Vietnamese - Thai Lexicon for Machine Translation

Dang Ngoc Huy and Pusadee Seresangtakul

Natural Language and Speech Processing Laboratory (NLSP),
Department of Computer Science,
Faculty of Science, Khon Kaen University, Khon Kaen, Thailand
dangngochuy87@gmail.com, pusadee@kku.ac.th

Abstract

This paper proposes the design and development of a Vietnamese-Thai lexicon for a Vietnamese to Thai machine translation. The Vietnamese-Thai lexicon in this research has of five main features: Vietnamese words, Thai words, parts of speech, sub-parts of speech, and Thai meanings. The lexicon consists of 25,000 Vietnamese words. Moreover, Vietnamese to Thai transcription rules are also proposed. These rules were applied to words that do not appear in the lexicon, for example specific names and places.

Keywords: Vietnamese, Thai, lexicon, machine translation

1 Introduction

In the near future, the ASEAN Economic Community (AEC) will provide chances for the participating ASEAN countries, which includes both Vietnam and Thailand, to open their doors for freeing trade, services, investment, skilled workers and the capital market. In addition, this will bring the exchanges in various areas among the ASEAN countries. At present, Thailand and Vietnam have strong ties in areas such as tourism, economy and culture. These ties have increased the number of Thai people who want to study the Vietnamese language. Many Thai universities actually offer B.A. programs in the Vietnamese language and Vietnamese language for tourism. However, the choice of tools needed to help learn Vietnamese are quite limited with a lack of electronic tools for Vietnamese language translation being one of them. To help overcome this limitation, a Vietnamese–Thai machine translation will be developed. In order to develop translation system, one of the most important tools is a lexicon. This paper proposes the creation of a Vietnamese–Thai lexicon as a database in machine translation for any future work and a transcription module for unknown words.

2 Related literature and work

2.1 Vietnamese Language

Vietnamese is the national and official language of Vietnam. It is the native language of Vietnamese people and of about three millions overseas Vietnamese. It is a part of the Austroasiatic language family. The Vietnamese alphabet in use today is a Latin alphabet with additional diacritics for tones and certain letters. Vietnamese consists of initial consonants, vowels, final consonants and diacritics for tone as follows.

2.1.1 Vietnamese initial consonant:

There are 17 single consonants and 11 mixed consonants or digraph [1][2][3] in Vietnamese. Examples of the initial consonants are shown in Table 1.

Table 1. Examples of Vietnamese consonant

Vietnamese	Thai	Phonetic	Vietnamese	Thai	Phonetic
Single consonant					
b	บ	/b/	n	น	/n/
c	ก	/k/	p	ป	/p/
d	ย	/z/	q	ว	/k/
đ	ด	/d/	r	ร	/z/
g	ก	/g/	s	ซ	/ʒ/
h	ฮ	/h/	t	ต	/t/
k	ก	/k/	v	ว	/v/
l	ล	/l/	x	ซ	/ʂ~ɕ/
m	ม	/m/			
Mixed consonant					
ch	จ	/ʈ/	ngh	ง	/ ŋ /
gh	ก	/g/	nh	ย	/ɲ /

2.1.2 Vietnamese vowels

The vowels in Vietnamese consist of 12 single vowels and 41 mixed vowels [1][2][3]. Table 2 shows examples of the Vietnamese vowels.

Table 2. Examples of Vietnamese vowel

Vietnamese	Thai	Phonetic	Vietnamese	Thai	Phonetic
		Single vowel			
a	อา	/ɑ /	o	ออ	/ɔ /
ă	อ๊า	/a/	ô	โอ	/o/
â	เอ๊อ	/ə /	ơ	เออ	/ ɣ /
e	แอ	/ɛ /	u	อู	/u/
ê	เอ	/e/	ư	อือ	/ɯ /
i	อี	/i/	y	อี	/iː/
		Mixed vowels			
ai	อาย	/ aï /	âu	เอิว	/əʉ/
ay	อัย	/ ɛi /	eo	แอว	/ɛʉ/
ây	เอ็ย	/ei/	êu	เอว	/eʉ/
oi	ออย	/ ɔi /	iu	อิว	/iʉ/
ôi	โอย	/oi/	ưu	อืว	/ ɯʉ/
ia	เอีย	/iə/	oa	วา	/wɑ/
ua	อัว	/uə/	oe	แว	/wɛ/
ơi	เอย	/ ɣï /	oai	อวาย	/wɑï/
ui	อูย	/u ï/	ươi	เอือย	/wəi/
au	เอา	/aʉ/	iêu	เอียว	/iəʉ/

2.1.3 Vietnamese final consonant

There are eight final consonants in Vietnamese [1][2][3] as shown in Table 3.

Table 3. Vietnamese final consonant

Vietnamese final consonants	Corresponding Thai alphabet	Phonetic
c	ก	/k/
m	ม	/m/
n	น	/n/
ch	ด	/t/
ng	ง	/ŋ/
nh	น	/n/
t	ด	/t/
p	ป	/p/

2.1.4 Vietnamese tone

Vietnamese has six tones. The tone is indicated by diacritics, which are written above or below the vowel as shown in Table 4.

Table 4. The Vietnamese tones

Vietnamese Tone	Tone Marker	Example
Mid Tone		a
Low Tone	\	à
High Tone	/	á
Rising Tone	’	ả
High Rising Tone	˜	ã
Semi-vowel Tone	•	ạ

2.2 Related works

Mahahing S. and Seresangtakul P. [6] presented a Korean-Thai Lexicon for Natural Language Processing. The Korean-Thai lexicon that they created consists of 7 parts: Korean words, Korean Revised Romanization, parts of speech, sub parts of speech, special characteristics, Thai meaning and descriptions of the meaning of the Korean transcription.

Rajan, R , Sivan, R., Ravindran, R. and Soman, K.P., [7] proposed Rule Based Machine Translation from English to Malayalam. They used two types of rules namely, transfer link rules and morphological rules. The Part of Speech (POS) of the source words is obtained with the help of a parser. The source is assigned for each word by using POS. Then, the transfer link rule file is used to generate the target structure. Finally, the target sentence is generated by using a morphological dictionary and a word dictionary.

Lakkhawannakun P. and Seresangtakul P. [13] proposed an Isarn Dharma Alphabet to Thai Language Translation using Augmented Transition Networks (ATNs). They modified an Isarn Dharma – Thai dictionary, which proposed by Phaiboon N. and Seresangtakul P. [14]. The renew Isarn Dharma – Thai dictionary consists of Isarn words, Thai words, phoneme, word type, Thai meaning, sub parts of speech, special characteristics, English meaning and Thai description.

Quoc Hung Ngo and Winiwarter W. [15] presented an English – Vietnamese bilingual corpus for machine translation. In their work, the bilingual corpus was tagged with linguistic information, such as part of speech, chunks, and bitext alignment at the word level.

3 Method

3.1 Creating the Vietnamese-Thai Lexicon

In order to create the Vietnamese-Thai lexicon, SQLite was selected as the database management system because of its speed, small memory consumption and easy to move database. Moreover, it supports the Vietnamese alphabet with nvarchar data type. In the Vietnamese language, one word may have several meanings and be a different part of speech. To support our Vietnamese-Thai translation system in the future, a Vietnamese-Thai lexicon was constructed. The lexicon structure consists of the following 5:

1) Vietnamese words

2) Thai word: This attribute is used to store the translation of a Vietnamese word in Thai**.**

3) Parts of Speech (POS): Vietnamese and Thai Parts-of-Speech are similar. This work follows the Parts-of-Speech Tagged Corpus of Thai Text from NECTEC [12] and adds new parts of speech for the Vietnamese language.

4) Sub-Parts of Speech: This attribute stores sub parts of the POS. Its content follows the Parts-of-Speech Tagged Corpus of Thai Text from NECTEC [12]. It also had new sub-parts of speech added for the Vietnamese language.

5) Thai Meaning: this attribute defines definition of Vietnamese in Thai connotation.

Examples of the lexicon content are shown in Table 6.

3.2 Vietnamese to Thai transcription

It is not possible to keep all words in the dictionary, especially specific names such as personal names. This paper proposes a Thai transcription based on linguistic rules in order to convert Vietnamese pronunciation into Thai pronunciation for unknown words or words that do not appear in the dictionary. There are 85 rules. These rules can be grouped into 28 patterns as shown in Table 5.

As shown in Table 4, Vietnamese has a semi-vowel tone, which does not exist in the Thai tone system. In order to transcript the semi-vowel tone, the Thai writing rules were applied by combining the Thai consonant with the corresponding Vietnamese vowel. The tone will be applied by considering the class of the initial consonant of a syllable (high, medium or low), the type of syllable (live or dead), and the length of the vowel (long, or short) [16]. For example, a word “cạ” in Vietnamese will convert to “กะ” (/kà/) in Thai.

Table 5. Vietnamese to Thai transcription rules

Rule	Example		Rule	Example	
	Vietnamese	Thai		Vietnamese	Thai
CVS	cam	กาม	CVVV	hoai	ฮวอย
CVSS	mang	มาง	CCV	nga	งา
CCVS	khan	คาน	CCCV	nghi	งี
CCCVS	nghin	งิน	CCVV	ngay	งัย
CVVS	hoan	ฮวาน	CCCVV	nghia	เงีย
CVVVS	huyên	เฮวียน	CCVVV	ngoai	งวอย
CCVVS	nhiên	เยียน	VS	an	อาน
CCVVVS	nhuyên	เยวียน	VSS	ung	อูง
CCVSS	ngang	งาง	VVS	oan	อวาน
CVVSS	cương	เกือง	VVVS	uyên	เอียน
CCVVSS	chương	เจือง	VVSS	ương	เอือง
CCCVVVSS	nghiêng	เงียง	V	u	อู
CV	ho	ฮอ	VV	ai	อาย
CVV	hay	ฮัย	VVV	yêu	เอียว

Table 6. An example of lexicon content

Vietnamese word	Thai word	POS	Sub-POS	Thai meaning
Nhà	บ้าน	NOUN	NCMN	บ้านพัก, บ้านเรียน
Ăn	กิน	VERB	VACT	กินข้าว, กินอาหาร, กินยา

4 Experimental Results

In the study, a Vietnamese-Thai lexicon was constructed, which comprised of 25,000 words. The words were collected from dictionaries and textbook, which cover general daily word usage. Furthermore, Vietnamese-Thai transcription rules to transcript Vietnamese pronunciation to Thai pronunciation were proposed. In order to evaluate the transcription rules, Vietnamese newspapers and novels were transcribed by the system. The experimental results are shown in Table 7.

The results show that most of the errors occur from the Vietnamese words with semi-vowel tones and words with mixed vowels, which do not exist in Thai pronunciations. Table 8 shows examples of the transcription results.

Table 7. Transcription accuracy results

Topic	Number of word	Number of correct word	Accuracy
Newspaper	845	725	85.80%
Novel	1,010	877	86.83%
Total	1,855	1,602	86.36%

Table 8. Transcription results

Vietnamese	Thai pronunciation
Đỗ xanh là một loại thực phẩm dưỡng sinh giúp thanh nhiệt, giải độc, nhưng không phải ai cũng thích hợp với ăn đỗ xanh.	โด๋ซานหล่าโหม่ดหล่วายถึ่กฝึมเหยืองซีนยู้ปทานเหยี่ยดหยายโด่กยึงโคงฝายอายกุ๋ง ทึ้ดเห่อปเว้ยอันโด๋ซาน
Nhiều nhà làm phim đã chọn khai thác đề tài những nhân vật có thật trong lịch sử.	เหยี่ยวหย่าหล่ามฟีมด๋าจ่อนคายท้ากเด่ต่ายหยึงยันหวั่ดก๊อถั่ดตองหลี่ดสือ

5 Conclusion and Future Work

To date, a Vietnamese-Thai lexicon was created, which covered a large number of Vietnamese vocabularies. The lexicon is used in a Vietnamese to Thai machine translation. In addition, Parts of Speech (POS) and Sub-Parts of Speech will be used to distinguish the correct meaning of a word according to the purpose of translation. Not only was the Vietnamese-Thai lexicon but also Rule Based transcription was proposed for the Vietnamese to Thai machine translation. Future work will focus on the Vietnamese to Thai machine translation system.

References

[1] Dao M.T, Dao T.M.N, Dao, Nguyen M.V, Le K.N, Le T.H, Nguyen P.T and Do B.L. Sentence structure of Vietnamese language, Vietnam, 2010.

[2] Wanrotsaphak M. Intensive course of Vietnamese. Retrieved November 24, 2012, from http://www.e4thai.com.

[3] Vietnamese language. Retrieved May 30, 2012, from http: //en.wikipedia.org/wiki/ Vietnamese_language.

[4] Yaprom N. and Seresangtakul P. Lanna to Thai Language Translation. *Proceedings of the 11th National Computer Science and Engineering Conference (NCSEC2007)*, Bangkok, pp. 180-189, 2007.

[5] Vo T.H and Fafiotte G. UVDict – a machine translation dictionary for Vietnamese language in UNL system. *International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS)*, Seoul, pp. 310 – 314, 2011.

[6] Mahahing S. and Seresangtakul P., Korean-Thai Lexicon for Natural Language Processing. *The 4th International Conference on Information Science and Applications (ICISA)*, Pattaya, Thailand, pp. 1-4, 2013.

[7] Rajan, R , Sivan, R., Ravindran, R. and Soman, K.P. Rule Based Machine Translation from English to Malayalam. *International Conference on Advances in Computing, Control, & Telecommunication Technologies (ACT '09)*, Trivandrum Kerala, pp. 439 – 441, 2009.

[8] Nguyen C.T. Thai – Vietnamese Dictionary, Vietnam, 2000.

[9] Vietnamese grammar. Retrieved May 30, 2012, from http:// en.wikipedia.org/wiki/ Vietnamese_grammar.

[10] Thai language. Retrieved May 30, 2012, from http://en.wikipedia.org/wiki/Thai_language

[11] Rule–based machine translation. Retrieved December 7, 2012, from http://en.wikipedia.org/wiki/Rule-based_machine_translation.

[12] ORCHID: Part-of-Speech Tagged Corpus of Thai Text. Retrieved December 7, 2012, from http://culturelab.in.th/files/orchid.html.

[13] Lakkhawannakun P. and Seresangtakul P. Isarn Dharma Alphabets to Thai Language Translation by ATNs. *International Symposium on System Engineering and Computer Simulation (SECS-2013)*, Danang, Vietnam, 2013. [accepted]

[14] Phaiboon N. and Seresangtakul P. Isarn Dharma Alphabet phoneme transcription. *Proceedings of the 13th National on Computer Science and Engineering Conference (NCSEC2009)*, pp.287-292. King Mongkut's University of Technology Thonburi, Bangkok, 2009.

[15] Quoc Hung Ngo and Winiwarter W. Building an English-Vietnamese Bilingual Corpus for Machine Translation. *2012 International Conference on Asian Language Processing (IALP)*, Hanoi, Vietnam, pp. 157-160, 2012.

[16] David Smyth. Thai an Essential Grammar. *Routledge Taylor & Francies Group*, London and New York, 2002.