KC4Align: Improving Sentence Alignment Method for Low-resource Language Pairs

Ha Nguyen Tien
Multilingual Translation Project KC4.0
VNU - UET, Hanoi, Vietnam
tienhapt@gmail.com

Dat Nguyen Huu
Multilingual Translation Project KC4.0
VNU - UET, Hanoi, Vietnam
cs.nguyenhuudat@gmail.com,

Huong Le Thanh
Hanoi University of Science and Technology
Hanoi, Vietnam
huonglt@soict.hust.edu.vn

Vinh Nguyen Van and Minh Nguyen Quang
Multilingual Translation Project KC4.0
VNU - UET, Hanoi, Vietnam
vinhnv2000@gmail.com and
nqminh@vnu.edu.vn

Abstract

Bilingual corpus is an important resource for many applications of natural language processing (NLP). About low-resource language pairs, it is more necessary to build them automatically. In this paper, we propose a novel method that uses an embedding model with a sentence length ratio to align sentences from bilingual documents for low-resource language pairs. Our method is inspired by Vecalign, the state-of-the-art method, and overcomes its limitation. Experiments on the low-resource Vietnamese-Lao language pair show that our proposed method achieves higher precision and recall than other good methods even on noisy data.

1 Introduction

Low-resource languages (LRLs)can be understood as less studied, resource scarce, less computerized, less privileged, less commonly taught, or low density, among other denominations (Magueresse et al., 2020), such as Lao, Khmer, Urdu, etc. It is difficult to build large datasets of such languages for the training task. Because of that, systems that apply machine learning or deep learning approaches often give low results or fail to work with such languages.

Machine translation has achieved significant improvement due to the application of deep learning models. This approach only works well with a large parallel corpus for training the model. To create such a corpus, we need good resources of parallel text, as well as a mechanism to align sentences in bilingual text pairs.

The sentence alignment can be done by creating all candidate sentence pairs from bilingual documents and then computing the semantic similarity between these sentence pairs. The simplest method to measure their semantic similarity is to count the common words between the two documents using bilingual dictionary. Some researches translate the two documents to a third language which is a rich-resources and popular one such as English, then use a similarity score such as Levenstein or cosine to compute their similarity (Abdul-Rauf and Schwenk, 2009) (Sennrich and Volk, 2011). Recent works (Grégoire and Langlais, 2018) (Artetxe and Schwenk, 2019a) (Thompson and Koehn, 2019a) (Chousa et al., 2020) apply a multi-lingual language model such as BERT or LASER to represent their sentence embedding vectors and compute the cosine similarity between these vectors. However, this method only provides good results for the language pairs that have enough training data for the language model. This method cannot apply to the language pairs which one of them is not supported by the multilingual language model. Vietnamese Laos is one of such language pairs. There is not any language embedding model support Laos language. LRLs also suffer from this problem.

As far as we know, there is no Vietnamese Laos dataset available yet. VecAlign - a famous sentence alignment tool cannot provides good results for this language pair. To deal with this problem, we propose to choose a machine translation system that can provide a good translation from Laos to Vietnamese. The reason for this language selection is that the translation system can use different words

to express the same meaning. Therefore, to measure the similarity between two text pairs, a semantic score should be used to measure the similarity between their semantic vectors, which are created by a language embedding model. In addition, we propose several techniques to align sentences from two documents in the same language.

Our contribution as follows:

1. The idea of using machine translation to translate text from low-resource language to another language that is supported by a language embedding model;

2. Propose a method to limit the error propagation in sentence alignment and to reduce the candidate sentence alignment pairs to be considered;

3. Improve Vecalign to overcome its limitations on pairs of sentences that are highly similarity but are not translations of each other.

Besides, we also public a sentence alignment tool for Vietnamese - Laos and a sentence segmentation tool.

The rest of the paper is laid out as follows: Section 2 presents the related works; Section 3 our proposed method; Section 4 presents experiment results, and we conclude and present future work with Section 5.

2 Related works

Sentence alignment is the task that automatically extracts parallel sentences from noisy parallel documents to build a sentence-aligned bilingual corpus that is used in many applications of NLP. In the past, there have been many works proposing methods of automatic sentence alignment that use some language features such as sentence length, dictionaries, document structure, etc.

Ma (2006) proposed a method that is called Champollion, which performs sentence alignment based on lexicon. Champollion increases the robustness of the alignment by assigning greater weights to less frequent translated words. It is designed for robust alignment of potential noisy parallel text. The disadvantage of this method is that it requires a bilingual dictionary during implementation. The decision of this method depend on the size and quality of bilingual dictionary.

Varga et al. (2007) proposed a method that is called Hualign. Its input is tokenized and sentence-segmented text in two languages. In the simplest case, its output is a sequence of bilingual sentence pairs. This method solves the sentence alignment problem as follows: In the presence of a dictionary, hunalign uses it, combining this information with Gale-Church sentence-length information. In the absence of a dictionary, it first falls back to sentence-length information, and then builds an automatic dictionary based on this alignment. Then it realigns the text in a second pass, using the automatic dictionary.

Kaufmann (2012) proposed a method that is called JMaxAlign. it uses a maximum entropy classifiers to detect parallel sentences between any language pairs with small amounts of training data. The accuracy of this method depends on the quality and size of the training data.

Ha et al. (2018) proposed the improvement of an language - independent sentence alignment method (Huyn and Rossignol, 2006) for VietnameseEnglish bilingual texts that called viXAlign. viXAlign extends to m-to-n alignment (m,n are sentence numbers in source and target language, respectively) and proposes a suitable penalty value in DTW algorithm for the English-Vietnamese language pair.

Recently, most of the works focus on using deep learning network for bilingual sentence alignment:

Thompson and Koehn (2019b) proposed a novel bilingual sentence alignment method which is linear in time and space that called Vecalign. Vecalign based on similarity of sentence embeddings and a DP(Dynamic Programming) approximation. It works in about 100 languages, without the need for a machine translation system or lexicon. It uses similarity of multilingual sentence embeddings to judge the similarity of sentences and an approximation to Dynamic Programming based on Fast Dynamic Time Warping which is linear in time and space with respect to the number of sentences being aligned. Experiments show that this method has state-of-the art accuracy in high and low resource settings and improves downstream machine translation quality.

Artetxe and Schwenk (2019b) proposed a method for sentence alignment that based on multilingual sentence embeddings. In contrast to previous ap-

proaches, which rely on nearest neighbor retrieval with a hard threshold over cosine similarity, their proposed method accounts for the scale inconsistencies of this measure, considering the margin between a given sentence pair and its closest candidates instead. Their experiments show large improvements over existing methods. Their method's result outperform the best published results on the BUCC mining task and the UN reconstruction task by more than 10 F1 and 30 precision points, respectively.

Chousa et al. (2020) proposed a novel method of automatic sentence alignment from noisy parallel documents. Firstly, they formalize the sentence alignment problem as the independent predictions of spans in the target document from sentences in the source document. Then they introduce a total optimization method using integer linear programming to prevent span overlapping and obtain non-monotonic alignments. They implement cross-language span prediction by fine-tuning pre-trained multilingual language models based on BERT architecture and train them using pseudo-labeled data obtained from unsupervised sentence alignment method. While the baseline methods use sentence embeddings and assume monotonic alignment, their method can capture the token-to-token interaction between the tokens of source and target text and handle non-monotonic alignments.

Luo et al. (2021) proposed an unsupervised sentence alignment method and explores features in training biomedical neural machine translation systems. They use a simple but effective way to build bilingual word embeddings to evaluate bilingual word similarity and transferred the sentence alignment problem into an extended earth mover's distance problem. This method achieved high accuracy in both 1-to-1 and many-to-many cases. Pretraining in general domain, the larger in-domain dataset and n-to-m sentence pairs benefit the neural machine translation model. Fine-tuning in domain corpus helps the translation model learns more terminology and fits the in-domain style of text.

Although methods using deep learning network have shown superior performance compared to previous ones, it requires large training data. This is the biggest difficulty when applying deep learning approaches to low-resource language pairs. In addition, the learned models sometimes do not cover all languages features. Besides the above weakness, the state-of-the-art method VecAlign also has other limitations. First, it cannot align pairs of sentences that are located far apart in the source and target documents. Second, two sentences that are not translations of each other but have a highly similarity still be aligned by VecAlign. In this paper, we propose a sentence alignment method that overcomes the above-mentioned limitations by taking advantages of machine translation tools and improving VecAlign.

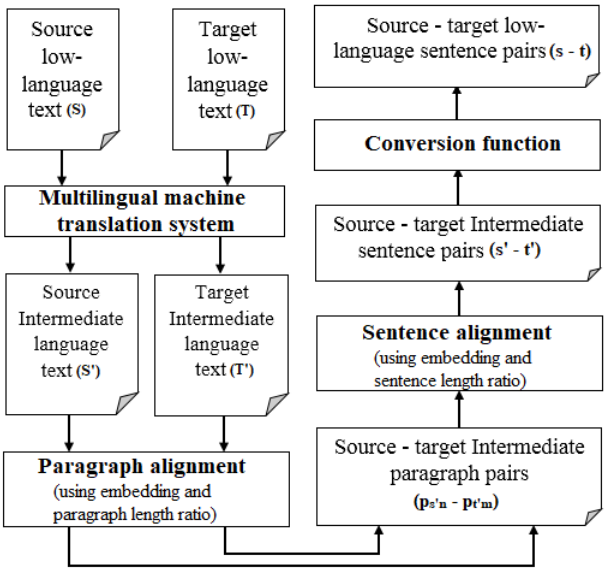

Figure 1: Sentence Alignment for Low-resource Language Pairs

3 Our proposed method

Given parallel or comparable documents in two languages, the task is to match minimal groups of sentences that are translations of each other. The comparable, or semi-parallel documents, are the ones in two languages containing similar information. Some research generates candidate pairs by mapping all possible sentences from the two documents, then evaluate the similarity between them to get the final bilingual sentence pairs. This approach has two problems. First, the computational cost of the alignment task is high when pairing all possible sentence pairs. Second, an alignment error can propagate from one pair to another, since the sentence that should be in one pair is moved to another pair. To

Lao sentence: ກອງທະຫານເຄື່ອນທີ່ໄປແລະ ສົບຂອງຊາຍຄົນນັ້ນບໍ່ເຄ

Translated by GoogleTranslator: Quân đội di chuyển và thi thể của người đàn ông không bao giờ được tìm thấy. (*The army moved and the man's body was never found.*)

Reference sentence: Quân đội sau đó lại tiếp tục hành quân và xác của người đàn ông thì mãi mãi không thể tìm lại được. (*The army then resumed its march and the man's body was never found.*)

Figure 2: The translation version from Laos to Vietnamese using Deep-translator

solve the two problems mentioned above, we carry out paragraph alignment before sentence alignment. Only sentences within each bilingual paragraph pair are used to generate candidate sentence pairs. This method will reduce computational cost and limit alignment errors that propagate from this bilingual paragraph pair to another.

The architecture of our sentence alignments system is shown in Figure 1.

To perform the alignment task for low-resource language pair, a machine translation system is used to translate the input texts to an intermediate language supported by an embedding model. The purpose is to project two input texts into the same embedding space for future similarity comparison purposes. Another condition for choosing the intermediate language is that the MT system works well for those translation tasks. LASER, a sentenceàembedding model that has been pretrained on 93 languages, is used in our system for sentence representation.

If one of the source and target languages is not included in these 93 languages, the remaining language should be chosen as the intermediate language. If both languages are not in these 93 languages, the chosen intermediate language is English because it is known to be the most rich-resource language.

To carry out bilingual sentence alignment for Vietnamese - Lao language pairs, since LASER has been pretrained for Vietnamese language, we translate Laos documents to Vietnamese and then carry out sentence alignment between two Vietnamese documents. After detecting all sentence alignment pairs in the intermediate language, we recover them to their original language by mapping original sentences and sentences in the text of intermediate language.

As analyzed in the Introduction section, we evaluate the similarity between two text spans by the cosine similarity between their embedding vectors. Sentence embedding similarity has been shown effective at filtering out non-parallel sentences (Hassan et al., 2018) (Chaudhary et al., 2019). However, in some situations, although two sentences have a high embedding similarity, they are not the paraphrasing of each other. Such situations often happen when two text spans have different lengths, sometimes a text span is just a part of the other. To remedy this issue, we propose to use textual embedding similarity with the ratio of text length to find pairs of parallel text.

Each module of our sentence alignments system will be analyzed in detailed below:

3.1 The multilingual machine translation system

The machine translation (MT) system is an important factor to decide the accuracy of the sentence alignment system. The MT system should produce output sentences which are understandable and contain main ideas of the input sentences. There are several good machine translators from big IT companies such as Google, Microsoft, IBM, Amazon, ... Google translator is capable of translating more than 100 languages, including LRLs such as Lao, Urdo, Zulu, etc. Some machine translation APIs are shared for the research community. After investigating several machine translators, we have chosen Deep-translator[1] to translate from Laos to Vietnamese. The Deep-translator is a free and unlim-

[1]https://github.com/nidhaloff/deep-translator

itedàtool to translate between different languages in a simple way using multiple translators, including Google and Microsoft ones. The translation quality from Laos to Vietnamese is quite good. Figure 2 shows a translation version from Laos to Vietnamese using Deep-translator.

3.2 Paragraph alignment

Bilingual paragraph alignment is the task of finding paragraph pairs that are translations of each other from the input source text (S) and the target one (T). This task becomes monolingual paragraph alignment after source and target documents have been translated to the intermediate language.

The fact shows that, some document pairs have crossing paragraph alignments such as the example shown in Figure 3. Existing alignment methods often do not consider these cases, resulting in propagating alignment errors from one paragraph to the others. Our proposed method remedies this issue and reduce sentence alignment candidates when removing paragraph alignment 1-0, 0-1,2-0,0-2.

Our paragraph alignment method uses the Cosine similarity of two paragraphs and length ratio of them by character. We use a combination of both conditions because there exist paragraph pairs with high similarity but they are not translations of each other.

Our proposed paragraph alignment method as follows:

The input documents after being translated to the intermediate language, S and T, are segmented into paragraphs based on new line symbols. These paragraphs are represented as paragraph embedding vectors by the LASER library.

Then we find out candidates for paragraph alignments by using a dredging algorithm. Two paragraphs are aligned if they satisfy the following conditions:

1. The Cosine similarity of two paragraphs in the pair is greater than a threshold θ;
2. The length ratio of these paragraphs is within a limit (α, β).

The thresholds value θ depends on each language pair and is determined manually by experimenting on the sample data set for that language pair. This value is chosen as 0.8 for Vietnamese-Lao language.

Input: Document pair are translations of each other S', T'.
Output: Paragraph pair are translations of each other ps', pt'.

```
Begin
ps'[1,...,n] = segment(S');
pt'[1,...,m] = segment(T');
vps'[1,...,n] = LASER(ps'[1,...,n]);
vpt'[1,...,m] = LASER(pt'[1,...,m]);
for i=1 to n do
    for j=1 to m do
        if (mk([i],[j])and(si[i][j] > θ) and
        (α < ra[i][j] < β) then
            export(ps'[i],pt'[j]);
            mk(ps'[i],pt'[j])=false; break;
        else
            if
             mk([i],[j+1])and(si[i][j+1] >
             0.8)
            and(α < ra[i][j+1] < β) then
                export(ps'[i],pt'[j+1]);
                mk(ps'[i],pt'[j+1])=false;break;
            else
                if mk([i+1],[j])and(si[i+
                 1][j] > 0.8)
                and(α < ra[i+1][j] < β)
                 then
                    export(ps'[i+1],pt'[j]);
                    mk(pt'[i+1],pt'[j])=false;break;
                else
                    if mk([i+1],[j+1])and
                    (si[i+1][j+1] > 0.8)and
                    (α < ra[i+1][j+1] < β)
                     then
                        export(ps'[i+1],pt'[j]);
                        mk(pt'[i+1],pt'[j+1])=false;
                        break;
                    end
                end
            end
        end
    end
end
End
```

Algorithm 1: Paragraph alignment

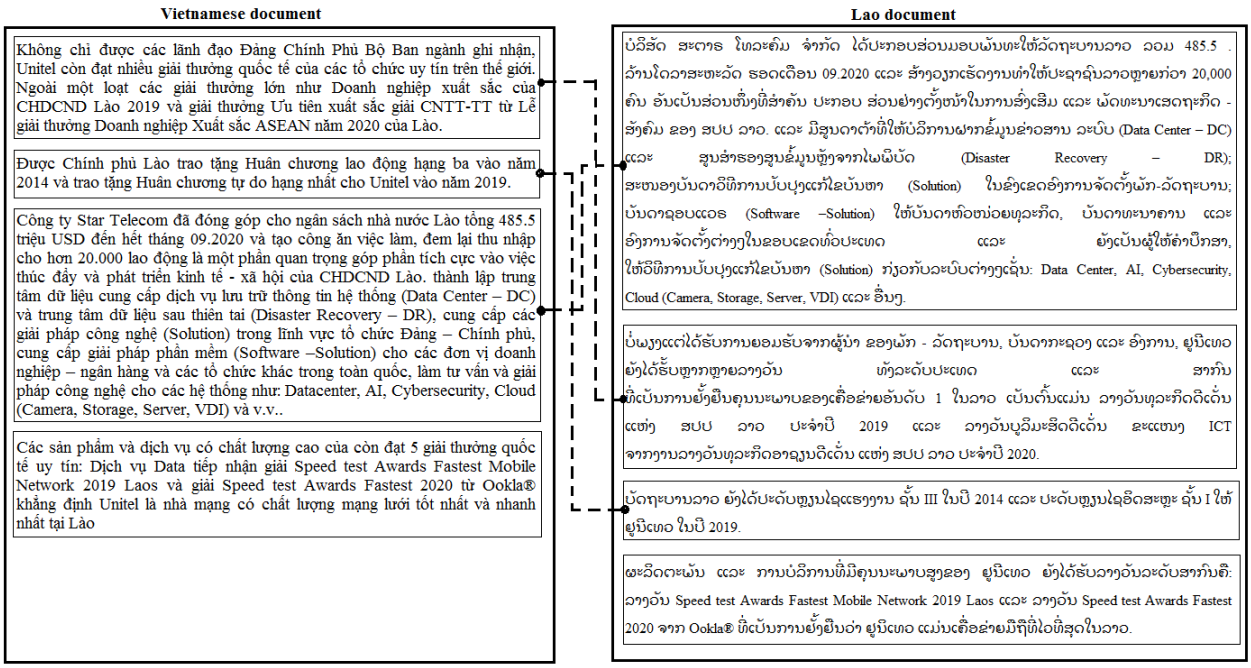

Figure 3: Across paragraph alignment in the Vietnamese - Lao document pair

α, β are the smallest and maximum character-based length ratio between source sentence and target sentence, respectively. They are estimated based on statistics on the given sentence-aligned bilingual corpus.

The Cosine similarity of two paragraphs is computed as follows:

$$si[i+x][j+y] = cosine(\sum_{p=0}^{x} vps'[i+p], \sum_{q=0}^{y} vpt'[j+q]);$$

Where:

- si[i+x][j+y] is the Cosine similarity of the paragraphs from the positions i^{th} to $(i+x)^{th}$ in the source text and the paragraphs from the positions j^{th} to $(j+y)^{th}$ in the target one;
- $\sum_{p=0}^{x} vps'[i+p]$ is the sum of the paragraph embedding vectors from the positions i^{th} to $(i+x)^{th}$ in source text;
- $\sum_{q=0}^{y} vpt'[j+q])$ is the sum of the paragraph embedding vectors from the positions j^{th} to $(j+y)^{th}$ in target text;
- Function cosine() is calculated as follows: cosine(A,B) = Dot(A,B)/Norm(A)*Norm(B).

The length ratio of two paragraphs is computed as follows:

$$ra[i+x][j+y] = \frac{len(\sum_{p=0}^{x} ps'[i+p])}{len(\sum_{q=0}^{y} pt'[j+q])}$$

Where:

- ps'[i] is the character-based length of the ith paragraph in source text.
- pt'[j] is the character-based length of the jth paragraph in target text.

Algorithm of ours proposed paragraph alignment method is showed in Algorithm 1. The formulas and symbols used in this algorithm are as described above. Besides:

Function mk() is used to mark paragraphs that have been aligned before:

$mk(ps'[i+x], p[j+y]) = true$ if $(ps'[i] = true, ..., ps'[i+x] = true)$ and $(pt'[j] = true, ..., pt'[j+y] = true)$ and vice versa.

Function $len(s)$ is used to calculate the length of the string s by characters.

3.3 Sentence alignment

Our sentence alignment is developed from Vecalign - a fast sentence alignment tool that works with 100 languages, in conjunction with LASER. Vecalign uses cosine similarity to measure the similarity between texts. The original Vecalign algorithm gets errors in aligning some sentence pairs that have high Cosine similarity, but are not translations of each other. The pair of sentences in Figure 4 is an example.

We propose to remedy this issue by using ratio of sentence length between them. If the sentence pair (s[u],t[u]) aligned by Vecalign that has the sentence length ratio in $(\alpha; \beta)$, we accepted this sentence alignment. Otherwise, this alignment is rejected.

Our sentence alignment as follows:

Given two paragraphs (ps'; pt') that are aligned together, the system segments them into sentences (s'[1],..., s'[n]]; t'[1], ..., t'[m]). Then these sets of sentences are used as input of Vecalign to extract sentence alignment pairs. The length ratio of each output sentence pair is used to filter out alignments that are not correspondence in the length criteria.

4 Experiment Results

As far as we know, there is no sentence-aligned bilingual corpus for Vietnamese - Laos. Therefore, in this research, we concentrate on building a sentence alignment tool for this language pair, using our proposed method. To evaluate our system, we built two test sets:

- Test set 1: This set is constructed as parallel documents, including 150 Vietnamese - Lao sentence pairs in which 100 pairs that are translations of each other. These sentences are put randomly in each document. For example, the first sentence in the Vietnamese document can align with the 64^{th} sentence in the Laos document. The Vietnamese sentences length in this set is in the range from 17 characters to 352 ones and from 21 characters to 332 ones for Lao sentence' length. We designed Test set 1 for comparing the quality of sentence alignment tools as they face the problem in finding out the pairs of sentence alignment that the source and target sentences are located far apart in the source and target documents.

- Test set 2: This set is constructed as 7 pairs of comparable documents in Vietnamese and Laos language. Each document has 18 sentences in average. Each sentence has 198 characters in average. We designed Test set 2 for comparing the quality of sentence alignment tools on common bilingual document pairs.

The above test sets are used in our experiments with our proposed method and other existing sentence alignment tools including Champolion[2], Hualign[3], and the original version of Vecalign[4] to compare their accuracy. Since these tools do not support LRLs such as Laos, we translate from Laos to Vietnamese after the sentence segmentation step to do the alignment task.

Underthesea[5] is used to segment Vietnamese texts into sentences. Since there is no sentence segmentation tool for Laos available, we implement it by using end-of-sentence symbols. It includes '.', '?', '!'. Especially for the mark '.', we consider cases where it is not used to end sentences, such as when it is used to separate digits or used in acronyms, web or email addresses, etc. Our sentence segmentation tool for Laos gets the accuracy of 95%.

After splitting text into sentences, Deep-translator is used to translate from Laos to Vietnamese. Given two documents in Vietnamese, we carried out steps of paragraph alignment, sentence alignment, then recovering text in resulting sentence pairs into their original languages. The source codes for our Laos sentence segmentation and Vietnamese-Laos sentence alignment are published on github[6].

For Vecalign, since it can't work directly for Lao, we also use Deep-translator to call Google translate api to translate from Laos to Vietnamese.

[2]https://github.com/LowResourceLanguages/champollion
[3]https://github.com/danielvarga/hunalign
[4]https://github.com/thompsonb/vecalign
[5]https://github.com/undertheseanlp/underthesea
[6]https://github.com/NHDat2/UET.PJKC4.0

Vecalign's similarity	Vietnamese sentence	Laos sentence
0.81	Merchant muốn tăng trưởng doanh thu bền vững thì cần thiết phải thực sự hiểu khách hàng của mình (*Merchant want sustainable revenue growth, they need to really understand their customers*)	ຮ້ານຄ້າຕ້ອງການເພີ່ມລາຍໄດ້ຢ່າງຍືນຍົງນັ້ນ ຈຳເປັນຕ້ອງເຂົ້າໃຈລູກຄ້າຂອງຕົນເອງຢ່າງເລິກເຊິ່ງ, ຕ້ອງຮູ້ລັກສະນະພຶດສະດຳລູກຄ້າ ແລະ ຕ້ອງຮູ້ເຖິງພຶດຕິກຳການບໍລິໂພກຂອງພວກເຂົາ ຈາກຂໍ້ມູນເຫຼົ່ານັ້ນ ແມ່ນສາມາດນຳມາໃຊ້ໃນການປັບປຸງຜະລິດຕະພັນສິນຄ້າ, ລາຄາ, ວິທີການບໍລິການ ຫຼື ການສ້າງໂປຣໂມຊັ່ນກະຕຸ້ນຍອດຂາຍໃຫ້ເໝາະສົມກັບລູກຄ້າໃຫ້ຫຼາຍທີ່ສຸດ. (*Stores want sustainable revenue growth, they need to really understand their customers, it is about the private characteristics of customers and their consumption behavior.*)

Figure 4: Pair of sentences are aligned by vecalign

We use $precision$ and $recall$ measure to evaluate the quality of alignment tools. Results on two test sets are shown in Tables 1 and 2.

Table 1: Alignment results on testset 1

	$Precision$	$Recall$	F_1
Champolion	2.00%	2.00%	2.00%
Hunalign	2.00%	3.00%	2.40%
Vecalign	2.70%	4.00%	3.22%
Our's method	**95.74%**	**90.00%**	**92.78%**

Table 2: Alignment results on testset 2

	$Precision$	$Recall$	F_1
Champolion	45.45%	38.13%	41.47%
Hunalign	72.30%	79.66%	75.80%
Vecalign	87.80%	91.52%	89.62%
Our's method	**99.15%**	**97.48%**	**98.31%**

Experimental results on Table 1 and Table 2 show that our proposed system provides best results among existing sentence alignment tools, for both cases of parallel documents and comparable documents.

Here we review the alignment results of our proposed method with vecalign, the method has state-of-the art accuracy in high and low resource settings and improves downstream MT quality and it got better than Champolion, Hualign in both Test sets.

The Champolion and Hunalign methods get bad results because these alignment methods require some additional language information which depends on each language pair, while the Vietnamese-Lao low-resource language pair does not have enough additional linguistic information for them.

- In Testset 1, the performance of Vecalign and other alignment tools are deeply reduced because some pairs of sentences that are translations of each other appear in the input document pair located far apart. Our method works well in those cases since we used a paragraph alignment algorithm which is able to find and align paragraphs correctly even if pairs of paragraphs that are translations of each other appear in the input document pair located far apart (In this case Each sentence is considered as a paragraph by our method). As a result, sentence alignment achieves higher accuracy. An example of this case is shown in figure 5.

- In Testset 2, our method gets better result than Vecalign in $precision$ because we have eliminated incorrected alignment cases of Vecalign relating to sentence lengths. An example of this case is shown in Figure 4. For $recall$, our method also gets better result than Vealign because there are some document pairs that have across paragraph alignments as showed in Figure 3. It is the cause of the incorrected sentence alignments of Vecalign

Our experiments prove that our sentence alignment tool is efficient and reliable enough to be used in automatically generating large sentence-alignment corpora in low-resource language pairs for the machine translation task.

5 Conclusions and Future Work

In this paper, we have proposed an approach to align sentences from bilingual documents of LRLs. We created two datasets for testing sentence alignment tools. The proposed approach has been applied to build the tool for Vietnamese-Laos language sentence alignment. As far as we know, this is the first

Source Vietnamese sentence (s)	Line index of s in source document	Target Laos sentence (t)	Line index of t in target document	Vecalign	Our method
Các phi công có thể đã cố hạ cánh khẩn cấp trên mặt nước.(*Maybe the pilots have tried to make an emergency landing in the water.*)	38	ນັກບິນອາດພະຍາຍາມທີ່ຈະລົງຈອດສຸກເສີນທີ່ໜ້ານ້ຳ. (*Maybe the pilots have tried to make an emergency landing in the water.*)	89	Failed	True
Một quả bom đã phát nổ vào đêm thứ Ba tại Colombo, thủ đô của Sri Lanka. (*A bomb exploded on Tuesday night in Colombo, the capital of Sri Lanka.*)	53	ເກີດລະເບີດຂຶ້ນໃນຄືນວັນອັງຄານທີ່ໂຄລໍາໂບ ເມືອງຫລວງຂອງສີລັງກາ. (*A bomb exploded on Tuesday night in Colombo, the capital of Sri Lanka.*)	21	Failed	True
Tôi đã ký một sắc lệnh về tình trạng chiến tranh. (*I signed a state of war decree.*)	51	ຂ້ອຍໄດ້ລົງນາມໃນວັດຖະບັນຍັດລົງຄາມພາຍໃນປະເທດ (*I signed a state of war decree*)	36	Failed	True

Figure 5: Vecalign got mistakes because the line indexes is too different

work using sentence embeddings for doing bilingual sentence alignment for low-resource language pairs. The experimental results with two datasets have shown that our proposed approach achieve much higher results than existing researches. In the future, we will study a method of aligning sentences for low-resource language pairs from a set of text pairs without knowing in advance whether they are actually translations of each other and using our proposed method to build sentence-aligned bilingual corpus Vietnamese-Lao, Vietnamese-Khmer.

Acknowledgments

This work has been supported by Ministry of Science and Technology of Vietnam under Program KC 4.0, No. KC-4.0.12/19-25.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve smt performance. In *EACL 2009 - 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, pages 16–23, 01.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy, July. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy, July. Association for Computational Linguistics.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy, August. Association for Computational Linguistics.

Katsuki Chousa, Masaaki Nagata, and Masaaki Nishino. 2020. SpanAlign: Sentence alignment method based on cross-language span prediction and ILP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4750–4761, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Francis Grégoire and Philippe Langlais. 2018. Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1442–1453, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Nguyen Tien Ha, Nguyen Thi Minh Huyen, and Nguyen Minh Hai. 2018. Building a sentence-aligned vietnameseenglish bilingual corpus in tourism domain for machine translation. *JOURNAL OF RESEARCH AND DEVELOPMENT ON INFORMATION AND COMMUNICATION TECHNOLOGY*, V-1, number 39.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Nguyn Thi Minh Huyn and Mathias Rossignol. 2006.

A language-independent method for the alignement of parallel corpora. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pages 223–230, Huazhong Normal University, Wuhan, China, November. Tsinghua University Press.

Max Kaufmann. 2012. JMaxAlign: A maximum entropy parallel sentence alignment tool. In *Proceedings of COLING 2012: Demonstration Papers*, pages 277–288, Mumbai, India, December. The COLING 2012 Organizing Committee.

Shengxuan Luo, Huaiyuan Ying, and Sheng Yu. 2021. Sentence alignment with parallel documents helps biomedical machine translation.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *CoRR*, abs/2006.07264.

Rico Sennrich and Martin Volk. 2011. Iterative, mt-based sentence alignment of parallel texts. In *NODALIDA 2011, Nordic Conference of Computational Linguistics*. Northern European Association for Language Technology (NEALT), May. The 18th Nordic Conference of Computational Linguistics ; Conference date: 11-05-2011 Through 13-05-2011.

Brian Thompson and Philipp Koehn. 2019a. Vecalign: Improved sentence alignment in linear time and space. 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)*, pages 1342–1348, Hong Kong, China, November. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2019b. Vecalign: Improved sentence alignment in linear time and space. 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)*, pages 1342–1348, Hong Kong, China, November. Association for Computational Linguistics.

Dòniel Varga, Põter Halòcsy, Andròs Kornai, Viktor Nagy, Lòszlú Nõmeth, and Viktor Trún. 2007. Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing IV*, pages 247–258. John Benjamins.