VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Adaptation in Statistical Machine Translation for low-resource domains in English-Vietnamese language

Luan-Nghia Pham[1*], Vinh-Van Nguyen[2]

[1] *Haiphong University, Haiphong, Vietnam*

[2] *University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam*

Abstract

In this paper we propose a new domain adaptation method in Statistical Machine Translation for low-resource domains in English-Vietnamese language pair. Specifically, our method only uses monolingual data in target language to adapt the phrase translation table, our system brings improvements over the SMT baseline system. We use two steps to improve the quality of SMT system: (i) classify phrases on target side of the phrase translation table use the probability classifier model, and (ii) adapt to the phrase translation table by recomputing the direct translation probability of phrases.

Our experiments are conducted with translation direction from English to Vietnamese on two very different domains those are legal domain *(out-of-domain)* and general domain *(in-of-domain)*. The English-Vietnamese parallel corpora is provided by the IWSLT 2015 organizers and the experimental results showed that our method significantly outperformed the baseline system. Our system improved on the quality of machine translation in the legal domain up to 0.9 BLEU scores over baseline system.

Keywords: Machine translation, Domain Adaptation.

1. Introduction

Statistical Machine Translation (SMT) systems [1] are usually trained on large amounts of bilingual data and monolingual data in target side. In general, these corpora may include quite heterogeneous topics and these topics usually define a set of terminological lexicons. Terminologies need to be translated taking into account the semantic context in which they appear.

The Neural Machine Translation (NMT) approach [2] has recently been proposed for machine translation. However, the NMT method requires a large amount of parallel data and it has some characteristics such as NMT system is too computationally costly and resource, the NMT system also requires much more training time than SMT system [3]. Therefore, SMT systems are still being studied for specific domains in low-resource language pairs.

Monolingual data are usually available in large amounts, parallel data are low-resource for most language pairs. Collecting sufficiently large high-quality parallel data is hard, especially on domain-specific data. For this reason, most of languages in the world

* Corresponding author. Email: luanpn@dhhp.edu.vn

are low-resource for statistical machine translation, including the English-Vietnamese language pair.

When SMT system is trained on the small amount of specific domain data leading to narrow lexical coverage which again results in a low translation quality. On the other hand, the SMT systems are trained, tuned on specific-domain data will perform well on the corresponding domains, but performance deteriorates for out-of-domain sentences [4].

Therefore, the SMT systems often suffer from domain adaptation problem during practical applications. When the test data and the training data from the same domains, the SMT systems can achieve good translation quality. Otherwise, the translation quality degrades dramatically. Therefore, domain adaptation is of significant importance to developing translation systems which can be effectively transferred from one domain to another.

In recent years, domain adaptation problem in SMT becomes more important [5] and is an active field of research in SMT with more and more techniques being proposed and put into practice [6]; [5]; [7]; [8]; [9]; [10]; [11]; [12]; [13]. The common techniques used to adapt two main components of contemporary state-of-the-art SMT systems: the language model and the translation model. In addition, there are also some proposals for adapting the Neural Machine Translation (NMT) system to a new domain [14]; [15]. Although the NMT system has begun to be studied more, domain adaptation for the SMT system still plays an important role, especially for low-resource languages.

This paper presents a new method to adapt the phrase translation table of the SMT system. Our experiments were conducted for the English-Vietnamese language pair in direction from Enghlish to Vietnamese. We use domain corpora comprising of two specific domains: 'Legal' and 'General'. The data has been collected from documents, dictionaries and the IWSLT 2015 organisers for the English-Vietnamese translation task.

In our works, we train a phrase translation table with paralell corpus in general domain, then we train a probability classifier model with monoligual corpus in legal domain. We use the classification probability of phrase on target side of phrase translation table to recompute the direct translation probability of the phrase translation table. This is the first adaptation method for the phrase translation table of SMT system, especially for low-resource language pairs as English-Vietnamese language pair. For comparison, we train a baseline SMT system and a Neural Machine Translation system (NMT) to compare with our method. Experimental results showed that our method significantly outperforms the baseline system. Our system improved the translation quality of machine translation system on the out-of-domain data *(legal domain)* up to 0.9 BLEU points compared to baseline system. Our method has also been accepted for presentation at the 31st Asia Pacific conference on language, information and computation.

The paper is organized as follows. In the next section, we present related works on the problem of adaptation in SMT; Section 3 describes our method; Section 4 describes and discusses the experimental results. Finally, we end with a conclusion and the future works in Section 5.

2. Related works

Domain adaptation for machine translation is known to be a challenging research problem that has substantial real-world application and this has been one of the topics of increasing interest for the recent years. Recent work on machine translation domain adaptation has focused on data centric or model centric of the SMT system.

Some authors used monolingual out-of-domain data and adapted the language model. The main advantage of language model adaptation in contrast to translation model adaptation is that monolingual out-of-domain data is needed.

For many language pairs and domains, no new-domain parallel training data is available. In [16] machine translate new-domain source language monolingual corpora and use the synthetic parallel corpus as additional training data by using dictionaries and monolingual source and target language text.

In [5] build several specific domain translation systems, then train a classifier model to assign the input sentence to a specific domain and use the specific domain machine translation system to translate the corresponding sentence. They assume that each sentence in test set belongs to one of the already existing domains.

In [13] build the machine translation system for different domains, it trains, tunes and deploys a single translation system that is capable of producing adapted domain translations and preserving the original generic accuracy at the same time. The approach unifies automatic domain detection and domain model parameterization into one system.

In [17] used a source classification document to classify an input document into a domain. This work makes the translation model shared across different domains.

Above related works automatically detected the domain and the classifier model works as a “switch” between two independent MT decoding runs.

To adapt a translation model trained from one domain data to another domain, previous works more attention to the studies of parallel corpus while ignoring the out-of-domain monolingual corpora which can be obtained more easily.

Our method have some differences from above methods. For adapt to the phrase translation table of SMT system, we build a probability classifier model to estimate the classification probability of phrases on target side of the phrase translation table. Then we use these classification probabilities to recompute the direct phrase translation probability $\phi(\mathrm{e}|\mathrm{f})$.

3. Our method

In the phrase-based SMT, the quality of SMT system depends on training data. SMT system are usually trained on large amounts of parallel corpus. Currently, high quality parallel corpora of sufficient size are only available for a few language pairs. Furthermore, for each language pair the sizes of the domain specific corpora and the number of domains available are limited. The English-Vietnamese is low-resource language pair and thus domain data in this language pair are limited, for the majority of domain data, only few or no parallel corpora are availale. However, monolingual corpora for the domain are availale, which can also be leveraged.

The main idea in this paper is leveraging out-of-domain monolingual corpora in target language for domain adaptation for MT. In phrase-table of SMT system, a phrase in the source language may have many translation hypotheses with different probability. We use out-of-domain monolingual corpora to recompute the scores of translation probability of these phrases which are defined in out-of-domain.

There are many studies of domain adaptation for SMT, which can be mainly divided into two categories: data centric and model centric. Data centric methods focus on either selecting training data from out-of-domain parallel corpora based a language model or generating parallel data. These methods can be mainly divided into three categories:

- Using monolingual corpora.
- Synthetic parallel corpora generation.
- Using out-of-domain parallel corpora:

multi-domain and data selection.

Most of the related works in section 2 use monolingual corpora to adapt language model, or to synthesize parallel corpora, or models selection which are trained with different domains. The English-Vietnamese is low-resource parallel corpora, thus we proposed a new method which only uses monolingual corpora to adapt the translation model by recomputing the score of phrases in the phrase-table and to update the phrase's direct translation probability.

In this section, We first give a brief introduction of SMT; Next, we proposed new method for domain adaptation in SMT.

3.1. Overview of phrase-based statistical machine translation

The figure 1 illustrates the process of phrase-based translation. The input is segmented into a number of sequences of consecutive words (so-called phrases). Each phrase is translated into an Vietnamese phrase, and Vietnamese phrases in the output may be reordered

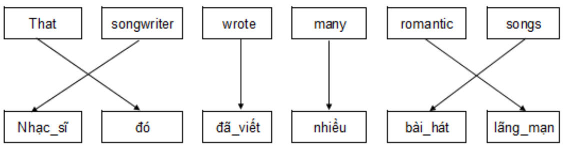

Figure 1. Example illustrates the process of phrase-based translation

The phrase translation model is based on the noisy channel model [1]. It uses Bayes rule to reformulate the translation probability for translating a foreign sentence f into English e. The best translation for a foreign sentence f is as equation 1:

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

The above equation consists of two components: a language model assigning a probability p(e) for any target sentence e, and a translation model that assigns a conditional probability p(e|f). The language model is trained using a monolingual corpora in the target language, the translation model is trained using parallel corpora, the parameters of the translation model are estimated from a parallel corpora, the best English output sentence ebest given a foreign input sentence f according to our model is

$$e = arg\max_{e} p(e|f) \tag{2}$$

$$= arg\max_{e} \sum_{m=1}^{M} \lambda_m h_m(e, f) \tag{3}$$

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

The figure 2 describes the architecture of phrase-based statistical machine translation system. There is some translation knowledge that can be used as language models, translation models, etc. The combination of component models (language model, translation model, word sense disambiguation, reordering model...)

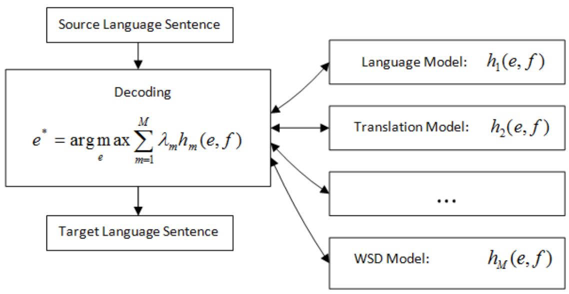

Figure 2. Architecture of Phrase-based Statistical Machine Translation

3.2. Translation model adaptation based on phrase classification

One of the essential parts of our experiments is the classifier used to identify the domain of a target phrase in the phrase-table, the accuracy of the classifier is very important in the final translation score of the sentences from the test set data. The Maximum Entropy was chosen to be the classifier for our experiments.

In this section, we first give a introduction of the maximum entropy classifier; Next, we describe our method for domain adaptation in SMT.

3.2.1. The Maximum Entropy classifier

To build a probability classification model, we use the Stanford classifier toolkit[1] with standard configurations. This toolkit uses a maximum entropy classifier with character n-grams features,... The maximum entropy classifier is a probabilistic classifier which belongs to the class of exponential models. The maximum entropy is based on the principle of maximum entropy and from all the models that fit training data, selects the one which has the largest estimate probability. The maximum entropy classifier is usually used to classify text and this model can be shown in the following formula:

$$p(y|x) = \frac{\exp(\sum_k \lambda_k f_k(x,y))}{\sum \exp(\sum_k \lambda_k f_k(x,z))} \qquad (4)$$

where λ_k are model parameters and f_k are features of the model [18].

We trained the probability classification model with 2 classes which are Legal and General. After training, the classifier model was used to classify list of phrases in the phrase-table in target side, we consider these phrases to be in the general domain at the begining. Output of the classification task is probability of phrase in each domain *(P(legal) and P(general))*, some results of the classification task as in the figure 3.

[1]https://nlp.stanford.edu/software/classifier.html

phrases in p-table	P(legal)	P(general)
tội_phạm	0.991	0.009
hợp_pháp	0.551	0.449
pháp_lý	0.891	0.109
biểu_tượng	0.519	0.481
bộ_phận	0.688	0.312
cảnh_sát	0.986	0.014
hiệp_hội	0.633	0.367
hậu_quả	0.977	0.023
thẩm_quyền	0.986	0.014
tinh_vi	0.870	0.130
tình_huống	0.642	0.358
ảnh_hưởng	0.742	0.258
xử_lý	0.996	0.004
buộc_tội	0.951	0.049
hạn_chế	0.840	0.160
tình_huống	0.642	0.358
bất_hợp_pháp	0.996	0.004
phạm_pháp	0.930	0.070
trái_phép	0.690	0.310
thực_thi	0.976	0.024
thi_hành	0.938	0.062

Figure 3. Some result of the classification task

3.2.2. Phrase classification for domain adaptation in SMT

State-of-the-art SMT systems use a log-linear combination of models to decide the best-scoring target sentence given a source sentence. Among these models, the basic ones are a translation model P(e|f) and a target language model P(e).

The translation model is a phrase translation table; we have a table of the probabilities of translating a specified source phrase f into a specified target phrase e, including phrase translation probabilities in both translation directions, the example about structure of phrase translation table as figure 4.

In figure 4 the phrase translation probability distributions ϕ(f|e) and ϕ(e|f), lexical weighting for both directions. Currently, four different phrase translation scores are computed:

1. inverse phrase translation probability ϕ(f|e)
2. inverse lexical weighting lex(f|e)

```
confirm ||| khẳng_định ||| 0.0571429 0.0238095 0.769 0.142857
confirm ||| xác_minh được không ||| 1 0.0370849 0.2 0.000728863
confirm ||| xác_nhận ||| 0.0625 0.09375 0.2 0.214286
confirm ||| xác_nhận được ||| 1 0.0469315 0.2 0.0306122
consequences ||| hậu_quả ||| 0.240506 0.259494 0.965 0.465909
consequences ||| hệ_quả ||| 0.151515 0.206897 0.106383 0.136364
consequences ||| kết_quả ||| 0.00448431 0.0088889 0.0425532 0.0909091
copyright ||| bản_quyền ||| 0.543478 0.532258 0.994 0.66
copyright ||| bảo_hộ ||| 0.142857 0.037037 0.027027 0.02
crime ||| phạm_tội ||| 0.454545 0.181818 0.949 0.0930233
crime ||| tội_lỗi ||| 0.0238095 0.0178571 0.0149254 0.0116279
crime ||| tội_phạm ||| 0.381579 0.275862 0.994 0.372093
crime ||| tội_ác ||| 0.40625 0.348837 0.19403 0.174419
```

Figure 4. Example of phrase translation scores in phrase-table

3. direct phrase translation probability ϕ(e|f)
4. direct lexical weighting lex(e|f)

In this paper we only conduct the experiments with translation direction from English to Vietnamese, thus we only investigate the direct phrase translation probability ϕ(e|f) of the phrase translation table, the translation hypothesis is higher probability ϕ(e|f) value, that translation hypothesis will often be chosen more than another, so we use the probability classification model to determine the classification probability of a phrase in the phrase translation table, then we recompute the phrase translation probability ϕ(e|f) of this hypothesis base on the classification probability.

Our method can be illustrated as figure 5 and summarised by the follow:

1. Build a probability classification model *(using the maximum entropy classifier with two classes, legal and general)* with monolingual data on legal domain in Vietnamese
2. Trainning the baseline SMT system with parallel data on general domain with translation direction from English to Vietnamese
3. Extract phrases on target side in phrase translation table of baselise SMT system and using the probability classification model for these phrases
4. Recompute the direct translation probability ϕ(e|f) of phrases in the phrase translation table for phrases are classified into legal label.

4. Experimental Setup

In this section, we describe experimental settings and report empirical results.

4.1. Data sets

We conduct experiments on the data sets of the English-Vietnamese language pair. We consider two different domains that are legal domain and general domain. Detailed statistics for the data sets are given in table 1

Out-of-domain data: We use monolingual data on legal domain in the Vietnamese language, this data set is collected from documents, dictionaries,... consists of 2238 phrases, manually labeled, including 526 in-of-domain phrases *(in legal domain and label is lb_legal)* and 1712 out-of-domain phrases *(in general domain and label is lb_general)*. This data set is used to train the probability classification model by the maximum entropy classifier with 2 classes, legal and general.

Additionally, we use 500 parallel sentences on legal domain in English-Vietnamese pair for test set.

In-of-domain data: We use the parallel corpora sets on general domain to training SMT system. This data sets are provided by the IWSLT 2015 organisers for the English-Vietnamese translation task, consists of 122132 parallel sentences for tranning set, 745 parallel sentences for development set and 1046 parallel sentences for test set.

Preprocessing: Data preprocessing plays a very important role in any data-driven method. We carried out preprocessing in two steps:

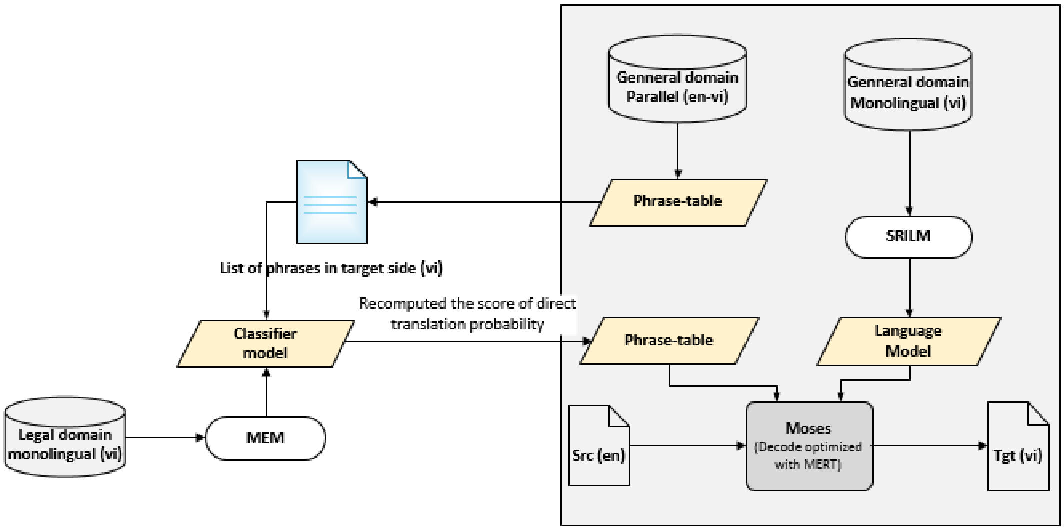

Figure 5. Architecture of the our translation model adaptation system

- ***Cleaning Data:*** We performed cleaning in two phases, *phase-1:* following the cleaning process described in [19] and *phase-2:* using the corpus cleaning scripts in Moses toolkit [20] with minimum and maximum number of tokens set to 1 and 80 respectively.

- ***Word Segmentation:*** In English, whitespaces are used to separate words [21] but Vietnamese does not have morphology [22] and [21]. In Vietnamese, whitespaces are not used to separate words. The smallest meaningful part of Vietnamese orthography is a syllable [23]. Some examples of Vietnamese words are shown as follows: *Single words:* "nhà" - house, "nhặt" - pick up, "mua" - buy and "bán" - sell. *Compound words:* "mua_bán" - buy and sell, "bàn_ghế" - table and chair, “cây_cối” - trees, “đường_xá” - street, "hành_chính" - administration. Thus, a word in Vietnamese may consist of several syllables separted by whitespaces.

  We used vntokenizer toolkit [24] to segment for Vietnamese data sets, this is the quite popular toolkit for Vietnamese segmentation and we used tokenizer script in Moses to segment for English data sets.

4.2. Experiments

We performed experiments on the Baseline_SMT and Adaptaion_SMT systems:

- The Baseline_SMT is a SMT baseline system. This system is the phrase-based statistical machine translation with standard settings in the Moses toolkit [2][20], this is a state-of-the-art open-source phrase-based SMT system. In our systems, the weights of feature funtions were optimized using MERT [25]. The Baseline_SMT is trained on the general domain *(in-of-domain)* data set and the Baseline_SMT system is evaluated sequentially on the General_test and Legal_test data sets.

- The Adaptation_SMT is based on the Baseline_SMT system after being adapted to the translation model by recomputing the direct translation probability $\phi(e|f)$ of phrases in the phrase

[2]http://www.statmt.org/moses/

Data Sets		Language	
		English	**Vietnamese**
Training	Sentences	**122132**	
	Average Length	15.93	15.58
	Words	1946397	1903504
	Vocabulary	40568	28414
Dev	Sentences	**745**	
	Average Length	16.61	15.97
	Words	12397	11921
	Vocabulary	2230	1986
General_test	Sentences	**1046**	
	Average Length	16.25	15.97
	Words	17023	16889
	Vocabulary	2701	2759
Legal_test	Sentences	**500**	
	Average Length	15.21	15.48
	Words	7605	7740
	Vocabulary	1530	1429

Table 1. The Summary statistical of data sets: English-Vietnamese

translation table, the Adaptaion_SMT is evaluated on the Legal_test data set.

We train a language model with 4-gram and Kneser-Ney smoothing was used in all the experiments. We used SRILM[3] [26] as the language model toolkit. For evaluate translation quality of the Baseline_SMT system and Adaptaion_SMT system, we use the BLEU score [27].

For comparison we also build a Neural Machine Translation (NMT) system use the OpenNMT toolkit[4] [28], this is a state-of-the-art open source neural machine translation system. Our the NMT system is trained with the default model, which consists of a 2-layer LSTM with 500 hidden units on both the encoder/decoder and the general attention type of Thang Luong [29].

4.2.1. Results

The table 3 showed that the baseline systems *(the SMT and NMT system)* is trained on the general domain data set, if the test data set *(here is the General_test data set)* is in the same domain as the training data, the BLEU score will be 31.3 for the Baseline_SMT system and 30.1 for the Baseline_NMT system. If the test data set is on the legal domain *(here is the Legal_test data set)*, the BLEU score will be 28.8 for the Baseline_SMT system and 20.9 for Baseline_NMT system.

The table 3 also showed that the SMT system is trained on the general domain if the test domain is different with the training domain, the quality of translation will be down. In this experiments, the BLEU score was reduced 2.5 points from 31.3 to 28.8. The Adaptaion_SMT system is adapted by our technique will improve the quality of the translation system. In this experiments, the BLEU score is improved to 0.9 points from 28.8 up to 29.7.

The experiment results also showed that the low-resource translation domains of english-vietnamese language pair such as legal domain and some other domains, the SMT system has better results than NMT system.

[3]http://www.speech.sri.com/projects/srilm/
[4]http://opennmt.net/

Source sentences ***(on the Legal_domain)***	**Target sentences**				**Reference sentences**
	Baseline_SMT system	**Adaptation_SMT system**	**NMT system**	**Google Translate**	
the working party ***took note*** of this commitment .	bữa tiệc làm_việc ***nốt_nhạc*** đã cam_kết này .	nhóm làm_việc ***ghi_nhận*** cam_kết này .	buổi tiệc đã nhận được sự cam_kết về những khiếm_khuyết của sự cam_kết này .	nhóm làm việc đã ***lưu ý*** về cam kết này.	ban công tác đã ***ghi_nhận*** cam_kết này .
according to the ***general statistical office*** , services had accounted for 37.98 percent of vietnam 's gdp in 2004	theo ***tổng_quát văn_phòng thống_kê*** , dịch_vụ đã chiếm 37.98% của việt_nam là gdp vào năm 2004 .	theo ***tổng_cục thống_kê*** , dịch_vụ đã chiếm 37.98% của việt_nam là gdp vào năm 2004 .	theo các ***văn_phòng thống_kê*** , dịch_vụ đã có những phần_trăm trong số gdp của việt nam trong sự kìm_kẹp của gdp ở dorset .	Theo ***cơ quan thống kê chung***, các dịch vụ đã chiếm 37,98% gdp của Việt Nam năm 2004.	theo ***tổng_cục thống_kê*** , dịch_vụ chiếm 37,98% gdp năm 2004 của việt_nam .
renewable certificates valid for five years were granted by the construction departments of cities and provinces .	giấy chứng_nhận ***tái_tạo*** có giá_trị trong 5 năm qua là cấu_trúc của các thành_phố và đại_lục .	giấy_phép ***gia_hạn*** có giá_trị trong 5 năm bởi cấu_trúc của các thành_phố và đại_lục .	những mẩu giấy ***tái_tạo*** có giá_trị trong 5 năm là do các nhà hoạt_động xây_dựng của các thành_phố và các quận biến_động .	Giấy chứng nhận ***tái tạo*** có giá trị trong năm năm được cấp bởi các sở xây dựng của thành phố và các tỉnh.	sở xây dựng các tỉnh và thành_phố cấp giấy_phép hành_nghề có hiệu_lực 5 năm và các giấy_phép này có_thể được ***gia_hạn*** .
the economic police received specialized training on intellectual property ***enforcement*** .	cảnh_sát kinh_tế được đào_tạo chuyên về ***cơ_quan*** sở_hữu_trí_tuệ .	cảnh_sát kinh_tế được đào_tạo chuyên về ***thực_thi*** quyền_sở_hữu_trí_tuệ .	cảnh_sát được đào_tạo chuyên về các ngăn ***vi_phạm*** sở_hữu_trí_tuệ .	cảnh sát kinh tế được đào tạo chuyên môn về ***thực thi*** sở hữu trí tuệ.	cảnh_sát kinh_tế được đào_tạo chuyên_sâu về ***thực_thi*** quyền_sở_hữu trí_tuệ .
administrative measures only applied to ***acts*** of low ***gravity*** .	***đo hành_chính*** chỉ áp_dụng cho ***hành_động*** của ***trọng_lực*** thấp .	***các biện_pháp hành_chính*** chỉ áp dụng cho ***các hành_vi nghiêm_trọng*** thấp .	các ***biện_pháp quản_lý*** chỉ áp_dụng vào ***những hành_động thấp*** của những thiên_thể thấp .	***biện pháp hành chính*** chỉ áp dụng cho các ***hành vi trọng lực*** thấp.	***các biện_pháp hành_chính*** chỉ áp_dụng với ***những hành_vi*** có tính ***nghiêm_trọng*** thấp .
evidence collected during an ***administrative procedure*** could be used by the civil court if necessary in accordance with ***civil procedure code*** of 2004 .	bằng_chứng thu_thập được trong một ***ca hành_chính*** có_thể được sử_dụng bởi những toà dân_sự nếu cần_thiết theo ***thủ_tục dân_sự mã*** của năm 2004 .	bằng_chứng thu_thập được trong ***thủ_tục hành_chính*** có_thể được sử_dụng bởi toà dân_sự nếu cần_thiết theo ***bộ_luật thủ_tục dân_sự*** của năm 2004 .	bằng_chứng được thu_thập trong một ***thủ_tục quản_lý*** có_thể được sử_dụng bởi toà_án dân_sự nếu cần_thiết trong ***hệ_thống dân_sự*** của năm 2004 .	bằng chứng thu thập trong ***một thủ tục hành chính*** có thể được sử dụng bởi tòa án dân sự nếu cần thiết theo ***quy tắc tố tụng dân sự*** năm 2004.	chứng_cứ thu được trong quá_trình ***xử_lý hành_chính*** sẽ được sử_dụng tại tòa dân_sự nếu thấy cần_thiết theo ***bộ_luật tố_tụng dân_sự*** năm 2004 .

Table 2. Some examples in our experiments

SYSTEM	BLEU SCORE
Baseline_SMT *(General_test)*	**31.3**
Baseline_SMT *(on Legal_test)*	**28.8**
Adaptaion_SMT *(on Legal_test)*	**29.7**
Baseline_NMT *(on General_test)*	**30.1**
Baseline_NMT *(on Legal_test)*	**20.9**

Table 3. The experiment results of Baseline_SMT and Adaptaion_SMT systems

4.2.2. Analysis and discussion

Some examples in the table 2 when translating source sentences in legal domain from english to vietnamese language. In the third sentence, the phrase ***"renewable"*** in context *"**renewable** certificates valid for five years were granted by the construction departments of cities and provinces" (source sentence column)* should been translated into ***"gia_hạn"*** as reference sentence but the Baseline_SMT system has translated the phrase ***"renewable"*** into ***"tái_tạo"***, the NMT system has translated that phrase into ***"tái_tạo"***, the Google Translate has translated that phrase into ***"tái tạo"*** and the Adaptaion_SMT system has translated the phrase ***"renewable"*** into ***"gia_hạn"*** like reference sentence.

The first, the Baseline_SMT system has translated the phrase ***"renewable"*** into ***"tái_tạo"*** because the direct translation probability *(4th column in figure 6)* of this phrase in phrase-table of Baseline_SMT system is highest *(0.454545)*, and the direct translation probability into ***"gia_hạn"*** is lower *(0.0909091)*. Therefore, when the SMT system combines component models as formulas 1, the ability to translate into ***"tái_tạo"*** will be higher ***"gia_hạn"***.

Later, apply the phrase classification model to compute the probability of

```
renewable ||| có_thể tái_tạo có ||| 1 0.0330019 0.0909091 0.00108034 ||| 0-0 0-1 ||| 1 11 1 ||| |||
renewable ||| có_thể tái_tạo ||| 0.0666667 0.0330019 0.0909091 0.0443787 ||| 0-0 0-1 ||| 15 11 1 ||| |||
renewable ||| gia_hạn ||| 0.25 0.0525026 0.0909091 0.0295858 ||| 0-0 0-1 ||| 4 11 1 ||| |||
renewable ||| năng_lượng sạch ||| 0.0625 0.00366395 0.0909091 0.00147929 ||| 0-0 0-1 ||| 16 11 1 ||| |||
renewable ||| tái khôi_phục ||| 1 0.0160714 0.0909091 0.00147929 ||| 0-0 0-1 ||| 1 11 1 ||| |||
renewable ||| tái khôi_phục được ||| 1 0.0160714 0.0909091 3.90215e-05 ||| 0-0 0-1 ||| 1 11 1 ||| |||
renewable ||| tái_tạo ||| 0.0641026 0.0657895 0.454545 0.384615 ||| 0-0 ||| 78 11 5 ||| |||
```

Figure 6. Examples about the direct translation probability of this phrase in phrase-table

"gia_hạn" and ***"renewable"*** phrase in legal domain, the probability of ***"gia_hạn"*** is higher than that, then update this value to phrase-table and the direct translation probabilitys $\phi(e|f)$ of phrase are recomputed. Therefore, the Adaptation_SMT has translated ***"renewable"*** phrase into ***"gia_hạn"***

Some other examples in the table 2 showed that translation quality of Adaptaion_SMT system is better than the Baseline_SMT system and with low-resource translation domains in English-Vietnamese language, the SMT system has more advantages than the NMT system.

5. Conclusions and Future Works

In this paper, we presented a new domain adaptation method in Statistical Machine Translation for low-resource domains in English-Vietnamese language. Our method only uses monolingual out-of-domain data to adapt the phrase translation table by recomputing the phrase's direct translation probability $\phi(e|f)$. Our system obtained a improved on the quality of machine translation in the legal domain up to 0.9 BLEU points over baseline. Experimental results show that our method is effective in improving accuracy of the translation.

In future works, we intend to study this problem in the other domains, the benefits of word embeding in phrase classification and integrate automatically our technique to decode of SMT system.

References

[1] P. Koehn, F. J. Och, , D. Marcu, Statistical phrase-based translation., In Proceedings of HLT-NAACL 2003, pages 127–133. Edmonton, Canada, 2003.

[2] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean, Google's neural machine translation system: Bridging the gap between human and machine translation, CoRR abs/1609.08144.
URL `http://arxiv.org/abs/1609.08144`

[3] L. Bentivogli, A. Bisazza, M. Cettolo, M. Federico, Neural versus phrase-based machine translation quality: a case study, 2016.

[4] B. Haddow, P. Koehn, Analysing the effect of out-of-domain data on smt systems., In Proceedings of the Seventh Workshop on Statistical Machine Translation, pages 422–432, 2012.

[5] P. Banerjee, J. Du, B. Li, S. K. Naskar, A. Way, J. van Genabith, Combining multi-domain statistical machine translation models using automatic classifiers., In Proceedings of AMTA 2010., 2010.

[6] G. Foster, R. Kuhn, Mixture-model adaptation for smt, Proceedings of the Second Workshop on Statistical Machine Translation, pages 128–135, Prague. Association for Computational Linguistics, 2007.

[7] G. Foster, B. Chen, R. Kuhn, Simulating discriminative training for linear mixture adaptation in statistical machine translation, Proceedings of the MT Summit, 2013.

[8] D. Dahlmeier, H. T. Ng, S. M. Wu4, Building a large annotated corpus of learner english: The nus corpus of learner english., In Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Appli- cations., 2013.

[9] B. Chen, R. Kuhn, G. Foster, Vector space model for adaptation in statistical machine translation, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1285–1293, 2013.

[10] E. Hasler, P. Blunsom, P. Koehn, B. Haddow, Dynamic topic adaptation for phrase-based mt., In Proceedings of the 14th Conference of the European Chapter of The Association for Computational Linguistics, pages 328–337., 2014.

[11] R. Masumura, T. Asam, T. Oba, H. Masataki, S. Sakauchi, A. Ito, Hierarchical latent words language models for robust modeling to out-of domain tasks, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1896–1901, 2015.

[12] L. Onrust, A. van den Bosch, H. V. hamme, Improving cross-domain n-gram language modelling with skipgrams, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 137–142, Berlin, Germany, 2016.

[13] H. Cuong, K. Sima'an, I. Titov, Adapting to all domains at once: Rewarding domain invariance in smt, Proceedings of the Transactions of the Association for Computational Linguistics (TACL), 2016.

[14] M. Freitag, Y. Al-Onaizan, Fast domain adaptation for neural machine translation, 2016.

[15] C. Chu, R. Dabre, S. Kurohashi, An empirical comparison of simple domain adaptation methods for neural machine translation, 2017.

[16] H. Wu, H. Wang, C. Zong, Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora., In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK, pp.993–1000., 2008.

[17] J. Xu, Y. Deng, Y. Gao, H. Ney, Domain dependent statistical machine translation, In Proceedings of the MT Summit XI, pages 515–520, 2007.

[18] A. Berger, S. D. Pietra, V. D. Pietra, A maximum entropy approach to natural language processing, Computational Linguistics, 22, 1996.

[19] S. Pal, S. Naskar, J. van Genabith, Uds-sant: English–german hybrid machine translation system, In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 152–157, Lisbon, Portugal, September. Association for Computational Linguistics, 2015.

[20] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, Moses: open source toolkit for statistical machine translation., In ACL-2007: Proceedings of demo and poster sessions, Prague, Czech Republic, pp.177-180., 2007.

[21] M. Aronoff, K. Fudeman, What is morphology, vol. 8. john wiley and sons., 2011.

[22] L. C. Thompson, The problem of the word in vietnamese., Word journal of the International Linguistic Association, 19(1):39–52, 1963.

[23] B. N. Ngo., The vietnamese language learning framework, Journal of Southeast Asian Language Teaching 10. Pages 1–24, 2001.

[24] L. H. Phuong, N. T. M. Huyen, A. Roussanaly, H. T. Vinh, A hybrid approach to word segmentation of vietnamese texts, 2008.

[25] F. J. Och, Minimum error rate training in statistical machine translation., In Proceedings of ACL 2003, pages 160-167, 2003.

[26] A. Stolcke, "srilm - an extensible language modeling toolkit," in proceedings of international conference on spoken lan-guage processing, 2002.

[27] Papineni, Kishore, S. Roukos, T. Ward, W. Zhu, Bleu: A method for automatic evaluation of machine translation, ACL, 2002.

[28] G. Klein, Y. Kim, Y. Deng, J. Senellart, A. M. Rush, OpenNMT: Open-Source Toolkit for Neural Machine Translation, ArXiv e-prints`arXiv:1701.02810`.

[29] T. Luong, H. Pham, C. D. Manning, Effective approaches to attention-based neural machine translation, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421,, 2015.