Grammatical error correction for Vietnamese using Machine Translation

Nghia Luan Pham[1,3], Tien Ha Nguyen[2], and Van Vinh Nguyen[3]

[1] Hai Phong University, Haiphong, Vietnam
luanpn@dhhp.edu.vn
[2] VNU University of Science, Hanoi, Vietnam
tienhapt@gmail.com
[3] VNU University of Engineering and Technology, Hanoi, Vietnam
vinhnv@vnu.edu.vn

Abstract Correction of Vietnamese grammatical errors plays an important role in Natural Language Processing. In this paper, we propose a new method using Machine Translation. We consider the grammatical error correction problem like machine translation problem with source language as grammatical wrong text and target language as grammatical right texts, respectively. Additionally, we carry out pre-processing step with grammatical wrong text using spelling checker such as MS Word spelling tool before using Machine translation model.
Our experiments based on the state-of-the-art Machine Translation systems combining with pre-processing step. Experimental results achieved 84.32 BLEU score with Vietnamese grammatical error correct based on SMT architecture and 88.71 BLEU score system based on NMT architecture, which indicates that our method achieves promising results.

Keywords: Vietnamese Grammatical error correction · Statistical Machine Translation · Neural Machine Translation.

1 Introduction

Nowadays, correction of grammatical errors is an active research topic, this topic based on Machine Translation has been applied to English, but there is not any research which uses Machine Translation for Vietnamese.

Vietnamese is not easy to learn, even both Vietnamese people and Vietnamese learners usually make grammatical errors in the text. There are several types of error, such as spelling mistakes, using wrong words. A Vietnamese grammatical error correction (GEC) system will have the benefit for Vietnamese and Vietnamese learners. Also, the GEC models can be applied to Natural Language Processing systems. The difference in our method is that we apply the model to Vietnamese, which is much harder than English. As the increasing number of information, we have a chance to access to the valuable source of knowledge about potential customers. Information extraction from Vietnamese online text, however, is a critical natural language understanding. This is the most challenge.

We propose a new method for Vietnamese grammatical error correction. It is useful for a non-native Vietnamese learner and for a native speaker. Our presentation is structured: Section 2 summarizes the related work. Section 3 described our method. Section 4 presents the experiments. Finally, conclusions are presented in Section 5.

2 Related work

As we mentioned above, the correction of grammatical errors is an active research topic. Therefore, many studies have been published. In this section, we present some approaches to correct grammatical errors in recent years.

In [8], Courtney Napoles and Chris Callison-Burch presented an investigation about components of a statistical machine translation pipeline then authors customized for grammatical error correction. They showed that extending the translation grammar with generated rules for spelling correction can improve the Max-Match metric score by as much as 20%.

In [1], Kai-Fu proposed an approach to grammatical error correction using neural machine translation for Chinese. Their staged approach includes: first they remove the surface errors. Then they built the grammatical error correction system using neural machine translation.

In [2], authors proposed the method that combines two popular approaches *(SMT and NMT)* to build a system for automated grammatical error correction. This combination system gains new results on the CoNLL-2014 and JFLEG benchmarks.

The methods above are most related to our method, but our method is different from these methods as some points:

1. We carry out pre-processing step using spelling checker with the Vietnamese input text, then put it in the machine translation system to correct remaining grammatical errors.
2. We also solve grammatical errors correction in Vietnamese language using Machine Translation. According to our understanding, this is the research that applying Machine Translation for Vietnamese grammatical errors correction, the first time.

3 Our method

We treat the Vietnamese grammar detection and correction problem like machine translation problem, so this task, we propose a method using machine translation. In particular, wrong grammar and right grammar texts are considered like source and target language respectively. Machine translation model detect and correct grammar errors.

3.1 Machine Translation

Phrase-based Statistical Machine Translation: The input texts are segmented into a number of sequences of words or phrases. Each phrase in the source sentence is translated into the target language. The translation model is built on the noisy channel model [4]. This model uses Bayes rules to reformulate translation probabilities to translate a foreign sentence f into e. The best translation for a foreign sentence f is the equation 1:

$$e = arg\max_{e} p(e)p(e|f) \tag{1}$$

The above equation consists of two main components: the language model p(e) and the translation model p(e|f). Monolingual data in the target side is used for training language model and parallel data is used for training translation model, parameters are estimated from parallel data, the best output sentence e for the input sentence f according to the equation

$$e = arg\max_{e} p(e|f) = arg\max_{e} \sum_{m=1}^{M} \lambda_m h_m(e, f) \tag{2}$$

where h_m is a feature function such as language model, translation model and λ_m corresponds to a feature weight.

Neural Machine Transaltion: Given a sentence in source side $x = (x_1, ..., x_m)$ and its corresponding sentence in target side $y = (y_1, ..., y_n)$. In paper, we use the attentional NMT architecture proposed by [6]. In their work, the encoder, which is a bidirectional recurrent neural network, reads the source sentence and generates a sequence of source representations $h = (h_1, ..., h_m)$. The decoder is another recurrent neural network, produces the target sentence at a time. The log conditional probability thus can be decomposed as follows:

$$\log p(y|x) = \sum_{i=1}^{n} \log p(y_t|y_{<t}, x) \tag{3}$$

where $y_{<t} = (y_1, ..., y_{t-1})$. As described in Equation 4, the conditional distribution of $p(y_t|y_{<t}, x)$ like a function of the previously predicted output y_{t-1}, the hidden state of the decoder s_t, and the context vector c_t.

$$p(y_t|y_{<t}, x) \propto \exp\{g(y_{t-1}, s_t, c_t)\} \tag{4}$$

The context vector c_t is used to determine the relevant part of the source sentence to predict y_t. It is computed as the weighted sum of source representations $h_1, ..., h_m$. Each weight α_{ti} for h_i implies the probability of the target symbol y_t being aligned to the source symbol x_i:

$$c_t = \sum_{i=1}^{m} \alpha_{ti} h_i \tag{5}$$

Given a parallel data of size N, the parameter θ of NMT model is trained to maximize the probabilities for all sentence pairs $\{(x^n, y^n)\}_{n=1}^{N}$:

$$\theta^* = arg \max_{\theta} \sum_{n=1}^{N} \log p(y^n|x^n) \tag{6}$$

where θ^* is the optimal parameter.

3.2 Our method for Vietnamese Grammatical error correction

Each language has its own characteristics, and so is Vietnamese. To correct Vietnamese grammatical errors, we must recognize as much error types as possible. Generally, the grammatical error types in Vietnamese can be divided into two groups, as below:

Errors in sentence structure: These errors include errors such as sentence components missing, overlapping sentence components and sentences components wrongly ordering.

- *Missing sentence component:* there is a lot of shortened sentences which have only component subject or predicate, thus it makes the sentence meaning ambiguous.
- *Overlapping sentence component:* These errors are often caused by learner's unclear ideas or their limited language ability.
- *The sentence components are in the wrong order:* Unlike English, in Vietnamese, the order of components in a sentence is very important. When we make this kind of error, it makes the sentence meaningless or ambiguous.

Errors in punctuation: punctuation in the text is very important because it defines the grammatical structure and expresses the meaning of the sentence. Therefore, errors in punctuation can negatively affect the learners' purpose, which can lead to serious misunderstandings.

The main idea of this paper is correction grammatical errors be considered like translation problem, so the input text in the source language as Vietnamese grammatical wrong and output text is Vietnamese grammatical right as the target language. To solve this problem, we proposed a new method which is described in Figure 1.

A key advantage of the machine translation is that errors are learned from parallel data automatically. To evaluate the effect of our method, we conduct experiments on the state-of-the-art Machine Translation systems: Statistical Machine Translation (SMT) and Neural Machine Translation (NMT).

4 Experiments

4.1 Dataset

We first collect 317,596 Vietnamese sentences from news sites like dantri.com.vn; vnexpress.net and then cleaning and make grammatical error types from the to

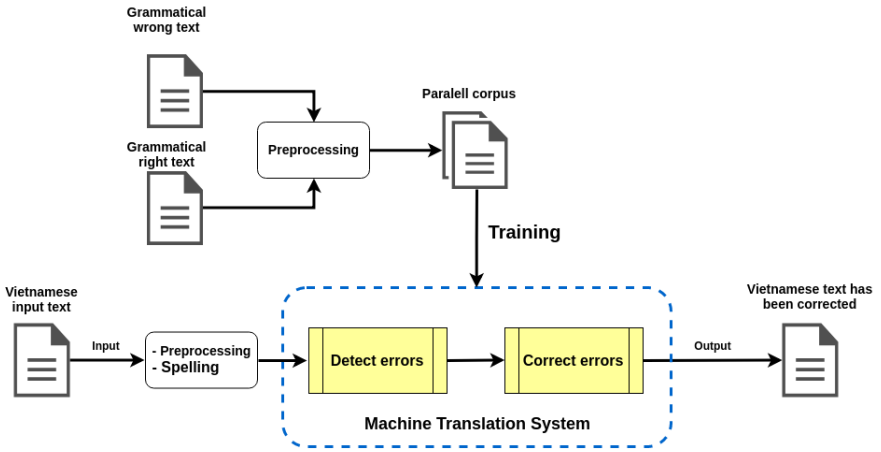

Figure 1. Illustration for our method. A parallel corpus is collected from grammatical wrong text and grammatical right text, this parallel corpus is used to build a Vietnamese GEC system using Machine Translation (SMT - NMT)

build about 271,822 parallel sentence pairs for training, 29,895 sentences pairs for validation, and 15,879 sentences pairs for the test. The table 1 is the data statistics for training our Vietnamese GEC systems.

4.2 Settings

We used Moses[4] and OpenNMT[5][3] to training our Vietnamese GEC systems.

The NMT system is trained Long Short-Term Memory (LSTM) network [5], we use 2-layer, 500 hidden units on the encoder/decoder and the general attention type of Thang Luong [7].

To evaluate the quality of our Vietnamese GEC, we use the BLEU score that standard metric to evaluate the quality of translation systems.

4.3 Results and Discussions

We trained two Vietnamese grammatical error correction systems based on SMT and NMT with the same parallel corpus, they are called Vietnamese GEC_SMT and Vietnamese GEC_NMT. We evaluate the quality of these two systems with two types of input text:

- ***None-Spelling:*** Vietnamese input text is pre-processed, do not carry out the spelling check step *(Vietnamese GEC_SMT and NMT)*;

[4] http://statmt.org/moses/
[5] https://github.com/OpenNMT/OpenNMT-py

Data Sets		Vietnamese language	
		Wrong grammar	**Right grammar**
Training	Sentences	**271,822**	
	Average Length	21.1	20.8
	Words	5,735,444	5,653,897
Validation	Sentences	**29,895**	
	Average Length	21.9	21.8
	Words	654,700	651,711
Test	Sentences	**15,879**	
	Average Length	21.8	21.6
	Words	346,162	342,986

Table 1. The data statistics for training our Vietnamese GEC systems.

– ***Spelling:*** Vietnamese input text is pre-processed and carry out the spelling check step *(Spell+Vietnamese GEC_SMT and NMT).*

We measured by BLEU score with the same data set for test, experimental results are described as in the Figure 2.

SYSTEM	BLEU SCORE
Vietnamese GEC_SMT	83.73
Spell + Vietnamese GEC_SMT	**84.32**
Vietnamese GEC_NMT	87.51
Spell + Vietnamese GEC_NMT	**88.71**

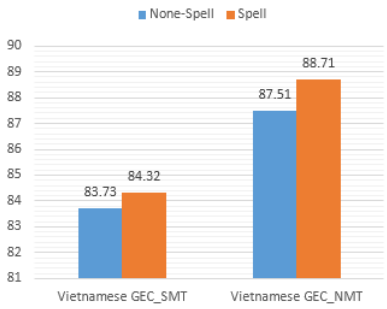

Figure 2. The BLEU score: Vietnamese GEC_SMT vs Vietnamese GEC_NMT

In the Figure 2 show experiemental results of the Vietnamese Grammatical error correction systems, the BLEU score achieved **83.73** points for the Vietnamese GEC_SMT system and **87.51** points for the Vietnamese GEC_NMT system. If the input text is pre-processed and spelling correction before appling Machine Translation models, our systems get better results: the BLEU score achieved **84.32** points for the Spell+Vietnamese GEC_SMT system and **88.71** points for the Spell+Vietnamese GEC_NMT system.

The Figure 3 shows some example outputs of our systems. From these results, it shows that the NMT system is better than SMT system in Vietnamese grammatical error correction. Both the Vietnamese GEC_SMT system and the

Vietnamese GEC_NMT system are restricted in correcting errors that sentence is lacked of characters, rhythm, etc.

The Vietnamese GEC_NMT system correct unk errors *(errors that it unknown)* are not good, but it can correct grammatical errors well. We could get better results when we carry out pre-processing step with the input text using spelling checker tool before using Machine Translation model.

Input sentence	Reference sentence	Output sentence of system			
		Vietnamese GEC_SMT	Vietnamese GEC_NMT	Spell+Vietnamese GEC_SMT	Spell+Vietnamese GEC_NMT
Lập doanh trại ở làng dgân nhất .	Lập doanh trại ở làng gần nhất	Lập doanh trại ở làng dgân nhất	**Lập doanh trại ở làng gần nhất .**	**Lập doanh trại ở làng gần nhất .**	**Lập doanh trại ở làng gần nhất .**
rMặt ndó sưng lên vì ăng đau.	Mặt nó sưng lên vì răng đau .	Mặt sưng lên vì nó tăng đau	Mặt nó sưng lên vì đau đau .	Mặt sưng lên vì nó răng đau .	**Mặt nó sưng lên vì răng đau .**
tôi vừa nhìn . thấy được , một hình người trong bóng tối .	Tôi vừa nhìn thấy được một hình người trong bóng tối .	**Tôi vừa nhìn thấy được một hình người trong bóng tối .**	**Tôi vừa nhìn thấy được một hình người trong bóng tối .**	**Tôi vừa nhìn thấy được một hình người trong bóng tối .**	**Tôi vừa nhìn thấy được một hình người trong bóng tối .**
Nhìn bao quát vù uê từ trên đỉnh đồ .	Nhìn bao quát vùng quê từ trên đỉnh đồi .	Nhìn bao quát vù quê từ trên đỉnh đồi .	Nhìn bao quát vù quê từ trên đỉnh đồ .	**Nhìn bao quát vùng quê từ trên đỉnh đồi .**	**Nhìn bao quát vùng quê từ trên đỉnh đồi .**
Có s traão chính trị giữa haai .	Có sự trao đổi tù chính trị giữa hai nước .	Có sự traão chính trị giữa hai .	Có lẽ chính trị giữa hai .	Có sự trao chính trị giữa hai .	Có sự chính trị giữa hai .
Qua bản báo cao cho ta thấy được thực trạng ô nhiễm môi trường hiện nay.	Bản báo cáo cho ta thấy được thực trạng ô nhiễm môi trường hiện nay.	Bản báo Qua cao cho ta thấy được thực hiện trạng ô nhiễm môi trường nay .	Bản báo cao cho ta thấy được thực trạng ô nhiễm môi trường hiện nay .	**Qua bản báo cáo cho ta thấy được thực hiện trạng ô nhiễm môi trường nay .**	**Bản báo cáo cho ta thấy được thực trạng ô nhiễm môi trường hiện nay .**

Figure 3. Some outputs of Vietnamese grammatical error correction systems

5 Conlustion and future work

In this paper, we presented a new method for Vietnamese grammatical errors correction. We have investigated the effectiveness of models trained with SMT model and NMT model *(the state-of-the-art MT now)* when we applied to solve this GEC problem for Vietnamese. The experimental results show that the quality of grammatical errors correction is promising and could apply this method in real-world.

In the future, we will focus on improving quality. First, we can use the bigger amount of data to train our GEC system, bigger training data is, the more accurate model is. Second, we will use a hybrid SMT and NMT system for GEC system. Finally, we also will focus on collecting and analyzing data, as long as creating more quality data to improve the system.

Acknowledgments

This work is funded by the project: Building a machine translation system to support translation of documents between Vietnamese and Japanese to help managers and businesses in Hanoi approach Japanese market, under grant number TC.02-2016-03 and the project of VNU University of Engineering and Technology, Hanoi, Vietnam.

References

1. Kai Fu, Jin Huang, and Yitao Duan. Youdao's winning solution to the nlpcc-2018 task 2 challenge: A neural machine translation approach to chinese grammatical error correction. In *inproceedings*, 2018.
2. Roman Grundkiewicz and Marcin Junczys-Dowmunt. Near human-level performance in grammatical error correction with hybrid machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018.
3. Kim Y. Deng Y. Senellart J. Klein, G. and A. M. Rush. Opennmt: Open-source toolkit for neural machine translation. arXiv preprint arXiv:1701.02810, 2017.
4. Philipp Koehn. Statistical machine translation. Cambridge University Press, 2010.
5. Pham H. Luong, M.-T. and C. D. Manning. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025, 2015.
6. Hieu Pham Minh-Thang Luong and Christopher D Manning. Effective approaches to attention based neural machine translation. arXiv preprint arXiv:1508.04025, 2015.
7. Hieu Pham Minh-Thang Luong and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In Proc of EMNLP, 2015.
8. Courtney Napoles and Chris Callison-Burch. Systematically adapting machine translation for grammatical error correction. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 345–356 Copenhagen, Denmark, September 8, 2017. c 2017 Association for Computational Linguistics*, 2017.