Hindawi
Complexity
Volume 2021, Article ID 5515407, 10 pages
https://doi.org/10.1155/2021/5515407

WILEY | Hindawi

Research Article

Improving Transformer-Based Neural Machine Translation with Prior Alignments

Thien Nguyen,[1] **Lam Nguyen**,[1] **Phuoc Tran**,[1] **and Huu Nguyen**[2]

[1]*Natural Language Processing and Knowledge Discovery Laboratory, Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam*
[2]*Faculty of Information Technology, Ho Chi Minh City University of Food Industry, Ho Chi Minh City, Vietnam*

Correspondence should be addressed to Thien Nguyen; nguyenchithien@tdtu.edu.vn

Received 30 January 2021; Accepted 30 April 2021; Published 8 May 2021

Academic Editor: Dr Shahzad Sarfraz

Transformer is a neural machine translation model which revolutionizes machine translation. Compared with traditional statistical machine translation models and other neural machine translation models, the recently proposed transformer model radically and fundamentally changes machine translation with its self-attention and cross-attention mechanisms. These mechanisms effectively model token alignments between source and target sentences. It has been reported that the transformer model provides accurate posterior alignments. In this work, we empirically prove the reverse effect, showing that prior alignments help transformer models produce better translations. Experiment results on Vietnamese-English news translation task show not only the positive effect of manually annotated alignments on transformer models but also the surprising outperformance of statistically constructed alignments reinforced with the flexibility of token-type selection over manual alignments in improving transformer models. Statistically constructed word-to-lemma alignments are used to train a word-to-word transformer model. The novel hybrid transformer model improves the baseline transformer model and transformer model trained with manual alignments by 2.53 and 0.79 BLEU, respectively. In addition to BLEU score, we make limited human judgment on translation results. Strong correlation between human and machine judgment confirms our findings.

1. Introduction

There was a long period of time when statistical machine translation (SMT) was a dominant translation paradigm. The most effective SMT model is phrase-based. Phrase-based SMT is interpretable, intuitive, and reminiscent of the human translation process. It consists of several separate steps of processing concatenating together in a sequence. For example, a famous phrase-based SMT system with the name Moses created by Koehn [1] contains 9 separate steps including token alignment, lexical translation table creation, and phrase-table creation. The explicitly modular architecture of phrase-based SMT has both advantages and disadvantages. It allows us to easily modify any module to improve the overall system, but it requires us to study multiple modules to create an effective phrase-based SMT system. State-of-the-art neural machine translation (NMT) based on deep learning, on the other hand, adopts an end-to-end approach different from traditional SMT. The whole NMT model is represented as a large neural network consisting of millions of trained parameters, taking as input a sequence of source tokens and returning a sequence of target tokens. NMT does not require us to study each stage of translation separately since it can function as a black box, i.e., if we enter a source sentence, then it will perform some complex numerical operations and return a predicted target sentence for us. Nevertheless, it has been reported that different parts of SMT actually improve NMT models. Han et al. [2] concatenated source token embeddings with their corresponding lexical translation embeddings as an additional input feature. Their experiments show the improvement in translation accuracy for the Chinese-English language pair. Song et al. [3] replaced source phrases with their

corresponding one-to-one target phrases in a phrase table. Their experiments on Chinese-English and English-Russian language pairs demonstrate that hybrid source sentences consistently lead to better translations. Chen et al. [4] proposed the use of prior alignments to guide NMT models. Their experiments with recurrent NMT models in translating from German to English and from English to French reveal large gains in translation quality of recurrent NMT models trained with prior alignments. Garg et al. [5] proposed an adjustment to the state-of-the-art transformer NMT model [6, 7], making the model capable of learning statistical prior alignments. Their experiments for the three language pairs German-English, Romanian-English, and English-French exhibit that the adjusted transformer model consistently produces better posterior alignments, compared with the baseline transformer model. However, an improvement in translation quality does not materialize. There are two possible reasons that the improvement does not occur. First, their statistical prior alignments are perhaps not good enough. Second, the studied language pairs are rich resources; consequently, the state-of-the-art transformer NMT model successfully captures their properties without the help of prior alignments. Nonetheless, there are many machine translation tasks without the luxury of available rich resources. The problem of translating news articles from Vietnamese into English that we are interested in is one of those tasks. Vietnamese-English is a low-resource language pair, and fortunately, a Vietnamese-English bilingual dataset with manually annotated prior alignments is publicly available by Ngo and Winiwarter [8, 9]. Based on these conditions, in this work, we first verify whether manual prior alignments (MA) improve the translation quality for the Vietnamese-English transformer-based NMT model. Second, we experiment different Vietnamese-English transformer-based NMT models trained with statistical prior alignments (SAs), with the objective of approaching the quality of the model trained with manual prior alignments.

The rest of the paper is divided into six sections. The first section reviews related works. The second section introduces the proposed transformer-based neural machine translation models guided by prior alignments. The third section presents the raw material and the preprocessing steps applied on it to get datasets for our study. The fourth section describes the experiments and discussion on their results. The fifth section unveils a limitation of the proposed models and a future work on improvement. The final section gives conclusions from this work.

2. Related Works

In this section, we briefly review works related to our study on improving transformer-based neural machine translation with prior alignments.

2.1. Token Alignments. Token alignments for a pair of sentences are a relation from the set of token positions in the source sentence to a set of token positions in the target sentence. An alignment can be intuitively represented in Pharaoh format [10] as a tuple $(j-i)$, where the first element indicates j-th source token and the second element indicates i-th target token. Preparing token alignments is a crucial part of the traditional SMT models. The most popular token alignment tool is Giza++ [11], which is used by default in the famous SMT system Moses [1]. Giza++ implements the IBM Model 4 [12]. In addition to Giza++, there is another efficient token alignment tool fast_align by Dyer et al. [13], which effectively implements the IBM Model 2 [12]. Dyer et al. reported that the fast_align tool provides alignment as well as Giza++ does, while running significantly faster. Based on the efficiency and alignment quality, in this study, we prefer fast_align to Giza++ for statistically aligning source and target tokens.

2.2. Recurrent NMT Models Trained with Prior Alignments. While modern NMT models outperform SMT models in terms of translation quality, the task of token alignment is still dominant by traditional statistical tools [5]. Chen et al. [4] combined the advantages of two approaches by using statistical prior alignments to train recurrent NMT models. For German-English and English-French tasks, they experiment two recurrent NMT models trained with prior alignments which have been generated with Giza++ [11]. Their experiment results show that the proposed models significantly improve over baseline recurrent NMT models. Chen et al. also introduced alignment cost for the mismatch between prior alignments and computed single-head attention mechanism of the recurrent models. Further developments on using prior alignments to improve recurrent NMT models can be found in [14–17]. Moreover, a recurrent neural network model trained with prior alignments has also been proved effective in speech synthesis task [18], which has sequence-to-sequence pattern similar to machine translation task.

2.3. Baseline Transformer Model. Recently, a novel deep neural network model, transformer [6], with an innovative multihead attention mechanism has been introduced. It has become the state-of-the-art model for many artificial intelligence tasks, including machine translation [19–22]. In comparison with other NMT models, including recurrent ones, transformer not only provides better translation results but also can be trained in a shorter period of time [6]. In this work, we use the transformer model as the baseline translation system. The transformer model is composed of encoder and decoder modules. The output probability distribution $p_t = (p_{t1}, p_{t2}, \ldots, p_{t\Psi})$ of the decoder is then used to predict the next target token.

Given a reference target sentence containing T tokens, the mathematical formulation of the optimization criterion for training the transformer model is presented in equation (1), revised from the one provided by Muller et al. [23]:

$$\mathscr{L}_1 = -\frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{\Psi}\left(r_{ij} \times \log\left(p_{ij}\right)\right). \quad (1)$$

In equation (1), the symbol r_{ij} indicates whether j-th token in the dictionary is the true value at the i-th position in the target sentence.

2.4. Transformer Model Guided by Prior Alignments. Garg et al. [5] altered the state-of-the-art transformer NMT model [6, 7] for joint alignment and translation tasks, making use of prior alignments in training the model. The revised transformer model has the same architecture as the baseline transformer model with a slightly different training procedure. They replace the optimization criterion with a modified one including prior alignments. Specifically, for a pair of source and target sentences of length K and T, respectively, and a prior alignment set $\mathscr{A} \subseteq \{(j-i): j = 1, \ldots, K; \quad i = 1, \ldots, T\}$, they randomly take the output of just a head (n can be any number from 1 to 8) of the fifth decoder layer and then project it into a sequence of T probability distributions $(q_{ij})_{i=1,j=1}^{i=T,j=K}$ over K tokens of the corresponding source sentence for every target token. They compare the probability distributions q_{ij} with the reference probability generated from prior alignments via cross-entropy:

$$\mathscr{L}_2 = -\frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{K}\left(a_{ij} \times \log\left(q_{ij}\right)\right). \tag{2}$$

In equation (2), the symbol a_{ij} indicates the probability of whether the i-th target token is correctly aligned with the j-th source token.

Taken together, the optimization criterion for the transformer-M model is the sum of cross-entropy for tokens and a weighted cross-entropy for alignments between source and target sentences in the training dataset:

$$\mathscr{L} = \mathscr{L}_1 + 0.05\mathscr{L}_2. \tag{3}$$

2.5. Proposed Transformer-Based Models Trained with Prior Alignments. In experiments for German-English, Romanian-English, and English-French translation tasks, Garg et al. used prior alignments created with Giza++ to train the revised transformer models. The models generate better posterior alignments but do not provide better translations. Motivated by the improvement in translation quality of recurrent NMT models trained with prior alignment [4], we experiment training transformer models with manually constructed alignments (transformer-M) for our Vietnamese-English translation task. The availability of manual token alignments $\mathscr{A}_M$ allows us to assess the statement on whether prior alignments help us to build a better transformer model. Unfortunately, the approach is labor-consuming and does not provide us the freedom to make a choice of token type other than the one used in manual token alignments. Consequently, aside from the transformer-M model, we build other transformer models trained on statistically constructed prior alignments (transformer-S). Transformer-S models employ different token types and are trained on statistically constructed prior alignments instead of manually annotated prior alignments, while keeping the same architecture and training procedure as for the transformer-M model.

2.6. Syllable-to-Word Transformer Model. The first transformer-S model (transformer-S1) is guided by alignments $\mathscr{A}_{S1}$ constructed with the fast_align token aligner in the place of Giza++ as in the study by Garg et al. [5]. In addition to the change of aligner, we adapt their procedure for constructing statistical alignments to suit the Vietnamese-English translation task. The adapted procedure is presented as Algorithm 1.

2.7. Word-to-Subword Transformer Model. Influenced by the work of Nguyen et al. [25] for Russian-Vietnamese NMT, we create the second transformer-S model (transformer-S2). While utilizing the same architecture, training procedure, and procedure to construct statistical alignments (Algorithm 1) of Transformer-S1 model, we tokenize the sentences differently in the transformer-S2 model. On the Vietnamese source side, we segment sentences into words, and on the English source side, we divide the sentences into subwords. We decide to adopt this mixed model due to the difference in linguistic morphology between Vietnamese and English. While Vietnamese is a noninflectional language, English is an inflectional language although not as morphologically rich as Russian. We use the VnCoreNLP tool developed by Vu et al. [26] and further improved by Nguyen et al. [27] to segment Vietnamese sentences into words. There is a popular phenomenon that, in Vietnamese, a syllable appears in many different words; therefore, these syllables are ambiguous to recognize by classifiers. We deploy segmentation of Vietnamese sentences into words to reduce ambiguity and, consequently, to enhance the quality of the transformer-S2 model. An example of a Vietnamese sentence and the result of its segmentation into words are presented in Table 1.

The VnCoreNLP tool employs character "_" to inform that neighboring syllables are concatenated into a word. In Table 1, two syllables "lãnh" and "thổ" are concatenated into a word "lãnh_thổ."

On the English target side, we divide sentences into subwords with BPE tool proposed by Sennrich et al. [28]. An example of an English sentence and the result of its segmentation into subwords are presented in Table 2.

The BPE tool uses a pair of characters "@@" to inform that a containing token is a subword and should be concatenated with the next token to form a word in the inference phase of the transformer-S2 model. For some words, segmentation into subwords is interpretable, such as the word "personally" is divided into 2 subwords "person" and "ally" (Table 2). Subword "person" is the root part of many other words, such as "personal," "personalize," and "personality." The segmentation actually has some grammatical meaning. A similar meaningful segmentation can be found for the word "ignorant" divided into "ignor" and "ant." Meanwhile, there are other words where their segmentation is not

(1) We tokenize both Vietnamese source sentences and English target sentences. We apply the types of tokens in the Transformer-S1 model as in the case of the Transformer-M model. A token in both source and target sentences is a sequence of characters delimited by spaces. Linguistically, Vietnamese-English Transformer-M and Transformer-S1 models are syllable-to-word models since spaces in Vietnamese delimit syllables and spaces in English delimit words.
(2) We construct many-to-one alignments from Vietnamese to English, using the fast_align token aligner.
(3) We repeat step 2 in the reverse direction from English to Vietnamese.
(4) We merge the bidirectional alignments generated in steps 2 and 3, following grow-diagonal heuristics proposed by Koehn et al. [24].

Algorithm 1: Procedure to construct statistical alignments $\mathscr{A}_{S1}$.

Table 1: A Vietnamese sentence and the result of its segmentation into words.

Feature	Example
Vietnamese sentence	"lãnh thổ Trung Quốc rộng bao nhiêu và diện tích đất của nó đứng hàng thứ mấy ?"
Segmentation into words	"lãnh_thổ Trung_Quốc rộng bao_nhiêu và diện_tích đất của nó đứng hàng thứ mấy ?"

Table 2: An English sentence and the result of its segmentation into subwords.

Feature	Example
English sentence	"I personally like to call them mob youth or ignorant angry youth."
Segmentation into subwords	"I person@@ ally like to call them mo@@ *b* youth or ignor@@ ant angry youth."

understandable. In Table 2, the word "mob" is divided into two meaningless subwords "mo" and "b."

Overall, the transformer-S2 model is a variant of the transformer-S1 model with different token representations on the source and target side.

Moreover, the imperfect segmentation of English sentences into subwords stimulates us to propose a novel transformer-S3 model without the use of English subwords, which puts more focus on the linguistic aspects of machine translation, such as the use of lemmas.

2.8. Hybrid Word-to-Word Transformer Model Trained with Statistical Word-to-Lemma Alignments. The transformer-S3 model can be seen as a hybrid of transformer-S1 and transformer-S2 models. Specifically, the transformer-S3 model is a word-to-word model. On the Vietnamese source side, we segment sentences into words, such as in the transformer-S2 model, while on the English target side, we choose to divide sentences into words, such as in the transformer-S1 model. Nevertheless, in preparing prior alignments $\mathscr{A}_{S3}$, we revise the procedure to construct statistical alignments (Algorithm 1), replacing English words with their lemmas. Step-by-step procedure to construct $\mathscr{A}_{S3}$ alignments is presented as Algorithm 2.

In Algorithm 2, we replace English words with their lemmas, using Stanza tool created by Qi et al. [29]. A word is a surface form of a lemma according to its grammatical role in sentences. For example, words "life" and "lives" are inferred from the same lemma "life," depending on the grammatical number. An example of an English sentence and the result of its lemmatization are shown in Table 3.

We adopt lemmatization of English words to reduce the size of vocabulary of the training dataset. The English side of the training dataset contains 36672 distinct tokens inflected from a smaller number of 28583 lemmas. We hope that a reduced vocabulary and an unchanged number of tokens will allow the fast_align aligner to produce better alignments and, consequently, lead to a better translation model trained on them. The relation between English words and their lemmas is one-to-one (see index sequences in Table 4); therefore, Vietnamese-word-to-English-lemma alignments can be employed in training the word-to-word transformer-S3 model.

We want to restate an important characteristic of the transformer-S3 model. The lemmatization of English target words is only applied in the construction of statistical alignments. We still use English words in the translation model.

3. Materials

In this work, we use English-Vietnamese Word Alignment Corpus (EVWACorpus) provided by Ngo et al. [9]. The dataset consists of 1000 news articles with 45,531 sentence pairs. These sentence pairs are already tokenized and manually aligned at the token level. A token is a sequence of characters delimited by spaces.

We apply the following processing procedures to the original EVWACorpus so that it fits our study.

3.1. True-Cased Corpus. First, we use true-case sentences in the dataset with Moses tool of Koehn et al. [1]. The term "true-case" means to convert a token to its most possible case. For example, the true-cased form of the token "The" is "the." An example of a sentence in its natural form and its converted true-cased form is presented in Table 5.

(1) We tokenize both Vietnamese source sentences and English target sentences into words
(2) We replace English words with their lemmas
(3) We construct many-to-one alignments from Vietnamese words to English lemmas, using the fast_align token aligner
(4) We repeat step 2 in the reverse direction from English lemmas to Vietnamese words
(5) We merge the bidirectional alignments generated in steps 2 and 3, following grow-diagonal heuristics proposed by Koehn et al. [24]

Algorithm 2: Procedure to construct statistical alignments $\mathcal{A}_{S3}$.

Table 3: Overview of the datasets.

Vietnamese/English	Training	Validation	Testing
Number of news articles	930	40	30
Number of sentences	42,026	1,482	1,527
Average sentence length	26.2/19.2	24.5/17.8	28.3/20.6
Alignments per sentence	22.4	20.8	23.1
Number of unique tokens	16441/36672	2720/4981	3462/6211
Number of alignments	942001	30821	35291
Number of tokens	1099205/806456	36276/26315	43286/31513

Table 4: An English sentence and the result of its lemmatization

Feature	Example
English sentence	"teaching English to primary students is very different from secondary or high school students, thus training teachers at primary schools needs careful attention, said John A. Scacco, at the US embassy in Bangkok."
Result of its lemmatization	"teach English to primary student be very different from secondary or high school student, thus training teacher at primary school need careful attention, say John A. scacco, at the US embassy in bangkok."
Word index sequence	"1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37"
Lemma token index sequence	"1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37"

True-casing procedure focuses on capitalized tokens (in Table 1, they are "The," "Fenqing," and "China." Based on the frequency calculated from the corpus, these tokens will be converted to lower-cased form or stay unchanged.

3.2. Filtered Corpus. Secondly, we leave some sentence pairs out of our work. We filter out wrongly aligned sentence pairs. Sentence pairs are considered wrongly aligned if the indices of tokens are greater than the length of sentences. Due to the computational reasons, we also remove sentence pairs containing any sentence of length greater than 80 tokens. Moreover, we transform the alignment representation in EVWACorpus into Pharaoh format for later use. Finally, we get 45,035 sentence pairs with manually annotated alignments. An example of a sentence pair in the filtered corpus is presented in Table 6.

3.3. Datasets Extracted from Filtered Corpus. We divide the filtered corpus into three datasets: training, validation, and testing dataset for training and evaluating different translation models. We apply a dividing procedure similar to the one used by Nguyen et al. [30]. Specifically, we randomly take 1,527 sentence pairs from 30 news articles and use them as the testing dataset. Then, we randomly take the other 1,482 sentence pairs from the other 40 news articles and use them as the validation dataset. The remaining 42,026 sentence pairs from 930 news articles form the training dataset. Overview of the datasets is shown in Table 3.

4. Experiments and Discussion

Google Brain team releases an implementation of the transformer model in the Tensor2tensor library [7]. The library is now replaced by its successor Trax (download at https://github.com/google/trax). The transformer model is implemented in other popular NMT libraries, such as opennmt [31, 32] and Fairseq [33] of Facebook AI Research team. To carry out our experiments, we choose to use Fairseq library because it allows us to build both transformer models trained with/without prior alignments.

Following the architecture and training procedure for transformer models presented in previous sections, we apply Adam optimizer with learning rate 0.0002 to train them in 10,000 steps of 3200 tokens. After completing each epoch of the training dataset, we save the model. Among all saved models, we choose the one with the best performance in the validation dataset.

We use the testing dataset to evaluate the translation models. Each model translates all Vietnamese sentences from the testing dataset, deploying a beam search of size 5. The predicted English sentences are then compared with the corresponding reference English sentences from the testing dataset via BLEU score [34]. We apply the script multi-

Table 5: An example of a sentence in its natural form and its converted true-cased form.

Sentence	Example
In natural form	"The fact that most Fenqing are ignorant of many things determines their opinions and views of the world and China."
In true-cased form	"the fact that most Fenqing are ignorant of many things determines their opinions and views of the world and China."

Table 6: A Vietnamese-English sentence pair with manually annotated alignments.

Feature	Example
Vietnamese sentence	"cô con gái đưa sự việc ra tòa để cố gắng lấy lại số tiền."
English sentence	"the daughter took the case to court in an effort to recover the funds."
Alignments	"0-1 1-1 2-1 3-2 4-4 5-4 6-5 8-7 9-9 10-9 11-11 12-11 13-13 14-13"

bleu.perl (download at https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl) in Moses program [1] to calculate the score. Since BLEU score is a statistical metric designed to be applied on the dataset level, we also make complementary human judgments on the sentence level. Specifically, we randomly take 5 Vietnamese-English sentence pairs from the testing dataset, where the source sentence is composed of 10, 15, 20, 25, and 30 tokens, sequentially. We then make human judgment on the selected sentence pairs to complement the automatic machine judgment in the form of BLEU scores.

Figure 1 shows BLEU scores of the translation results of the testing dataset by the transformer models. We can find that the transformer-M model trained with manual prior alignments significantly outperforms the baseline transformer model by $16.26 - 14.52 = 1.74$ BLEU ($\approx 12\%$) on the overall dataset level. The first question of our work already has an answer. Prior alignments actually help improve the translation quality of the transformer model.

Figure 1 also reveals a surprising result. Performance of the statistical transformer-S3 model is even better than expected. It not only outperforms other statistical models but also exceeds our expectation of approaching the result of the manual transformer-M model by giving the highest BLEU score. The statistical transformer-S3 model improves the manual transformer-M model by $17.05 - 16.26 = 0.79$ BLEU. This can be explained by the fact that the quality of manual alignments relies on human, and human does not always provide correct alignments. It is worth to notice that it is difficult to manually align tokens between the source and target sentences. This language-related task is generally ambiguous, which is stated by Lambert et al. [35]. Moreover, the highest BLEU score of the translation result by the transformer-S3 model demonstrates the power of the statistical approach and its flexibility.

We now examine whether human judgment on translation results is correlated with automatic machine judgment on the sentence level. Here are five testcases which we randomly take and study.

Table 7 shows the translation results of a Vietnamese sentence comprising 10 tokens by transformer models. Clearly, the three presented translation models fail to translate the Vietnamese source sentence. However, from the semantic standpoint, the transformer-S3 model is better than others, successfully translating the subject "siêu nhân" of the Vietnamese source sentence into the reference target word "Superman." Nevertheless, from the technical standpoint, the baseline transformer model performs better by providing the most number of reference target tokens "only," "can," "do," while the transformer-S3 model misunderstands the source phrase "làm được" and translates them into a passive verb phrase "are done." This incorrect translation is very interesting because Vietnamese token "được" is mostly used in passive voice. Thus, the transformer-S3 model does make the same mistake as foreign learners of Vietnamese usually do.

Table 8 presents the translation results of a Vietnamese sentence consisting of 15 tokens by transformer models. This test case actually proves the superiority of the transformer-S3 model in comparison with other models. Translation by the transformer-S3 model bears the most resemblance in meaning to the full English reference target sentence. Nevertheless, the transformer-S3 model chooses a wrong tense of the verb "stop." Instead of the reference verb phrase of the past perfect tense "had stopped," the transformer-S3 model uses the verb of simple present tense "stop." It is understandable, considering the fact that Vietnamese verbs, such as "ngừng" in the source sentence, usually do not appear in tense; hence, translation models or even human translators find it difficult to translate Vietnamese verbs.

Table 9 shows the translation results of a Vietnamese sentence consisting of 20 tokens by transformer models. All three presented translation models perform pretty well in this case. Their translations generally reflect the meaning of the source sentence. Still, the translation by the transformer-S3 model is semantically closest to the reference target sentence. The transformer-S3 model translates the key phrase "nhàm chán" into the correct target word "boring." However, it repeats the error of translating Vietnamese verbs as in testcase 2. It mistranslates the source verb phrase "không biết" into the target verb phrase of the present simple tense "don't know," while the reference target phrase "didn't know" is of past simple tense. At the same time, the baseline transformer model correctly identifies the tense, producing the target phrase "didn't know."

Translations of a Vietnamese sentence comprising 25 tokens are presented in Table 10. This test case unveils the positive effect of the flexibility of the statistical alignment approach. We can apply a statistical aligner to different kinds of tokens without limiting ourselves to a preselected kind of

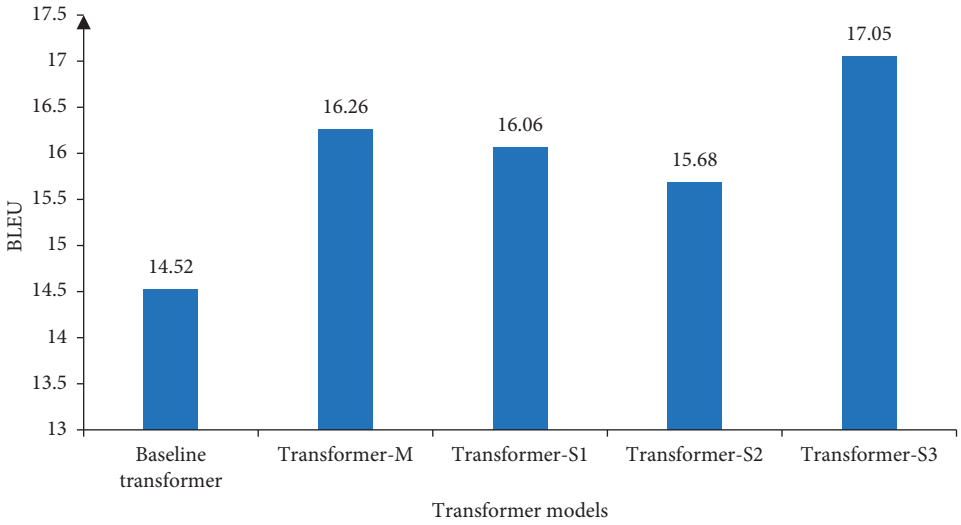

Figure 1: Translation results by transformer models.

Table 7: Translations of a Vietnamese sentence comprising 10 tokens.

	Testcase 1
Vietnamese source	"chỉ có siêu nhân mới làm được như vậy."
English reference	"only Superman can do that."
Translation by transformer	"only a new scan can do so."
Translation by transformer-M	"only an ultrasound is done as well."
Translation by transformer-S3	"only Superman are done."

Table 8: Translations of a Vietnamese sentence comprising 15 tokens.

	Testcase 2
Vietnamese source	"do tác dụng phụ nên 10% bệnh nhân ngừng uống thuốc trong hai năm."
English reference	"as a result of the side effects, 10% of the patients had stopped taking the drug within two years."
Translation by transformer	"due to the effects of 10% of the patient who took 10% of drinking pills for two years."
Translation by transformer-M	"due to side side effects should stop 10% of the patient who stopped the medicine for two years."
Translation by transformer-S3	"due to side effects, 10% of patients stop taking drugs in two years."

Table 9: Translations of a Vietnamese sentence comprising 20 tokens.

	Testcase 3
Vietnamese source	"buổi đêm ở đây khá nhàm chán vì chúng tôi không biết phải làm gì trước khi đi ngủ."
English reference	"night can be quite boring because we didn't know what to do before sleeping."
Translation by transformer	"it was pretty tired because we didn't know what to sleep before bed."
Translation by transformer-M	"tonight is quite tired because we don't know what to sleep before going to sleep."
Translation by transformer-S3	"this night is quite boring because we don't know what to do before going to bed."

tokens as in the case of manual alignments. Specifically, the transformer-S3 model successfully produces the target word "appearance," having concatenated two neighboring syllables "ngoại" and "hình" into one word "ngoại_hình" (see Table 10). This happens due to the fact that we choose to build the transformer-S3 model as a linguistics-informed word-to-word model, while the baseline transformer model and transformer-M model are syllable-to-word models. These models require tokenization of Vietnamese sentences into syllables and English sentences into words.

Table 11 displays the translations of a Vietnamese sentence comprising 30 tokens. All three presented translation models fail to translate the key phrases of the source sentence. The subject "bang Gujarat" (meaning: the state of Gujarat) of the source sentence is mistranslated into different things: "federal federal government," "the federal states," and "the state of states." Nevertheless, the translation by the baseline transformer model is smoother, consisting of many reference tokens. Unfortunately, it misses two key words "illegal" and "toxic"; therefore, its meaning is totally

Table 10: Translations of a Vietnamese sentence comprising 25 tokens.

	Testcase 4
Vietnamese source	"ngoại hình của chúng ta là yếu tố đầu tiên mà mọi người để ý đến và giúp họ hình dung về chúng ta."
English reference	"our appearance is the first thing people notice, and it gives them an idea of who we are."
Translation by transformer	"our foreign image is the first factor that people will take attention to their ideas and help them figure out."
Translation by transformer-M	"our foreign form is the first factor that people come to attention and help them figure out us."
Translation by transformer-S3	"our appearance is the first factor that people notice and help them figure out of us."

Table 11: Translations of a Vietnamese sentence comprising 30 tokens.

	Testcase 5
Vietnamese source	"tuần trước, bang Gujarat đưa ra một điều luật mới quy định rằng việc sản xuất và bán rượu độc bất hợp pháp sẽ bị phạt tử hình."
English reference	"last week, the state of Gujarat brought in a new law making the illegal manufacture and sale of toxic alcohol there punishable by death."
Translation by transformer	"last week, the federal federal government issued a new law that the production of manufacturing and selling alcohol would be charged with death."
Translation by transformer-M	"last week, the federal states issued a new law that production and illegal consumption would be charged with death."
Translation by transformer-S3	"last week, the state of states introduced a new law that the production of production and illegal alcohol will be charged with death."

different from the reference. While the transformer-S3 model delivers stutters ("state of states" and "production of production"), it yields a correct key word "illegal," making the translation result better resemble the reference in meaning.

On the whole, human judgment is in line with automatic machine judgment on the quality of the translation models. From the semantic point of view, the transformer-S3 model is the best model. Moreover, we discover that the transformer-S3 model does not succeed at handling the verb tenses.

4.1. Limitation and Future Work. Despite many advantages of training transformer-based NMT models with prior alignments, especially statistical ones, it still has a noticeable disadvantage. The models trained with them poorly handle verb tenses in translation. Translations of the best transformer-S3 model may reflect the meaning of the source sentences; however, they do not guarantee a high BLEU score since they generate verbs in an incorrect tense.

This work is the first step towards enhancing translation quality of transformer-based NMT models trained with prior alignments. Future work will address the pitfall of the word-to-word transformer-S3 model trained with statistical word-to-lemma alignments. Research into solving this problem is in progress. We will explore the selection of a head in the multihead attention mechanism, whose output is compared with prior alignments.

5. Conclusions

In this study, we have proved that prior alignments help better train the Vietnamese-English transformer-based neural machine translation model. Experiment results show the improvement of translation quality in terms of BLEU score. Moreover, to free ourselves from dependence on costly manual alignments, we have proposed a novel hybrid word-to-word transformer model trained on statistical word-to-lemma alignments. Unlike strict manual alignments, the flexible statistical aligner allows us to construct word-to-lemma alignments, representing a Vietnamese source sentence as a sequence of words and the corresponding English target sentence as a sequence of lemmas. Statistically constructed word-to-lemma alignments are then used to train a word-to-word transformer-S3 model instead of word-to-word alignment. Experiments have demonstrated that the novel word-to-word transformer-S3 model trained with statistical word-to-lemma alignments outperforms the transformer-M model trained with manual alignments in terms of BLEU score. In addition to machine judgment, we have made limited human judgments on translation results. Strong correlation between human and machine judgment has validated our findings.

Based on the experiment results, we recommend the use of statistical prior alignments in training the transformer-based neutral machine translation models at least in the context of low-resource translation tasks.

Data Availability

Readers can obtain the datasets used in this work by contacting the corresponding author Thien Nguyen via e-mail: nguyenchithien@tdtu.edu.vn.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors truly appreciate Ms. Trang Nguyen, a translator and scientist. She provided the authors invaluable recommendations and encouragement when they prepared the manuscript and chose a journal to submit our work to.

References

[1] P. Koehn, "Moses: open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the ACL on interactive Poster and Demonstration Sessions*, pp. 177–180, Stroudsburg, PA, USA, June 2007.

[2] D. Han, J. Li, Y. Li, M. Zhang, and G. Zhou, "Explicitly modeling word translations in neural machine translation," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 19, no. 1, pp. 1–17, 2020.

[3] K. Song, Y. Zhang, H. Yu, W. Luo, K. Wang, and M. Zhang: Code-switching for enhancing NMT with pre-specified translation, https://www.aclweb.org/anthology/N19-1044.pdf.

[4] W. Chen, E. Matusov, S. Khadivi, and J. T. Peter, "Guided alignment training for topic-aware neural machine translation," in *Proceedings of the AMTA 2016: 12th Conference of the Association for Machine Translation in the Americas*, pp. 121–134, Austin, Texas, USA, October 2016.

[5] S. Garg, S. Peitz, U. Nallasamy, and M. Paulik, "Jointly learning to align and translate with transformer models," 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)*, pp. 4453–4462, Hong Kong, China, November 2019.

[6] A. Vaswani, "Attention is all you need," in *Advances in Neural information Processing Systems*, pp. 5998–6008, MIT Press, Cambridge, MA, USA, 2017.

[7] A. Vaswani, "Tensor2tensor for neural machine translation," in *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas*, pp. 193–199, Boston, MA, USA, March 2018.

[8] Q. H. Ngo and W. Winiwarter, "Building an English-Vietnamese bilingual corpus for machine translation," in *Proceedings of the 2012 International Conference on Asian Language Processing*, pp. 157–160, Hanoi, Vietnam, November 2012.

[9] Q. H. Ngo, W. Winiwarter, and B. Wloka, "EVBCorpus-a multi-layer English-Vietnamese bilingual corpus for studying tasks in comparative linguistics," in *Proceedings of the 11th Workshop on Asian Language Resources*, pp. 1–9, Nagoya, Japan, October 2013.

[10] P. Koehn, "Pharaoh: a beam search decoder for phrase-based statistical machine translation models," in *Machine Translation: From Real Users to Research*, pp. 115–124, Springer, Berlin, Germany, 2004.

[11] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[12] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[13] C. Dyer, V. Chahuneau, and N. A. Smith, "A simple, fast, and effective reparameterization of ibm model 2," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 644–648, Atlanta, GA, USA, June 2013.

[14] J.-T. Peter, A. Nix, and H. Ney, "Generating alignments using target foresight in attention-based neural machine translation," *The Prague Bulletin of Mathematical Linguistics*, vol. 108, no. 1, pp. 27–36, 2017.

[15] T. Alkhouli and H. Ney, "Biasing attention-based recurrent neural networks using external alignment information," in *Proceedings of the Second Conference on Machine Translation*, pp. 108–117, Florence, Italy, August 2017.

[16] J. Zeng, "Multi-domain neural machine translation with word-level domain context discrimination," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 447–457, Brussels, Belgium, October 2018.

[17] L. Liu, M. Utiyama, A. Finch, and E. Sumita, "Neural machine translation with supervised attention," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3093–3102, Osaka, Japan, December 2016.

[18] X. Zhu, Y. Zhang, S. Yang, L. Xue, and L. Xie, "Pre-alignment guided attention for improving training efficiency and model stability in end-to-end speech synthesis," *IEEE Access*, vol. 7, pp. 65955–65964, 2019.

[19] J. Zhang, "Improving the transformer translation model with document-level context," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 533–542, Brussels, Belgium, November 2018.

[20] I. V Tetko, P. Karpov, R. Van Deursen, and G. Godin, "State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis," *Nature Communications*, vol. 11, no. 1, pp. 1–11, 2020.

[21] N. Parmar, "Image transformer," in *Proceedings of the International Conference on Machine Learning*, pp. 4055–4064, Jinan, China, May 2018.

[22] T. Wolf, "Transformers: state-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Suzhou, China, November 2020.

[23] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in *Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, pp. 4696–4705, Vancouver, Canada, December 2019.

[24] P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot, "Edinburgh system description for the 2005 IWSLT speech translation evaluation," in *Proceedings of the IWSLT-2005*, Pittsburgh, PA, USA, 2005.

[25] T. Nguyen, H. Nguyen, and P. Tran, "Mixed-level neural machine translation," *Computational Intelligence and Neuroscience*, vol. 2020, Article ID 8859452, 7 pages, 2020.

[26] T. Vu, D. Q. Nguyen, M. Dras, and M. Johnson, "VnCoreNLP: A Vietnamese natural language processing toolkit," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 56–60, New Orleans, LA, USA, June 2018.

[27] D. Q. Nguyen, T. Vu, M. Dras, and M. Johnson, "A Fast and Accurate Vietnamese Word Segmenter," in *Proceedings of the 11th International Conference on Language Resources and Evaluation, LREC 2018*, pp. 2582–2587, Miyazaki, Japan, May 2019.

[28] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725, Berlin, Germany, August 2016.

[29] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning: Stanza: A {Python} Natural Language Processing Toolkit for Many Human Languages, 2020, https://nlp.stanford.edu/pubs/qi2020stanza.pdf..

[30] T. Nguyen, H. Le, and V.-H. Pham, "Source-word decomposition for neural machine translation," *Mathematical Problems in Engineering*, vol. 2020, Article ID 4795187, 10 pages, 2020.

[31] G. Klein, Y. Kim, Y. Deng et al., "Neural machine translation toolkit," in *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas*, pp. 177–184, Boston, MA, USA, March 2018.

[32] G. Klein, F. Hernandez, V. Nguyen, and J. Senellart, "The OpenNMT neural machine translation toolkit: 2020 edition," in *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020)*, pp. 102–109, October 2020.

[33] M. Ott, "Fairseq: a fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 48–53, Minneapolis, MN, USA, June 2019.

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, PA, USA, July 2002.

[35] P. Lambert, A. Gispert, R. Banchs, and J. B. Mariño, "Guidelines for word alignment evaluation and manual alignment," *Language Resources and Evaluation*, vol. 39, no. 4, pp. 267–285, 2005.