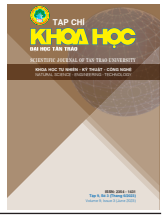
TẠP CHÍ KHOA HỌC ĐẠI HỌC TÂN TRÀO

ISSN: 2354 - 1431

http://tckh.daihoctantrao.edu.vn/



BUILDING A GOOD QUALITY BILINGUAL CORPUS FOR A LOW-RESOURCE LANGUAGE PAIR

Tien-Ha Nguyen[1,*], *Hung-Cuong Nguyen*[1], *Van-Vinh Nguyen*[2]

[1] *Hung Vuong University, Vietnam*

[2] *VNU University of Engineering and Technology*

**Email address: Tienhapt@gmail.com*

DOI: 10.51453/2354-1431/2023/962

Article info

Received: 06/12/2022
Revised: 12/3/2023
Accepted: 16/5/2023

Keywords:
Data mining, Big data, Bilingual corpus, Sentence alignment.

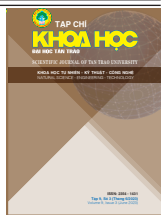
Abstract:

In natural language processing (NLP), a good quality bilingual corpus is very important in some applications, such as machine translation, building bilingual dictionaries, cross-language retrieval, etc. For low-resource language pairs, for example, the Vietnamese-Lao pair, it is very difficult to build a good quality bilingual corpus because bilingual resources are rare. In this paper, we presented the process of building a good quality bilingual corpus for a low-resource language pair and proposed a novel method of sentence alignment that takes advantage of pre-trained modern models for rich-resource languages. In our experiments on aligning sentences and building a bilingual corpus for the Vietnamese-Laos language pair, we achieved higher precision and recall than other good sentence alignment methods and a good quality sentence-aligned Vietnamese-Laos bilingual corpus.

TẠP CHÍ KHOA HỌC ĐẠI HỌC TÂN TRÀO

ISSN: 2354 - 1431

http://tckh.daihoctantrao.edu.vn/



BUILDING A GOOD QUALITY BILINGUAL CORPUS FOR A LOW-RESOURCE LANGUAGE PAIR

Tien-Ha Nguyen[1,*], *Hung-Cuong Nguyen*[1], *Van-Vinh Nguyen*[2]

[1] *Hung Vuong University, Vietnam*

[2] *VNU University of Engineering and Technology *Email address: Tienhapt@gmail.com*

DOI: 10.51453/2354-1431/2023/962

Article info

Received: 06/12/2022
Revised: 12/3/2023
Accepted: 16/5/2023

Keywords:

Data mining, Big data, Bilingual corpus, Sentence alignment.

Abstract:

In natural language processing (NLP), a good quality bilingual corpus is very important in some applications, such as machine translation, building bilingual dictionaries, cross-language retrieval, etc. For low-resource language pairs, for example, the Vietnamese-Lao pair, it is very difficult to build a good quality bilingual corpus because bilingual resources are rare. In this paper, we presented the process of building a good quality bilingual corpus for a low-resource language pair and proposed a novel method of sentence alignment that takes advantage of pre-trained modern models for rich-resource languages. In our experiments on aligning sentences and building a bilingual corpus for the Vietnamese-Laos language pair, we achieved higher precision and recall than other good sentence alignment methods and a good quality sentence-aligned Vietnamese-Laos bilingual corpus.

In order to create such corpora, we need good resources of parallel text, as well as a mechanism to align sentences in bilingual text pairs. Besides, to have a good quality bilingual corpus, it requires a construction process that has to include strict human control.

For low-resource languages, it is very difficult to collect parallel texts because they are really rare.The sentence alignment tools apply for them to get bad results because they have not untitled advances of deep learning models. Annotates and experts in these language pairs are not easy to find for controlling the parallel corpora constructions.

To overcome the above limitations. In this work, we propose the process of building a good quality bilingual corpus for a low-resource language pair and a novel method of sentence alignment for a low-resource language pair that takes advantage of pre-trained modern models for rich-resource languages.

The rest of the paper is laid out as follows: Section 2 presents the related works; Section 3 presents our proposed process of building a good quality bilingual corpus for a low-resource language pair; Section 4 presents experiment results, and we conclude and present future work with Section 5.

2 RELATED WORKS

The bilingual corpus is an electronic collection of texts in two languages put together in a principled way for the purpose of comparative linguistic studies and prepared in electronic form for search and analysis by computer [6]. Because it's important in NLP, there has been much research on building and extending these ones.

J.Tiedemann, 2016 [1] built an OPUS that is a freely available sentence-aligned parallel corpora. OPUS covers over 200 languages and language variants with a total of about 3.2 billion sentences. It is collected from various sources and domains. Each sub-corpus in it is provided in common data formats to make it easy to integrate them in research and development.

The Oslo Multilingual Corpus (OMC) is a product of the interdisciplinary research project Languages in Contrast (SPRIK), which is a collaboration between researchers at the Faculty of Humanities, University of Oslo.[1]. It is an extension of the ENPC that is a bidirectional translation corpus consisting of original English texts and their translations into Norwegian, and Norwegian original texts and their translations into English was built in the 1990s. The OMC contains many sub-corpora that differ in composition with regard to languages and number of texts included. It is mainly the languages Norwegian, English, French, and German that are represented in the sub-corpora, but some of the corpora include Dutch and Portuguese texts. In addition, there are related parallel corpora for English-Swedish and English-Finnish, compiled in Sweden and Finland, which are accessible from the same site.

E.Salesky et al., 2021 [2] released the Multilingual TEDx corpus to facilitate speech recognition and speech translation research. This corpus is a collection of audio recordings from TEDx talks in 8 source languages. They segment transcripts into sentences and align them to the source language audio and target-language translations. It is built to support speech recognition and speech translation research across many non-English source languages.

Most of the corpora built and shared are not good quality and are in rich-resource language pairs, but they are rare in low-resource language pairs.

In the process of building a bilingual corpus, a high-quality automatic sentence alignment tool reduces costs. Sentence alignment is the task that automatically extracts parallel sentences from noisy parallel documents to build a sentence-aligned bilingual corpus. 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.

X.Ma, 2006 [7] 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.

D.Varga et al., [8] proposed a method that is

[1]https://www.hf.uio.no/ilos/english/services/knowledge-resources/omc/

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.

N.T.Ha et al., 2018 [9] proposed the improvement of an language - independent sentence alignment method [10] for Vietnamese–English 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:

B.Thompson and P.Koehn [11] 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.

K.Chousa et al., 2020 [12] 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.

S.Luo et al,, 2021 [13] 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 language's features.

3 OUR PROPOSED METHOD

Our proposed method for building a good-quality corpus for a low-resource language pair is illustrated in Figure 1

3.1 Collecting bilingual texts

- Step 1: Search and crawl bilingual websites. For the Vietnamese-Laos language pair, we found it: vovworld.vn; www.qdnd.vn; wikipedia.or; vietlao.vietnam.vn; vnanet.vn "; "tapchilaoviet.org",
- Step 2: Only the text from HTML should be extracted. Depending on the posted date and categories of these html pages, in order to organize and store the collected text pairs on both the source and destination sides. Figure 2 illustrates the tree that stores text crawled from Vietnamese - Laos websites. Similar trees were created for the other lo-resource

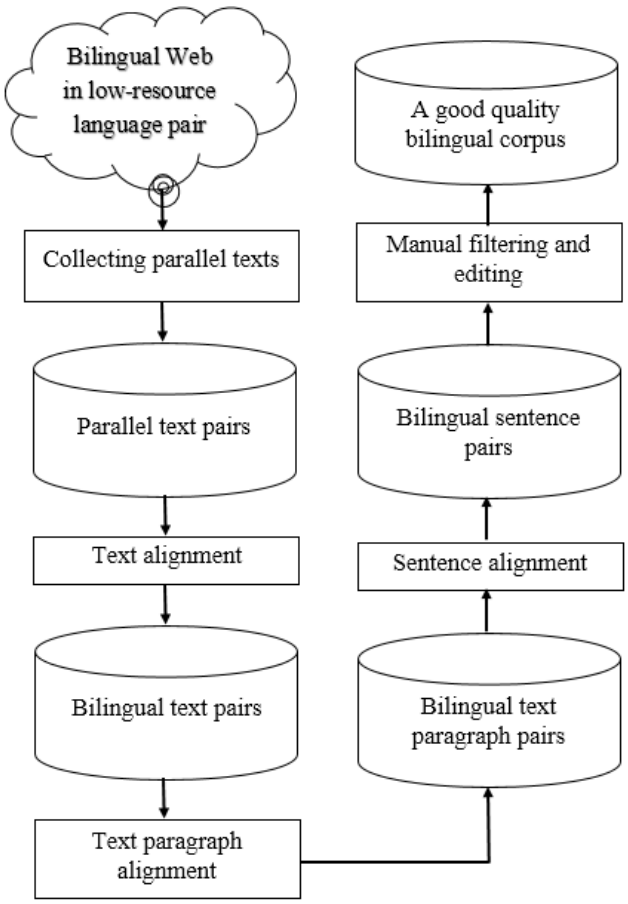


Figure 1: The process of building a good-quality bilingual corpus for a low-resource language pair

language pairs.

Many news texts are not translated into another language on the same day, but they are mostly posted in the same month instead. Therefore, when we crawl them from bilingual websites, they are stored by posted month. We only compare texts within the same month for each language pair, in order to reduce the complexity but not lose potential candidate document pairs.

3.2 Text alignment

Given a folder pair containing texts, so that texts from each language are in the same folder, text alignment is the process of matching two texts that are translations of each other. Our proposed text alignment method as follows:

1. Step 1: Text is segmented into sentences using segmentation tools. I use *Underthesea*[2] for Vietnamese and our tool, *Lao Sentence Tokenize*[3] for Laos.

2. Step 2: Using LASER3[4] to embed sentences into the sentence embedding vector. For languages that are not supported by LASER, I will use machine translation to translate them into languages that LASER supports.

3. Step 3: Calculate the value of the text embedding vector as follows: $v_D = \sum_{i=1}^{n} v_S$

   Where:

   v_D is a text embedding vector.

   v_S is a Sentence embedding vector.

   n is The text's sentence total.

4. Step 4: Using the exhaustive algorithm to compare the similarity between two text embedding vectors in the respective folders of the two languages, if it is greater than the ϵ

[2]https://underthesea.readthedocs.io/en/latest/readme.html
[3]https://github.com/NHDat2/UET.PJKC4.0
[4]https://github.com/facebookresearch/LASER

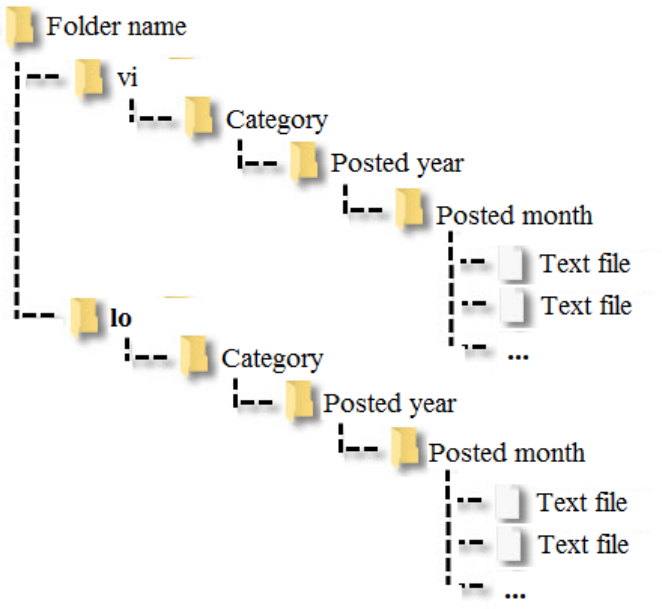


Figure 2: The text stored tree for Vietnamese - Laos pair

threshold, they are selected as bilingual text pairs.

3.3 Sentence alignment

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 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 alignment's system is shown in Figure 3.

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 sentence embedding 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 ef-

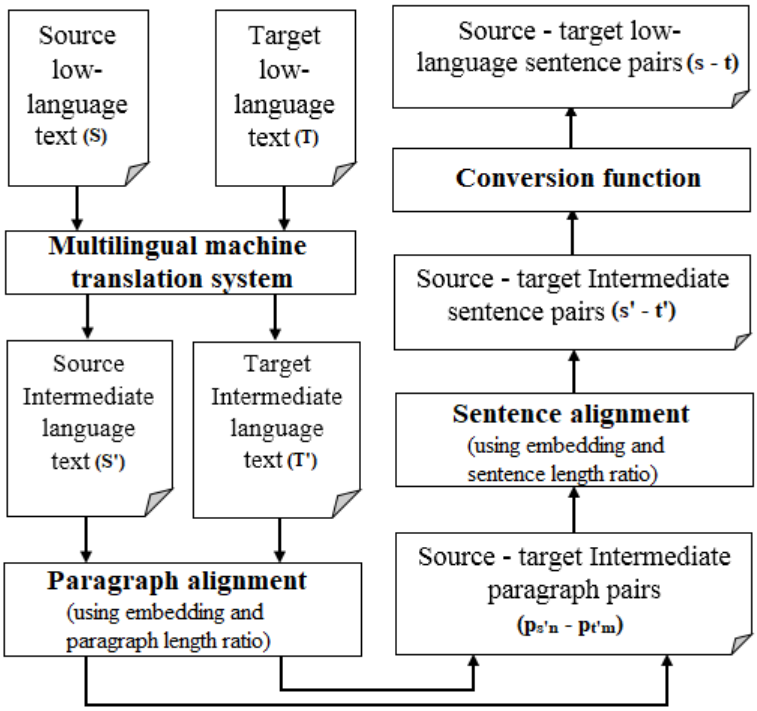


Figure 3: Sentence Alignment for Low-resource Language Pairs

fective at filtering out non-parallel sentences [14] [15]. 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 alignment's system will be analyzed in detailed below:

3.3.1 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[5] to translate from Laos to Vietnamese. The Deep-translator is a free and unlimited tool 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 4 shows a translation version from Laos to Vietnamese using Deep-translator.

3.3.2 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 5. 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

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

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

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 threshold's 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.

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.

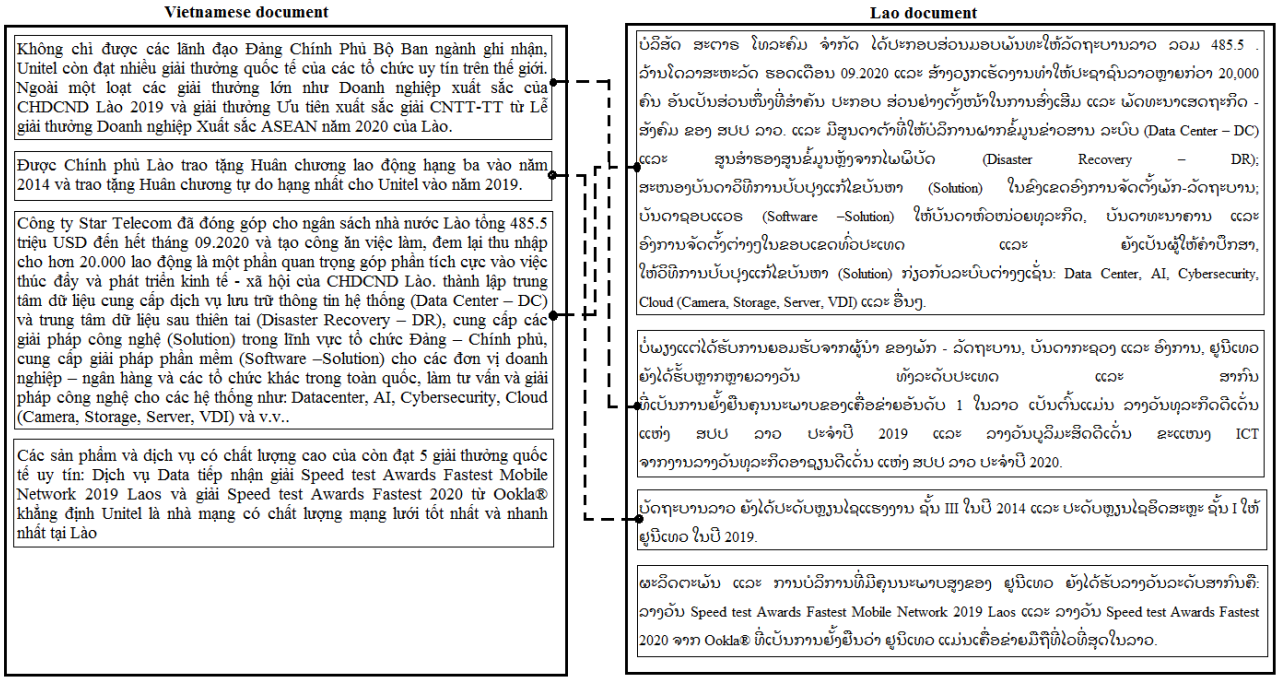


Figure 5: Across paragraph alignment in the Vietnamese - Lao document pair

Algorithm 1: Paragraph alignment

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 $(mk([i],[j]) and (si[i][j] > \theta)\ and\ (\alpha < ra[i][j] < \beta)$ **then**

      export(ps'[i],pt'[j]); mk(ps'[i],pt'[j])=false; break;

    else

      if $mk([i],[j+1]) and (si[i][j+1] > 0.8)\ and\ (\alpha < ra[i][j+1] < \beta)$ **then**

        export(ps'[i],pt'[j+1]); mk(ps'[i],pt'[j+1])=false;break;

      else

        if $mk([i+1],[j]) and (si[i+1][j] > 0.8)\ and\ (\alpha < ra[i+1][j] < \beta)$ **then**

          export(ps'[i+1],pt'[j]); mk(pt'[i+1],pt'[j])=false;break;

        else

          if $mk([i+1],[j+1]) and (si[i+1][j+1] > 0.8) and (\alpha < ra[i+1][j+1] < \beta)$ **then**

            export(ps'[i+1],pt'[j]);mk(pt'[i+1],pt'[j+1])=false; break;

End

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, ຕ້ອງຮູ້ລັກສະນະພຶດສະດຂອງລູກຄ້າ ແລະ ຕ້ອງຮູ້ເຖິງພຶດຕິກຳການບໍລິໂພກຂອງພວກເຂົາ ຈາກຂໍ້ມູນເຫຼົ່ານັ້ນ ແມ່ນສາມາດນຳມາໃຊ້ໃນການປັບປຸງພະລິດຕະພັນສິນຄ້າ, 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 6: Pair of sentences are aligned by Vecalign

3.3.3 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 6 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.

3.4 Manual filtering and editing

We built the online tool[6] for manual data reviewing for automatic aligned bilingual sentences. Figure 7 showed the reviewing interface for the Vietnamese-Laos pair.

Each bilingual sentence pair is designed with two reviewing levels, including:

- Good level: The bilingual sentence pair is selected at this level by annotators if they are exactly translation of each other. For bilingual sentence pairs that are easy to modify to get good pairs, annotators will select good level for it after modifying them.
- Bad level: The bilingual sentence pair is selected at this level by annotators if they are not translation of each other or difficult to modify to get good pairs.

If a bilingual sentence pair is evaluated as good, it will be added to the corpus. It will be removed otherwise. Our goal is to build a good-quality bilingual corpus, so we have chosen good annotators for manual data reviewing and use the best expert to randomly review 10% of the manually reviewed data every month for fine quality control of the data added to the corpus.

We recruited 15 annotators, who manually reviewed bilingual sentence pairs from the automatically collected ones. These annotators speak and write fluently in Vietnamese-Laos language pair. They were final-year undergraduate students, graduate students, or teachers at universities.

Firstly, all annotators were trained to use an online tool for manually reviewing and knowing how to choose good or bad for each sentence pair. Each annotator was provided with an account to review daily. On average, each day, one annotator spent about two hours reviewing the data, and it took one year to complete this corpus.

4 EXPERIMENT RESULTS

4.1 Sentence alignment for the Vietnamese-Lao language pair

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

[6]http://nmtuet.ddns.net:3000

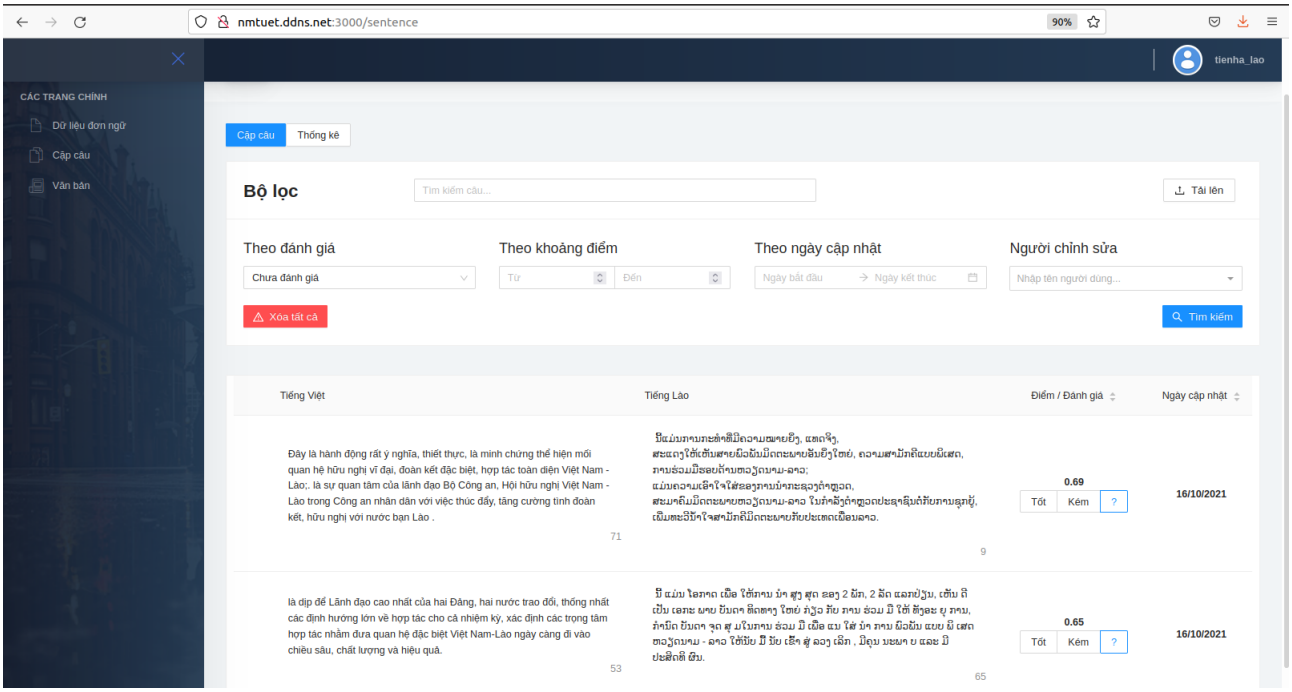


Figure 7: The online tool for manual Vietnamese-Laos data reviewing

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[7], Hualign[8], and the original version of Vecalign[9] 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[10] 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[11].

[7]https://github.com/LowResourceLanguages/champollion
[8]https://github.com/danielvarga/hunalign
[9]https://github.com/thompsonb/vecalign
[10]https://github.com/undertheseanlp/undertheseа
[11]https://github.com/NHDat2/UET.PJKC4.0

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.

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.

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 8.

- 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 6. 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 5. 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.

4.2 Evaluation of the quality of our Vietnamese-Laos bilingual corpus

Using our proposed method, We built a sentence-aligned bilingual corpus for the Vietnamese-Lao pair, which includes 150K sentence pairs for train set, 1000 sentence pairs for valid set, and 2018 sentence pairs for test set. We trained a NMT model.

Test set: Our test set consists of 2,018 bilingual sentence pairs for each language pair, built from 1,018 sentence pairs in the ALT test set combined with 1,000 pairs done manually by independent language experts. Manual data ensure the following criteria on the Vietnamese side for ensures coverage of test data. (1) Least 15 syllables per sentence; (2) Contains proper names such as person names, place names, organization names, and new terms; and (3) Includes political, economic, cultural, social, and sports fields.

Preprocessing: All bilingual texts were tokenized and truncated using sentence piece scripts, and then they are applied to Sennrich's BPE [16]. We explore 32000 operators learned to generate BPE codes. For Vietnamese, we only use Moses's scripts for tokenization and true-casing.

We trained our Transformer model[12] using the number of encoder 12, decoder layers are 6, 8 head is used, d_{model} is 512, dropout value is 0.1, batch size of 64, learning rate value is 0.4 with the aid of Adam optimizer. The learning rate has warmup updates by 8000 steps and the label smoothing value is 0.1. We utilized the best model to decode the test data for comparison purposes of our experiments.

Results The experiment results are shown in the Table 3.

Table 3 shows that our corpus is of good quality because when we add more data from our corpus, the quality of the machine translation system improves visibly. Despite having only 150k Vietnamese-Laos pairs for training, we got a 25.74 BLUE score.

[12]https://arxiv.org/pdf/2112.15272.pdf

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%**

5 CONCLUSIONS AND FUTURE WORK

In this paper, we propose the process of building a good-quality bilingual corpus for a low-resource language pair. Applying our proposed method, we have built a good-quality Vietnamese-Lao bilingual corpus that includes 150,000 sentence pairs. built automatic text and sentence alignment tools for low-resource language pairs and an automatic sentence segment tool for Laos. In the future, we will expand the size of the Vietnamese-Lao bilingual corpus and build more bilingual corpora in some other low-resource language pairs. In addition, we also study the use of our corpus to improve some natural language applications.

REFERENCES

[1] J. Tiedemann, "OPUS – parallel corpora for everyone," in Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products, Baltic Journal of Modern Computing, Riga, Latvia, 2016.

[2] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, M. Post, "The Multilingual TEDx Corpus for Speech Recognition and Translation," CoRR, **abs/2102.01757**, 2021.

[3] S. Siripragada, J. Philip, V. P. Namboodiri, C. V. Jawahar, "A Multilingual Parallel Corpora Collection Effort for Indian Languages," CoRR, **abs/2007.07691**, 2020.

[4] L. Doan, L. T. Nguyen, N. L. Tran, T. Hoang, D. Q. Nguyen, "PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation," 2021.

[5] A. Magueresse, V. Carles, E. Heetderks, "Low-resource Languages: A Review of Past Work and Future Challenges," CoRR, **abs/2006.07264**, 2020.

[6] N. Dash, A. Selvaraj, Limitations of Language Corpora, 259–272, 2018, doi:10.1007/978-981-10-7458-5_15.

[7] X. Ma, "Champollion: A Robust Parallel Text Sentence Aligner," in Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC06), European Language Resources Association (ELRA), Genoa, Italy, 2006.

Table 3: BLEU score of the system when trained with our corpus

The size of the training set	**20k**	**50k**	**80k**	**100k**	**150k**
BLEU	8.92	15.48	19.30	23.84	25.74

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	ນັກບິນອາດພະຍາຍາມທີ່ຈະລົງຈອດສຸກເສີນທົ່ງພື້ນ້ຳ. (*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	ຂ້ອຍໄດ້ລົງນາມໃນລັດຖະບັນຍັດສົງຄາມພາຍໃນປະເທດ (*I signed a state of war decree*)	36	Failed	True

Figure 8: Vecalign got mistakes because the line indexes is too different

[8] D. Varga, P. Halácsy, A. Kornai, V. Nagy, L. Németh, V. Trón, "Parallel corpora for medium density languages," in Recent Advances in Natural Language Processing IV, 247–258, John Benjamins, 2007.

[9] N. T. Ha, N. T. M. Huyen, N. M. Hai, "Building a sentence-aligned Vietnamese–English bilingual corpus in tourism domain for machine translation," JOURNAL OF RESEARCH AND DEVELOPMENT ON INFORMATION AND COMMUNICATION TECHNOLOGY, **V-1, number 39**, 2018.

[10] N. T. M. Huyen, M. Rossignol, "A language-independent method for the alignement of parallel corpora," in Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation, 223–230, Tsinghua University Press, Huazhong Normal University, Wuhan, China, 2006, doi: http://hdl.handle.net/2065/29065.

[11] B. Thompson, P. Koehn, "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), 1342–1348, Association for Computational Linguistics, Hong Kong, China, 2019, doi:10.18653/v1/D19-1136.

[12] K. Chousa, M. Nagata, M. Nishino, "SpanAlign: Sentence Alignment Method based on Cross-Language Span Prediction and ILP," in Proceedings of the 28th International Conference on Computational Linguistics, 4750–4761, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, doi:10.18653/v1/2020.coling-main.418.

[13] S. Luo, H. Ying, S. Yu, "Sentence Alignment with Parallel Documents Helps Biomedical Machine Translation," 2021.

[14] H. Hassan, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang, M. Zhou, "Achieving Human Parity on Automatic Chinese to English News Translation," CoRR, **abs/1803.05567**, 2018.

[15] V. Chaudhary, Y. Tang, F. Guzmán, H. Schwenk, P. Koehn, "Low-Resource Corpus Filtering Using Multilingual Sentence Embeddings," in Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), 261–266, Association for Computational Linguistics, Florence, Italy, 2019, doi:10.18653/v1/W19-5435.

[16] R. Sennrich, B. Haddow, A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1715–1725, Association for Computational Linguistics, Berlin, Germany, 2016, doi: 10.18653/v1/P16-1162.