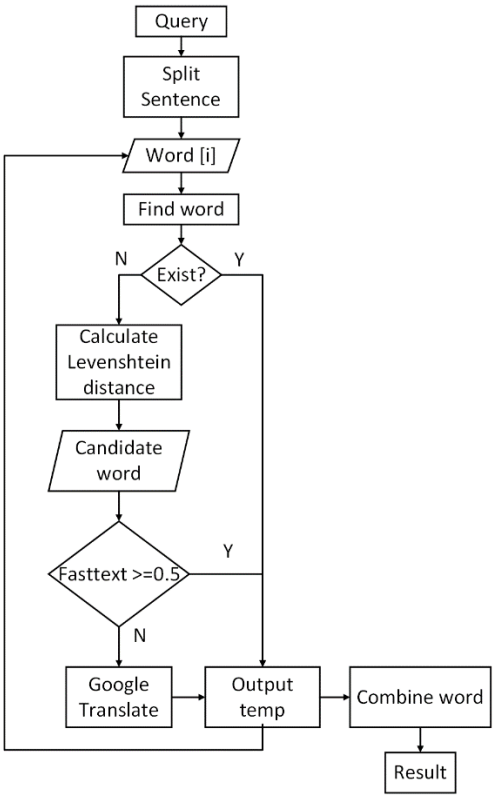

Fig. 2. Translation Process

Calculating the minimum distance will be checked for the similarity of the words using FastText [15]. FastText was chosen because it is fast in training data and can create vectors of new words that did not appear in the training data [16]–[18]. FastText is trained using all article Indonesian

Wikipedia to ensure various words in training. The result of similarity FastText uses a threshold of 0.5 as the minimum similarity of words. FastText result more than 0.5 will be selected and considered to be a word similar to the query. Otherwise, when the similarity result is less than 0.5, the words are considered not similar and the word will be translated using google translate. The similarity threshold of 0.5 is obtained from several experiments from several accurate numbers to get the word's similarity threshold. The similarity experiment of more than 0.5 was too high. There were similar words, but because the threshold was too high, it was considered not similar and vice versa. Similarity experiments of less than 0.5 also cause dissimilar words to be similar. From these experiments, the appropriate threshold used in this paper is 0.5. All the words in the query will be translated and the results will be combined to become the final machine translation result. The results of this machine translation will be used for relevant document search.

III. Result and Experiment

In this study, the document search problem (CLIR) uses query input Indonesian and document output using Arabic language. This research is expected to support Indonesian people who often seek reference sources for Arabic documents to solve socio-religious problems. The dictionary resource in this study using is the Al-Munawwir dictionary which is often used in Indonesia. The dictionary is often used because the writes created a dictionary based on vocabulary in the Arabic document reference. The query input used is in the form of socio-religious questions obtained from piss-ktb.com.

The evaluation used is BLEU (Bilingual Evaluation Understudy) score. BLEU score is calculated based on the correctness of words in the results compared with n-grams of words according to expert results. We use the translation results from two relevant experts as a reference in document search. The two results were also carried out to find the diversity of words were often used by experts in document searches. The expert's results show that the translation results are the same in some cases, and the document does use this word in writing documents. In other cases, experts have different results but have the same meaning. The translation results of the two experts are relevant in document searching, and increased due to variations in the translation word.

TABLE I. Result of Experiment

No	Proposed Method	Google Translate	No	Proposed Method	Google Translate
1	**0,44**	0,29	36	**0,44**	0,28
2	0,50	0,50	37	**0,50**	0,44
3	**0,50**	0,36	38	**0,50**	0,47
4	0,38	**0,48**	39	**0,52**	0,50
5	0,47	**0,52**	40	**0,27**	0,19
6	**0,52**	0,42	41	0,33	**0,38**
7	0,41	**0,46**	42	**0,47**	0,40
8	**0,62**	0,53	43	**0,42**	0,47
9	**0,56**	0,37	44	**0,48**	0,39
10	0,33	**0,48**	45	**0,37**	0,30
11	0,42	**0,44**	46	0,40	**0,56**
12	**0,45**	0,36	47	0,50	**0,55**
13	**0,55**	0,47	48	**0,42**	0,39
14	0,48	**0,60**	49	**0,54**	0,44
15	0,34	**0,42**	50	**0,41**	0,26
16	**0,54**	0,49	51	**0,41**	0,32
17	0,42	**0,50**	52	**0,42**	0,29
18	**0,61**	0,48	53	**0,55**	0,52
19	**0,52**	0,46	54	0,30	**0,36**
20	**0,65**	0,50	55	**0,47**	0,41
21	0,36	**0,45**	56	0,31	**0,44**
22	0,53	**0,55**	57	**0,61**	0,55
23	**0,69**	0,56	58	**0,40**	0,32
24	**0,57**	0,39	59	**0,67**	0,57
25	**0,59**	0,48	60	**0,48**	0,39
26	0,54	**0,55**	61	**0,57**	0,48
27	**0,46**	0,43	62	**0,48**	0,28
28	0,34	**0,42**	63	0,52	**0,64**
29	**0,36**	0,33	64	**0,62**	0,40
30	**0,57**	0,50	65	**0,52**	0,34
31	**0,40**	0,29	66	0,41	**0,49**

No	Proposed Method	Google Translate	No	Proposed Method	Google Translate
32	**0,47**	0,41	67	**0,41**	0,35
33	0,41	**0,44**	68	0,35	**0,44**
34	**0,39**	0,37	69	0,46	**0,50**
35	**0,46**	0,34	70	**0,44**	0,38
Average				**0,47**	0,43

Table 1 is a comparison between the proposed methods and other resource (google translate). The highest average BLEU score is the proposed method with a value of 0.47, which is higher than google translate score of 0.43. The higher BLEU results show which the proposed method results are closer to the translation results by experts. Fig. 3 shown that the proposed method mostly has a higher BLEU score than Google translate score in the experimental case. In the 70 cases that were tested, the results of 48 cases using the proposed method were better than Google translate. Meanwhile, 21 cases of google translate results are better and the results are the same in 1 case. The proposed method provides a more precise word translation than google translate resource. Dictionary is a reliable resource and has been confirmed by experts in translating words. A comparison of resources using a dictionary is not perform because the results of word translation are incomplete. The resource dictionary is not powerful enough to become a machine translator because the translation results are incomplete due to dictionary words and standard languages' limitations. Not all words are recognized, especially queries that use modern words or colloquial words. Translation using the proposed method tries to translate all words using a combination of a dictionary and google translate.

In this study, the dictionary is the primary resource used to translate words. When the word in the dictionary does not exist, the word is translated with google translate resource. However, the lack of understanding of the right words in the context of the sentence causes this study's results to be classified as low because the choice of words in some cases is not correct according to the words used by experts. In terms of meaning, the translation results of all combined resources and google translate are correct.

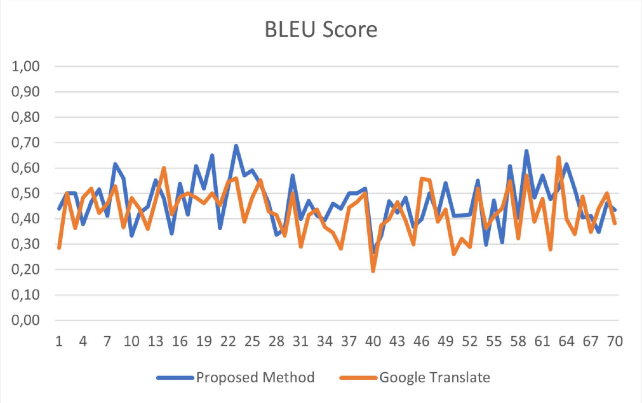

Fig. 3. Bleu Score of experiment result

In terms of sentence context, the method cannot provide an appropriate word. Google translate, in some cases, can understand the context of words such as cases 4, 5, 17, 68 but google translate sometimes uses words that are not usually used by experts so that the results are considered incorrect and lower than others. The results that are not usually used by experts are not used in the document, and this causes the translation not provide a relevant document even though it has same meaning. Experts translation uses words frequently used in documents, and the result is more accurate when searching for documents.

IV. Conclusion

In this paper, we proposed machine translation approach for Arabic-Indonesian Cross-lingual Information Retrieval by adapting google translate using dictionary and word embedding. The proposed method gets the highest average BLEU score with a value of 0.47 which is higher than google translate score with value 0.43. These results indicate that the proposed method has better translation results than Google translate. However, the proposed method does not consider the position of the word in the context of the sentence. Understanding the context of a sentence is important because the same word in a sentence but different contexts can use different words in the translation problem. Future work in this research is understanding the context of words in sentences. Translation of the words to ensure that the word is correct can be done, but it is necessary to check whether the word is correct or not in the context of the word in the sentence.

Acknowledgment

We would like to express our gratitude to the Ministry of Research and Technology, with grant number 3/E1/KP.PTNBH/2020.

References

[1] M. N. Asim, M. Wasim, M. Usman, G. Khan, N. Mahmood, and W. Mahmood, "The Use of Ontology in Retrieval : A Study on Textual , Multilingual , and Multimedia Retrieval," *IEEE Access*, vol. 7, pp. 21662–21686, 2019.

[2] S. Saleh and P. Pecina, "Term Selection for Query Expansion in Medical Cross-lingual Information Retrieval," in *European Conference on Information Retrieval*, 2019, pp. 507–522.

[3] S. Saleh and P. Pecina, "Document Translation vs . Query Translation for Cross-Lingual Information Retrieval in the Medical Domain," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6849–6860.

[4] M. Niyogi, G. Kripabandhu, and B. Arnab, "Learning Multilingual Embeddings for Cross-Lingual Information Retrieval in the Presence of Topically

Aligned Corpora," *arXiv Prepr. arXiv1804.04475*, 2018.

[5] M. Madankar, M. B. Chandak, and N. Chavhan, "Information Retrieval System and Machine Translation : A Review," in *Procedia Computer Science*, 2016, vol. 78, pp. 845–850.

[6] C. Poornima and V. Dhanalakshmi, "Rule based Sentence Simplification for English to Tamil Machine Translation Rule based Sentence Simplification for English to Tamil Machine Translation System," *Int. J. Comput. Appl.*, vol. 25, no. 8, pp. 38–42, 2011.

[7] A. Esan, J. Oladosu, C. Oyeleye, and I. Adeyanju, "Development of a Recurrent Neural Network Model for English to Yorùbá Machine Translation," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 5, 2020.

[8] M. K. Vathsala and H. Ganga, "RNN based machine translation and transliteration for Twitter data," *Int. J. Speech Technol.*, pp. 1–6, 2020.

[9] Q. Wang *et al.*, "Learning Deep Transformer Models for Machine Translation," *arXiv Prepr. arXiv 1906.01787.*, 2019.

[10] D. Ringel, G. Lavee, I. Guy, and K. Radinsky, "Cross-Cultural Transfer Learning for Text Classification," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3864–3874.

[11] M. Hossain, F. Labib, A. S. Rifat, A. K. Das, and M. Mukta, "Auto-correction of English to Bengali Transliteration System using Levenshtein Distance," in *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, 2019, no. June, pp. 1–5.

[12] C. Zhao and S. Sahni, "String correction using the Damerau-Levenshtein distance," *BMC Bioinformatics*, vol. 20, no. 11, p. 277, 2019.

[13] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys. Dokl.*, vol. 10, no. 8, pp. 707–710, 1966.

[14] C. Whitelaw, B. Hutchinson, G. Y. Chung, and G. Ellis, "Using the Web for Language Independent Spellchecking and Autocorrection," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009, pp. 890–899.

[15] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," *arXiv Prepr. arXiv1607.01759.*, 2016.

[16] J. Choi and S. Lee, "Improving FastText with inverse document frequency of subwords ☆," *Pattern Recognit. Lett.*, vol. 133, pp. 165–172, 2020.

[17] J. Wu, M. Wen, R. Lu, B. Li, and J. Li, "Toward efficient and effective bullying detection in online social network," *Peer-to-Peer Netw. Appl.*, pp. 1–10, 2020.

[18] F. Tajaddodianfar, J. W. Stokes, and A. Gururajan, "Texception: A Character/Word-Level Deep Learning Model for Phishing URL Detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2857–2861.