information

MDPI

Article

Chinese–Vietnamese Pseudo-Parallel Sentences Extraction Based on Image Information Fusion

Yonghua Wen [1,2], Junjun Guo [1], Zhiqiang Yu [1,2] and Zhengtao Yu [1,*]

[1] Faculty of Information Engineering and Automation, Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China; wenyonghua@ymu.edu.cn (Y.W.); junjunguo@kust.edu.cn (J.G.); yzqyt@ymu.edu.cn (Z.Y.)
[2] School of mathematics and computer science, Yunnan Minzu University, Kunming 650500, China
* Correspondence: ztyu@hotmail.com

Abstract: Parallel sentences play a crucial role in various NLP tasks, particularly for cross-lingual tasks such as machine translation. However, due to the time-consuming and laborious nature of manual construction, many low-resource languages still suffer from a lack of large-scale parallel data. The objective of pseudo-parallel sentence extraction is to automatically identify sentence pairs in different languages that convey similar meanings. Earlier methods heavily relied on parallel data, which is unsuitable for low-resource scenarios. The current mainstream research direction is to use transfer learning or unsupervised learning based on cross-lingual word embeddings and multilingual pre-trained models; however, these methods are ineffective for languages with substantial differences. To address this issue, we propose a sentence extraction method that leverages image information fusion to extract Chinese–Vietnamese pseudo-parallel sentences from collections of bilingual texts. Our method first employs an adaptive image and text feature fusion strategy to efficiently extract the bilingual parallel sentence pair, and then, a multimodal fusion method is presented to balance the information between the image and text modalities. The experiments on multiple benchmarks show that our method achieves promising results compared to a competitive baseline by infusing additional external image information.

Keywords: neural machine translation; pseudo-parallel sentence extraction; image information fusion

Citation: Wen, Y.; Guo, J.; Yu, Z.; Yu, Z. Chinese–Vietnamese Pseudo-Parallel Sentences Extraction Based on Image Information Fusion. *Information* **2023**, *14*, 298. https://doi.org/10.3390/info14050298

Academic Editor: David Martins de Matos

Received: 11 April 2023
Revised: 18 May 2023
Accepted: 19 May 2023
Published: 21 May 2023

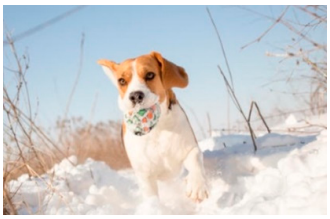
1. Introduction

Parallel sentences are crucial for various NLP tasks, particularly for cross-lingual tasks such as machine translation. The performance of neural machine translation (NMT) positively correlates with the scale and quality of parallel sentence pairs [1]. Unfortunately, large-scale parallel data are available only for a relatively small number of language pairs. Therefore, constructing large-scale high-quality parallel sentence pairs is one of the key tasks in low-resource machine translation [2]. Due to the time-consuming and laborious nature of manual construction, many low-resource language pairs, such as Chinese and Vietnamese, still lack large-scale parallel sentence pairs [3]. Therefore, finding out how to automatically extract and construct parallel sentence pairs from existing internet resources is of great significance to ameliorate low-resource machine translation.

Parallel sentence extraction aims to automatically identify sentence pairs in different languages that convey similar meanings. During the initial phases of parallel sentence extraction, classifiers were developed utilizing maximum entropy classifiers, Bayesian neural networks, and support vector machines [2,4–7]. These methods relied heavily on parallel data, which is unsuitable for low-resource scenarios. While transfer learning and unsupervised learning based on cross-lingual word embeddings and multilingual pre-trained models are the current mainstream research direction for low-resource languages [8–13], they may not be effective for languages with substantial differences.

There is a large amount of multimodal data on the Internet. Image and text are two different types of modalities with great differences in data form and semantic representation. Specifically, image is language-independent modal information, which can interact semantically with multiple languages simultaneously and provide language-independent semantic representation. Mining the natural image text alignment information in Internet data is helpful for the extraction of bilingual parallel sentence pairs by improving the semantic alignment of multilingual languages. As shown in Table 1 below, by fusing the image information associated with the text, the dog information in the image helps to align the entities "Nó" and "dog" in the bilingual text, and the fused image information assists in the extraction of parallel sentence pairs.

Table 1. Examples of Chinese–Vietnamese sentence pairs aligned with text and image.

Image	Description of Image Content	
	Chinese	小狗在雪地里玩耍。 (The dog is playing in the snow.)
	Vietnamese	Nó đang chơi trong tuyết.

Conventional parallel sentence extraction models are mostly limited to the discrimination task of pure text mode. Bilingual parallel sentence pairs are judged based on statistical or deep semantic information. For the multimodal image and text data widely existing on the Internet, how to build a unified deep representation network to represent and align the image and text information in the public semantic space, furthermore, to use language-independent image information to realize the determination of bilingual parallel sentence pairs and improve the quality of parallel sentence pair extraction is one of the key problems to be solved in this paper.

To address the multimodal parallel sentence pair extraction problem, this paper proposes a parallel sentence pair extraction model based on cross-modal interactive fusion. The image information associated with bilingual candidate texts is amplified by mining Internet image–text alignment data, and the candidate images are selectively fused based on the multi-modal interactive fusion module so as to improve the representation of bilingual sentence pairs and determine whether the extracted sentence pairs based on it are highly reliable. The contributions of the paper are as follows:

1. A pseudo-parallel sentence extraction model based on the adaptive fusion of visual and text features is proposed. The performance of text parallel sentence pair extraction is improved by adaptive fusion of candidate image information.
2. A multimodal fusion method based on text selective gating is proposed. Based on multimodal gating, the effective fusion of text and candidate sentences is realized, and the representation ability of text sentence pairs is improved.
3. The experimental results based on multimodal data sets show the effectiveness of the proposed method. By fusing the information of image modality, the ability of extracting parallel sentence pairs is effectively improved.

2. Related Work

The primary objective of parallel sentence extraction is to automatically identify sentence pairs in different languages that convey similar meanings. Previous research has primarily focused on extracting pseudo-parallel sentence pairs in text modality. These methods can be broadly categorized into two categories: (i) utilizing a classifier constructed with maximum entropy classifiers, Bayesian neural networks, and support vector machines and

(ii) utilizing transfer learning and unsupervised learning based on cross-lingual word embeddings and multilingual pre-trained models. Both methods have shown promise in extracting parallel sentences for certain language pairs.

Francis Gregoire et al. [2] encode the source language and the target language based on the bidirectional recurrent neural network, and then judge whether the source sentence and the candidate target sentence are parallel based on the classifier. Jason Smith et al. [4] used IBM Model 1 and a maximum entropy classifier for mining bilingual parallel sentence pairs. Akbar Karimi et al. [5] proposed the use of a twin bidirectional recurrent neural network (BiRNN) to construct the most advanced classifier and use the classifier to detect parallel sentences. Marie et al. [6] propose a new method for extracting pseudo-parallel sentences from a pair of large monolingual corpora. Without relying on any document-level information, the proposed method exploits word embeddings to efficiently evaluate trillions of candidate sentence pairs and use a classifier to find the most reliable sentences. Bonet et al. [7] identified parallel sentence pairs from comparable corpora by measuring the similarity and semantic correlation between translated sentences and then combining the context and similarity measurement information. While these methods have shown promising results, they require bilingual supervision such as a bilingual lexicon or parallel sentences, which may not be available for many low-resource language pairs.

The success of cross-lingual word embeddings and multilingual pre-trained models without the use of parallel corpora has made it possible to apply transfer learning and unsupervised learning to parallel sentence extraction for low-resource language pairs. HangYa et al. [8] first proposed an unsupervised parallel sentence pair extraction method based on the embedding of bilingual words, which relies on word similarity and uses the average value of word vectors to evaluate sentence similarity. XiaYang et al. [9] effectively improved the quality of extracted sentence pairs by using a very small seed lexicon (about hundreds of entries) during the process of learning cross-lingual word representations. ShaoLin et al. [10] proposed a new unsupervised method for obtaining parallel sentence pairs by mapping bilingual word embeddings through postdoc adversarial training and introducing a new cross-domain similarity adaption. YuSun et al. [11] proposed an approach based on transfer learning to mine parallel sentences in an unsupervised setting, which utilizes bilingual corpora of rich-resource language pairs to mine parallel sentences without bilingual supervision of low-resource language pairs. Kvapilíková et al. [12] and Tran et al. [13] used multilingual pre-trained models to improve cross-lingual sentence representations and achieved positive results in parallel sentence pair extraction tasks. These methods can improve the accuracy and recall of mining parallel sentence pairs, but their performance is better for similar languages and not ideal for languages with substantial differences due to the lack of additional alignment information.

In recent years, multimodal machine translation has received widespread attention from researchers [14–16]. Inspired by this idea, we propose a pseudo-parallel sentence extraction method based on image information fusion. To the best of our knowledge, this is the first time that image information has been applied to parallel sentence pair extraction tasks for languages with significant differences, such as Chinese and Vietnamese. By using images as additional alignment information, our method can effectively improve the quality of pseudo-parallel sentence pair extraction and the performance of machine translation for low-resource languages.

3. Method

To handle pseudo-parallel sentence extraction efficiently, our method uses a new image encoder to expand the source sentence encoder and uses the attention mechanism to attach the image representation to the text representation so that the model can focus on the text and image features together. The detail of the model is shown in Figure 1.

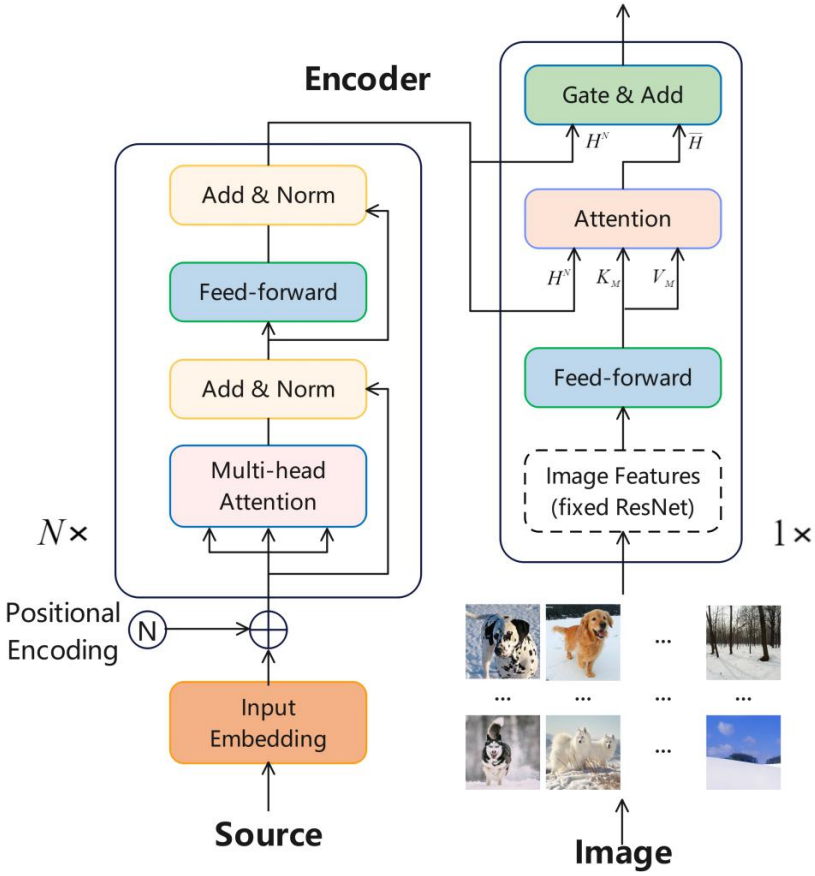


Figure 1. Encoder model of image information fusion. For simplicity, we only list the model of the source language. The encoding mode of the target language is the same as that of the source language.

3.1. Image Retrieval

Since the input of the model is a sentence–image pair, we convert the existing sentence–image pair into a subject–image lookup table, which assumes that the subject words in the sentence should be related to the paired images. Therefore, by searching the subject–image lookup table, a sentence can have a set of images. In order to better focus on the main part of the sentence and reduce noise such as stop words and low-frequency words, we designed a filtering method based on the word frequency inverse document frequency (TF-IDF) [17] to extract the "subject" words in the sentence. Specifically, given an original input sentence $x = \{x_1, x_2, \ldots, x_I\}$ of length I and its paired image e, x is first filtered through a stop word list, and then, the sentence is processed as a document g. Then, the TF-IDF value $TFIDF_{i,j}$ is calculated for each word x_i in g as follows:

$$TFIDF_{i,j} = \frac{o_{i,j}}{\sum_k o_{k,j}} \times \log \frac{|g|}{1 + |j : x_i \in g|} \tag{1}$$

where $o_{i,j}$ denotes the number of times the word x_i appears in the input sentence g, $|g|$ denotes the total number of source sentences in the training data, and $|j : x_i \in g|$ denotes the number of source sentences which contain the word x_i in the training data. Then, we select the first w words with the highest TF-IDF score as the new image description $t = \{t_1, t_2, \ldots, t_w\}$ of the input sentence x. After preprocessing, each filtered sentence t is paired with an image e, and each word $t_i \in t$ is regarded as the subject word of the image e. After processing the whole corpus, a topic–image lookup table Q is formed, in which each topic word t_i will be paired with dozens of images.

Finally, for the input sentence, we first obtain its subject words according to the text preprocessing method described above. Then, the relevant images of each subject word

are retrieved from the lookup table Q, and all the retrieved images are combined to form an image list L. Since an image may be associated with more than one subject word, it appears more than once in the list. We finally rank the images according to the frequency of occurrence and select M images for each sentence.

3.2. Sentence Information Representation

As shown in Figure 1, we use the transformer [18] to encode the source sentence. Given the source language sentence $x = \{x_1, x_2, \ldots, x_I\}$ of length I, a self-attention encoder composed of stacks of the same layers is used to represent it, in which each layer includes two sublayers. The first sublayer is a self-attention module, while the second sublayer is a position-based fully connected feedforward neural network. A residual connection is applied between the two sublayers, and then, layer normalization is performed. Formally, the source sentence is described as follows:

$$\overline{H}^n = LN\left(ATT^n\left(Q^{n-1}, K^{n-1}, V^{n-1}\right) + H^{n-1}\right) \tag{2}$$

$$H^n = LN\left(FNN^n\left(\overline{H}^n\right) + \overline{H}^n\right) \tag{3}$$

where $ATT(\cdot)$, $LN(\cdot)$, and $FFN^n(\cdot)$ denote the attention module, layer normalization, and the feedforward network, respectively. Q^{n-1}, K^{n-1}, and V^{n-1} are the query, key, and value obtained from the update of layer H^{n-1}. After N times stacking, we obtain the final representation H^N of the source sentence.

3.3. Fusion of Text and Image Representation

To achieve the fusion of text and image representation, we paired each original sentence $x = \{x_1, x_2, \ldots, x_I\}$ with images $e = \{e_1, e_2, \ldots, e_m\}$ retrieved from the subject–image lookup table Q. First, the source sentence $x = \{x_1, x_2, \ldots, x_I\}$ is fed to the encoder to generate the representation H^n. Then, the image list $e = \{e_1, e_2, \ldots, e_m\}$ is fed to the pretrained ResNET [19], followed by the feedforward neural network layer to obtain the representation of the original image $textM \in R^{m\times 2048}$. Then, the attention mechanism is used to attach the image to the text representation:

$$\overline{H} = ATT_M\left(H^N, K_M, V_M\right) \tag{4}$$

where $\{K_M, V_M\}$ is the overall representation of the M source images.

Intuitively, the pseudo-parallel sentence extraction model aims to fuse the image information into the encoder and increase the semantic information of the source sentence and the target sentence instead of generating a group of images. In other words, we need the image information to introduce directive information into the matching process. Therefore, we set a weight parameter $\lambda \in [0, 1]$ to calculate the importance of the image to each source word:

$$\lambda = sigmoid\left(W_\lambda \overline{H} + U_\lambda H^N\right) \tag{5}$$

where W_λ and U_λ are model parameters; the final representation of the source sentence is

$$H = H^N + \lambda \overline{H} \tag{6}$$

The encoding process of the target sentence is the same as that of the source sentence. After encoding, the target sentence representation t of the fused image information is obtained. It should be noted that there is an aggregation layer in the model to fuse image and text information.

3.4. Prediction Module

After encoding the source sentence and the target sentence by using the transformer model, their matching information is captured by using their element product and absolute element difference, and then, the matching vector is fed to the feedforward neural network with a sigmoid output layer to estimate the conditional probability of sentence parallelism:

$$U_i^{(1)} = H \odot T \tag{7}$$

$$U_i^{(2)} = \left| H - T \right| \tag{8}$$

$$U_i = \tanh(W^{(1)} U_i^{(1)} + W^{(2)} U_i^{(2)} + b) \tag{9}$$

$$p(y_i = 1|U_i) = \sigma(vU_i + b) \tag{10}$$

where $\sigma(\cdot)$ is sigmoid function, $W^{(1)} \in R^{d_f \times d_h}$, $W^{(2)} \in R^{d_f \times d_h}$, $v \in R^{d_f}$, and $b \in R^{d_f}$ are model parameters, and d_f is the dimension of the hidden layer in the feedforward neural network.

For prediction, if the probability of a sentence pair is greater than or equal to the decision threshold ρ, it is classified as parallel:

$$\hat{y}_t = \begin{cases} 1 \; if \; p(y_i = 1|U_i) \geq \rho, \\ 0 \; otherwise. \end{cases} \tag{11}$$

4. Experimental Setting

In this section, we primarily present our experimental settings and describe the datasets used.

4.1. Dataset

4.1.1. Pseudo-Parallel Sentences

In order to evaluate the effectiveness of our method in extracting parallel sentence pairs, we collected about 10,000 Chinese–Vietnamese pseudo-parallel sentence pairs. These pseudo-parallel sentence pairs were used to build a text–image aligned corpus and also served as positive samples for the evaluation of sentence pair extraction experiments.

4.1.2. Text–Image Aligned Corpus

In order to construct the subject–image lookup table proposed in Section 3.1, we used the CLIP [20] tool for text–image retrieval to match each pseudo-parallel sentence pair with an image and built a Chinese–Vietnamese text–image aligned corpus. These images were mainly used to construct the subject–image lookup table and also provided additional multimodal alignment information for the extraction of pseudo-parallel sentence pairs.

4.1.3. Monolingual Corpus

In order to evaluate the performance of the pseudo-parallel sentence pair extraction method, we extracted about 2 million monolingual Chinese and Vietnamese sentences from the CC-100 corpus [21]. We primarily selected sentences that had a strong correlation with the images in the text–image aligned corpus to improve the quality of the extracted corpus.

4.2. Experiment Setup

To observe the effect of our method on different neural network structures, we implement our model on Bi-RNN, Bi-LSTM, and transformer. For Bi-RNN and Bi-LSTM, the model encoder is set to 6 layers, and the word embedding dimension in the experiment is set to 512. For transformer, the model encoder is set to 6 layers, the dimensions of all input and output layers are set to 512 and 1024, the number of heads of the multi-head attention

is set to 8, and the filter size is set to 2048. In the training, we use the Adam [22] optimizer to adjust the model parameters.

To verify the effect of the sentences extracted by the model on machine translation performance, we extracted 200,000 and 300,000 pseudo-parallel sentence pairs from the Chinese and Vietnamese monolingual corpus using three different methods: the pure-text extraction method [2], the method based on multilingual pre-trained language models [13], and our proposed method. We take the extracted sentence pairs as the machine translation training data set and evaluate the translation quality on the trained model. We select vanilla transformer as the NMT system, the layers of encoder and decoder are set to 6, the dimensions of input and output layers are set to 512 and 1024, and Adam is selected as the optimizer of the model.

5. Results and Discussion

5.1. Classifier Accuracy

We compare our proposed method to different baseline models in terms of precision, recall, and F1 score. As shown in Table 2, the baseline model and experimental results are reported.

Table 2. Evaluation results of our method implemented on different neural network structures (dataset: Chinese–Vietnamese text–image dataset).

Model	P(%)	R(%)	F1(%)
Bi-RNN	82.62	70.80	76.26
Bi-LSTM	85.25	73.64	79.04
Transformer	**87.54**	**75.01**	**80.79**

The experimental results show that the proposed method implemented on transformer achieves better performance compared to Bi-RNN and Bi-LSTM implementation on the Chinese–Vietnamese benchmark. One possible reason is that the self-attention mechanism can conduct efficient use of image information. Apart from this, for the three implementations, we achieved the expected evaluation performance in terms of precision, recall, and F1 score, which also proves that our method can reduce noise and extract effective image representation information.

5.2. Machine Translation Quality

We conducted a comparison of the BLEU scores obtained by our proposed method and the baseline. Table 3 displays the BLEU scores obtained by the NMT system trained with varying numbers of Chinese–Vietnamese pseudo-parallel sentences extracted by the extraction models.

Table 3. BLEU scores obtained on machine translation tasks by different numbers of Chinese–Vietnamese parallel sentences extracted by the baseline system and our method.

Pseudo-Parallel Sentences	Extraction Model	BLEU
200 k	Grégoire et al. [2]	18.04
	Tran et al. [13]	16.34
	Our method	**18.91**
300 k	Grégoire et al. [2]	19.02
	Tran et al. [13]	17.02
	Our method	**19.85**

According to the experimental results, we observed that, compared with the baseline models, our method can achieve significant translation performance improvement when parallel sentence pairs of the same scale are extracted. The results show that the image information can indeed help to improve the performance of the pseudo-parallel sentence

extraction model and improve the quality of the Chinese–Vietnamese NMT model. In addition, when the scale of extracted data increased from 20 k to 30 k, the translation quality is further improved, which shows that our model is effective for noise control and the extraction process is robust.

The main motivation of the proposed method is to incorporate image information as additional alignment cues to improve the quality of extracting pseudo-parallel sentence pairs under unsupervised conditions. Although the pure-text extraction method (Baseline1) was proposed earlier, its performance is better than the current best unsupervised method (Baseline2). We believe there are two main reasons for this. Firstly, Baseline1 uses parallel corpora, which can improve the performance of the method. Secondly, the differences between Chinese and Vietnamese are quite significant, making it difficult for unsupervised methods to effectively learn the alignment relationship between the languages. However, the experimental results show that by utilizing image information, our method achieves the current best results under unsupervised conditions, surpassing traditional pure-text matching methods as well.

Although we only validated the method for Chinese–Vietnamese sentence pair extraction, we believe that this method can be applied well to other low-resource languages with significant differences, which can greatly benefit the development of machine translation for these languages.

5.3. Parameter M

In the extraction procedure, we select M images for each sentence. M is an important parameter in sentence pair extraction. If it is set too small, it is difficult to retrieve useful image information. Otherwise, noise will be introduced when M is set too loosely. We conduct an experiment to study the setting of M, and Figure 2 shows the result.

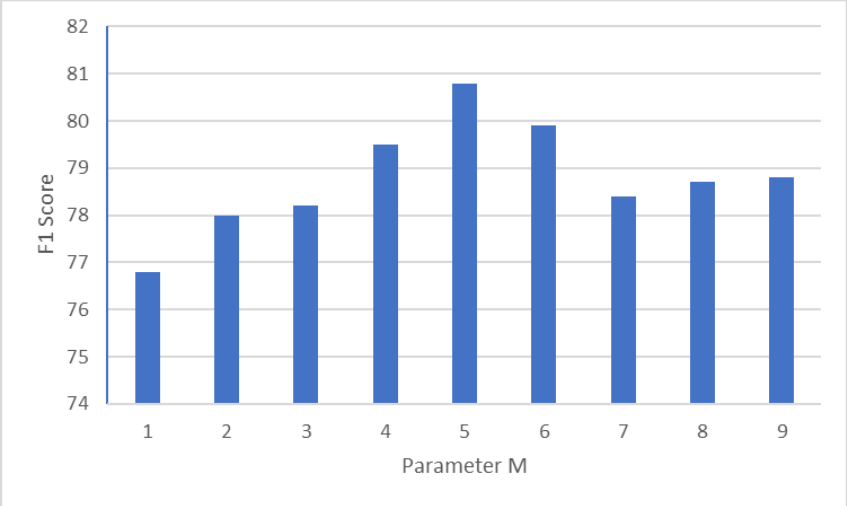


Figure 2. Influence of different values of M on classifier accuracy.

It can be observed from the results that the performance of the pseudo-parallel sentence extraction model is best when M is set to 5. In other words, noise is impaired while the introduction of image information during encoding is maximized.

5.4. Case Study

We conducted a case study on the extracted pseudo-parallel sentences. On the extracted corpus, we carried out example learning. It can be observed from Table 4 that our model can better extract high-quality pseudo-parallel sentence pairs with the help of images.

Table 4. Chinese-Vietnamese pseudo-parallel sentences extracted by our method.

Chinese	根据统计局的预测，今年上半年国内经济增长迅速。(According to the forecast of the Bureau of Statistics, the domestic economy grew rapidly in the first half of this year.)
Vietnamese	Theo báo cáo của tổng thống, nền kinh tế đang phát triển nhanh chóng. (According to the report, the economy is growing rapidly.)
Chinese	在这段时间里经常下雨。 (It rains frequently during this time.)
Vietnamese	Thời gian này sẽ mở ra một vòng mưa. (There will be a round of rain during this time.)

6. Limitations

While the pseudo-parallel sentence pairs we extracted demonstrate strong semantic similarity, they also contain a lot more noise compared to the parallel corpora used in traditional machine translation. This noise can have a negative impact on translation accuracy and make it challenging to achieve high-quality results. Therefore, it is crucial to take into account the noise level when utilizing pseudo-parallel sentence pairs for machine translation.

7. Conclusions

We proposed a novel pseudo-parallel sentence extraction method. The traditional parallel sentence pair extraction task is typically text-based, limiting the available information sources and ignoring the potential benefits of multimodal information. To address this problem, we explored the integration of multimodal information (i.e., text and image) into the attention-based transformer encoder architecture so that the model can focus on text and image information together, rather than relying solely on pure text. We first extracted visual features and incorporated them into the encoding procedure as one of the steps, enabling the model to pay attention to both text and image information simultaneously. Then, we designed a simple and effective attention layer to fuse the visual and source sentence representations and enrich the semantic representation of the sentence. Finally, we calculated the similarity between the source and target sentences using a prediction module, which improves the sentence extraction performance of the Chinese–Vietnamese data pair. The experimental results demonstrate that our method significantly improved the sentence alignment ability for sentence pairs with substantial language differences. Although we have only validated the method for Chinese–Vietnamese sentence pair extraction, we believe that the proposed method can be effectively applied to other low-resource languages with significant differences, which can greatly benefit the development of machine translation for these languages.

Author Contributions: Conceptualization, Y.W. and J.G.; methodology, Z.Y. (Zhiqiang Yu); software, Y.W.; validation, Y.W. and J.G.; formal analysis, Y.W., J.G. and Z.Y. (Zhiqiang Yu); investigation, Y.W., J.G. and Z.Y. (Zhiqiang Yu); resources, Y.W. and J.G.; data curation, Z.Y. (Zhengtao Yu); writing—original draft preparation, Z.Y. (Zhengtao Yu). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant Nos. 62266028, 62241604), Fundamental Research Project of Yunnan Province, China (Grant Nos. 202001AT070046, 202301AT070015), and Yunnan Key Research Projects (Grant No. 202202AE090008-3).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Koehn, P.; Knowles, R. Six Challenges for Neural Machine Translation. In Proceedings of the First Workshop on Neural Machine Translation, Vancouver, BC, Canada, 20 July–4 August 2017; Association for Computational Linguistics: Toronto, ON, Canada; pp. 28–39.
2. Grégoire, F.; Langlais, P. Extracting Parallel Sentences with Bidirectional Recurrent Neural Networks to Improve Machine Translation. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1442–1453.
3. Tran, P.; Thien, N.; Dinh, H.V.; Huu-Anh, T.; Bay, V. A Method of Chinese-Vietnamese Bilingual Corpus Construction for Machine Translation. *IEEE Access* **2022**, *10*, 78928–78938. [CrossRef]
4. Smith, J.R.; Quirk, C.; Toutanova, K. Extracting parallel sentences from comparable corpora using document level alignment. In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies, Los Angeles, CA, USA, 2–4 June 2010; pp. 403–411.
5. Karimi, A.; Ansari, E.; Bigham, B.S. Extracting an English-Persian parallel corpus from comparable corpora. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; pp. 3477–3482.
6. Marie, B.; Fujita, A. Efficient Extraction of Pseudo-Parallel Sentences from Raw Monolingual Data Using Word Embeddings. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2): Short Papers, Vancouver, BC, Canada, 20 July–4 August 2017; pp. 392–398.
7. España-Bonet, C.; Varga, C.; Barrón-Cedeño, A.; Genabith, V. An Empirical Analysis of NMT-Derived Interlingual Embeddings and Their Use in Parallel Sentence Identification. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1340–1350. [CrossRef]
8. Hangya, V.; Braune, F.; Kalasouskaya, Y. Unsupervised Parallel Sentence Extraction from Comparable Corpora. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 1224–1234.
9. Xiayang, S.; Ping, Y.; Xinyi, L.; Chun, X.; Lin, X. Obtaining Parallel Sentences in Low-Resource Language Pairs with Minimal Supervision. *Comput. Intell. Neurosci.* **2022**, *2022*, 5296946.
10. Shaolin, Z.; Chenggang, M.; Tianqi, L.; Yang, Y.; Chun, X. Unsupervised Parallel Sentences of Machine Translation for Asian Language Pairs. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* **2023**, *22*, 64.
11. Sun, Y.; Zhu, S.; Yifan, F.; Mi, C. Parallel sentences mining with transfer learning in an unsupervised setting. In Proceedings of the 2021 Conference of North American Chapter of the Association for Computational Linguistics (NAACL'2021), Mexico City, Mexico, 6–11 June 2021; pp. 136–142.
12. Kvapilíková, I.; Artetxe, M.; Labaka, G.; Agirre, E.; Bojar, O. Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'2020), Online, 5–10 July 2020; pp. 255–262.
13. Tran, C.; Tang, Y.; Li, X.; Gu, J. Cross-lingual retrieval for iterative self-supervised training. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 2207–2219.
14. Baltrušaitis, T.; Ahuja, C.; Morency, L. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 423–443. [CrossRef] [PubMed]
15. Yao, S.; Wan, X. Multimodal Transformer for Multimodal Machine Translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'2020), Online, 5–10 July 2020; pp. 4346–4350.
16. Caglayan, O.; Kuyu, M.; Amac, M.S.; Madhyastha, P.; Erdem, E.; Erdem, A.; Lucia, S. Cross-lingual Visual Pre-training for Multimodal Machine Translation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 1317–1324.
17. Chen, K.; Wang, R.; Utiyama, M.; Sumita, E.; Zhao, T. Neural machine translation with sentence-level topic context. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1970–1984. [CrossRef]
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, 4–9 December 2017; Red Hook: New York, NY, USA, 2017; pp. 6000–6010.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
20. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML'2021), Online, 18–24 July 2021.

21. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 18 July–2 August 2019.
22. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3th International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.