Transliterating Nôm Scripts into Vietnamese National Scripts using Statistical Machine Translation

Dien Dinh[1], Phuong Nguyen[2], Long H. B. Nguyen[*3]
University of Science, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam

***Abstract*—Nôm scripts were used as the Vietnamese writing system from the 10th century to the early 20th century. During this period, Nôm scripts were the means to record a broad range of historical events, literary works, medical knowledge, as well as wisdom of many other domains. Unfortunately, since hardly any native Vietnamese speaker can read Nôm scripts nowadays, these valuable documents have not been fully harnessed. To address this gap, it is necessary to build an automatic transliteration system that can support us in decoding the ancient scripts and gaining knowledge of our Vietnamese ancestors. This study focuses on categorizing and reviewing the current progress on the Statistical Machine Translation (SMT) approaches to transliterate Nôm scripts into Vietnamese national scripts. In this paper, we discuss the differences between Nôm scripts and Vietnamese national scripts, systematically compare SMT models in transliterating Nôm scripts into Vietnamese national scripts, as well as having a thorough outlook on several promising research directions.**

***Keywords*—Statistical machine translation; automatic transliteration; Nôm Script (chữ Nôm); vietnamese national script (chữ Quốc ngữ)**

I. Introduction

Transliteration is a type of conversion of a text from one script to another, in the same language. For instance, the Cyrillic scripts of the Russian language, "Путин", is transliterated into the Latin scripts as "Putin". This transliteration is relatively straightforward, because there is only one correspondence in the Latin scripts for most of the letters in the Cyrillic scripts. Since both scripts are based on alphabets that contain a limited number of graphemes (strokes) to represent speech, transliteration can be done by looking up the mapping table. Table I is a portion of the mapping table from Cyrillic scripts to Latin scripts.

Table I. A Portion of the Mapping Table from Cyrillic Scripts to Latin Scripts

Cyrillic letter	Latin letter
А а	A a
Б б	B b
И и	I i
Н н	N n
П п	P p
Т т	T t
У у	U u

On the contrary, transliteration from Nôm scripts to Vietnamese national scripts is challenging because they do not belong to the same writing system. While Nôm scripts belong to the logographic writing system, Vietnamese national scripts belong to the alphabetic writing system. In other words, Nôm - Vietnamese national scripts is the one-to-many relationship. For instance, the Nôm character 𠊛 can be transliterated into *nghĩ* or *nghỉ*. Due to differences between the two writing systems, the mapping table method presented in the aforementioned Russian language example is not applicable when transliterating from Nôm scripts into Vietnamese national scripts.

The one-to-many mapping from Nôm scripts to the Vietnamese national scripts causes difficulties in transliterating process because people have to simultaneously read Nôm text and guess the appropriate meaning. Successful Nôm-transliteration also requires extra-linguistic knowledge about the culture, history, geography, dialects, specialized terminologies of ancient Vietnam. In recent years, rich-resource languages have gained success in applying machine translation. Chinese [1] and many European languages, including German [2], Greek [3], English [4], Spanish [5], French [6], Finnish [7], Italian [8], Dutch [9], and Portuguese [10] are some of those rich-resource languages. Besides, research in low-resource Southeast Asian languages such as Indonesian [11], Khmer [12], Lao [13], Malay [14], Myanmar [15], Philippines [16], and Thai [17], also yields significant results, which motivates us to apply machine translation in transliterating Nôm scripts into Vietnamese national scripts. Two state-of-the-art approaches in machine translation are Statistical Machine Translation (SMT) and Neural Machine Translation (NMT). However, NMT requires a large amount of data [18], which is impractical for the low-resource language pair Nôm - Vietnamese national scripts. Therefore, we apply SMT for the transliterating task in this study.

Given the mechanism, the larger the manually transliterated training data are given to the computer, the more accurate the transliteration the computers generate. Besides, the machine can also improve the transliteration accuracy if humans supervise and manually revise the incorrect results that the computers previously produce. The more times we repeat the supervising and revising loop, the better the transliteration results become.

In this paper, we present the automatic transliteration from Nôm scripts to Vietnamese national scripts using Statistical Machine Translation. Our research steps are as the following: (1) collect and clean (i) the Nôm-Vietnamese national parallel corpus as the training data for the translation model and (ii) the monolingual Vietnamese national scripts as the training data for the language model, (2) classify corpora according to literary forms and domains, (3) experiment, and (4) analyze the experimental results.

Our main contributions are:

- Providing detail background about Nôm scripts

- Systematically comparing Nôm scripts with Chinese scripts
- Experimenting to show significance of the translation models in transliterating Nôm scripts into Vietnamese national scripts

The remaining of the paper is organized as follows: in Sections II and III, we provide an overview of Nôm scripts and of related studies, respectively. Then, we present our proposed model in Section IV and discuss the experimental results in Section V. Section VI concludes the study.

II. Overview of Nôm Scripts

Nôm scripts were created based on Chinese characters, which results in various similarities between Nôm scripts and Chinese scripts. Different from all phonological recording systems, Chinese scripts are the only logographical writing system currently used in the world [19]. Regarding the phonological writing system, there are symbols that record phonemes of a language. Meanwhile, Chinese characters are used to mark morphemes, ideas, basic concepts such as the sun (日), moon (月), tree (木), human (人), water (水), and heart (心). These basic elements are called radicals (部首). Radicals are the building blocks from which Chinese characters (Hanzi - 汉字) are built. According to the Han dictionary Shuowen (说文), there are six methods (六书) of constructing Chinese characters, including:

- Pictograms (象形): 日 (sun), 月 (moon), 木 (tree), etc.
- Ideogram (指事): 上 (above), 一 (one), 本 (root), etc.
- Combined ideogram (会意): 信 (trust), 林 (woods), 森 (forest), etc.
- Ideogram plus phonetic (形声): 妈 (mother) with 马(/mǎ/) as phonetic element and 女(/nǚ/) as ideographic element, etc.
- Derivative cognates (转注): 少 (a few - thiểu/to lack - thiếu), etc.
- Rebus (假借): 自 (oneself) which loans from the character 鼻(nose), etc.

Among these six types of characters, 90 percent belong to the ideogram-plus-phonetic category [19], i.e., each character is a *morpheme-syllable* compound. Meanwhile, Vietnamese language is constituted by *morpho-syllables*, which means *units* that constitute the two writing systems are equivalent. In the Chinese language, morphemes are radicals because a radical is the smallest meaningful unit of the Chinese writing system. Radicals are also the basis to arrange entries in Chinese dictionaries. For instance, to look up for the Chinese character 妈 (mother), we first search for the radical 女 (woman), since the character 妈 contains the radical 女. Then, we look up the remaining component, 马, by the number of strokes, which is three.

According to [20], while there are about 10,000 distinct pure morpho-syllables (not including transliterated morpho-syllables of loan words or scripts of ethnic languages) in Vietnamese, there are approximately 13,000 distinct Chinese characters (not including ancient characters, characters used for translitering loan words) in Chinese. Also from [20], each Chinese character has its own Unicode; the Chinese Unicode Charset is constructed based on various Chinese encoding charsets such as Big5 and GB; these encoding systems are gathered and aggregated into Unicode CJK charset; the first version of CJK was released in 1980 with roughly 13,000 Chinese characters; the number of encoded Chinese characters has grown over the years and reached 80,000 in 2018.

Most of the Nôm scripts were also created in the form of semantic (meaning)-phonetic (sound) compounds. The ancient Vietnamese usually borrowed two elements - one element for meaning and the other for the sound - from Chinese character collection to construct a Nôm character. For instance, the Nôm character 𠄩 means *number three*. In the Nôm character 𠄩, the Chinese character 巴, which has pinyin /bā/, denotes the sound, while the Chinese character 三 indicates the meaning. Similarly, in the Nôm character 爸, which means *father*, the Chinese character 巴 signifies the sound, while the Chinese character 父 expresses the meaning. Apart from the aforementioned semantic-phonetic compounds, there are a number of Nôm characters created by other methods, such as rebus, repetition, transfer, and diacritics adding. These methods signify the phonological difference between Nôm and Chinese characters [21].

Because Nôm scripts are mainly built on the semantic-phonetic compound method, there are cases in which one Nôm character is mapped to two or more Vietnamese national scripts. This typically happens when the national scripts have similar pronunciation and indicate synonymous meanings. This phenomenon can be explained by linguistic characteristics. While Vietnamese and Chinese languages are both tonal languages, they do not have the same number of tones. In particular, while there are six tones corresponding to six diacritics in Vietnamese, there are only four tones in Chinese. Moreover, different script creation methods, regional dialects, and Sino-Vietnamese variants due to different times of adoption also account for the one-to-many mapping between Nôm scripts and Vietnamese national scripts. For instance, the Nôm character 味 has two corresponding national scripts. The first one corresponds to *mùi* (smell), as it was adopted before the Tang Dynasty. Meanwhile, the second one corresponds to *vị* (flavor) as it was adopted from the Tang Dynasty onwards [22]. In the Nôm-Vietnamese national scripts dictionary[1], a considerable number of Nôm characters are **polyphonic** (a *polyphonic* Nôm character has more than one corresponding Vietnamese national script). For example, character 折 (Unicode code 6298h) has 19 corresponding national scripts (chệch, chét, chết, chẹt, chiết, chít, chịt, díp, gãy, gẩy, giẹp, giết, giỡn, nhét, nhít, siết, trét, triếp, xiết). This is also the one with the highest number of meanings in the Nôm-Vietnamese national scripts dictionary. In contrast, each **monophonic** Nôm character has only one corresponding Vietnamese national script. Table II are examples of polyphonic Nôm characters.

From the Nôm-Vietnamese national scripts dictionary, which includes 22,264 entries, we can classify Nôm characters according to the number of Vietnamese national scripts of each Nôm character, details are in Table III. According to the first row of the Table III, monophonic Nôm characters account for 76.654 percent of all dictionary entries. So, the remaining

[1]Hanosoft3. [Online]. Available: https://hanosoft-3-0-hanokey-2010.soft112.com. Accessed Jan 2019.

TABLE II. EXAMPLES OF POLYPHONIC NÔM CHARACTERS

Nôm script	Vietnamese national scripts	Quantity of Vietnamese national scripts
一	nhất, nhứt	2
丁	đinh, đĩnh	2
丐	cái, gái	2
万	muôn, vàn, vạn	3
与	dữ, dự, đử	3
丑	sấu, xấu, sửu	3
且	thả, vã, vả	3
世	thá, thé, thế, thể	4
中	đúng, trong, trung, trúng, truồng	5
丞	bậy, chăng, chẳng, phi, phỉ, vầy, vậy	7

23.346 percent are polyphonic Nôm characters. The proportion of polyphonic Nôm characters in comparison with monophonic Nôm characters is presented in Fig. 1.

TABLE III. FREQUENCY OF NÔM CHARACTERS ACCORDING TO THE NUMBER OF THEIR CORRESPONDING VIETNAMESE NATIONAL SCRIPTS

Quantity of corresponding national scripts	Quantity of Nôm character (Frequency)	Proportion
1	11,610	76.654%
2	1,907	12.591%
3	787	5.196%
4	384	2.535%
5	209	1.380%
6	94	0.621%
7	66	0.436%
8	30	0.198%
9	20	0.132%
10	16	0.106%
11	11	0.073%
12	3	0.020%
13	3	0.020%
14	2	0.013%
15	1	0.007%
16	1	0.007%
18	1	0.007%
19	1	0.007%
Total	**15146**	**100%**

Choosing the suitable Vietnamese national script for a given Nôm character is a difficult problem not only for the machine but also for the transliterators. Consider the Nôm character 𠊛 (Unicode code 2025Dh), which appears in the 12th sentence of Tale of Kieu in Fig. 2. 𠊛 might be transliterated into two national scripts as *nghĩ* (to think) or *nghỉ* (a pronoun used to indicate an old man in ancient Vietnamese) [23]. Scholars have been debating for over 50 years on which national script is correct in the given situation. Both sides provide various arguments, historical evidence, and literary evidence, etc. to demonstrate why one out of the two national scripts would be more suitable than the other. Therefore, requiring a computer to generate a 100-percent accurate transliteration output is impracticable, at least at present time and in near future.

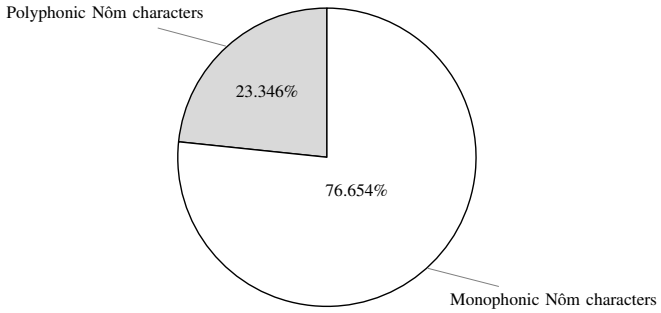

Figure 1. Proportion Polyphonic Nôm Characters versus Monophonic Nôm Characters (based on Table III).

家	資	𠊛	拱	常	常	堛	中
Gia	tư	**nghĩ /nghỉ**	cũng	thường	thường	bậc	trung

Figure 2. The 12th Sentence in Tale of Kieu by Nguyen Du.

III. RELATED WORKS

The digitization of Nôm scripts has been proposed and implemented since the 1990s by Ngo Thanh Nhan, Nguyen Quang Hong, among other scholars[2]. Thanks to these contributors, most of the common Nôm characters have become a part of the Unicode encoding system. This significant work is a solid foundation for lateral digitizing steps, such as storage, lookup, processing, automatic transliteration, etc.

Moreover, *Việt Hán Nôm 2002*, a software developed by Phan Anh Dung[3], allows us to type and look up both Chinese and Nôm characters. Another software, Hanosoft, developed by Tong Phuoc Khai[4], also includes several utilities for looking up and transliterating from Chinese characters into Nôm characters. In the aforementioned software, the authors have developed a tool to automatically transcribe Chinese characters into Sino-Vietnamese, Chinese characters into pinyin, and Nôm characters into national scripts. However, the central issue of the problem, which is choosing the proper National script for a given polyphonic Nôm character, has not yet been addressed. The software just randomly selects a Sino-Vietnamese phonetic transcript or a phonetic transcript among all possibilities. Besides, the website of the Vietnamese Nôm Preservation Foundation[5] includes a Chinese character-Nôm lookup tool and a digital library of Nôm documents, most of

[2]http://dir.vietnam.online.fr/home/vnChuNom.htm. Accessed May 2014.

[3]http://www.hannom.org.vn/detail.asp?param=507&Catid=363. Accessed Jun 2006.

[4]https://hanosoft-3-0-hanokey-2010.soft112.com. Accessed Jan 2019.

[5]http://www.nomfoundation.org. Accessed Oct 2019.

which are images of hand-written Nôm. Some literary works have also been digitized.

The work that is most closely related to our study is the Nôm converter[6], which is a toolkit used to automatically transliterate Nôm scripts to national scripts and vice versa. The system applies Statistical Machine Translation (SMT) approach and is based on Moses [24]. The data sets used to train Moses are parallel corpora. These corpora are 22 manually transliterated texts corresponding to 3,234 lines in total. The tool works fine, except for some cases in which input contains strange untrained Nôm scripts. For those cases, Nôm converter just ignores the strange untrained scripts and transliterates the rest of the input scripts as normal. Nôm converter has a rather high rate of choosing the correct national script when compared with the referenced transliteration version carried out by humans. To the best of our knowledge, it may be considered as the first automatic Nôm-Vietnamese national script transliteration tool that utilizes machine learning technology. Our approach is similar to Nôm converter, but with new modifications and improvements to address the limitations of the existing system.

IV. Proposed Model

In our proposed model, we customized a Statistical Machine Translation model (SMT) and improved the transliteration accuracy based on our work in automatic translation from English into Vietnamese [25]. Instead of following the Nôm converter system's approach in transliterating both directions (from Nôm scripts into Vietnamese national scripts and vice versa), we only focused on one-way transliteration from Nôm scripts into Vietnamese national scripts. Our core aim is to harness the Vietnamese ancient Nôm text, and the transliteration from national scripts to Nôm scripts does not imply as much practical significance. Besides, focusing on a one-way transliteration from Nôm scripts into national scripts allows us to invest more in improving national script output through various language models.

To overcome the shortage of parallel corpora for training as in the Nôm converter system, we added a Sino-Vietnamese dictionary into the phrase table of the Moses system. To improve the accuracy, we also added more manually transliterated literary works that Nôm converter has not yet included. Our major contribution is categorizing the Nôm script input data and providing language models for the Vietnamese national script output. The most challenging issue that we observed in transliterating Nôm script into Vietnamese national script was choosing the correct national script among all possibilities. This selection depends on context, form, domain, and even on the chronology of the input data. Nôm converter merely selects the national scripts according to the context in the training dataset, which is mixed in terms of form, domain, and time. Therefore, we classified the training dataset and language models by form and domain in our proposed model.

Because each form has its own rules for choosing the national script output, we classified the form into two categories: verse (such as Tale of Kieu, Tran Te Xuong's poems, etc.) and prose (The legend of Quynh, Biography of Phan Boi Chau, etc.). That is, these two forms required different language models. Besides, we also built corpora for three different domains, which were literature, history, and religion. New domains will be added into the current list of domains if we constructed and developed more corpora. Since each domain has its own terminologies, determining the domain to which the input scripts belong helped us narrow down the domain of possible national script output to improve the possibility of selecting the correct national scripts, especially for the cases in which the input is polyphonic Nôm scripts.

The final step was to build language models in the target language which was Vietnamese national scripts. The principle of machine learning is that the more training data we feed into the model, the better the transliterating accuracy will become. Due to this reason, not only did we utilize the national script dataset available in the parallel corpora that were used to train Moses in the previous step, but we also provided additional national script data that were already categorized by form and domain. This step improved the accuracy of the proposed model significantly since we included hundreds of thousands of sentences to build language models, compared to only thousands of sentences in the parallel corpora. A larger dataset of N-gram language models also allows the machine to generate the most linguistically natural transliteration output.

Later, when our proposed model is put into use, users will be able to select the form and domain of input data they want to transliterate from a menu. According to users' selection, computers will use the corresponding knowledge they have been trained to fit the form and domain of input data.

Let n be the source language sentence (Nôm scripts) and q be the target language sentence (Vietnamese national scripts), we have the following equation of the SMT model:

$$\hat{q} = \underset{q}{\operatorname{argmax}}\, P(q)P(n|q) \tag{1}$$

Through equation (1), the SMT model's working process is as follows:
i) estimate probability of seeing target string q language models $P(q)$;
ii) estimate probability that the source string n is the translation of the target string q given the translation model $P(n|q)$;
iii) choose the sentence q so that the value of product $P(q)P(n|q)$ is the maximum.
Fig. 3 shows a complete statistical translation system.

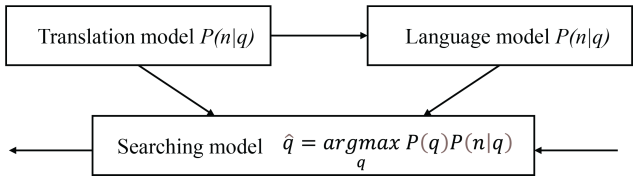

Figure 3. Transliteration Model

V. Experimental results and development directions

In this section, we present experimental results on our proposed model and compare the transliteration output with the baseline system Nôm converter. Then, we show limitations of

[6]Nôm converter. [Online]. Available: https://chunom.org/pages/moses/. Accessed Oct 2019.

our proposed model and corresponding development directions to improve those limitations.

A. Experiments and Results

In this sub-section, we describe the training data and experimental results of the proposed model.

1) Training, testing, and tuning datasets: We used single-character dictionaries listed in Table IV and compound-character dictionaries listed in Table V. In single-character dictionaries, each entry is one morpho-syllable such as 乂 - một (one), 仡 - là (to be). Meanwhile, entries in compound-character dictionaries have at least two morpho-syllables such as 乂匹 - một vài (several), 祝㦖辭潤 - chúc mừng năm mới (happy new year).

TABLE IV. SINGLE-CHARACTER DICTIONARIES USED FOR TRAINING

ID	Description	Size (Entries)	Source
1	Nôm - National Script dictionary	22264	Hanosoft3
2	Sino - Vietnames dictionary	16402	Hanosoft3
3	Sino (simplified) - Vietnames dictionary	10758	Hanosoft3
4	Sino (traditional) - Vietnames dictionary	11285	Hanosoft3
5	Nôm - National Script dictionary	32838	www.hannom-rcv.org
6	Nôm - National Script dictionary	879	www.chunom.org
7	Tale of Kieu (1902) dictionary	3353	www.nomfoundation.org
Total		97779	
After removing duplicates		**38897**	

TABLE V. COMPOUND DICTIONARIES USED FOR TRAINING

ID	Description	Size (Entries)	Source
1	Nôm - National Script dictionary	1951	www.chunom.org
2	Nôm - National Script dictionary	4520	www.hannom-rcv.org
Total		6471	
After removing duplicates		**6205**	

The Tale of Kieu (1902) dictionary in Table IV is not a publicly available dictionary. We observed there are Nôm characters in Tale of Kieu that have not been included in the other six single-character dictionaries. Therefore, we manually created the Tale of Kieu (1902) dictionary by listing all distinct pairs of Nôm and Vietnamese national scripts. We then utilized a computer program to aggregate all dictionaries listed in Table IV and Table V and removed the duplicate entries afterwards.

Regarding parallel sentences in the training and testing datasets, we used the corpus documents available on the websites chunom.org[7], Vietnamese Nôm Preservation Foundation[8], Việt Hán Nôm[9], and han-nom.org[10]. Details about domain and literary form of those sentences are listed in Table VI.

TABLE VI. NÔM-NATIONAL SCRIPT PARALLEL TEXTS IN TRAINING, TESTING, AND TUNING DATASETS

ID	Domain	Form	Size (Sentences)
1	Literature	Verse	7232
2	Literature	Prose	521
3	Religion	Verse	46
4	History	Verse	121
Total			7920

Currently, we have not utilized the domain information yet because the majority of the dataset belongs to the Literature domain. Classification of data is only useful if we have a considerably large corpus of various domains. Although we are not using categorizing information at the moment, we still include it into the program as a foundation for future work. We also collected corpora written in Vietnamese national scripts on the websites Gác Sách [11], Sách Phật giáo [12], and Ô Cửa Sổ [13]. Domain and form of monolingual corpora are listed in Table VII.

TABLE VII. MONOLINGUAL CORPORA USED TO TRAIN LANGUAGE MODEL

ID	Domain	Form	Size (Sentences)
1	History	Prose	257269
2	Religion	Prose	83535
3	Literature	Verse	29383
Total			370187

The training, testing, and tuning datasets are splitted by the ratio 1:1:8 as follows: for each text in Table VI, roughly 1/10 is distributed to the testing set, 1/10 is distributed in the tuning set, the remaining 8/10 is for the training set.

2) Experimental results: Using Moses SMT system [24], we conducted experiments on the corpora previously discussed and yielded the results in the Table VIII. From henceforth, Experiment 1 is abbreviated as Exp1, and Experiment 2 as Exp2, etc.

In Exp1, we measured the impact of the parallel corpus and the monolingual language model corpus on transliteration results with a single-character dictionary. With only a

[7]https://www.chunom.org. Accessed Oct 2019.
[8]http://www.nomfoundation.org. Accessed Oct 2019.
[9]http://hannom.huecit.vn. Accessed Oct 2019.
[10]http://www.han-nom.org. Accessed Oct 2019.
[11]http://www.gacsach.com. Accessed Jan 2020.
[12]http://www.sachphatgiao.net. Accessed Jan 2020.
[13]http://www.ocuaso.com. Accessed Jan 2020.

single-character dictionary equipped, the transliteration system behaves like a human with dictionaries, looking up a Nôm character in the dictionary and writing down a corresponding Vietnamese national script of that Nôm character. No linguistic knowledge was used and barely any understanding of Nôm script was required. The difference lies in the time it takes to transliterate Nôm script to national script. Humans might take days or weeks or even months to manually look up all Nôm characters in 786 lines of Nôm-script. However, machines take less than an hour if Moses is run on a high-spec machine. Since no linguistic knowledge was applied to the transliteration, the BLEU score [26] was 13.32. The results were acceptable, given the ambiguous nature of Nôm script. Because the model could not determine the context surrounding the input Nôm scripts, it could not choose the correct national scripts to generate a fluent output.

In Exp2, the data was the same as in Exp1, but the model has been tuned for better transliteration quality. The resulting BLEU score was 14.56, which was slightly better than the previous experiment's result.

In Exp3, we measured the impact of the language model by adding 370,817 lines of Vietnamese national script to train the model instead of using the default national scripts extracted from the parallel corpus. This made a significant difference, as the BLEU score increased from 13.32 to 63.89. This was because the language model supported phrase-based translation and provided context for the transliteration model to choose the most likely national script for a given Nôm script.

In Exp4, we tuned the model from Exp3, and the BLEU score increased from 63.89 to 65.94.

In Exp5, we added 6205 entries of compound dictionaries, growing the parallel corpus compared to that of Exp1. Consequently, the BLEU score increased from 13.32 to 36.82.

In Exp6, we tuned the model from Exp5, and the BLEU score increased from 36.82 to 44.24.

In Exp7, we added 370,817 lines of Vietnamese national script to train language model. Compared to the results in Exp5, the BLEU score in this experiment increased from 36.82 to 67.19.

In Exp8, we tuned the model from Exp7, and the BLEU score increased from 67.19 to 69.16.

In Exp9, we added 6,348 pairs to the parallel corpus. Compared to Exp1 and Exp5, the BLEU score increases from 13.32 to 36.82 to 80.50.

In Exp10, we tuned the model from Exp9, and the BLEU score increased from 80.50 to 80.83, which was not a considerable difference.

In Exp11, we added 370,817 lines of Vietnamese national script to train language models. Compared to Exp9, the BLEU score increased from 80.50 to 82.30. In this case, since we already had parallel corpus with long sentences of national script, adding a language model corpus did not yield a significant difference as in Exp3 and Exp7.

In Exp12, we tuned the model from Exp11, and the BLEU score increased from 82.30 to 85.38.

In the last four experiments, from Exp13 to Exp16, we used the parallel corpus without dictionaries to train the model and got acceptable results. However, there was similarity between the training corpus and the testing corpus, so the BLEU score was quite high, ranging from 75.71 to 79.40.

We hypothesized that if the testing data contain Nôm scripts from other domains such as medicine or agriculture, which have not been in the training data set yet, then the models with dictionaries will work better. This was based on the assumption that even though lacking the context, dictionaries cover a broader scope of vocabularies. However, that missing context could be made up by the additional language model as in Exp3 and Exp4, where the BLEU scores were 63.89 and 65.94 respectively. Those results were acceptable given that we trained the model only with dictionaries and additional language model data, without any parallel sentences. At the moment, we did not have data to verify our hypothesis. Verifying this hypothesis will be put in our future work, when we collect data from various other domains.

The corpora we used to train and test the transliteration system included single-character dictionaries, compound-character dictionaries, and parallel pairs of Nôm-Vietnamese national script sentences, whose model was trained by dictionaries. Parallel sentences were separated with the ratio of train:tune:test as 8:1:1 (tune here refers to the data used to tune the model, that is, to find the optimal parameters for the transliteration model).

In the third column of Table VIII, "Default" refers to the monolingual national scripts extracted from the parallel corpus in the second column. In experiments with "Default" monolingual corpus, we did not use additional language model corpora in Table VII. In the fourth column of Table VIII, BLEU stands for Bi-Lingual Evaluation Understudy, a metric used to measure quality of machine translation output in comparison to human-generated output. Format of BLEU score is **overall**, uni-gram/2-gram/3-gram/4-gram. The fifth column signifies whether an experiment was tuned or not. As mentioned previously, the purpose of tuning is to find optimal parameters for the transliteration model, and thereby generating better transliteration output, compared to the untuned model.

After training the model, we chose 10 percent of the sentences in the testing data to evaluate the proposed model. Exp12 yielded the highest BLEU score, which was 85.38. Therefore, we selected some sentences in the testing set of this experiment to compare to the corresponding output generated by Nôm converter. 12 sentences from Tale of Kieu (version 1902) were tested, and the results are presented in Table IX.

We use different typefaces to distinguish between correct and incorrect transliteration. The differences are explained as follows:

- **Compared transliteration**
- Correct transliteration
- *Synonymical transliteration*
- ***Incorrect/un-handled transliteration***

TABLE VIII. EXPERIMENT RESULTS

Exp ID	Training Data		BLEU	Tuned
	Parallel Corpus	Monolingual Corpus		
1	Table IV	Default	**13.32**, 44.6/19.6/8.9/4.0	No
2	Table IV	Default	**14.56**, 47.1/22.1/9.6/4.5	Yes
3	Table IV	Table VII	**63.89**, 84.4/69.3/57.6/49.6	No
4	Table IV	Table VII	**65.94**, 83.4/71.3/60.6/52.5	Yes
5	Table IV, Table V	Default	**36.82**, 67.5/45.6/29.7/20.1	No
6	Table IV, Table V	Default	**44.24**, 71.9/52.1/37.5/27.3	Yes
7	Table IV, Table V	Table VII	**67.19**, 85.0/72.3/61.4/54.0	No
8	Table IV, Table V	Table VII	**69.16**, 85.3/74.2/64.1/56.5	Yes
9	Table IV, Table V, 6348 sentence-pairs	Default	**80.50**, 91.3/84.1/76.7/71.4	No
10	Table IV, Table V, 6348 sentence-pairs	Default	**80.83**, 91.5/84.5/77.2/71.6	Yes
11	Table IV, Table V, 6348 sentence-pairs	Table VII	**82.30**, 92.2/85.4/79.1/73.7	No
12	Table IV, Table V, 6348 sentence-pairs	Table VII	**85.38**, 93.3/88.0/82.8/78.3	Yes
12.2.1	Table IV, Table V, 6348 sentence-pairs	Table VII, 6348 sentences	**85.76**, 93.9/88.5/83.1/78.5	No
12.2.2	Table IV, Table V, 6348 sentence-pairs	Table VII, 6348 sentences	**85.71**, 93.6/88.4/83.1/78.6	Yes
13	6348 sentence-pairs	Default	**77.04**, 89.3/81.0/73.0/66.9	No
14	6348 sentence-pairs	Default	**77.01**, 89.2/80.9/73.0/66.8	Yes
15	6348 sentence-pairs	Table VII	**75.71**, 88.7/79.7/71.6/65.0	No
16	6348 sentence-pairs	Table VII	**79.40**, 90.2/82.9/75.9/70.2	Yes

TABLE IX. TRANSLITERATION OUTPUT OF PROPOSED MODEL IN COMPARISON WITH NÔM CONVERTER

Nôm input sentences	Referenced transliteration	Proposed model	Nôm converter
𤾓𢆥𥪝𡎝𠊛些	trăm năm trong cõi người ta	trăm năm trong cõi người ta	trăm năm trong cõi người ta
𡨸才𡨸命窖罗恄饒	**chữ** tài **chữ** **mệnh** khéo **là** ghét nhau	chữ tài chữ mệnh khéo là ghét nhau	*𡨸* tài *𡨸* *mạng* khéo *罗* ghét nhau
𣦆戈沒局𣷭𣘃	**trải** qua **một** cuộc bể **dâu**	trải qua một cuộc bể dâu	*𣦆* qua *沒* cuộc bể *𣘃*
仍調𥉫𧡊㐌疠疸𢚸	những điều trông thấy đã đau **đớn** lòng	những điều trông thấy đã đau đớn lòng	những điều trông thấy đã đau *疸* lòng
𨔍咦彼嗇斯豐	**lạ** gì **bỉ sắc tư phong**	lạ gì bỉ sắc tư phong	*罗* gì *彼嗇斯豐*
𡗶撐悁貝𦟐紅打慳	trời xanh quen với má hồng đánh ghen	trời xanh quen với má hồng đánh ghen	trời xanh quen với má hồng đánh ghen
稿蒼吝拱𩙻畑	**cảo** thơm lần **giở** trước đèn	cảo thơm lần giở trước đèn	*稿* thơm lần *拱* trước đèn
風情固錄羣傳史撐	phong tình có **lục** còn **truyền** sử xanh	phong tình có lục còn truyền sử xanh	phong tình có *錄* còn ***truyện*** sử xanh
浪𢆥嘉靖朝明	rằng năm gia **tĩnh triều** minh	rằng năm gia tĩnh triều minh	rằng năm gia *靖* ***chiều*** minh
𦊚方滂𤅶𠄩京凭鐄	bốn phương phẳng **lặng** hai kinh vững vàng	bốn phương phẳng lặng hai kinh vững vàng	bốn phương phẳng ***láng*** hai kinh vững vàng
固茄員外户王	có nhà viên ngoại họ vương	có nhà viên ngoại họ vương	có nhà viên ngoại họ vương
家資俌拱常常堛中	gia tư **nghĩ** cũng thường thường bậc trung	gia tư *nghỉ* cũng thường thường bậc trung	gia tư nghĩ cũng thường thường bậc trung

The BLEU score in Exp12.2.1 was the highest score. However, the way we separated data into training set and testing set previously might cause biased results because of the similarity between the training set and testing set. That is, it may not be practical to distribute each poem (text) in both training and testing sets with the ratio of 8:1, because in real world situations, users might want to transliterate an unseen Nôm text, completely different from the one used to train our model. Therefore, we applied k-fold cross validation to better evaluate the skills of our proposed model. There are 7,920 lines of parallel Nôm-Vietnamese national scripts in total. We shuffled the data and then distributed parallel text into 10 folds (parts). After equally distributing all sentence-pairs into 10 folds, each fold contained 792 pairs of sentences. Based on the experiment results presented in Table VIII, we observed that Exp12.2.1 set-up generated the highest transliteration quality. Consequently, we implemented k-fold cross validation using dictionaries and an additional language model for model training as in Exp12.2.1. Then 10 previously separated folds were

distributed into training, tuning, and testing sets as follows: eight folds for training the model, one fold for tuning, and the one remaining for testing. This time, the data used for language model training was slightly different from that of Exp12.2.1. In addition to the data as in Exp12.2.1, we also extracted and used the national scripts from eight folds of the parallel corpus to feed more data into the model, and thereby improving the transliteration quality generated from the model. We conducted the experiment 10 times, with the corresponding BLEU evaluations presented in Table X. Averaging BLEU

TABLE X. EXPERIMENTS AND RESULTS OF K-FOLD CROSS VALIDATION

Exp ID	BLEU	Exp ID	BLEU
1	**81.88**, 92.0/85.2/78.8/73.1	6	**83.32**, 92.6/86.5/80.3/75.2
2	**83.47**, 92.5/86.5/80.7/75.3	7	**83.83**, 92.6/86.6/81.0/76.1
3	**83.85**, 92.6/86.7/81.0/76.2	8	**83.00**, 92.3/86.2/80.0/74.5
4	**83.65**, 92.7/86.6/80.7/75.6	9	**82.35**, 92.0/85.5/79.2/74.1
5	**83.11**, 92.4/86.4/80.2/74.7	10	**82.67**, 92.1/85.9/79.7/74.6
Average		**83.10**	

scores of 10 experiments, we got 83.10, which was quite close to our best result of 85.76 in Exp12.2.1, Table VIII. We conclude that the experiments carried out in Table VIII were relatively fair, as they were not biased due to the data distribution among the training set and the testing set.

B. Limitations and Development Directions

Based on the test results presented in S ection V-A, we observe that our proposed model still has limitations in choosing the correct national script for a given Nôm script input. While our goal to resolve this difficulty remains, it is unlikely to attain 100-percent accurate transliteration output since even humans argue over which national script should be used for a given Nôm character.

To overcome the aforementioned limitations, we will continue to collect and build a larger parallel corpus for the translation model as well as a monolingual corpus for language models. We will also categorize input data into domains to improve the transliteration quality. We will keep collecting corpora from some other domains such as medicine and ideology. In addition, we will conduct more experiments and train our proposed model with new machine learning models.

VI. CONCLUSION

In this paper, we have presented an automatic transliteration from Nôm scripts into Vietnamese national scripts using the SMT paradigm in computational linguistics. Our proposed model demonstrates significant improvements compared with the existing transliteration system, Nôm converter. Not only does the model recognize a broader range of Nôm scripts, it can also choose the national script for a given Nôm character with higher accuracy according to the context of the input Nôm scripts. Our finding of the distinct characteristic of the language pair Nôm - Vietnamese national scripts and our contribution in building a separate corpus for the language model beside the default language model extracted from the parallel corpus lead to a high result in the SMT approach.

In the future, we will build domain-specific language models and integrate linguistic knowledge to improve transliteration accuracy. Moreover, we can conduct manual post-editing to introduce further improvement. Our proposed model, therefore, will be able to generate more accurate transliteration results. This automatic transliteration system will bridge the gap between our past and our present, stemming the differences in our two writing systems, the historical Nôm scripts and our current national scripts. Thanks to this system, the priceless treasure of our ancestors in history, literature, religion, geography, and traditional medicine will be explored and harnessed effectively. Scholars can now browse and understand the main ideas of a Nôm text without having to invest an immense amount of time to manually work on the ancient scripts.

REFERENCES

[1] S. Liu, L. Wang, and C.-H. Liu, "Chinese-portuguese machine translation: A study on building parallel corpora from comparable texts," 04 2018.

[2] J. Marx, N. Smith, and Staudinger, "Some problems in the evaluation of the russian-german machine translation system miroslav," 02 2021.

[3] O. Nikolaenkova, "Applying clp to machine translation: A greek case study," *Journal of Applied Linguistics and Lexicography*, vol. 1, pp. 69–78, 09 2019.

[4] Y. Eytani, A. Lavie, E. Peterson, K. Probst, and S. Wintner, "Hebrew to english machine translation," 02 2021.

[5] M. Crespo and M. Sánchez-Saus Laserna, "Graded acceptance in corpus-based english-to-spanish machine translation evaluation," 01 2016.

[6] F. Bouzit and M. T. Laskri, "Arabic to french machine translation system based on dcf approach," 02 2021.

[7] G. Tang, R. Sennrich, and J. Nivre, "Understanding pure character-based neural machine translation: The case of translating finnish into english," 12 2020.

[8] R. Tse, S. Mirri, T. Su-Kit, G. Pau, and P. Salomoni, "Building an italian-chinese parallel corpus for machine translation from the web," 09 2020, pp. 265–268.

[9] R. Cornet, C. Hill, and N. de Keizer, "Comparison of three english-to-dutch machine translations of snomed ct procedures," *Studies in health technology and informatics*, vol. 245, pp. 848–852, 01 2017.

[10] Y. Li, C. Pun, and F. Wu, "Portuguese-chinese machine translation in macao," 02 2021.

[11] M. Dwiastuti, "English-Indonesian neural machine translation for spoken language domains," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 309–314. [Online]. Available: https://www.aclweb.org/anthology/P19-2043

[12] Y. Kyaw Thu, V. Chea, A. Finch, M. Utiyama, and E. Sumita, "A large-scale study of statistical machine translation methods for Khmer language," in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China, Oct. 2015, pp. 259–269. [Online]. Available: https://www.aclweb.org/anthology/Y15-1030

[13] A. Srithirath and P. Seresangtakul, "An approach to lao-english rule based machine translation," *Proceedings of the 2015-7th International Conference on Knowledge and Smart Technology, KST 2015*, pp. 93–98, 02 2015.

[14] S. Ab, N. Abdul Rahman, and N. Aziz, "Improving word alignment in an english – malay parallel corpus for machine translation," 02 2021.

[15] M. Zin, T. Racharak, and N. Le, "Construct-extract: An effective model for building bilingual corpus to improve english-myanmar machine translation," 01 2021, pp. 333–342.

[16] N. Oco and R. Roxas, "A survey of machine translation work in the Philippines: From 1998 to 2018," in *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*. Boston, MA: Association for Machine Translation in the Americas, Mar. 2018, pp. 30–36. [Online]. Available: https://www.aclweb.org/anthology/W18-2204

[17] S. Lyons, "A review of thai–english machine translation," *Machine Translation*, vol. 34, 09 2020.

[18] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, Aug. 2017, pp. 28–39. [Online]. Available: https://www.aclweb.org/anthology/W17-3204

[19] H. Rogers, *Writing Systems: A Linguistics Approach*, 1st ed. Blackwell, 2005.

[20] D. Dinh, *Từ điển học tính toán*. VNU-HCM, 2019.

[21] T.-C. Nguyen, *Diên cách cấu trúc chữ Nôm Việt*. VNU-Hanoi, 2012.

[22] K. D. Le, *Từ vựng gốc Hán trong tiếng Việt*. VNU-HCM, 2002.

[23] H. T. Le, "Nghĩ về một số từ khó hiểu trong Truyện Kiều," *Kiến thức ngày nay*, Jan 2016.

[24] 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, and E. Herbst, "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*. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 177–180. [Online]. Available: https://www.aclweb.org/anthology/P07-2045

[25] D. Dinh, K. Hoang, and E. Hovy, *BTL: an Hybrid Model in the English – Vietnamese Machine Translation System*. Proceedings of the MT Summit IX, 2003.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://www.aclweb.org/anthology/P02-1040