Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 4795187, 10 pages
https://doi.org/10.1155/2020/4795187

Research Article

Source-Word Decomposition for Neural Machine Translation

Thien Nguyen, Hoai Le, and Van-Huy Pham

Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam

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

Received 20 August 2020; Revised 1 November 2020; Accepted 3 December 2020; Published 16 December 2020

Academic Editor: Yakov Strelniker

End-to-end neural machine translation does not require us to have specialized knowledge of investigated language pairs in building an effective system. On the other hand, feature engineering proves to be vital in other artificial intelligence fields, such as speech recognition and computer vision. Inspired by works in those fields, in this paper, we propose a novel feature-based translation model by modifying the state-of-the-art transformer model. Specifically, the encoder of the modified transformer model takes input combinations of linguistic features comprising of lemma, dependency label, part-of-speech tag, and morphological label instead of source words. The experiment results for the Russian-Vietnamese language pair show that the proposed feature-based transformer model improves over the strongest baseline transformer translation model by impressive 4.83 BLEU. In addition, experiment analysis reveals that human judgment on the translation results strongly confirms machine judgment. Our model could be useful in building translation systems translating from a highly inflectional language into a noninflectional language.

1. Introduction

Neural machine translation (NMT) is an active research field with a lot of newly published works [1–4]. They study different aspects of NMT in order to improve it. In [1], the authors proposed a single model to translate from multiple source languages to multiple target languages. In [2], the author proposed using adversarial input to train the model. Adversarial input is generated from the original input with a small perturbation. In [3], the authors proposed a mechanism to adapt NMT models to new languages and domains. In [4], the authors proposed enriching the training dataset with the predicted sentences. Although these works develop in different directions, all of them are based on end-to-end NMT. End-to-end NMT is a universally applicable translation paradigm. It is complicated from the technical point of view, but very simple from the point of view of linguistics. In contrast to building an effective statistical machine translation system, building a NMT system does not require specialized knowledge of the applied language pair. For all language pairs, regardless of their characteristics, end-to-end NMT takes a sequence of source words from a fixed dictionary, processes, and then returns a sequence of target words from another fixed dictionary. The effectiveness and simplicity in application lead to widespread use of NMT [5–7]. NMT has become a dominant translation paradigm. In ideal circumstances where all words of languages frequently appear in a large training dataset and computational capacity to train translation models is unlimited, end-to-end NMT will work perfectly. In practice, such ideal circumstances do not take place, so the performance of end-to-end NMT is worsened by rare words and out-of-vocabulary problem, which takes place in all translation tasks, especially for low-resource language pairs. To make NMT capable of translating rare words and out-of-vocabulary words, in [8–10], the authors proposed novel NMT models representing words as sequences of subwords, which occur more frequently than the words themselves. For example, according to the byte pair decoding method [8], the Russian word "Цризывают" (meaning: call) is segmented into two subwords "Цризыва@@" and "ют." In the subword-based NMT source, sentences as sequences of source subwords are processed, and then, sequences of target subwords are generated. Generated sequences of target subwords are concatenated to form target sentences, based on characters "@@" informing that containing subwords should be

attached to the following subwords. Their experiment results showed that subword-based NMT delivered a substantial improvement, compared with word-based NMT for high-resource English-German and English-Russian language pairs. The success of subword representation in NMT was further confirmed in the studies [11, 12] for several high-resource language pairs, such as English-Spanish, English-French, and English-Chinese. Recently, a revolutionary NMT model called transformer [13, 14] with the self-attention mechanism significantly outperformed the best previously reported translation models. In cooperation with subword representation, transformer has established itself as the state-of-the-art translation paradigm.

Although subword-based NMT models are able to work well without considering linguistic characteristics of languages, we wonder whether linguistic knowledge helps NMT systems to work more efficiently. Motivated by works [15, 16] in speech recognition with a similar sequence-to-sequence pattern where an original raw input data in the form of discrete speech signal overtime can be represented as a sequence of features, such as log-Mel frequencies, Mel frequency cepstrum, and the knowledge of morphological rich Russian source language and analytic Vietnamese target language, in this work of building a Russian-Vietnamese machine translation system, we experiment the idea of representing each source word in sentence as a combination of features: lemma, grammatical role in sentence, part-of-speech, and morphological features. The decomposition is only deployed on the Russian source side, but not in the Vietnamese target side. The idea comes to mind, since Russian is a morphological rich language, while Vietnamese is an analytic language which lacks morphological marking of case, gender, number, and tense. A Russian sentence consists of tokens infected from lemmas based on their grammatical roles and part-of-speech tags. Inflected tokens are usually called words. By replacing words in the source sentence by a combination of features, we actually increase their appearance frequency in the training dataset; therefore, reduce the severity of rare word and out-of-vocabulary problem in inference. For example, in the training dataset, we have two Russian words, "Пришёл" (meaning: arrived) and "Полюблю" (meaning: fall in love), and in the testing dataset, we have two other words "Приду" (meaning: will arrive) and "Полюбил" (meaning: fell in love). In this case, end-to-end NMT systems are going to recognize two words in the testing dataset as unknown. However, there are close linguistic relationships between words in the training and testing datasets. Both Russian words "Пришёл" and "Приду" are inflected forms of the same lemma "Прийти." The training Russian word "Пришёл" is inflected from the lemma "Прийти," since it is a verb of muscular gender, in singular number and in the past tense, while the testing Russian word "Приду" is inflected from the lemma "Прийти," as it is a verb in singular number and in the future tense. A relationship is also found for two words "Полюблю" and "Полюбил." The training Russian word "Полюблю" is inflected from the lemma "Полюбить," since it is a verb in singular number and in the future tense, while the testing Russian word "Полюбил" is inflected from the lemma "Полюбить," as it is a verb of muscular gender in singular number and in the past tense. If we decompose all these words into features, then in the testing phase, we will have lemmas and grammatical features which are well-known in regard to the training dataset.

In total, this work is dedicated to building a novel transformer-based NMT model taking a sequence of vectors of linguistic features from source words and predicting a sequence of target words.

The rest of this paper is organized as follows. A brief overview of related works is given in the following section. The third section outlines a novel methodology of source-word decomposition for neural machine translation. The fourth section describes materials and methods used in the work. The fifth section shows and analyses experiment results. The final section lists our conclusions from this work.

2. Related Works

This section briefly examines a variety of translation unit representation methodologies used in machine translation systems for several language pairs containing at least one inflectional language, such as Russian, Czech, German, and English.

Before the emergence of NMT, phrase-based SMT used to be very popular. There is a wide range of literature studying phrase-based SMT. Among these studies, factored phrase-based SMT models [17–19] are the most worthy to mention in our work. In factored phrase-based SMT linguistic features, such as lemma, part-of-speech and morphological features are integrating into the surface form of word. The factored phrase-based SMT systems improve translation quality over standard SMT systems for multiple language pairs, such as English-German, English-Spanish, and English-Czech. The approach gained further popularity, after it had been implemented in the famous SMT tool called Moses [20]. In [21, 22], the authors continued to apply and develop the approach and achieved good results. Although the approach belongs to a group of obsolete statistical translation paradigms, its success in integrating linguistic features inspires us to take advantage of linguistic information in redefining the translation unit in modern NMT paradigm.

In recent years, there has been growing interest in integrating linguistic features into NMT architectures. In [23], the authors proposed a novel factored subword-based neural model based on recurrent neural networks that learns source translation unit embeddings, leveraging subword embedding, subword-tag embedding, lemma embedding, part-of-speech embedding, dependency label embedding, and morphological label embedding. They used many different linguistic features in addition to subword itself to take advantage of high-resource characteristic of the English-German language pair. They found that the factored subword-based neural model notably improved translation for the high-resource English-German language pair. Our preliminary experiments with the state-of-the-art transformer NMT model confirm their finding. Our subword-based transformer model combining linguistic-feature

embeddings with subword embedding outperforms a standard subword-based transformer model for the Russian-Vietnamese language pair. However, we believe that we can further improve the system, considering our context of the low-resource and linguistically distant language pair. Due to totally different morphology of Russian and Vietnamese, in the training dataset, the number of unique words of highly inflectional Russian in the source side is multiple times larger than the number of unique words of noninflectional Vietnamese in the target side, which leads to a great probability that a Russian word in reference phase is unseen in the training dataset. To solve the problem, instead of integration, we suggest to use replacement for training from Russian into Vietnamese. Specifically, we calculate source translation unit embedding using only linguistic-feature embeddings without word or subword embedding.

Word representation in [24] bears a close resemblance to our translation unit representation. The authors used a combination of lemma and part-of-speech tag to represent a word in translation for multiple language pairs: English-German, English-Turkish, English-Czech, and English-Latvian. The main difference from our technique lies in the side of translation. They applied their technique in the target side of a NMT system, while we redefine the source-side translation unit. According to their method, a source word is translated into a vector of lemma and part-of-speech tag. Based on that vector, the system predicts a surface form of target word. Obviously, their approach is geared towards translation into an inflectional language. In our case of translation from Russian into noninflectional Vietnamese, we could not take advantage of their technique.

Recently, in [25], the authors introduced an approach of modeling word formation in transformer-based NMT for the English-German language pair. They segmented both English source words and German target words as sequences of vectors of subwords and linguistic subword tags. They reported an improvement over a standard system. Unfortunately, their approach is language-specific, since they deployed a morphological analyzer for English-German language pairs only. Although their approach is interesting, it is not applicable for our task, as we have not found any similar subword taggers for our Russian-Vietnamese language pairs.

3. Source-Word Decomposition for Neural Machine Translation

3.1. Base Transformer Model. Our feature-based transformer model is based on the original transformer model [13], which is the state-of-the-art NMT model. The transformer model has the encoder-decoder architecture. In this work, we make a novelty by modifying the embedding representation in the encoder; therefore, in the following, we describe that part of the encoder in more detail. If you are interested in the general architecture of the transformer model, you can read the original paper [13].

The encoder of the model maps a sequence $\mathbf{x} = \{x_i, \text{for } i = 1, \ldots, n\}$ of n source words x_i from a fixed dictionary $x_i \in \Psi_x$ into a sequence $\mathbf{C} = \{\mathbf{c}_i, \text{for } i = 1, \ldots, n\}$ of individual embeddings of a fixed size $\mathbf{c}_i \in \mathbb{R}^d$. The process of mapping is as follows. First, the encoder looks up each source word x_i in a dictionary of embeddings and retrieves its embedding vector $\mathbf{e}_{x_i}$ of a fixed size $\mathbf{e}_{x_i} \in \mathbb{R}^d$. Next, the encoder looks up the position i of x_i in another dictionary of positional embeddings and retrieves a positional embedding $\mathbf{e}_i$, where $\mathbf{e}_i \in \mathbb{R}^d$. Finally, the encoder adds embedding $\mathbf{e}_i$ with $\mathbf{e}_{x_i}$ weighted by a factor $\sqrt{d}$. Applying the mapping process for all source words, the encoder generates a sequence $\mathbf{C}$ of combined embeddings $\mathbf{c}_i$.

Considering the relationship between words in the sentence with the self-attention mechanism, the encoder transforms the sequence of individual embeddings into a sequence $\mathbf{Z} = \{\mathbf{z}_i, \text{for } i = 1, \ldots, n\}$ of context-aware continuous representations $\mathbf{z}_i \in \mathbb{R}^d$.

Based on the sequence $\mathbf{Z}$, the decoder of the model generates a sequence $\mathbf{y} = \{y_i, \text{for } i = 1, \ldots, k\}$ of target words y_i from another fixed dictionary $y_i \in \Psi_y$. The decoder generates one target word at a time, using previously generated target words as additional input. Mathematically, the transformer model can be represented as a composition of functions as follows.

$$\begin{aligned}
\mathbf{e}_{x_i} &= \phi_x[x_i], \quad \text{for } i = 1, \ldots, n,\\
\mathbf{e}_i &= \phi_i[i], \quad \text{for } i = 1, \ldots, n,\\
\mathbf{c}_i &= \sqrt{d} \times \mathbf{e}_{x_i} + \mathbf{e}_i,\\
\mathbf{C} &= \{\mathbf{c}_i, \quad \text{for } i = 0, \ldots, n\},\\
\mathbf{Z} &= g(\mathbf{C}),\\
y_i &= f(\mathbf{Z}, y_1, y_2, \ldots, y_{i-1}), \quad \text{for } i = 1, \ldots, k.
\end{aligned} \tag{1}$$

In the above equations, notations g and f stand for trainable functions, while notations ϕ_x and ϕ_i are the dictionaries of trainable embeddings with dimension d.

3.2. Source-Word Decomposition. Unlike the basic transformer model, which take source words x_i from a fixed dictionary $x_i \in \Psi_x$ as input, our proposed transformer model takes tuples of linguistic features:

(1) Lemma a_i from a fixed dictionary $a_i \in \Psi_a$

(2) Dependency label b_i from a fixed dictionary $b_i \in \Psi_b$

(3) Part-of-speech tag u_i from a fixed dictionary $u_i \in \Psi_u$ and

(4) Morphological features v_i from a fixed dictionary $v_i \in \Psi_v$

In the place of corresponding source words x_i, consider the fact that the Russian source word is, in fact, a surface form of a lemma, which is inflected on the basis of its grammatical role and part-of-speech. In the other words, a source word can be viewed as a combination of lemma, dependency label, part-of-speech tag, and morphological features.

The grammatical role of a word in sentence is expressed through dependency label assigned to the word. Dependency labels are presented in the study [26]. Part-of-speech types

and corresponding tags of words are listed alphabetically in Table 1, customized for use with Russian from Version 2 of Universal Dependency (https://universaldependencies.org/u/pos/index.html).

Each part-of-speech follows its own morphology rule. For instance, from a lemma "любовь," which is an inanimate noun of feminine gender, we can generate many surface forms according to grammar rules for a noun which has 6 cases (nominative, accusative, genitive, dative, instrumental, and prepositional), two numbers (singular and plural). A noun in Russian also has gender and animate features. Each gender (masculine, feminine, and neuter) has its own inflection rule. Similarly, each animate feature (animate and inanimate) has its own inflection rule. All surface forms inflected from lemma "любовь" are presented in Table 2.

Replacing source words with a combination of the features, we replace the input dictionary Θ_x of the cardinality $|\Theta_x|$ with the tuple of input dictionaries $\Theta_a, \Theta_b, \Theta_u, \Theta_v$ of the summarized cardinality $|\Theta_a| + |\Theta_b| + |\Theta_u| + |\Theta_v|$. From the above analysis, we can see that the summarized cardinality is many times smaller than the cardinality of dictionary of source words—combinations of the features. In the extreme counting, the cardinality of dictionary of source words as combinations of the features is the product $|\Theta_a| \times |\Theta_b| \times |\Theta_u| \times |\Theta_v|$. The reduced input dictionary actually helps to reduce the severity of rare word and out-of-vocabulary problem in inference.

Example of applying source-word decomposition to the Russian word "Последние" (meaning: last) in a sentence is given in Table 3. The application results in a vector of linguistic features: "Последний" (lemma), "amod"(dependency label: an adjectival modifier of a noun), "ADJ" (part-of-speech tag: adjective) and "Animacy = Inan, Case = Acc, Degree = Pos, Number = Plur" (morphological features: inanimate, accusative case, positive degree of comparison, and plural number).

3.3. Feature-Based Transformer Model. In order to decompose source words into tuples of features, we make changes in the encoder of the transformer model, so that it takes vectors of linguistic features as inputs in place of source words. The modified encoder requires four sequences of linguistic features including lemma a_i, dependency label b_i, part-of-speech u_i, and morphological features v_i, for $i = 1, \ldots, n$. Each linguistic-feature tag j is considered as a string in a corresponding dictionary Θ_j. Trainable embeddings $\mathbf{e}_j$ of all linguistic-feature tags are looked up in corresponding dictionaries Θ_j by the modified encoder. As proposed in [23], we apply concatenation operation () on embeddings of all linguistic-feature labels of each source word. The concatenation results in a concatenated embedding corresponding to each source word in sentence. Positional embedding of each source word is then added to the concatenated embedding to form the final embedding representing source word. Given a sequence of final embeddings representing source words in a sentence, following steps of the modified encoder are essentially the same as in the standard encoder.

Table 1: Part-of-speech types and corresponding tags of Russian source words.

Tag	Meaning
ADJ	Adjective
ADP	Adposition
ADV	Adverb
AUX	Auxiliary
CCONJ	Coordinating conjunction
DET	Determiner
INTJ	Interjection
NOUN	Noun
NUM	Numeral
PART	Particle
PRON	Pronoun
PROPN	Proper noun
PUNCT	Punctuation
SCONJ	Subordinating conjunction
SYM	Symbol
VERB	Verb

Table 2: Russian lemma любовь (love) in different cases and numbers.

	Singular number	Plural number
Nominative case	"любовь"	"любови"
Genitive case	"любви"	"любовей"
Dative case	"любви"	"любовям"
Accusative case	"любовь"	"любови"
Instrumental case	"любовью"	"любовями"
Prepositional case	"любви"	"любовяч"

Mathematically, the feature-based transformer model can be represented as a composition of functions as follows.

$$\begin{aligned}
\mathbf{e}_{a_i} &= \phi_a[a_i], \quad \text{for } i = 1, \ldots, n,\\
\mathbf{e}_{b_i} &= \phi_b[b_i], \quad \text{for } i = 1, \ldots, n,\\
\mathbf{e}_{u_i} &= \phi_u[u_i], \quad \text{for } i = 1, \ldots, n,\\
\mathbf{e}_{v_i} &= \phi_v[v_i], \quad \text{for } i = 1, \ldots, n,\\
\mathbf{e}_i &= \phi_2[i], \quad \text{for } i = 1, \ldots, n,\\
\mathbf{c}_i &= \sqrt{d} \times \left(\mathbf{e}_{a_i} \| \mathbf{e}_{b_i} \| \mathbf{e}_{u_i} \| \mathbf{e}_{v_i}\right) + \mathbf{e}_i,\\
\mathbf{C} &= \{\mathbf{c}_i, \quad \text{for } i = 0, \ldots, n\},\\
\mathbf{Z} &= g(\mathbf{C}),\\
y_i &= f(\mathbf{Z}, y_1, y_2, \ldots, y_{i-1}), \quad \text{for } i = 1, \ldots, k.
\end{aligned} \tag{2}$$

In the above equations, notations ϕ_j for $j \in \{a, b, u, v\}$ are the dictionaries of trainable embeddings. It is worth to mention that the sum of dimensions of the embeddings is equal to d to make the concatenated embeddings compatible with the dimension of positional embeddings $\mathbf{e}_i$.

4. Materials and Methods

4.1. Materials. We created our corpus very much in the same way as indicated in the works [27, 28], which are dedicated to study another low-resource language pair in the form of

Table 3: Applying source-word decomposition to Russian word "Цоследние" in a sentence.

	Tag	Meaning
Word	"Цоследние"	Last
In sentence	"в Цоследние Ґоды у Цолитическич Цартий была Цлочая реЦутация в Црессе, и для этоҐо есть веские" Цричины"	"In the last years, political parties had a bad reputation in the press, and there are good reasons for this"
Lemma	"Цоследний"	Last
Dependency label	"Amod"	An adjectival modifier of a noun
Part-of-speech tag	"ADJ"	Adjective
Morphological features	Animacy = Inan	Inanimate
	Case = Acc	Accusative case
	Degree = Pos	Positive degree of comparison
	Number = Plur"	Plural number

Chinese-Vietnamese. First, we picked 33,027 Russian sentences from News Commentary data (http://www.statmt.org/wmt13/training-parallel-nc-v8.tgz) of shared task: Machine Translation of ACL 2013 Eighth Workshop on Statistical Machine Translation. Next, we translated the Russian sentences into Vietnamese. The translation was carried out as follows. First, we used Google translate service to translate all the Russian sentences into Vietnamese. Then, we corrected the translation results, so they not only reflect the meaning of the Russian source sentences but also sound naturally, taking advantage of the fact that we are native speakers of Vietnamese and understand Russian. As a result, we had 33027 Russian-Vietnamese sentence pairs. We then randomly arranged the sentence pairs. From the shuffled corpus, we first took out 1500 sentence pairs to form the testing dataset. After that, we took another 1500 sentence pairs to form the development dataset. The remaining 30027 sentence pairs were used as the training dataset. Summary of the datasets is demonstrated in Table 4 [29]. The summary reveals the huge difference in the dictionary size between Russian and Vietnamese. The number of unique tokens in the Russian training dataset is over 8.5 times larger than the one in the Vietnamese side. The difference can be explained by the fact that Russian is a morphological rich language, while Vietnamese is a noninflectional language. On the other hand, the summary also highlights the difference in average sentence length between Russian and Vietnamese. The number of tokens per Vietnamese sentence is over 1.5 times larger than the one in the Russian side. In other words, to express the same idea, on average, we need to use more Vietnamese words than Russian words.

4.2. Methods. To evaluate feature-based NMT, we performed four experiments. In each experiment, we built and assessed a NMT model. Input and output of each model are presented in Table 5. We used deep learning library PyTorch [30] to build the NMT models with the required input and output by altering an implementation (https://github.com/bentrevett/pytorch-seq2seq) of the state-of-the-art transformer. The source codes of the proposed feature-based and baseline NMT models are provided at GitHub page (https://github.com/ThienCNguyen/Russian-Vietnamse) of the first author.

In the first experiment, we built the baseline W2W model which takes a sequence of Russian words as input and predicts a sequence of Vietnamese words as output. The W2W model is comprised of an encoder and a decoder. The encoder has a 256-dimensional embedding layer, 256-dimensional hidden states, three sublayers consisting of 8-head self-attention layer and 512-dimensional feedforward layer, and dropout layers with level = 10%. The decoder has a similar configuration as the encoder. We used tokenized training and development datasets to train the W2W model. We tokenized Russian sentences in the training and development datasets to produce corresponding sequences of Russian words by using space delimiters between Russian words. We tokenized Vietnamese sentences in the training and development datasets to produce corresponding sequences of Vietnamese words by using a tool provided in [31]. Using Adam optimizer with a learning rate = 0.0005 as reported in [32], we trained the W2W model in 20 epochs of the training dataset. Then, we chose the parameters of the model providing the least cross-entropy loss in the development dataset.

In the second experiment, we built the S2S model which takes a sequence of Russian subwords as input and predicts a sequence of Vietnamese subwords as output. The S2S model has the same configuration and optimization procedure as the baseline W2W model. To produce sequences of subwords for building the model, we tokenized sentences in the training and development datasets by using a tool provided in [33].

In the third experiment, we built the SnF2S model which takes a sequence of Russian subwords and their features (subword tag, lemma, dependency label, part-of-speech tag, and morphological label) as input and predicts a sequence of Vietnamese subwords as output. Subword tag is one of four types: *B, I, E*, and *O*, corresponding to four types: beginning part, inside part, ending part, and full word. Linguistic features of a subword are the same as the ones of containing words which are generated by a deep learning tool Stanza [34]. The SnF2S model is an improvement on the model proposed in [23]. We substituted recurrent neural networks with the state-of-the-art transformer. The SnF2S model also has a similar configuration and optimization procedure as the S2S model except for the encoder embedding layer and dimension of hidden states. The encoder embedding layer is

Table 4: Summary of the parallel datasets used in the study.

Number of	Russian			Vietnamese		
	Training	Development	Testing	Training	Development	Testing
Sentences	30,027	1,500	1,500	30,027	1,500	1,500
Tokens	438,875	21,820	21,941	693,681	34,436	34,651
Tokens per sentence	14.6	14.5	14.6	23.1	23.0	23.1
Unique tokens	46,789	7,520	7,450	5,402	1,985	2,058

Table 5: Input and output of NMT models.

Experiment	Model name	Input unit	Output unit
1	W2W	Word	Word
2	S2S	Subword	Subword
3	SnF2S	Subword and features	Subword
4	F2W	Features only	Word

composed of six embedding sublayers: 352-dimensional subword embedding, 7-dimensional subword-tag embedding, 117-dimensional lemma embedding, 12-dimensional dependency label embedding, 12-dimensional part-of-speech-tag embedding, and 12-dimensional morphological label embedding. We chose the dimension of embeddings, following the ratio recommended in [23]. We applied 512-dimensional hidden states to make them compatible with embedding dimension.

In the second and third experiments with sequences of Vietnamese subwords as output, we applied a postprocessing of concatenating subwords to form a sequence of words in the same way as proposed in [23].

In the fourth experiment, we built the proposed feature-based NMT model called S2F, which takes sequences of features of Russian source words (lemma, dependency label, part-of-speech tag, and morphological label) as input and predicts a sequence of Vietnamese words as output. In turn, the S2F model has a similar configuration and optimization procedure as the SnF2S model except for the encoder embedding layer and dimension of hidden states. The encoder embedding layer is composed of four embedding sublayers: 190-dimensional lemma embedding, 22-dimensional dependency label embedding, 22-dimensional part-of-speech-tag embedding, and 22-dimensional morphological label embedding. We chose the 256-dimension for hidden states to be compatible with embedding dimension.

In all experiments, we used the same assessment procedure for all NMT models. First, we fed Russian sentences of the testing dataset to the models. Then, we compared the predictions by the NMT models with reference Vietnamese sentences in the testing dataset in terms of the lowercase BLEU score which is calculated by the natural language toolkit NLTK [35].

5. Results and Analysis

Primary translation results are provided at GitHub page (https://github.com/ThienCNguyen/Russian-Vietnamse) of the first author. In this section, we analyze translation results for Russian-Vietnamese. We compare the performance of the proposed feature-based NMT with baseline NMT models. We also present human judgment of translation results.

5.1. Machine Judgment. Figure 1 shows the corpus-level BLEU scores of translation results from the testing dataset by the NMT models. We can observe that, among the baseline models, the SnF2S model yields the best result. In comparison with the word-based W2W model, the subword-based S2S model improves by 2.54 BLEU. Compared with the subword-based S2S model, the subword-based feature-added SnF2S model provides an improvement of 3.83 BLEU. This result suggests that we should compare it with the SnF2S model which is the strongest baseline model in order to prove the effectiveness of our proposed model. In comparison with the strongest baseline SnF2S model, our feature-based F2W model outperforms by an impressive 4.83 BLEU. Nevertheless, on the sentence level, the proposed F2W model does not always prove itself better than the SnF2S model. Among 1500 sentences in the testing dataset, the F2W model worsens the translation quality in 41.13% cases, while it improves the BLEU score in 57.6% cases. Detail of the comparison is presented in Figure 2.

5.2. Human Judgment. In addition to machine judgment, we also applied human judgment on translation results by NMT models. We made human analysis to have a more complete assessment on translation results. Specifically, we randomly picked 5 cases from the testing dataset. Here, we present the selected cases and human analysis on translation results. Description of each case consists of a Russian source sentence, its meaning in English, Vietnamese reference, translation results by NMT models, and their corresponding sentence-level BLEU scores.

Table 6 shows translation results by NMT models from a simple source sentence. Although the source sentence is simple, two models, W2W and SnF2S, give wrong translations. Their translations with the meanings "Europe is still in place of Barack Obama, Barack Obama has gone" and "Europe is still impressed with the tragedy of Barak Obama" are far from the initial meaning of the source sentence. On the other hand, two NMT models, S2S and the proposed F2W, perform pretty well for this source sentence. The meaning of the translation by S2S is "Europe is still impressed with the impression of the visit to Obama," which is close to the meaning of the source sentence. The result still has a flaw. Repeated phrase "ấn tượng"g(meaning: impression) in the translation result may make it more difficult

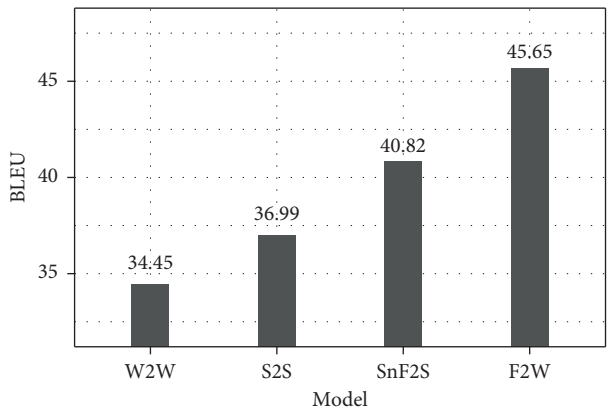

Figure 1: BLEU scores of predicted Vietnamese sentences by NMT models.

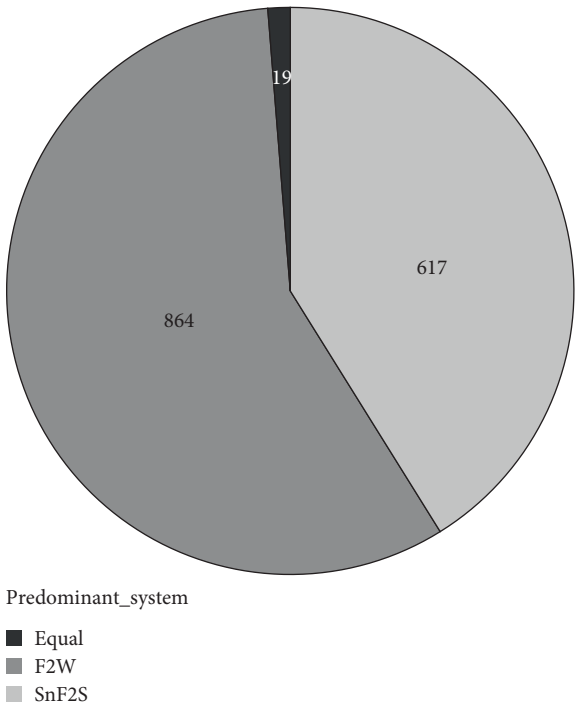

Figure 2: Comparison between F2W and SnF2S models on the sentence level BLEU score.

for us to catch the meaning. Compared to other models, the proposed F2W model gives the best translation result. The meaning of the translation is "Europe is still impressed with the visit of Barack Obama," which bears the closest resemblance with the meaning of the source sentence.

Table 7 shows translation results by NMT models from a more complicated source sentence where the subject has a singular form but plays a plural role. The quality of translation by NMT models is very different in this case. Three baseline models, W2W, S2S, and SnF2S, provide translation results with the wrong meanings "but most books and Stalin's red is a positive light in light," "But most of Stalin's book and press pretend," and "But most of Stalin's books and authors were light cakes under the light." At the same time, our proposed F2W model gives a translation identical to the gold reference.

Table 8 shows translation results by NMT models from a complex source sentence where an infinitive clause is used as subject complement. This long complex sentence is a challenge for NMT models. There is no translation model that gives a good enough result for this case. Furthermore, quality of the translation results by NMT models for this example is the perfect reflection of overall machine judgment on NMT models. Among the baseline models, the SnF2S model gives the best result. Specifically, it partly translates the key phrase "красныч чмеров" (meaning: Khmer Rouge) into "đỏ," while other baseline models mistranslate the phrase. Compared to the best baseline SnF2S model, the proposed F2W model also partly translates that key phrase and improves translation by successfully translating the other key phrase "диШломатическич усилий" (meaning: diplomatic efforts).

Table 9 shows translation results by NMT models from a sentence where the proposed F2W model slightly worsens the translation quality in terms of the BLEU score. In comparison with the best baseline SnF2S model (62.69 BLEU), the proposed F2W model provides a slower BLEU score (60.29 BLEU). From the human perspective, the meaning of the translation result by the F2W model (meaning: we operate with the private sector, not compete with it) is very close to the meaning of the translation result by the SnF2S model (meaning: we work with the private sector, not competing with it) and the reference itself.

Table 10 shows translation results by NMT models from a sentence where the proposed F2W model significantly worsens the translation quality in terms of the BLEU score. In comparison with the best baseline SnF2S model (55.39 BLEU), the proposed F2W model provides a far slower BLEU score (20.69 BLEU). Nevertheless, from the human perspective, the translation result by the F2W model (meaning: from the Persian Gulf, the oil and gas import region in the United States) partially reflects the meaning of the reference, while the SnF2S model mistranslates the Russian source sentence. The meaning of the translation result by the SnF2S model is "a third of oil exported only to the United States."

In overall, both machine and human judgments prove the superiority of the proposed feature-based transformer

Table 6: Translation results by NMT models from a simple source sentence.

Tag	Content
Russian	“евроПа все еще начодится Под вПечатлением визита барака обамы”
Meaning	“Europe is still under the impression of Barack Obama’s visit”
Reference	“châu âu vẫn còn ấn tượng bởi chuyến thăm của Barack Obama”
W2W (19.56)	“châu âu vẫn đang ở vị trí của Barack Obama, Barack Obama đã đi”
S2S (53.73)	“châu âu vẫn còn ấn tượng với sự ấn tượng của chuyến thăm Obama”
SnF2S (45.06)	“châu âu vẫn đang ấn tượng với bi kịch của Barak Obama”
F2W (53.04)	“châu âu vẫn đang ở ấn tượng với chuyến thăm của Barack Obama”

Table 7: Translation results by NMT models from a more complicated source sentence where the subject has a singular form but plays a plural role.

Tag	Content
Russian	“но большинство книҐ и авторов изображают сталина в Положительном свете”
Meaning	“But most books and authors portray stalin in a positive light”
Reference	“nhưng hầu hết các cuốn sách và tác giả miêu tả stalin trong một ánh sáng tích cực”
W2W (27.86)	“nhưng hầu hết các sách và màu đỏ của stalin là một ánh sáng tích cực trong ánh sáng”
S2S (36.85)	“nhưng hầu hết cuốn sách và giả vờ báo chí của stalin”
SnF2S (29.19)	“nhưng hầu hết cuốn sách và tác giả của stalin được bánh sáng dưới ánh sáng”
F2W (100)	“nhưng hầu hết các cuốn sách và tác giả miêu tả stalin trong một ánh sáng tích cực”

Table 8: Translation results by NMT models from a complex source sentence where an infinitive clause is used as subject complement.

Tag	Content
Russian	“официальное решение создать трибунал для красныч чмеров является большим достижением После десятилетия диПломатическич усилий”
Meaning	“The official decision to establish a tribunal for the Khmer Rouge is a great achievement after a decade of diplomatic efforts”
Reference	“quyết định chính thức thành lập toà án cho người Khmer đỏ là một thành tựu lớn sau một thập kỷ nỗ lực ngoại giao”
W2W (30.34)	“toà án chính thức để tạo ra toà án <unk> <unk> là một thành tựu lớn sau những nỗ lực ngoại giao ngoại giao”
S2S (16.12)	“quyết định chính thức tạo ra toà án chính thức cho những” gìn mederon là thành tựu lớn sau những nỗ lực ngoại giao
SnF2S (16.28)	“một quyết định chính thức để tạo ra toà án đỏ là thành tựu to lớn sau thập kỷ to”
F2W (38.92)	“quyết định chính thức để tạo ra một toà án đỏ cho các <unk> là thành tựu lớn sau nhiều nỗ lực ngoại giao”

Table 9: Translation results by NMT models from a sentence where the proposed F2W model slightly worsens the translation quality in terms of the BLEU score.

Tag	Content
Russian	“мы работаем с частным сектором, а не конкурируем с ним”
Meaning	“We work with the private sector, not compete with it”
Reference	“chúng tôi hợp tác với khu vực tư nhân, không cạnh tranh với nó”
W2W (27.53)	“chúng tôi đã đối phó với khu vực tư nhân, không phải là một <unk> với nó”
S2S (49.89)	“chúng tôi làm việc với khu vực tư nhân chứ không phải là một người cạnh tranh với họ”
SnF2S (62.69)	“chúng tôi làm việc với khu vực tư nhân, không phải cạnh tranh với nó”
F2W (60.29)	“chúng ta hoạt động với khu vực tư nhân, không phải cạnh tranh với nó”

Table 10: Translation results by NMT models from a sentence where the proposed F2W model significantly worsens the translation quality in terms of the BLEU score.

Tag	Content
Russian	“из ПерсидскоҐо залива ПостуПает лишь Пятая часть имПортируемой в соединенные штаты нефти”
Meaning	“Only one fifth of the oil imported into the United States comes from the Persian Gulf”
Reference	“chỉ một phần năm lượng dầu nhập khẩu vào hoa kỳ đến từ vịnh ba tư”
W2W (26.15)	“từ vịnh ba tư chỉ có một phần của hoa kỳ đã được hưởng dầu được hưởng lợi từ dầu”
S2S (29.85)	“trong phần trăm năm, hoa kỳ chỉ làm nhập khẩu dầu”
SnF2S (55.39)	“một phần ba của dầu chỉ xuất khẩu chỉ nhập khẩu hoa kỳ”
F2W (20.69)	“từ vịnh ba tư, khu vực nhập khẩu dầu khí ở hoa kỳ”

model in comparison to other available transformer translation models for translating from Russian into Vietnamese.

6. Conclusions and Perspectives

In this paper, we have successfully integrated linguistic knowledge into the state-of-the-art transformer translation model. We have introduced the feature-based transformer model, which replaces source words by combinations of their features comprising of lemma, dependency label, part-of-speech tag, and morphological label. We have empirically compared the proposed model with other baseline models. Experiment result for the Russian-Vietnamese language pair shows that our model outperforms other models by great distances.

Based on the translation results and our knowledge of the investigated Russian and Vietnamese languages and their relations to other languages, we strongly recommend the feature-based NMT model for building systems translating from highly inflectional synthetic Slavic languages including Russian, Belarusian, Ukrainian, Polish, Bulgarian, Czech, and Serbian into noninflectional analytic languages, such as Vietnamese and Chinese.

Data Availability

The dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] R. Aharoni, M. Johnson, and O. Firat, "Massively multilingual neural machine translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 3874–3884, Minneapolis, MN, USA, June 2019.

[2] Y. Cheng, L. Jiang, and W. Macherey, "Robust neural machine translation with doubly adversarial inputs," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4324–4333, Florence, Italy, August 2019.

[3] A. Bapna and O. Firat, "Simple, scalable adaptation for neural machine translation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 1538–1548, Hong Kong, China, November 2019.

[4] W. Zhang, Y. Feng, F. Meng, D. You, and Q. Liu, "Bridging the gap between training and inference for neural machine translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4334–4343, Florence, Italy, August 2019.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, May 2015.

[6] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proceedings of the Eighth Syntax, Semantics and Structure in Statistical Translation*, vol. 103, Doha, Qatar, October 2014.

[7] 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.

[8] 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*, vol. 1, pp. 1715–1725, Berlin, Germany, August 2016.

[9] K. Vylomova, T. Cohn, X. He, and G. Haffari, "Word representation models for morphologically rich languages in neural machine translation," in *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, vol. 103, Copenhagen, Denmark, September 2017.

[10] T. Kudo, "Subword regularization: improving neural network translation models with multiple subword candidates," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 66–75, Melbourne, Australia, July 2018.

[11] Y. Wu, M. Schuster, Z. Chen et al., "Google's neural machine translation system: bridging the gap between human and machine translation," *CoRR*, 2016.

[12] M. Johnson, M. Schuster, Q. V. Le et al., "Google's multilingual neural machine translation system: enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.

[13] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008, Long Beach, CA, USA, December 2017.

[14] A. Vaswani, S. Bengio, E. Brevdo et al., "Tensor2tensor for neural machine translation," in *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas*, vol. 1, pp. 193–199, Boston, MA, USA, March 2018.

[15] H. Sarma, N. Saharia, and U. Sharma, "Development and analysis of speech recognition systems for assamese language using htk," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 17, no. 1, pp. 1–14, 2017.

[16] S. Bhattacharya, D. Manousakas, A. G. C. P. Ramos, S. I. Venieris, N. D. Lane, and C. Mascolo, "Countering acoustic adversarial attacks in microphone-equipped smart home devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–24, 2020.

[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*, pp. 868–876, Prague, Czech Republic, June 2007.

[18] 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.

[19] P. Koehn and J. Schroeder, "Experiments in domain adaptation for statistical machine translation," in *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 224–227, Prague, Czech Republic, June 2007.

[20] P. Koehn, H. Hoang, A. Birch et al., "Moses: open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and*

Poster Sessions, pp. 177–180, Prague, Czech Republic, June 2007.

[21] D. Kolovratnık, N. Klyueva, and O. Bojar, "Statistical machine translation between related and unrelated languages," in *Proceedings of the Conference on Theory and Practice on Information Technologies*, pp. 31–36, Citeseer, Dolný Kubín, Slovakia, September 2009.

[22] S. Huet, E. Manishina, and F. Lefèvre, "Factored machine translation systems for Russian-English," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August 2013.

[23] R. Sennrich and B. Haddow, "Linguistic input features improve neural machine translation," in *Proceedings of the First Conference on Machine Translation*, vol. 1, pp. 83–91, Berlin, Germany, August 2016.

[24] M. García-Martínez, O. Caglayan, W. Aransa, A. Bardet, F. Bougares, and L. Barrault, "Lium machine translation systems for wmt17 news translation task," 2017, http://arxiv.org/abs/1707.04499.

[25] M. Weller-Di Marco and A. Fraser, "Modeling word formation in English–German neural machine translation," in *Proceedings of the 58th Annual Meeting of the Association For Computational Linguistics*, pp. 4227–4232, January 2020, https://www.aclweb.org/anthology/2020.acl-main.389/.

[26] M.-C. De Marneffe, T. Dozat, N. Silveira et al., "Universal stanford dependencies: a cross-linguistic typology," *Language Resources and Evaluation*, vol. 14, pp. 4585–4592, 2014.

[27] 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, 7 pages, 2016.

[28] 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.

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

[30] A. Paszke, S. Gross, F. Massa et al., "Pytorch: an imperative style, high-performance deep learning library," in *Proceedings of the Advances In Neural Information Processing Systems*, pp. 8024–8035, Vancouver, Canada, December 2019.

[31] T. Vu, D. Q. Nguyen, 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*, pp. 56–60, New Orleans, LA, USA, January 2018.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186, Minneapolis, MN, USA, June 2019.

[33] R. Sennrich, B. Haddow, and A. Birch, "Edinburgh neural machine translation systems for wmt 16," in *Proceedings of the First Conference on Machine Translation*, vol. 2, pp. 371–376, Berlin, Germany, August 2016.

[34] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," in *Proceedings of the 58th Annual Meeting of the Association For Computational Linguistics*, July 2020, https://www.aclweb.org/anthology/2020.acl-demos.14/.

[35] E. Loper and S. Bird, "NLTK: the natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies For Teaching Natural Language Processing and Computational Linguistics*, pp. 63–70, Philadelphia, PA, USA, July 2002.