Cross Language Information Retrieval Using Parallel Corpus with Bilingual Mapping Method

Rinaldi Andrian Rahmanda, Mirna Adriani, Dipta Tanaya
Faculty of Computer Science, Universitas Indonesia
Depok, Indonesia
rinaldi.andrian@ui.ac.id, mirna@cs.ui.ac.id, diptatanaya@cs.ui.ac.id

***Abstract*—This study presents an approach to generate a bilingual language model that will be used for CLIR task. Language models for Bahasa Indonesia and English are created by utilizing a bilingual parallel corpus, and then the bilingual language model is created by learning the mapping between the Indonesian model and the English model using the Multilayer Perceptron model. Query expansion is also used in this system to boost the results of the retrieval, using pre-Bilingual Mapping, post-Bilingual Mapping and hybrid approaches. The results of the experiments show that the implemented system, with the addition of pre-Bilingual Mapping query expansion, manages to improve the performance of the CLIR task.**

Keywords-- Cross Language Information Retrieval; parallel corpus; Bilingual Mapping; Language Model; Multilayer Perceptron

I. INTRODUCTION

Cross Language Information Retrieval deals with searching documents written in different languages with user's queries. It requires a translation process due to language difference between queries and documents. In general, two methods which could be used to solve this problem are query/documents translation and bilingual model approach [1]. Translation approach looks for matching words in both languages by using dictionary [2], machine translation [3], or parallel corpus [4]. In this approach, a high-quality bilingual dictionary or machine translator resource are needed to get good results. Many studies have been carried out using this translational approach. Ballesteros and Croft's research uses a bilingual dictionary that can be read by a computer (Machine Readable Dictionary) to translate queries [2]. Oard's study uses a bilingual dictionary and machine translation in a comparative study [3]. Research by [4] uses a parallel corpus to search for pairs of words in both languages to form queries for the target language.

Meanwhile, in the bilingual model approach, each document in both languages is mapped into a vector space, represented by word embedding. Word embedding is capable to capture word similarity, context of a word in a document, word relation, etc. By using word embedding, the similarity between query and documents can be measured by calculating the vectors similarity. This bilingual language approach has been studied by [5]. In this study, Vulic and Moens used the bilingual language model approach using word embedding for CLIR.

An excess potential of the bilingual language model approach using word embedding compared to the translational approach is how similar words will have high similarity values. This will be useful in cases where the words used in the query are translated differently from the words contained in the desired document, even though they have the same meaning. For example, in the case of an Indonesian-English CLIR, a query containing "*senang*" (in English: glad) was given. Meanwhile, documents that are expected to contain the word "glad". In the translation approach, the system will fail to search for documents if the word "*senang*" (in English: glad) is translated into words other than "glad", for example "happy". Fortunately, in the language model approach, this does not matter because of the high similarity of the two words, "*senang*" (in English: glad) and "glad".

The potential of the language model approach, and the success of Vulic and Moens' research [5], inspired this research. This study will try to apply language model approach in a CLIR system. To create a bilingual model that will be used by the CLIR system, the bilingual mapping method as done in [6]. First, we use a bilingual parallel corpus from Global Voices. the language model or language representation for both languages in the parallel corpus is formed using word embedding. Then, the bilingual model will map the representation of the language. This mapping system will use the neural network classifier to get the best transformation for all existing training data. Using this model, CLIR will be done by taking a query, constructing a vector representation, mapping the vector into English, then searching for the relevant documents.

Research on CLIR is important to do because of the gaps phenomenon in the availability of information in a specific language. For example, we can observe that the amount of information available on the internet in Indonesian is still less than the information available in English. The CLIR system can help overcome this problem by finding and displaying available information, even though the information is available in other languages. The research on Cross Language Information Retrieval is expected to help overcome the problem of information gap.

978-1-7281-5014-7/19/$31.00

II. Research Design

The following steps we employ to create bilingual model for CLIR task:

1. Preparing parallel corpus
2. Training word embedding as word representation
3. Learning bilingual mapping
4. Use the bilingual model for CLIR task
5. Improving CLIR results by query expansion

A. Parallel Corpus

Bilingual models for this system will be built with the help of a parallel corpus. The parallel corpus contains a collection of pairs of sentences in Indonesian and English. These pairs of sentences will be used in the training process of mapping words from Indonesian to English. We use a news collection published on the GlobalVoices website[1] in Indonesian and English, taken from the OPUS[2] open-source parallel corpus project [7]. The corpus has 11,488 sentence pairs, around 200,000 Indonesian language tokens and around 200,000 English token. Information taken from GlobalVoices is the parallel corpus as well as the pairs of Indonesian and English words found in the corpus. The dictionary corpus contains 4,851 pairs of words that can be found in the parallel corpus.

B. Making Word Representations from Parallel Corpus

We use word embedding to represent words in parallel corpus. Each word in the corpus will be converted into a vector representation in a certain dimension. Word vector representation is made using the Doc2Vec model from the gensim[3], a library used for performing several NLP and IR tasks such as modeling topics, document indexing and similarity retrieval [8]. The word embedding building includes the following stages:

1. Prepare the GlobalVoices parallel corpus containing 11.488 pairs of sentences.
2. Initialize the Doc2Vec model. The model is initialized by selecting the Distributed Bag of Words (DBOW) training algorithm, using the dbow_words option so that the model produces the word vector. As for the dimensions of vectors, we will try several size variations and look for the best vector size.
3. Train the Doc2Vec. The training process will produce a word embedding model where each word in the corpus has a corresponding vector. By using the Doc2Vec model, each sentence in the corpus will also have a corresponding vector, as well as one or more labels as the identifier of the document.

C. Bilingual Mapping

Bilingual language models are formed using the bilingual mapping method. This method uses word representation vectors of documents in both languages. Then, using corresponding word vectors from both languages, a transformation that can map vectors from the original language to the target language is formed. Mapping will be done using Multilayer Perceptron.

After doing word embedding training from parallel corpus, word vectors will be obtained. The next step is to use the corpus dictionary to get pairs of corresponding words in the corpus. For each pair of words on this dictionary, take the word vector for the word from both languages, then use these two vectors as the input and output expected from the training process. Through this process, a bilingual language model that can predict the target language word from the input word of the original language will be formed. For example, if we have the Indonesian word vector "apel" (in English: apple) [0.5, 0.3, 0.1, 0.9, 0.3] and the "apple" English word vector [0.1, 0.6, 0.8, 0.9, 0.8], the transformation system is expected to find a transformation which can map vectors [0.5, 0.3, 0.1, 0.9, 0.3] to [0.1, 0.6, 0.8, 0.9, 0.8]. According to [9], in making bilingual models based on bilingual mapping, the resulting bilingual model will be more optimal if the dimensions of the original language vector representation are two to four times greater than the target language vector representation. In this study, several variations of word embedding dimension models will be used to see the relationship between the dimensions of the model and the quality of the results of the CLIR system.

D. CLIR System

Cross Language Information Retrieval is done by using the bilingual language model. The process of retrieving relevant documents is done by comparing document vectors and query vectors. For this reason, the document embedding method is carried out on news documents from the corpus in order to obtain document vector from each document.

To do the cross language information retrieval, several steps were taken from adaptations of the steps used in [5]:

1. take the query,
2. using the existing model, get the target language version of the query,
3. get the closest documents from the query by measuring the similarity score between the query and the document,
4. sort the document based on the similarity score.

Retrieval in this experiment consists of two types, monolingual retrieval and bilingual retrieval. The results of monolingual retrieval will be the baseline of the CLIR system. The retrieval process will be carried out on two corpus documents:

1. GH95, a news corpus from the 1995 Glasgow Herald.

[1] http://globalvoices.org/

[2] http://opus.nlpl.eu/

[3] http://radimrehurek.com/gensim/

2. LAT94, a news corpus from the 1994 Los Angeles Times newspaper.

The query is taken from CLEF (Cross Language Evaluation Forum)[4]. We take 50 queries that have the Indonesian and English versions. Each query has a title section (topic), a description of what information is expected (description), and a brief narrative about the desired information (narration). From these sections, the title and description are taken only referring to previous research [5] [10].

E. Query Expansion

To improve the quality of the obtained relevant documents, we add query expansion to the retrieval process. According to [11], query expansion has the potential to improve the quality of information retrieval results because query expansion can overcome the main problem that decreases search quality, that is the problem when the words in the query do not match the expected document (vocabulary problem).

One of the query expansion methods that can be done is using the word embedding model. In this method, each word in the query is expanded using the words closest to the word in word embedding space. The similarity between words is measured using the cosine similarity value between word vectors. For example, if given a "car" query, and the closest words to the word "car" in word embedding space are "vehicles" and "automotive" then the query will be expanded to "automotive vehicle cars". This word embedding based method has been investigated in [12] and produce better results compared to queries that are not expanded.

In this study, we try 3 expansion query methods: Pre-BM Query Expansion, Post-BM Query Expansion, and a combination of both methods. Pre-BM Query Expansion expands the query before mapping proses to the target language is done. Post-BM Query Expansion expands the query after mapping query in Indonesian to English using a bilingual language model. The combined method combines Pre-BM Query Expansion and Post-BM Query Expansion.

F. Evaluation

IR system evaluation is done by comparing the retrieval results from the IR system with the gold standard that has been made based on the results of manual evaluations by humans. Evaluation is done by measuring the value of Mean Average Precision (MAP) from the results of the retrieval.

Average Precision is the average value of precision for each document that is considered relevant to a query. The formula for calculating Average Precision is as follows [13]:

$$AP = \frac{\sum_{d=1}^{n} P(d) x\ rel(d)}{number\ of\ relevant\ documents}$$

where n is the number of documents that should be considered relevant, d is the document number, $P(d)$ is the value of the precision query on the document and $rel(d)$ is 1 if the document is relevant; 0 if the document is not considered relevant.

For calculating MAP, we use formula as follows [13]:

$$MAP = \frac{\sum_{q=1}^{n} AP(q)}{n}$$

The formula calculates the mean of Average Precision from each query. The q variable represents a query, $AP(q)$ is Average Precision from the query, and n is the number of queries.

III. Results and Discussion

A. Language Model

The results of the word embedding training for each language are monolingual language models for both languages. From the use of the DBOW + dbow_words algorithm for this training, word vectors are obtained for each word in the corpus. The results of training data in English can be directly used for monolingual English document retrieval.

In this study, a bilingual language model was also produced. Similar to the monolingual language model, bilingual language models can search for words with a high level of similarity. The difference is that when an Indonesian word is entered as input, the system gets a vector representation of the word, and converts that representation into an English form. Then the system will look for similar vectors in English using this new vector. For example, the search for the word "perang" (in English: war). The system will get the words war, wars, unrest, apartheid, perestroika, strike, warfare, battle, recession, and riot.

B. Retrieval Results

Information retrieval is done in several scenarios, and the results are assessed using the trec_eval application. Performance is measured using MAP metrics, or Mean Average Precision. First, as a baseline for this language model approach, retrieval is done using Okapi BM-25. Retrieval is done twice, monolingual (English to English) and cross-language (Indonesian to English). Cross language retrieval is done by previously translating the query Indonesian into English using Google Translate. This is called the method Google Translate + Okapi BM-25. The results of this method can be seen in Table 1.

Table 1 Okapi BM-25 Evaluation (Baseline)

	Method	MAP
1	Okapi BM-25	0.3869
2	Google Translate+Okapi BM-25	0.3370

[4] http://clef.isti.cnr.it/

Then, retrieval using a monolingual language mode (English to English) is done, as a baseline for the performance of cross-language retrievals performed. The four dimensions of word embedding are used at this stage, namely 100, 200, 300, and 400. The results of this stage can be seen in table 2.

Table 2 shows the results of evaluation of monolingual information retrieval (English queries to English documents) using the language model of word embedding. In general the results obtained here are worse than those obtained from the monolingual Okapi BM-25 method.

Table 2 Monolingual (English-English) Retrieval Evaluation

Method	MAP	vs baseline	Method	MAP	vs baseline
WE-100	0.3623	-2.46%	WE-300	0.3654	-2.15%
WE-200	0.3693	-1.76%	WE-400	0.3697	-1.72%

Furthermore, cross-language retrieval is done. Several combinations of word embedding dimensions from both languages are used. The results of this stage in more detail can be seen in table 3. The following are the results of evaluation retrieval using the trec_eval application:

Table 3 Bilingual Retrieval Evaluation

	Method	MAP	vs baseline	vs mono-lingual
1	Monolingual (WE) -> WE-English-100	0.2967	-4.03%	-7.3%
2	WE-Ind-100 -> WE-English-100	0.2967	-4.03%	-7.3%
3	WE-Ind-200 -> WE-English-100	0.3328	-0.42%	-3.69%
4	WE-Ind-200 -> WE-English-200	0.2843	-5.27%	-8.54%
5	WE-Ind-300 -> WE-English-100	0.3353	-0.17%	-3.44%
6	WE-Ind-300 -> WE-English-200	0.3251	-1.19%	-4.46%
7	WE-Ind-300 -> WE-English-300	0.2798	-5.72%	-8.99%
8	WE-Ind-400 -> WE-English-100	0.3525	+1.55%	-1.72%
9	WE-Ind-400 -> WE-English-200	0.3369	-0.01%	-3.28%
10	WE-Ind-400 -> WE-English-300	0.3163	-2.07%	-5.34%
11	WE-Ind-400 -> WE-English-400	0.2651	-7.19%	-10.46%

Table 3 shows that the best results from the CLIR system are obtained when the system uses word embedding using an Indonesian language vector of 400 dimensions and an English language vector with a dimension of 100. The greater use of initial language vectors gives better retrieval results. However, large vector dimension results increasing training time.

Monolingual query is still better compared to the cross-language retrieval result. A possible cause of this phenomenon lies in the process of vector transformation between languages. In the process of bilingual mapping, the vector of an Indonesian word is not 100% mapped with English. For example, the word “perang” (in English: war) is not 100% the same as the word "war" in English (the similarity is 0.8491). This difference can cause changes in the value of the relevance of some documents and lead to the worse results.

The results of the developed CLIR system retrieval are better than the results of the Okapi BM-25 using Google Translate of 0.0155 or 1.55%. However, the results obtained using the monolingual language model are worse than the results of the Okapi BM-25 method, with lower MAP results of 0.0172 or 1.72%. The use of language models in information retrieval tasks like this certainly depends on the quality of the data used in the training steps. The data used must be large enough and cover the contents of the corpus that is the target of retrieval. The language style of training data may also have an effect. But the biggest factor why the use of this language model is still losing is the mapping done from documents in the form of words to vector space. By mapping words in a document into a vector, some information inherent in the document might be lost. This resulted in a decrease in retrieval performance that could be done, and resulted in the Okapi BM-25 algorithm as the baseline still leading in the monolingual retrieval aspect.

Using query expansion methods, the following results are obtained:

Table 4 Query Expansion Evaluation

	Method	MAP Non-QE	MAP QE Pre-BM	MAP QE Post-BM	MAP Pre-Post BM QE
1	WE-Ind-100 -> WE-English-100	0.2967	0.3119	0.3029	0.2987
2	WE-Ind-200 -> WE-English-100	0.3328	0.3538	0.3533	0.3589
3	WE-Ind-200 -> WE-English-200	0.2843	0.3018	0.3018	0.2977
4	WE-Ind-300 -> WE-English-100	0.3353	0.3564	0.3556	0.3554
5	WE-Ind-300 -> WE-English-200	0.3251	0.3446	0.3356	0.3398
6	WE-Ind-300 ->	0.2798	0.2952	0.2945	0.2938

	Method	MAP Non-QE	MAP QE Pre-BM	MAP QE Post-BM	MAP Pre-Post BM QE
	WE-English-300				
7	WE-Ind-400 -> WE-English-100	0.3525	0.3746	0.3739	0.3727
8	WE-Ind-400 -> WE-English-200	0.3369	0.357	0.3571	0.3559
9	WE-Ind-400 -> WE-English-300	0.3163	0.3343	0.3347	0.3345
10	WE-Ind-400 -> WE-English-400	0.2651	0.2807	0.2814	0.2803

Table 4 shows the results obtained from the use of query expansion methods, with three approaches. The table above shows an increase in MAP in all dimensions variations used. The results of Indonesia word embedding with dimensions 400 and English word embedding with dimension 100 by using the query expansion method Pre-BM managed to get MAP of 0.3746 which is higher than non-QE retrieval of 0.0221 or 2.21%. This result is 3.76% better from the baseline. This is the best result obtained by this system.

IV. Conclusion

In conducting Cross Language Information Retrieval, various methods can be applied. One of them is using a bilingual language model. Bilingual language model is created from the results of transformation of the original language word vector to the targeted word vector language.

In this study, an Indonesian-English CLIR system is developed using a bilingual language model approach. A parallel corpus containing documents in Indonesian language and English are used as the basis of the model. Word embedding and document embedding are carried out to get the vector representations of each language. Then, the Multilayer Perceptron is used to map vectors from Indonesian language to corresponding vectors in English to form a bilingual language model.

CLIR is then done by taking a query in Indonesian, getting the vector representation, mapping the vector into English, and finally looking for the documents that best match the query vector. Experiments were carried out by comparing the MAP values from the results obtained by the CLIR system to the Okapi-BM25 baseline and monolingual retrieval results using language models with several vector dimension combinations. The best results obtained showed a MAP value of 35.25%, 1.55% better than the baseline result, but 1.72% lower than the monolingual result.

To improve the search results obtained, query expansion methods are used to enrich search words. Three query expansion approaches are carried out: query expansion before query mapping in bilingual language model (Pre-BM), query expansion after query mapping (PostBM), and a combination of the two approaches.

The experiment was done to examine the quality of the query expansion. The experiments show increasing results for all vector combinations, with the Pre-BM approach giving better results than the other two approaches. The best result obtained at this stage is the MAP value of 37.46%, which is 3.76% better from baseline.

To conclude, a CLIR system was developed using a bilingual language model with a bilingual mapping approach. The method of query expansion is done to enrich the query with the aim of improving search results. The best result obtained by the system is MAP of 37.46%, an increase from the baseline of 3.76%.

After conducting experiments and analyzing the results obtained in this study, there are some suggestions that might be useful for further research:

1. In conducting training in making bilingual language models, the quality and quantity of training data is very important. For further research, a larger parallel corpus can be used with more words and documents so that the mapping process can be done better.
2. In the process of bilingual mapping, other types of classifier other than MLP can be used, for example Convoluted Neural Network (CNN).
3. This Bilingual Mapping method can be tried in other languages to see the difference in performance.
4. The bilingual mapping approach used in this study is only one-way, from Indonesian to English. For further research, the mapping process can be done in two-way.
5. Additional features can be explored to conduct information retrieval, such as Latent Semantic Indexing.

ACKNOWLEDGEMENTS

We gratefully thank the Universitas Indonesia for the International Publication Grants (*Hibah PITTA-B*)

References

[1] P. Sorg and P. Cimiano, "Cross-lingual information retrieval with explicit," in *In Working Notes for the CLEF 2008 Workshop*, 2008.

[2] L. Ballesteros and W. B. Croft, "Phrasal translation and query expansion," *ACM,* vol. 31, pp. 84-91, 1997.

[3] D. W. Oard, "A comparative study of query and document translation," in *Conference of the Association for Machine Translation in the Americas*, 1998.

[4] . J.-Y. Nie, M. Simard, P. Isabelle and Durand, "Cross-language information retrieval based on parallel texts and automatic mining of parallel texts

on the web," in *22nd annual international ACM SIGIR conference on Research and development in information retrieval*.

[5] I. Vulic and M. -F. Moens, "Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings," in *38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015.

[6] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," in *Workshop at ICLR*, 2013.

[7] J. Tiedemann, "Parallel data, tools and interfaces in opus," in *Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 2012.

[8] R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010.

[9] T. Mikolov, Q. V. Le and I. Sutskever, *Exploiting similarities among languages for machine translation,* CoRR, abs/1309.4168, 2013.

[10] I. Vulic, W. De Smet and M. F. Moens, "Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora," *Information Retrieval,* vol. 3, no. 16, pp. 331-368, 2013.

[11] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Computing Surveys (CSUR),* vol. 44, no. 1, 2012.

[12] D. Roy, D. Paul, M. Mitra and U. Garain, *Using word embeddings for automatic query expansion,* CoRR, abs/1606.07608, 2016.

[13] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," in *29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006.