Hindawi
Complexity
Volume 2021, Article ID 5935958, 9 pages
https://doi.org/10.1155/2021/5935958

Research Article

Sublemma-Based Neural Machine Translation

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

[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 14 May 2021; Revised 15 June 2021; Accepted 24 September 2021; Published 8 October 2021

Academic Editor: Shahzad Sarfraz

Powerful deep learning approach frees us from feature engineering in many artificial intelligence tasks. The approach is able to extract efficient representations from the input data, if the data are large enough. Unfortunately, it is not always possible to collect large and quality data. For tasks in low-resource contexts, such as the Russian ⟶ Vietnamese machine translation, insights into the data can compensate for their humble size. In this study of modelling Russian ⟶ Vietnamese translation, we leverage the input Russian words by decomposing them into not only features but also subfeatures. First, we break down a Russian word into a set of linguistic features: part-of-speech, morphology, dependency labels, and lemma. Second, the lemma feature is further divided into subfeatures labelled with tags corresponding to their positions in the lemma. Being consistent with the source side, Vietnamese target sentences are represented as sequences of subtokens. Sublemma-based neural machine translation proves itself in our experiments on Russian-Vietnamese bilingual data collected from TED talks. Experiment results reveal that the proposed model outperforms the best available Russian ⟶ Vietnamese model by 0.97 BLEU. In addition, automatic machine judgment on the experiment results is verified by human judgment. The proposed sublemma-based model provides an alternative to existing models when we build translation systems from an inflectionally rich language, such as Russian, Czech, or Bulgarian, in low-resource contexts.

1. Introduction

Many neural models have been introduced for machine translation [1–5]. Although they have different architectures, they all follow the sequence-to-sequence pattern. Source sentences are represented as sequences of source units. The source sequences are processed by the neural models; then the models generate corresponding sequences of target units. The target sequences are then concatenated to form target sentences. The most intuitive representation of source/target units is words. If the bilingual datasets used to train neural machine translation (NMT) models are large enough, the models will be able to learn reliable statistics of source/target words. Unfortunately, in practice, there are many cases of scarce data, such as Russian ⟶ Vietnamese translation tasks. The language pair is of low resource. Moreover, Russian is a highly inflectional language. A word can have different forms according to its grammatical role in sentences. The property leads to a high chance that we will meet word forms which do not occur frequently enough in humble-size training datasets.

The first attempt to solve the scarce data problem for Russian ⟶ Vietnamese translation tasks can be found in the work of Nguyen et al. [6]. The authors use a mixed-level representation system, where Russian source units are subwords, and Vietnamese target units are words. Due to the division of Russian words, rare words are replaced by more common subwords; therefore, the severity of the rare-word problem is reduced. Another solution to the scarce data problem for Russian ⟶ Vietnamese translation tasks is proposed by Nguyen et al. [7]. They decompose a Russian source word into a set of linguistic features: part-of-speech, morphology, dependency label, and lemma.

We have assessed the available approaches of unit representation on a Russian-Vietnamese bilingual data collected from TED talks [8]. Experiment results show that

the decomposition approach significantly outperforms mixed-level representation. Nevertheless, we still believe in the effectiveness of subword representation, which has become a default part of many NMT models [9–13]. Therefore, we experiment combining source-word decomposition and subword representation. Specifically, we perform a two-step procedure. First, we decompose a Russian source word into a set of features as in source-word decomposition approach. After that, we continue to divide lemmas into sublemmas using BPE algorithm [14]. Since many Russian lemmas are derived from the same root with different prefixes or suffixes, it makes sense to divide them into smaller parts. For example, the verbs "Цричодить" (to arrive), "вчодить" (to enter), "Црочодить" (to go by), "Цодчодить" (to approach), "вычодить" (to leave), "дочодить" (to reach), and "учодить" (to leave) have the same root part "чодить" (to go), with a prefix added to modify their meaning. Sometimes, both prefixes and suffixes modify the same root to create different lemmas; for example, the verbs "являть" (show), "Цоявлять" (appear), "являться" (to be shown), and "Цоявляться" (to be appeared) have the same root "являть."

In total, we propose a sublemma-based NMT model for Russian ⟶ Vietnamese translation. On the Russian source side, we represent a translation unit as a combination of part-of-speech tag, morphology, dependency labels, and a list of sublemmas with their corresponding tags informing that a sublemma is the beginning, middle, or final part of a lemma. On the Vietnamese target side, we tokenize sentences into sequences of subtokens with BPE algorithm. A token is a sequence of characters delimited by space. In Russian, a token is a word. In Vietnamese, there are few cases when a token is a word. Usually, a Vietnamese token is a syllable. In this work, we use the term "subtoken" to indicate a part of a token regardless of whether it is a word or syllable.

This work is composed of six sections. This first section introduces our study. The second section reviews related works. The third section describes our proposed sublemma-based NMT model revised from the state-of-the-art Transformer NMT model. The fourth section describes materials and methods. The fifth section presents the experiment results and analysis. Conclusions from this work are given in the final section.

2. Related Works

In this section, we briefly describe the approaches of translation unit representation in NMT models which influence our study.

While the use of linguistic features as part of a translation unit is widespread in traditional factored statistical models [15–18], it is only recently that Sennrich and Haddow [19] has applied it in a modern deep model. The authors complement a source word with its features. As a result, they represent a source unit as a combination of a source word and its linguistic input features. Their approach performs well for English ↔ German and English ⟶ Romanian translation tasks. For their Russian ⟶ Vietnamese translation system, Nguyen et al. [7] made a step further by removing source words in the list of features. They represented a source translation unit as a combination of linguistic features: part-of-speech, morphology tag, dependency label, and lemma. On the target Vietnamese side, they simply used words as translation units. Their NMT model with source-word decomposition outperformed baseline NMT models including the one by Sennrich and Haddow [19]. Their source-word decomposition is the first processing step in our two-stage procedure to represent a source translation unit.

To handle the rare-word problem, Kudo and Richardson [20] created a language-independent word segmentation algorithm, SentencePiece, to divide words into subwords. Their work comes from an intuition that smaller units of rare words, such as compounds, are easier to translate. They demonstrated the quality of their algorithm in an English ⟶ Japanese translation task. As in the work of Kudo and Richarson [20], Sennrich et al. [14] adapted byte pair encoding (BPE) algorithm originally used for compression to divide words into subwords. First, they considered characters as translation units. Considering words as sequences of translation units, they merged their frequent pairs to form new translation units. They repeated the merging process for a predefined number of times. Clearly, their approach is also language-independent. They reported improvements in translation quality for English ⟶ German and English ⟶ Russian translation tasks. In this work, we actually apply BPE algorithm for representing source translation units. Instead of word segmentation in the original work, we use the algorithm to divide lemmas into sublemmas, since we have already decomposed Russian words into features including the lemma in the first place.

Being language-independent tools, BPE and SentencePiece algorithms are really popular, since they can operate for all languages. However, these wonderful tools should not be utilized blindly. In a Russian ⟶ Vietnamese news translation task, Nguyen et al. [6] showed that an NMT model with mixed-level representation outperformed a baseline NMT model where BPE algorithm was applied on both translation sides. Influenced by a work on a traditional statistical machine translation model for Chinese ⟶ Vietnamese [21], the authors only applied BPE algorithm on Russian source side, while using words on Vietnamese target side, considering the different effects of BPE algorithm on each side of their bilingual corpus. Although their approach is interesting, it fails to take into account rare foreign named entities, which are commonly found in Vietnamese texts translated from a foreign language. Since our bilingual corpus contains many foreign named entities on both sides and we already apply BPE algorithm on the source side, we opt to use BPE method to tokenize Vietnamese target sentences into sequences of subtokens.

2.1. Sublemma-Based Transformer Model. Following the recommendation of Nguyen et al. [22], our sublemma-based NMT model is based on the state-of-the-art model Transformer [4]. The proposed model has a similar architecture

except for the embedding layer of the encoder of Transformer. In this section, we describe the source and target translation unit representation and the encoder of Transformer model which is revised to adopt the proposed translation unit representation.

2.2. Translation Unit Representation

2.2.1. Sublemma-Based Representation of Source Translation Units. We represent a source translation unit as a combination of sublemma-based features, following a two-step procedure.

In the first step, we transform a Russian source sentence into a sequence of linguistic features: part-of-speech (POS), morphology (MOR), dependency label (DEP), and lemma (LEM), following source-word decomposition approach [7]. The grammatical parsing is performed with the help of a natural language processing toolkit, Stanza [23]. Typical part-of-speech tags of Russian words are shown in Universal Dependencies treebank [24], such as nouns, pronouns, verbs, auxiliary, numerals, particles, determiners, adjective, and adverbs. Russian has a rich morphology. A Russian word is inflected from an original lemma, depending on its part-of-speech and grammatical role in sentence. A word's grammatical role in a sentence is denoted with a dependency label [25]. An example of a short Russian sentence being transformed into a sequence of linguistic features is presented in Table 1.

In the second step, we apply BPE method, segmenting lemmas into sublemmas. After the segmentation, the sequence of sublemmas is longer than its corresponding sequences of other features. Following the work of Sennrich and Haddow [19], we broadcast the sequences of other features, so that they have the same length as the sequence of sublemmas. Specifically, all sublemmas extracted from a lemma will have the same labels of features corresponding to the lemma. Moreover, using their subword notation, we assign a tag to each sublemma (TAG), depending on the position of the sublemma relative to the initial lemma. A sublemma can be the beginning (B), inside (I), ending (E), or the full lemma (O). In addition, the beginning and inside sublemmas are suffixed with characters "@@" to inform their roles. An example of sublemma-based sequences of linguistic features is shown in Table 2.

In total, we represent a Russian source sentence as a sequence of collections of sublemma-based features: sublemma, sublemma tag, part-of-speech tag, morphology label, and dependency label. Each source translation unit is represented as a collection of its features.

2.2.2. Target Translation Unit Representation. Applying BPE algorithm [14], we segment Vietnamese target sentences into sequences of subtokens. The algorithm appends characters "@@" to the beginning and inside subtokens for later merge operations. Sequences of target subtokens are used to train the translation model. Generated sequences of target subtokens are merged to form target sentences, based on the characters "@@." A Vietnamese sentence and its corresponding sequence of subtokens are shown in Table 3.

In Table 3, we can see that BPE algorithm focuses on tokens which are the foreign named entity "Geographic Society." It segments the entity into a sequence of subtokens "Geo@@ graphic So@@ ci@@ e@@ ty."

2.3. Embedding Layer in the Encoder of Sublemma-Based Transformer Model. As in [7, 19], we consider all features x_{ij} from the i-th source translation unit in a source sequence as strings in their respective domains $x_{ij} \in \mathcal{S}_j$, where $\mathcal{S}_j$, $j = 0, \ldots, 4$, is the set of sublemmas, sublemma tags, part-of-speech tags, morphology labels, and dependency labels, respectively. The trainable embedding $\mathbf{e}_{ij}$ of a feature j is extracted from a corresponding dictionary $f_j\colon x_{ij} \mapsto \mathbf{e}_{ij} \in \mathbb{R}^{d_j \times |\mathcal{S}_j|}$, where d_j is a predefined size of embeddings of the feature j (equation (1)).

$$\mathbf{e}_{ij} = f_j\left(x_{ij}\right), \tag{1}$$

and the embedding of a source translation unit is represented as the concatenation of embeddings of its features (equation (2)).

$$\mathbf{e}_i = \text{concat}\left(\mathbf{e}_{ij}, \quad \text{for } j = 0, \ldots, 4\right). \tag{2}$$

Since Transformer model does not leverage the order of translation units in its core layer, it deploys a positional embedding principle, such as sinusoidal positional embedding $\mathbf{p}_i$ [4]. In total, the i-th source translation unit in a source sequence has the overall embedding computed as in the following equation:

$$\mathbf{o}_i = \sqrt{d} \times \mathbf{e}_i + \mathbf{p}_i, \tag{3}$$

where $d = \sum_{j=0}^{4} d_j$.

3. Materials and Methods

3.1. Materials. To assess NMT models, we used a bilingual Russian-Vietnamese corpus consisting of sentence pairs of length in the range (10 tokens, 30 tokens) extracted from TED talks [8]. The chosen sentences are ended with a punctuation mark and contain only word characters and punctuation. As in [26–28], we randomly divide the corpus into three datasets: training, development, and testing datasets. Specifically, a set of 47750 sentence pairs are randomly selected from the corpus and used as the training dataset. Furthermore, a set of 1500 sentence pairs are selected from the left corpus and used as the development dataset. The remaining 1500 sentence pairs are used as the testing dataset. Statistical summary of the datasets is presented in Table 4.

In Table 4, we use the term "token" to denote a sequence of characters delimited by space. Linguistically, it can be a Russian word, a Vietnamese syllable, or a punctuation.

4. Methods

We compared the proposed sublemma-based Transformer model with three baseline Transformer models. These

Table 1: A short Russian sentence with its corresponding sequence of linguistic features.

Words	POS	MOR	DEP	LEM
Коґда	SCONJ	_	Mark	коґда
вода	NOUN	Animacy = Inan, Case = Nom, Gender = Fem, Number = Sing	Nsubj	вода
Ноднимается	VERB	Aspect = Imp, Mood = Ind, Number = Sing, Person = 3, Tense = Pres, VerbForm = Fin, Voice = Mid	Advcl	Нодниматься
,	PUNCT	_	Punct	,
Ноτом	ADV	Degree = Pos	Advmod	Ноτом
отстуНает	VERB	Aspect = Imp, Mood = Ind, Number = Sing, Person = 3, Tense = Pres, VerbForm = Fin, Voice = Act	Root	отстуНать
,	PUNCT	_	Punct	,
начодишь	VERB	Aspect = Imp, Mood = Ind, Number = Sing, Person = 2, Tense = Pres, VerbForm = Fin, Voice = Act	Conj	начодить
в	ADP	_	Case	в
нем	PRON	Case = Loc, Gender = Masc, Number = Sing, Person = 3	Obl	он
новые	ADJ	Animacy = Inan, Case = Acc, Degree = Pos, Number = Plur	Amod	новый
ракушки	NOUN	Animacy = Inan, Case = Acc, Gender = Fem, Number = Plur	Obj	ракушка
.	PUNCT	_	Punct	.

Table 2: Sublemma-based sequences of linguistic features

Sublemmas	TAG	POS	MOR	DEP	From lemma
коґда	O	SCONJ	_	Mark	коґда
вода	O	NOUN	Animacy = Inan, Case = Nom, Gender = Fem, Number = Sing	Nsubj	вода
Нодниматься	O	VERB	Aspect = Imp, Mood = Ind, Number = Sing, Person = 3, Tense = Pres, VerbForm = Fin, Voice = Mid	Advcl	Нодниматься
,	O	PUNCT	_	Punct	,
Ноτом	O	ADV	Degree = Pos	Advmod	Ноτом
от@@	**B**	**VERB**	**Aspect = Imp, Mood = Ind, Number = Sing, Person = 3, Tense = Pres, VerbForm = Fin, Voice = Act**	**Root**	отстуНать
стуНать	**E**	**VERB**	**Aspect = Imp, Mood = Ind, Number = Sing, Person = 3, Tense = Pres, VerbForm = Fin, Voice = Act**	**Root**	отстуНать
,	O	PUNCT	_	Punct	,
начодить	O	VERB	Aspect = Imp, Mood = Ind, Number = Sing, Person = 2, Tense = Pres, VerbForm = Fin, Voice = Act	Conj	начодить
в	O	ADP	_	Case	в
он	O	PRON	Case = Loc, Gender = Masc, Number = Sing, Person = 3	Obl	он
новый	O	ADJ	Animacy = Inan, Case = Acc, Degree = Pos, Number = Plur	Amod	новый
ра@@	**B**	**NOUN**	**Animacy = Inan, Case = Acc, Gender = Fem, Number = Plur**	**Obj**	ракушка
ку@@	**I**	**NOUN**	**Animacy = Inan, Case = Acc, Gender = Fem, Number = Plur**	**Obj**	ракушка
шка	**E**	**NOUN**	**Animacy = Inan, Case = Acc, Gender = Fem, Number = Plur**	**Obj**	ракушка
.	O	PUNCT	_	Punct	.

Table 3: A Vietnamese sentence and its corresponding sequence of subtokens.

Vietnamese sentence	"Vì vậy tôi bắt đầu làm việc với tạp chí National Geographic Society cùng các báo khác và dẫn các cuộc thám hiểm tới Nam Cực."
Sequence of subtokens	"Vì vậy tôi bắt đầu làm việc với tạp chí National Geo@@ graphic so@@ ci@@ e@@ ty cùng các báo khác và dẫn các cuộc thám hiểm tới nam Cực."

Table 4: Statistical summary of the datasets.

Russian/Vietnamese	Training	Development	Testing
Average sentence length	16.1/18.1	16.2/21.2	16.2/21.3
Unique tokens	73205/25939	7202/2646	7120/2692
All tokens	766446/866175	24257/31741	24363/31948

models are the foundations from which our model is derived. The first baseline model is mixed-level Transformer model [6]. The second baseline model is a subtoken-based Transformer model [14]. The third baseline model is Transformer model with source-word decomposition [7]. We create all models with an open-source library, OpenNMT-tf [29, 30]. The architecture and hyperparameters of the baseline models can be found in the

respective works. Here, we only describe how we build our proposed model.

As reported in the description of the proposed model, we use Stanza natural language processing tool [23] to decompose Russian words into sets of features. Then, we use BPE algorithm [14] with 10,000 merge operations to divide lemmas into sublemmas. We also use the algorithm to divide Vietnamese target sentences into sequences of subtokens. The number of items in each feature domain is presented in Table 5.

We apply the sizes of 179, 11, 22, 22, and 22 for embeddings of sublemmas, sublemma tags, part-of-speech tags, morphology labels, and dependency labels, respectively. In total, we use 256 dimensions for concatenated embeddings of source translation units.

On the Vietnamese target side, we also use 256 dimensions to represent the embeddings of target units.

In addition to embedding layers, the proposed sublemma-based Transformer model consists of 6 hidden layers. The hidden layers contain 8-head attention sublayers and feedforward neural networks of 512 dimensions. Hidden states of the model are comprised of 256 values. To prevent the overfitting problem, we apply a dropout of 0.1 in all hidden layers. To generate translations, the model contains an inference module implementing a beam search algorithm with beam width = 5 [31].

For all models, the training procedure is as follows:

(1) First, we train the model in 15,000 steps. In each training step, we use 64 sentence pairs from the training dataset to optimize the cross-entropy criterion described in the work of Muller et al. [32]. While there are many efficient algorithms for optimization, we choose to apply LazyAdam optimizer [33], as it is available in the chosen OpenNMT-tf library. We employ the optimizer with $\beta_1 = 0.9$ and $\beta_1 = 0.998$ and learning rate = 2.

(2) Second, we save the values of the model parameters when we complete $n \times 10^3$ training steps, where $n \in \{11, \ldots, 15\}$. We use the development dataset to validate the translation quality of all candidate values. The values giving the best translation quality in the development dataset are chosen for the model parameters.

We validate translation quality of the models with the BLEU score [34]. The BLEU scores are computed with the script multi-bleu.perl [35]. BLEU is the abbreviation of "bilingual evaluation understudy," measuring the similarity of candidate translations to their corresponding references. It is the geometric mean of constituent n-gram scores, where $n = 1, \ldots, 4$. All n-grams are extracted from the candidate translations. While unigrams are individual words, bigrams, trigrams, and four grams are phrases of two, three, and four neighboring words, respectively. We calculate a constituent n-gram score by dividing the number of the n-grams appearances in the references by the total number of the n-grams in the candidate translations.

Table 5: Size of vocabulary in sublemma-based Transformer model.

Language side	Vocabulary	Size
Source	Sublemmas	9417
Source	Sublemma tags	4
Source	Part-of-speech tags	15
Source	Morphology labels	484
Source	Dependency labels	38
Target	Subtokens	8628

After training the models, we assess their translation quality using the testing dataset. To have a complete assessment, we employ not only the automatic BLEU scores but also limited human judgment on translation results. We accompany the BLEU score with human judgment, since it has an obvious pitfall. It only measures total matching of n-grams in the candidate translations and the references regardless of their meaning. To solve the problem, we compare the meanings of the candidate translations and their references, considering synonyms, as well as the similarity of meanings. We do this for all levels, from individual words to phrases and complete sentences.

5. Results and Analysis

BLEU scores of the comparative Transformer models are shown in Figure 1.

Among the baseline models, the model with source-word decomposition provides the best scores of 13.52 and 13.84 BLEU in the development and testing datasets, respectively. Fortunately, our proposed sublemma-based Transformer model outperforms the best baseline model in both development and testing datasets, delivering improved BLEU scores of 14.46 and 14.81, respectively. The improvements of 0.94 and 0.97 BLEU are recorded.

The performance order of the models for the development dataset is maintained for the testing dataset: mixed-level model < subtoken-based model < model with source-word decomposition < the proposed sublemma-based model. This consistency makes us more confident about the effectiveness of our proposed sublemma-based model.

In addition to machine judgment with automatic BLEU scores, we semantically studied a limit number of translation results by the two best models: the model with source-word decomposition (from now on, we call it "baseline" model) and the proposed sublemma-based model (from now on, we call it "proposed" model). Five cases in the testing dataset were randomly chosen and studied.

Table 6 shows the source, its meaning, the target, and the predicted sentences by the baseline and proposed models in the first case. The first case seems easy, since both models provide correct translations. Although the models literally choose words different from the reference, the meanings are the same. For example, the verb phrase "phủ nhận" (negate) by the models is similar in meaning to the reference "chối bỏ" (deny).

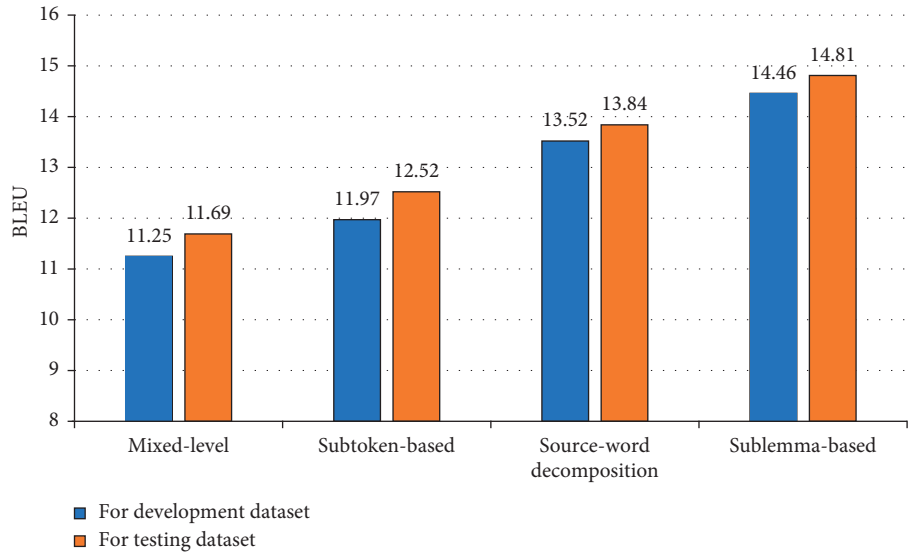

Figure 1: BLEU scores of comparative Transformer models.

Table 6: First case study with correct translations.

Source	“ Мы не можем отрицать факт тоҐо, что все Цотребление Цищи человечеством имеет Ґлобальные Цоследствия.”
Meaning	“We cannot deny the fact that all human consumption of food has global implications.”
Reference	“Không có cách gì để chối bỏ sự thật là những gì chúng ta ăn có ảnh hưởng đến toàn cầu.”
Baseline model	“Chúng ta không thể phủ nhận thực tế rằng tiêu thụ lương thực của loài người có tác động toàn cầu.”
Proposed model	“Chúng ta không thể phủ nhận rằng mọi thứ tiêu thụ thực phẩm trên toàn thế giới đều có hậu quả toàn cầu.”

Table 7 presents the second case study. The general meaning of the reference is found in the translations by the two models, except for a keyword “Phượng hoàng” (Phoenix). The corresponding source named entity “Феникс” (Phoenix) is a rare word; hence both models fail to translate the named entity. Nevertheless, the proposed model semantically performs better than the baseline model in this case. Although the phrase “các hòn đảo” (islands) by the proposed model and the phrase “hòn đảo”(island) by the baseline model are distinct from the reference “Quần đảo” (Archipelago), we think that the former translation is conceptually closer to the reference than the latter translation.

Table 8 demonstrates the third case study. Although the translations by the models contain many reference words, their meanings are not accurate. The key source phrase “с остальным миром” (with the rest of the world) is incorrectly translated into the phrase “với thế giới ngoài không gian” (with the world in outer space) and the phrase “với một thế giới khác” (with another world) by the baseline and proposed models, respectively. Comparing the models with each other, we think that the proposed model is better than the baseline model in this case. The phrase “trong không gian” (in space) by the proposed model better reflects meaning of the source “в Цространственном смысле” (in a spatial sense) than the phrase “ngoài không gian” (in outer space) by the baseline model.

Table 9 shows the fourth case study. Although the proposed model does not generate a translation completely reflecting the meaning of the source, it outshines the baseline model. It even successfully translates the rare named entity “Дубаи” (Dubai). At the same time, the baseline model completely fails in this case with an incorrect translation which contains unknown words <unk>.

Table 10 displays the fifth case study. This case again proves the power of the proposed model in translating rare words. It successfully translates the rare source word “биоразнообразия” (biodiversity) into the phrase “sự đa dạng sinh học” as in the reference. The rare word is a keyword in the source sentence. Due to the ability to handle rare words, the proposed model finds itself superior to the baseline model. The translation by the proposed model keeps the meaning of the source sentence. On the other hand, the baseline model misses the key source word and hence provides an incomplete translation.

After semantically studying the test cases, we found that the proposed sublemma-based model tends to provide longer and better translations than the best baseline model. The similarity between manual evaluation and automatic assessment consolidates our proposal of using the sublemma-based Transformer model in place of the model with source-word decomposition.

Table 7: Second case study with inaccurate translations.

Source	“ Но вернемся обратно к островам Феникс, которые являются темой сеҐодняшнеҐо выстуШления.”
Meaning	“But back to the Phoenix Islands, which are the topic of today’s talk.”
Reference	“Nhưng hãy quay lại với Quần đảo Phượng hoàng, đó là chủ đề của bài nói chuyện này.”
Baseline model	“Nhưng quay trở lại với hòn đảo, đó là chủ đề của bài thuyết trình hôm nay.”
Proposed model	“Nhưng hãy quay trở lại các hòn đảo Erex, chủ đề của buổi nói chuyện hôm nay.”

Table 8: Third case study with incorrect translations.

Source	“ И давайте сравним её с остальным миром в Цространственном смысле.”
Meaning	“And let’s compare it with the rest of the world in a spatial sense.”
Reference	“Và hãy so sánh nó với phần còn lại của thế giới theo giới hạn không gian.”
Baseline model	“Hãy so sánh nó với thế giới ngoài không gian.”
Proposed model	“Hãy so sánh nó với một thế giới khác trong không gian.”

Table 9: Fourth case study with a rare named entity.

Source	“Я Цереечал в Дубаи на Цост лидера разработки содержания ЦроҐрамм для ЗаЦадной телевизионной сети.”
Meaning	“I moved to Dubai as the content development leader for a Western television network.”
Reference	“Tôi chuyển đến Dubai với vai trò là người chịu trách nhiệm về nội dung cho một đài TV của phương Tây.”
Baseline model	“Tôi chuyển sang <unk> để nghiên cứu về các phần mềm ở <unk>.”
Proposed model	“Tôi chuyển tới gần Dubai, một nhà lãnh đạo những chương trình xây dựng chương trình tại Bờ Tây.”

Table 10: Fifth case study with a rare compound.

Source	“Лти места наиболее боҐаты с точки зрения биоразнообразия и наиболее важны с точки зрения функционирования экосистемы.”
Meaning	“These sites are the richest in terms of biodiversity and the most important in terms of ecosystem functioning.”
Reference	“Đó là những nơi giàu nhất trong đa dạng sinh học và là quan trọng nhất từ quan điểm chức năng hệ sinh thái.”
Baseline model	“Những nơi này giàu về sự đa dạng và quan trọng nhất so với cách hệ gen.”
Proposed model	“Những nơi này rất phong phú với sự đa dạng sinh học và quan trọng nhất với phương diện hoạt động của hệ sinh thái.”

6. Conclusions

In this study, we have proposed a sublemma-based Transformer model for translation from Russian into Vietnamese. It is a derivation from the model with source-word decomposition and models with subword representation. In the proposed model, a source unit is represented as a combination of a sublemma, its tag, part-of-speech tag, dependency label, and morphology label, while a target unit is a subtoken. Experimental results show that our proposed model surpasses all available models for Russian ⟶ Vietnamese translation task. Human judgment on the translation quality of the models has validated the comparison in terms of BLEU score.

Standing on the results of this study, we recommend our sublemma-based Transformer model for translation from a highly inflectional language, such as Russian, Bulgarian, or Czech.

Data Availability

The datasets used in this study are accessible upon request to the corresponding author Thien Nguyen via e-mail: nguyenchithien@tdtu.edu.vn.

Conflicts of Interest

The authors declare that there are no conflicts of interest in this paper.

References

[1] K. Cho, B. Merrienboer, C. Gulcehre et al., “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 1724–1734, Doha, Qatar, October 2014.

[2] M.-T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, September 2015.

[3] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, “Convolutional sequence to sequence learning,” in *Proceedings of the International Conference on Machine Learning*, pp. 1243–1252, Ho Chi Minh City, Vietnam, January 2017.

[4] A. Vaswani, N. Shazeer, N. Parmar et al., “Attention is all you need,” in *Proceedings of the Advances in neural information*

processing systems, pp. 5998–6008, Long Beach, CA, USA, December 2017.

[5] S. Garg, S. Peitz, U. Nallasamy, and M. Paulik, "Jointly learning to align and translate with transformer models," in *Proceedings of the EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, pp. 4453–4462, Hong Kong, China, November 2020.

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

[7] 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.

[8] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4512–4525, November 2020.

[9] S. Ding, A. Renduchintala, and K. Duh, "A call for prudent choice of subword merge operations in neural machine translation," in *Proceedings of the Machine Translation Summit XVII*, pp. 204–213, Dublin, Ireland, August 2019.

[10] Y. Wu and H. Zhao, "Finding better subword segmentation for neural machine translation," in *Proceedings of the Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pp. 53–64, Springer, Changsha, China, October 2018.

[11] C. Wang, K. Cho, and J. Gu, "Neural machine translation with byte-level subwords," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9154–9160, New York, NY, USA, February 2020.

[12] M. Pinnis, R. Krišlauks, D. Deksne, and T. Miks, "Neural machine translation for morphologically rich languages with improved sub-word units and synthetic data," in *Proceedings of the International Conference on Text, Speech, and Dialogue*, pp. 237–245, Prague, Czech Republic, August 2017.

[13] H. Deguchi, M. Utiyama, A. Tamura, T. Ninomiya, and E. Sumita, "Bilingual subword segmentation for neural machine translation," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4287–4297, Barcelona, Spain, September 2020.

[14] 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.

[15] S. Huet, E. Manishina, and F. Lefèvre, "Factored machine translation systems for Russian-English," 2013.

[16] A. Birch, M. Osborne, and P. Koehn, "CCG supertags in factored statistical machine translation," in *Proceedings of the second workshop on Statistical Machine Translation*, pp. 9–16, Prague, Czech Republic, June 2007.

[17] P. Koehn and H. Hoang, "Factored translation models," in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 868–876, Prague, Czech Republic, June 2007.

[18] Y. Wang, L. Wang, X. Zeng, D. F. Wong, L. S. Chao, and Y. Lu, "Factored statistical machine translation for grammatical error correction," in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 83–90, Baltimore, MD, USA, June 2014.

[19] R. Sennrich and B. Haddow, "Linguistic input features improve neural machine translation," 2016, http://arxiv.org/abs/1606.02892.

[20] T. Kudo and J. Richardson, "SentencePiece: a simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium, November 2018.

[21] P. Tran, D. Dinh, and H. T. Nguyen, "A character level based and word level based approach for Chinese-Vietnamese machine translation," *Computational Intelligence and Neuroscience*, vol. 2016, Article ID 9821608, 2016.

[22] T. Nguyen, H. Nguyen, and P. Tran, "Exploring neural machine translation on the Russian-Vietnamese language pair," in *Proceedings of the Advances in Intelligent Information Hiding and Multimedia Signal Processing*, pp. 393–400, Sendai, Japan, June 2021.

[23] 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.

[24] J. Nivre, M.-C. de Marneffe, F. Ginter et al., "Universal dependencies v1: A multilingual treebank collection," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1659–1666, Portorož, Slovenia, May 2016.

[25] M.-C. De Marneffe, T. Dozat, N. Silveira et al., "Universal Stanford dependencies: A cross-linguistic typology," *LREC*, vol. 14, pp. 4585–4592, 2014.

[26] P. Tran, D. Dinh, and L. H. B. Nguyen, "Word re-segmentation in Chinese-Vietnamese machine translation," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 16, no. 2, pp. 1–22, 2016.

[27] P. Tran, D. Dinh, T. Le, and L. H. B. Nguyen, "Linguistic-relationships-based approach for improving word alignment," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 17, no. 1, pp. 1–16, 2017.

[28] T. Nguyen, L. Nguyen, P. Tran, and H. Nguyen, "Improving transformer-based neural machine translation with prior alignments," *Complexity*, vol. 2021, 2021.

[29] G. Klein, Y. Kim, Y. Deng, V. Nguyen, J. Senellart, and A. M. Rush, "OpenNMT: 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.

[30] 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, Orlando, FL, USA, October 2020.

[31] M. Freitag and Y. Al-Onaizan, "Beam search strategies for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*, pp. 56–60, Melbourne, Australia, July 2017.

[32] 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, NeurIPS 2019*, pp. 4696–4705, December 2019, https://proceedings.neurips.cc/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html.

[33] D. P. Kingma and J. Ba, "Adam: {A} method for stochastic optimization," in *Proceedings of the 3rd International*

Conference on Learning Representations, {ICLR} 2015, San Diego, CA, USA, May 2015, http://arxiv.org/abs/1412.6980.

[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, Stroudsburg, PA, USA, July 2002.

[35] P. Koehn, H. Hoang, A. Birch, and C. Callison-Burch, "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.