Cross-Lingual Summarization: English - Bahasa Indonesia

Achmad F. Abka[1,2], Mahardhika Pratama[3], Wisnu Jatmiko[2]

[1] *Research Center for Informatics*, *National Research and Innovation Agency*, Indonesia
[2] *Faculty of Computer Science*, *Universitas Indonesia*, Indonesia
abka@ui.ac.id, wisnuj@cs.ui.ac.id
[3] *School of Computer Science and Engineering*, *Nanyang Technological University*, Singapore
mpratama@ntu.edu.sg

2021 6th International Workshop on Big Data and Information Security (IWBIS) | 978-1-6654-2451-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IWBIS53353.2021.9631862

Abstract—**Progress of abstractive summarization has been accelerated since the introduction of sequence-to-sequence neural networks. Summarization is no longer limited to selecting words or sentences that exist in the source document as in the extractive approach, but can generate completely new words or sentences that have never appeared in the source document. Big push came from machine translation research with the introduction of attention mechanisms. Attention mechanism is the key to the information bottleneck problem in encoder-decoder model. Cross-Lingual Summarization (CLS) is the task of generating a summary in target language from source document in different language. Traditional methods split this task into two steps: summarization and translation. This paper describes a study on CLS without explicitly using translator, thereby reducing one step as in existing method. We incorporate multilingual embeddings in sequence-to-sequence neural networks with attention mechanisms to handle this task. Multilingual embeddings are used to represent words as if the source language and the target language are the same language. Experiments show comparable performance between monolingual summarization and cross-lingual summarization in Amazon Fine Food review data indicated by ROUGE scores which are only 1-2 points apart.**

Keywords—cross-lingual summarization, multilingual word embeddings, sequence-to-sequence neural network, attention mechanism

I. INTRODUCTION

Summarization is one of important problems in natural language understanding. One approach in summarizing document is by taking pieces of text from the source document, usually in the form of whole sentences, that are considered important. This method is quite easy, because by taking snippets of text from the source document, one can be sure that the grammar and accuracy are good. This approach is known as extractive summarization. Another way that allows for more sophisticated capabilities such as: paraphrasing, generalization, or the use of knowledge in the real world can only be done with abstractive summarization approach. This approach seeks to produce a bottom-up summary so that words or phrases that were not previously present in the source document may appear.

Initially, extractive summarization was preferred because of its simpler approach [1], [2]. However, the success of Recurrent Neural Network (RNN) with the sequence-to-sequence model paved the way for the abstractive summarization approach [3]-[7]. This sequence-to-sequence model is used by adding attention mechanism that is inspired from attention-based neural machine translation model [8], [9]. This progress cannot be separated from the development of word representation based on distributed representation which is commonly known as word embeddings [10]-[13]. Word embeddings are also used in more modern extractive summaries which are reported to perform well [14], [15].

Cross-Language Information Retrieval (CLIR) is a subfield of information retrieval that focuses on the problem of retrieving relevant information/documents but written in a different language from the query given by the user [16]-[19]. In general, CLIR system has a translation mechanism, some use it for the queries [20] and others use it for the retrieved documents [21]. Another approach uses an intermediate "language" (interlingua) as a third space where documents and queries will be mapped together [22], [23].

Cross-Lingual Summarization (CLS) is the task of generating a summary in target language from source document in different language [24], [25]. Traditional methods split this task into two steps: summarization and translation. The problem is that the machine translation performance is still considered not good enough. These two steps approach also introduce the problem of error propagation. If in some way the translation process can be skipped, then the summarization time can be reduced. This is very helpful when processing very large data which is commonly encountered nowadays.

Studies on CLS have been done [26]-[28]. Traditional approaches to CLS are generally divided into two, namely extraction-based and compression-based. Most of the extraction-based are similar to the two previously discussed CLIR approaches. The first approach does the translation for the source document while the second approach does the translation for the target summary. Another approach utilizes source documents in two languages, both source and target languages, to produce summaries in the target language [29], [30]. Compression-based generally contains two steps, namely selection and compression. The selection process is carried out to obtain relevant information in the form of sentences or phrases (bilingual). The compression process is carried out by removing sentences or phrases that do not meet certain criteria (e.g., information content, readability, grammar/structure, etc.) [31]-[33]. In [34] an abstractive CLS framework has been introduced. Although it is said to be abstractive, the core of the framework is a pair of bilingual concepts and facts taken extractively from the source document. These pairs are further processed, graded, ranked,

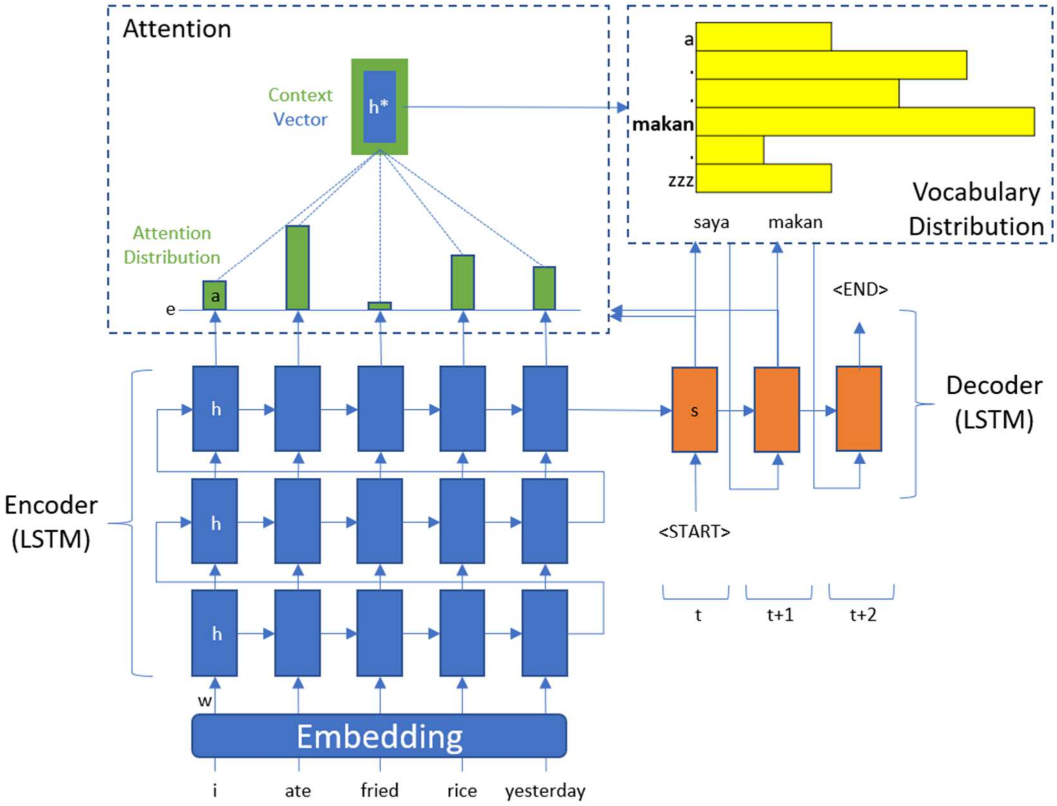

Fig. 1. Sequence-to-sequence model with attention. The model accepts input of English words in the source document and produces Bahasa Indonesia words in the generated summary.

and compiled as a summary. The abstractive process is carried out by combining these pairs and is considered paraphrasing.

In this paper we present an end-to-end abstractive cross-lingual summarization model that can produce Bahasa Indonesia summary from English document. We adapted an approach that is currently quite popular in summarization and machine translation. Although adapting the approach used in machine translation, our approach avoids using machine translators explicitly. We use multilingual embeddings to represent words in both languages. The input (source document) will be represented by the source language embeddings and the output (target summary) will be represented by the target language embeddings, but these two embeddings are already aligned into the same space. This alignment makes it seem as if the two languages are the same language, at least at word level. This paper shows that multilingual word embeddings can be effectively used for CLS problems.

II. Related Works

A. Neural Abstractive Summarization

Rush et al. succeeded in achieving state-of-the-art for abstractive summarization using neural networks [6]. The model adapts the attention mechanism used in machine translation model [8]. The sequence-to-sequence problem is handled using encoder-decoder architecture. The encoder extracts information from the source document (input). The decoder generates a word-by-word summary (output). The attention mechanism solves bottlenecks problem in the encoder by allowing the decoder to access information from the encoder (in the intermediate hidden state).

B. Multilingual Word Embeddings

Lample et al. worked on multilingual word embeddings that are aligned in a common space [35]. Supervised approach requires bilingual resources such as dictionaries or parallel corpus. Lample et al. proposes a method to map monolingual word embeddings in unsupervised way without bilingual data. The implementation can be accessed at MUSE: Multilingual Unsupervised and Supervised Embeddings[1].

III. Methods

The model used in this paper is similar to the model used by Nallapati et al. [5] and See et al. [7] which can be seen in Fig. 1. The differences lie in the encoder layers and the embedding layer. We use encoder with three Long Short-Term Memory (LSTM) layer and we use multilingual word embeddings provided by MUSE in the embedding layer. The pre-trained embeddings act like a lookup table when converting words into their vector representation.

Input in the form of words (w) enter the embedding layer. The words are then converted into their vector representation. This vector then goes to the encoder. The encoder will generate a hidden state (h). On the other hand, at each unit time (t) decoder will receive input:

[1] https://github.com/facebookresearch/MUSE

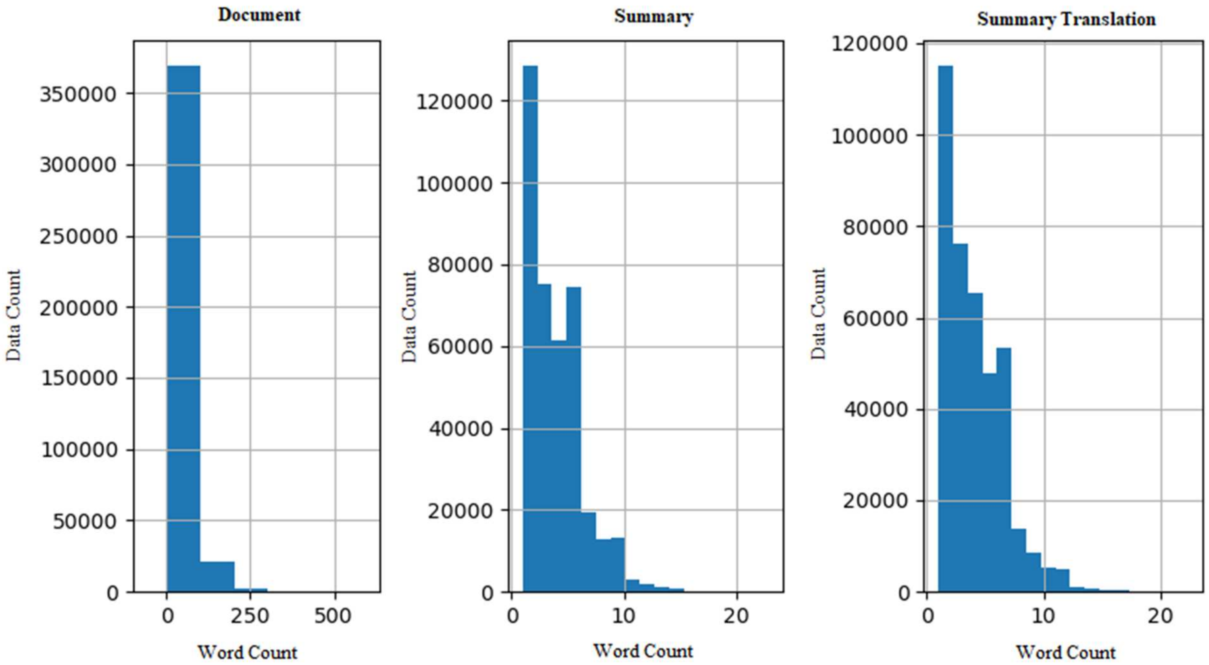

Fig. 2. Distribution of data by length (word count). The x-axis is the word count interval in a document or summary. The y-axis is the amount of data in the interval.

- If during training, then the input is the word embedding of the previous word in the summary, or
- If at test time, then the input is word embedding of the previous word from the decoder

then decoder will generate decoder state (s). Attention distribution (a) was obtained using the formula from Bahdanau et al. [8]:

$$e_i^t = v^t \tanh(W_h h_i + W_s s_t + b_{attn}) \quad (1)$$

$$a^t = softmax(e^t) \quad (2)$$

where v, W_h, W_s, dan b_{attn} are values that can be trained. Attention distribution (a) then used together with encoder hidden state (h) to calculate the context vector (h*) with the following formula:

$$h_t^* = \sum_i a_i^t h_i \quad (3)$$

the result is then concatenated with the state decoder (s) and fed into softmax function to produce vocabulary distribution. There is no separation of summarization and translation steps in the model.

IV. Experiments

A. Dataset

We used dataset of food review from Amazon Fine Food[2] [36], the data is in English. Data captured over a 10-year period, ~500,000 reviews as of October 2012. Reviews include product and user information, ratings, and review text. The detailed description of the data is as follows:

- Timeframe: Oct 1999 - Oct 2012
- 568,454 reviews
- 256,059 users
- 74,258 products
- 260 users with > 50 reviews

Specifically, each review has the full text of the review and its summary. The summary is then translated into Bahasa Indonesia using Google Translate[3]. The review, summary, and summary translation will be used in the experiment. The pre-processing includes: removing duplicate data, removing HTML tags, converting into lowercase, removing punctuation, removing stop words and removing special characters. The distribution of data by length (word count) can be seen in Fig. 2.

B. Training Setup

Experiments were carried out using the following configuration:

- Batch size: 64
- Encoding layer: 3
- Document length: maximum of 30 words
- Summary length: maximum of 8 words
- Word embedding size: 300
- LSTM hidden units: 300

The configuration of document length and summary length is determined by taking into account the distribution of data by length as can be seen in Fig. 2. There are 64,526 unique words in the data. The experiment was divided into two based on the amount of data used. The first experiment used 100,000 data and the second experiment used 500,000 data. Each experiment will train two models, the English to English summarization model (en-en) and English to Bahasa Indonesia summarization model (en-id). The performance of

[2] https://www.kaggle.com/snap/amazon-fine-food-reviews

[3] https://translate.google.com

978-1-6654-2451-6/21//$31.00© 2021 IEEE

TABLE 1. VARIATIONS OF "kabupaten" TOKEN

Token	Category
kabupaten	correct
kabupatennya	correct
sekabupaten	correct
#alihkabupaten	*rare/typo*
bkabupaten%	*rare/typo*
waringinkurungkabupaten	*rare/typo*
kabupaten/kota	incorrect
/kabupaten	incorrect
kabupaten/	incorrect
barat, kabupaten	incorrect

these two models will be compared. This was done because there were no similar studies with readily available data. We wanted to see the performance of cross-lingual compared to monolingual.

We found an issue that could reduce the performance of the model. The issue, especially in the context of Bahasa Indonesia, is the absence of an adequate language tool. In many cases, Bahasa Indonesia language tools are adopted from more established languages, such as English, but the implementation is still not perfect. One example is tokenizer error. Several variations of "kabupaten" token found in MUSE can be seen in Table 1. Token in rare or typo categories can be ignored (discarded) because their frequency is likely to be very low. However, token in incorrect category can be very influential because the frequency is likely to be quite high. The word "kabupaten" is a word that is used quite often. As a result, its position in the vector space may shift slightly.

C. Model Evaluation

To evaluate our models, we use Recall-Oriented Understudy for Gisting Evaluation (ROUGE). ROUGE is a set of metrics and software packages used to evaluate automatic summarization and machine translation [37]-[39]. This metric compares the results of an automatically generated summary or translation against a set of human-generated reference summaries. Some of the available measures are as follows:

- ROUGE-1: measures unigram overlaps between summary and reference results.
- ROUGE-2: measures bigram overlaps between summary and reference results.
- ROUGE-L: measures longest matching sequence of words.

There are other measurements, for more information about these evaluation metrics please refer to [37].

As an illustration, the following is an example of calculating ROUGE-1:

Generated summary: i ate fried rice yesterday

Reference summary: i ate rice yesterday

TABLE 2. ROUGE F1 SCORE

Data	Model	ROUGE		
		1	2	L
100K	en-en	0.11	0.02	0.11
	en-id	0.09	0.03	0.09
500K	en-en	0.12	0.02	0.12
	en-id	**0.14**	**0.04**	**0.14**

TABLE 3. ROUGE PRECISION SCORE

Data	Model	ROUGE		
		1	2	L
100K	en-en	0,11	0,02	0,10
	en-id	0,09	0,02	0,09
500K	en-en	0,12	0,02	0,12
	en-id	**0,15**	**0,05**	**0,15**

TABLE 4. ROUGE RECALL SCORE

Data	Model	ROUGE		
		1	2	L
100K	en-en	0,12	0,03	0,12
	en-id	0,11	0,03	0,11
500K	en-en	0,13	0,03	0,13
	en-id	**0,14**	**0,04**	**0,14**

ROUGE precision can be calculated using the following formula:

$$\text{Precision} = \frac{\text{number of overlapping words}}{\text{total words in generated summary}} \quad (4)$$

Using that formula, ROUGE precision of the example above is 4 divided by 5 or equal to 0.8. This means that 4 out of the 5 words in the generated summary were relevant or correct.

ROUGE recall can be calculated using the following formula:

$$\text{Recall} = \frac{\text{number of overlapping words}}{\text{total words in reference summary}} \quad (5)$$

Using that formula, ROUGE recall of the example above is 4 divided by 4 or equal to 1. This means that all the words in the reference summary have been captured by the generated summary.

Finally, using both scores to calculate ROUGE F1 score:

$$\text{F1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (5)$$

ROUGE F1 score is the harmonic mean of ROUGE precision and ROUGE recall.

V. Results

The experimental results can be seen in Table 2, 3, and 4. The reported ROUGE metrics are ROUGE-1, ROUGE-2, and

TABLE 5. EXAMPLES OF GENERATED SUMMARIES

No.	Source Document (without stopwords)	Reference Summary	Generated Summary
1	always used atora suet living uk makes best dumplings bought mainly use making christmas puddings perfect product	produk sempurna *perfect product*	Terbaik *best*
2	sprite favorite soft drink refreshing goes well kind food delicious	minuman lezat *delicious drink*	minuman yang enak *good drink*
3	drank tea baby started sleep night however could sleep love way relaxed great tea love add sugar	teh yang enak *good tea*	teh yang enak *good tea*
4	love granola met dr flynn outdoor market st petersburg fl bought one bag hooked glad find available amazon think healthy granola complaint think price high understand small producer	lezat *delicious*	granola terbaik yang pernah ada *best granola ever*
5	healthy snack food goes love heat thanks reviewers prepared color hardness ok first pound lasted week suspect addictive	sangat bagus *very good*	camilan enak *delicious snack*
6	son loves sauce however price way high compared price local stores looking buy case price	terlalu mahal *too expensive*	harga yang pantas *fair price*
7	dog loves pig ear strips tried brands really picky please therefore please us	anjing mencintai mereka *dogs love them*	anjing saya suka ini *my dog likes this*
8	recently switched formula looking cheapest price preferred formula cheapest price found anywhere including manufacturer local big box stores package came advertised packaging unlike reviews happy price product value	harga yang bagus *good price*	harga yang bagus *good price*
9	sent gift retired father gathering family proclaimed texas tasty	Keluarga saya menyukainya *My family loves it*	hadiah yang bagus *nice gift*
10	sardines flavorful pleasing tomato sauce salt plus much saturated fat make regular staple nice occasional treat	ikan sarden yang sangat bagus *very good sardines*	mustard yang luar biasa *awesome mustard*

ROUGE-L. The score is calculated using ROUGE 2.0[4]. It can be seen that the difference between monolingual and cross-lingual models is only around 1-2 points, in other words, the performance of the two is comparable. This shows that multilingual word embeddings can be used in CLS problems.

We would like to discuss the low ROUGE score. It should be noted that our model is purely an abstractive model. Several studies have addressed the problem with Bilingual Evaluation Understudy (BLEU) [40], [41]. BLEU and ROUGE have the same problem. Here are some of the problems:

- ROUGE does not pay attention to meaning. If the model returns words with the same meaning, such as synonyms, ROUGE will consider them different and return a bad score.
- ROUGE does not pay attention to sentence structure. A summary whose word order is scrambled, even if the meaning changes, is likely still get the same score.

From the two shortcomings above, it can be seen that ROUGE tends to be more suitable to be used to evaluate extractive summaries and less suitable to be used to evaluate abstractive summaries. Despite having a low ROUGE score, the model can produce meaningfully correct summaries. Examples of the summary can be seen in Table 5. It is necessary to explore and develop evaluation metrics that can assess the performance of abstractive summaries more accurately.

VI. Conclusion

In this paper, we present an end-to-end abstractive cross-lingual summarization (CLS). Experiments show comparable performance with monolingual summarization on the same data. Even though the ROUGE evaluation still gives a low score, reflecting on the state-of-the-art of monolingual summarization, improvement can be achieved by accommodating extractive mechanisms. The pure abstractive route can also be taken, considering that high-level abstraction is still an open problem.

In future work, we would like to investigate the performance of our model for longer and more complex data. We will also compare our end-to-end approach with the more traditional 2-step approach (summarization and translation) to better understand their strengths and weaknesses. Finally, due to the weakness of ROUGE evaluation, we want to perform a human evaluation to ensure that the resulting abstractive summaries are good.

Acknowledgment

The authors would like to express their gratitude to Mirna Adriani for her guidance and direction in this research. The authors would also like to thank Lab 1231 Fasilkom UI team, especially Machmud Roby Alhamidi and Grafika Jati, for their helpful feedback and discussion. Finally, the authors would like to acknowledge the contribution of LIPI high-performance computing facilities to the results of this research. URL - https://hpc.lipi.go.id.

References

[1] C. D. Paice, "Constructing literature abstracts by computer: techniques and prospects," Information Processing & Management, vol. 26, p. 171–186, 1990.

[2] J. Kupiec, J. Pedersen and F. Chen, "A trainable document summarizer," in Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, 1995.

[3] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems, 2014.

[4] S. Chopra, M. Auli and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.

[4] http://rxnlp.com/rouge-2-0

[5] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang and others, "Abstractive text summarization using sequence-to-sequence rnns and beyond," arXiv preprint arXiv:1602.06023, 2016.

[6] A. M. Rush, S. E. A. S. Harvard, S. Chopra and J. Weston, "A neural attention model for sentence summarization," in ACLWeb. Proceedings of the 2015 conference on empirical methods in natural language processing, 2017.

[7] A. See, P. J. Liu and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," arXiv preprint arXiv:1704.04368, 2017.

[8] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[9] B. Sankaran, H. Mi, Y. Al-Onaizan and A. Ittycheriah, "Temporal attention model for neural machine translation," arXiv preprint arXiv:1608.02927, 2016.

[10] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in Proceedings of the 25th international conference on Machine learning, 2008.

[11] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, "Natural language processing (almost) from scratch," Journal of machine learning research, vol. 12, p. 2493–2537, 2011.

[12] 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, 2013.

[13] R. Collobert, "Word embeddings through hellinger pca," in in Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014.

[14] G. Rossiello, P. Basile and G. Semeraro, "Centroid-based text summarization through compositionality of word embeddings," in Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres, 2017.

[15] M. Isonuma, T. Fujino, J. Mori, Y. Matsuo and I. Sakata, "Extractive summarization using multi-task learning with document classification," in Proceedings of the 2017 Conference on empirical methods in natural language processing, 2017.

[16] J. Wang, Matching meaning for cross-language information retrieval, University of Maryland, College Park, 2005.

[17] J. Wang and D. W. Oard, "Matching meaning for cross-language information retrieval," Information processing & management, vol. 48, p. 631–653, 2012.

[18] G. Grefenstette, Cross-language information retrieval, vol. 2, Springer Science & Business Media, 2012.

[19] G. Jena and S. Rautaray, "A Comprehensive Survey on Cross-Language Information Retrieval System," Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), vol. 14, pp. 127-134, 2019.

[20] J. Gao, J.-Y. Nie, E. Xun, J. Zhang, M. Zhou and C. Huang, "Improving query translation for cross-language information retrieval using statistical models," in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001.

[21] D. W. Oard and P. G. Hackett, "Document translation for cross-language text retrieval at the University of Maryland," 1997.

[22] M. Ruiz, A. Diekema, P. Sheridan and D. C. Plaza, "CINDOR conceptual interlingua document retrieval: TREC-8 evaluation," in TREC, 1999.

[23] K. Kishida and N. Kando, "A hybrid approach to query and document translation using a pivot language for cross-language information retrieval," in Workshop of the Cross-Language Evaluation Forum for European Languages, 2005.

[24] X. Wan, H. Li and J. Xiao, "Cross-language document summarization based on machine translation quality prediction," in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010.

[25] F. Boudin, S. Huet and J.-M. Torres-Moreno, "A graph-based approach to cross-language multi-document summarization," Polibits, p. 113–118, 2011.

[26] W. Ogden, J. Cowie, M. Davis, E. Ludovik, H. Molina-Salgado and H. Shin, "Getting information from documents you cannot read: An interactive cross-language text retrieval and summarization system," in Joint ACM DL/SIGIR workshop on multilingual information discovery and access, 1999.

[27] H. Saggion, D. R. Radev, S. Teufel, W. Lam and S. M. Strassel, "Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Cross-lingual Environment.," in LREC, 2002.

[28] L. Yu and F. Ren, "A study on cross-language text summarization using supervised methods," in 2009 international conference on natural language processing and knowledge engineering, 2009.

[29] X. Wan, "Using bilingual information for cross-language document summarization," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.

[30] X. Wan, F. Luo, X. Sun, S. Huang and J.-g. Yao, "Cross-language document summarization via extraction and ranking of multiple summaries," Knowledge and Information Systems, vol. 58, p. 481–499, 2019.

[31] J. G. Yao, X. Wan and J. Xiao, "Phrase-based compressive cross-language summarization," in Proceedings of the 2015 conference on empirical methods in natural language processing, 2015.

[32] E. L. Pontes, S. Huet and J.-M. Torres-Moreno, "A multilingual study of compressive cross-language text summarization," in Mexican International Conference on Artificial Intelligence, 2018.

[33] E. L. Pontes, S. Huet, J.-M. Torres-Moreno and A. C. Linhares, "Cross-language text summarization using sentence and multi-sentence compression," in International Conference on Applications of Natural Language to Information Systems, 2018.

[34] J. Zhang, Y. Zhou and C. Zong, "Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, p. 1842–1853, 2016.

[35] A. Conneau, G. Lample, M. Ranzato, L. Denoyer and H. Jégou, "Word translation without parallel data," arXiv preprint arXiv:1710.04087, 2017.

[36] J. J. McAuley and J. Leskovec, "From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews," in Proceedings of the 22nd international conference on World Wide Web, 2013.

[37] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, 2004.

[38] C. Y. Lin, G. Cao, J. Gao and J. Y. Nie, "An information-theoretic approach to automatic evaluation of summaries," in Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, 2006.

[39] K. Ganesan, "Rouge 2.0: Updated and improved measures for evaluation of summarization tasks," arXiv preprint arXiv:1803.01937, 2018.

[40] E. Sulem, O. Abend and A. Rappoport, "Bleu is not suitable for the evaluation of text simplification," arXiv preprint arXiv:1810.05995, 2018.

[41] E. Reiter, "A structured review of the validity of BLEU," Computational Linguistics, vol. 44, p. 393–401, 2018.