Study of Machine Translation to Build Pattern-based Model for Vietnamese – Cham

Van Ngoc Sang

Educational Technology , Universiti Teknology Malaysia, Johor Bahru, Malaysia
*Corresponding Author: *sangpodam@yahoo.com*

Abstract Cham script appeared in 4th century on stone stele in Tra Kieu Vietnam. It was considered as the first language in South East Asia. Due to many social and historical reasons, Cham language is faced with the risk of deformation. In this research, we found that Cham grammar structure is still unclear for the Vietnamese Cham machine translation. Hence, in order to preserve Cham language, we propose a Pattern-based model to translate Vietnamese into Cham language with the size of the bilingual dictionary is approximately 4,500 sentences, bilingual corpus 2500 pairs of Vietnam Cham sentence, 1950 pairs of translation sample, 324 function words, 768 vocabulary, 57 prepositions, and all are stored in text file. Initially is tested based on the basis of bilingual corpus, bilingual dictionaries and translation sample with limited resources, the results achieved were relatively satisfying and high quality if the input sentence matches the translation pattern and this translation pattern is correct. However, bilingual corpus, bilingual dictionaries and translation sample are usually made manually. For this reason, there are many costs in making a pattern-based for machine translation system. In this experiment, we obtained good results. Hence, construction pattern-based machine translation for Vietnamese to Cham is an essential research and development in order to preserve Cham language.

Keywords Cham Translation; Viet Cham Translation; Cham Machine Translation

1. Introduction

There are various efforts have been made in developing machine translation (MT) systems and many approaches on MT research: Pattern-based, transfer-based, interlingua-based, and etc. Pattern-based MT is a very traditional MT method that uses translation patterns and translation word (phrase). In this method the translation get high-quality translation results if the input sentence matches the translation pattern and this translation pattern is correct. Preliminary studies, we found that this method suitable for Vietnam Cham language MT, because now Cham grammar structure is unclear, hence, we need to create the bilingual corpus, bilingual dictionary and translation sample to support for this method.

2. Literature Review

In the pioneering work of Nagao [1,2], Machine translation (MT) is the translation of text by a computer with no human involvement. Today human with various efforts are involved in developing MT systems for practical use. A few different types of main techniques for machine version; Statistics Based Machine Translation (SBMT), Rules Based Machine Translation (RBMT), and Hybrid Systems which combine RBMT and SBMT [3], are being used. Each approach has its own branches and has different advantages and disadvantages. Among these approaches, we selected the pattern-based machine translation for this research.

Pattern-based MT is a traditional method that was proposed in the 1960s [4]. It has three resources; source pattern, target pattern and dictionaries word translation. In pattern-based MT, a translation pattern provides a word order. Then, if the input sentence matches a translation pattern, the translated sentence will be of high quality. However, this form of machine translation has disadvantages as well. It cannot translate input sentences that do not match any of the stored translation patterns. This means that to match many sentences, we either have to make many patterns or generalize these patterns [5].

3. Objective

In order to achieve this aim, the objectives of the research to be investigated are as follows:

(i) To build machine translation model for Vietnamese Cham language.
(ii) To build data source for bilingual corpus, bilingual dictionary, translation sample to support for this method.
(iii) To develop application program for proposed model.

4. Methodology

4.1 Structure of Vietnamese and Cham Simple Sentence

Many linguists classify that Cham language is a member of the Malayo-Polynesian branch of the Austronesian family and Cham language is polysyllabic word forms. Cham people use their own script. The script has a separate character system derived from Devanagari script of India. On the contrary, the Vietnamese language is a part of Austroasiatic language family of which it has by far the most speakers. Vietnamese alphabet in use today is a Latin alphabet with additional diacritics for tones and certain letters. Vietnamese belongs to the type of Isolating language as it is single syllabic. The similarities both Vietnamese and Cham language were relatively orderly, which is very significant and convenient for language translation. Because the different order of each word is able to make a difference in the meaning of that entire sentence or make a grammatical error. Grammatical structure consists of Subject (S), Verb (V) and Object (O). It can be divided into 6 difference patterns such as SVO, SOV, VSO, VOS. OVS and OSV. Almost all the language sentences are SVO or SOV [6]. However, the structure of Vietnamese and Cham language is similar: SVO. See Figure1.

No.	Bilingual Sentences			
(1)	Tom	nhớ	Jakei	Vietnamese
	Tom	*miss*	*Jakei*	
	[illegible]	[illegible]	[illegible]	Cham
	S	V	O	
(2)	Jakei	nhớ	Tom	Vietnamese
	Jakei	*miss*	*Tom*	
	[illegible]	[illegible]	[illegible]	Cham
	S	V	O	
(3)	Nhớ	Jakei	Tom	Vietnamese
	Miss	*Jakei*	*Tom*	
	[illegible]	[illegible]	[illegible]	Cham
	V	S	O	

Figure 1. Simple structure of Vietnamese and Cham languages

The sentence (1) and (2) have different meanings , because in (1) “Tom” is the subject and “Jakei” is the object, and vice versa in (2) that is “Jakei” is the subject and “Tom” is the object. It can be said that these two sentences have the same lexical items and number of words, but they have different orders and meanings. In case (3) the grammar and the syntactic rules of this sentence is incorrect and meaningless in Cham language.

4.2 Pattern-Based Machine Translation Vietnamese-Cham

The model of pattern-based MT system as proposed has three main resources: Bilingual corpus, bilingual Vietnamese-Cham dictionary, and translation sample. As described below:

Step 1. Prepare Bilingual corpus, Bilingual dictionary, and Translation sample.

Step 2. Input Vietnamese sentence.

Step 3. Search for a Vietnamese pattern that matches the input of step 2.

Step 4. Output Cham Pattern corresponding to the Cham pattern made in Step 3.

Step 5.Generate a Cham sentence by using Vietnamese-Cham in Bilingual corpus, Bilingual dictionary and Cham pattern in Translation sample.

These steps are described as shown in Figure 2

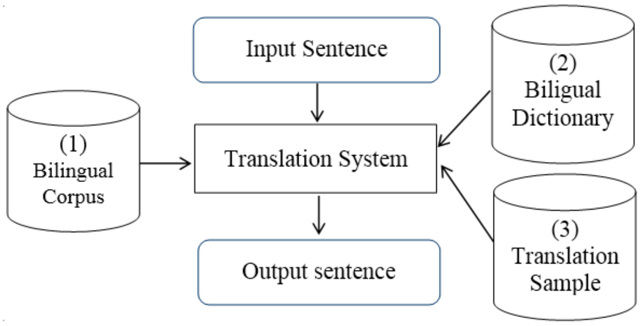


Figure 2: Pattern-based Vietnamese Cham Machine Translation Model

4.2.1 Bilingual Corpus

Bilingual corpus presented in this research is pairs of sentences in the Vietnamese-Cham, each pair of sentences are separated by a blank line to distinguish each pair and check the corpus contents. Bilingual corpus built in this research should be written with the sentences to be considered a standard that is grammatically correct and is widely accepted. This corpus does not contain the text or translation as personal, because it does not ensure the reality of the corpus. Particularly with bilingual corpus, to build a 1-1 translation of each other, were not translated gratification, summary, translation equivalent / synonym or translation styled explain, interpret [7]. Bilingual corpus sample is shown below:

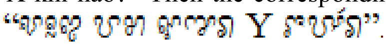
No.	Bilingual Sentence
1.	Anh ở đâu ?
	[illegible] ?
	← blank line
2.	Tôi ở Sài Gòn .
	[illegible] .

Figure 3. Vietnamese-Cham sample sentence construction

4.2.2 Construction of Translation Sample

In order to construct the translation sample, consider the two cases as follows:

+ Case sample sentence with a variable

Assuming exist this sentence in translation sample *"Anh học* chữ X khi nào?" Then the corresponding translation would be:  . As shown in Figure 4.

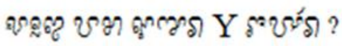
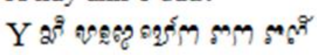
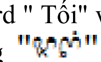
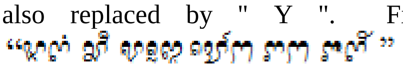
No.	Sentence one variable
1.	Anh học chữ X khi nào ? 
2.	X nay anh ở đâu? 

Figure 4. Sample sentence with a variable

Sentence (1), when users need to translate one sentence: "Anh học chữ Cham khi nào?" then the word "Cham" is replaced by "X" and its meaning  is also replaced by "Y" and the translation result is:

 (when did you learn Cham language?).

Other cases in sentence (2), when users need to translate a sentence "Tối nay anh ở đâu?" Then the word " Tối" will replace the" X" and at the same time its meaning  is also replaced by " Y ". Finally the result is  (where were you tonight?).

+ Case sample sentence with many variable

Assuming exist this sample in corpus translation sample "X1 thích X2" and the corresponding translation  . In case the user needs to translate the sentence "Tôi thích mèo" then "Tôi" is replaced by "X1" and its meaning  was replaced by "Y1". Similarly, "mèo" was replaced by "X2" and its meaning  was replaced by "Y2". Meaning of complete sentences was  . (I like cat).

Case sample sentence with many variables, we often extracted samples into two variables form or sample sentence with one variable. If the process of extracting cannot execute into sample sentence with one variable, then sample sentence with many variables will be excluded from the corpus translation sample. This sample with two variables is the general case of sample one variable during extraction. For example, sample sentence (V-C) is the general case of sample sentences (V1-C1) and sample sentences (V2-C2).

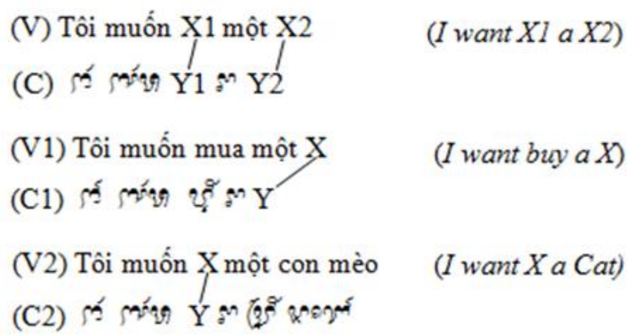


Figure 5. Sample sentence with many variable

Thus, in the form of translation sample containing one variable and samples containing two variables can be converted into sample with one variable as described above.

4.2.3 Bilingual Dictionary Construction

The bilingual dictionary of 5,000 entries is created in Dbase format for mapping equivalent entries of the input string. The Cham output word or sentence is generated. In order to build a Vietnamese-Cham pattern-based machine translation model, the first hard work needed is to build bilingual corpus, bilingual dictionary, translation sample, Vietnamese and Cham grammar structures, basic information about word-class and pattern mapping Vietnamese – Cham.

4.3 String Edit Distance

In order to extract the string, we use method of string edit distance (Computing Levenshtein distance) to measure distance between strings. The idea of this method is minimum number of "characters edit operations" needed to turn one sequence into the other. The operations include copy, substitute, insert and delete. For turning a word (a) = "SPAKE" into (b)= "PARK" , using the dynamic program table for string edit as shown in Figure 6.

		P	A	R	K
	c_{00}	c_{01}	c_{02}	c_{03}	c_{04}
S	c_{10}	c_{11}	c_{12}	c_{13}	c_{14}
P	c_{20}	subst c_{21}	delete c_{22}	c_{23}	c_{24}
A	c_{30}	insert c_{31}	???		
K					
E					

Figure 6. Dynamic Program Table for String Edit

For turning word (a) into word (b) at the minimum, we need three operations. First delete “S”, second insert “R” and third delete “E” in sentence (a). To calculate the value in each cell of the dynamic program table, we apply the following formula:

$$D_{ij} = \min \begin{cases} D(i,j) = \text{score of best alignment from s1..si to t1..tj} \\ D(i-1, j-1) \quad , \text{if si=tj} \quad // \text{ Copy} \\ D(i-1, j-1) + 1, \text{ if si=tj} \quad // \text{ Substitute} \\ D(i-1, j) + 1 \quad // \text{ Insert} \\ D(i, j-1) + 1 \quad // \text{ Delete} \end{cases}$$

Table dynamic program to calculate the distance between two words, the final score of aligning all of both strings as shown in Figure 7.

		P	A	R	K
	0	1	2	3	4
S	1	1	2	3	4
P	2	1	2	3	4
A	3	2	1	2	3
K	4	3	2	2	2
E	5	4	3	3	3

Figure 7. The final score of both strings edit distance

5. Result

In order to develop MT program, the software we use is Visual C++ 6.0 and run under Window 7 on a Pentium PC. The size of the bilingual dictionary is approximately 4,500 sentences, bilingual corpus is 2500 pairs of Vietnam Cham sentence, 1950 pairs of translation sample, 324 function words, 768 vocabulary, 57 prepositions, and all are stored in text file. See Table1.

Table 1. Size of Data Warehouse in Testing

No.	Data	Size
1.	Bilingual dictionary	4,500
2.	Bilingual corpus	2500
3.	Pattern mapping	1950
4.	Function words	325
5.	Vocabulary	768
6.	Preposition	57

From the experiment showed that, case input with a simple sentence if it exists in bilingual corpus, or the sentence with a variable and this variable is present in the bilingual dictionary, then the result translation is correct. Conversely, we found that the major issue of incorrect sentences is caused by a sentence with multiple meanings. That means a word has more than one meaning whereas the machine translation still lacks word-sense to select a proper meaning of the word to suit the context of the sentence. Or bilingual dictionary, bilingual corpus or translation sample is limited, so the result of the translation process is not as desired.

Sentence	Meaning	
1. Vietnamese	Tôi thích Mèo	
Cham	[illegible]	(correct)
	I like Cat	
2. Vietnamese	Tôi thích Mèo và Chó	
Cham	[illegible]	(correct)
	I like Cat and Dog	
3. Vietnamese	Cô ấy đẹp	
Cham	[illegible] / [illegible]	
	She is beautiful / She is nice female	
4. Vietnamese	Nhà này đắt	
Cham	[illegible]	(correct)
	This house is expensive	
	[illegible]	(incorrect)
	This house is high	

6. Discussion

Based on Table 2 the results showed that, with sentence (1) the sample sentence can be one variabe such as “Tôi thích X” or “X thích Mèo”, or can be two variables as “X thích Y”. This sentence translation is correct because the input sentence “Tôi thích Mèo” exists in bilingual corpus or these words exist in bilingual dictionary. Similar in sentence (2), this sentence translation is correct because it exists in bilingual corpus or these words exist in bilingual dictionary, and it can be generated into three variables such as “X thích Y và Z”, or two variables “Tôi thích X và Y” and these samples can be converted into sample with one variable as “X thích Mèo và Chó”, “Tôi thích X và Chó”, or “Tôi thích Mèo và X”. Sentence (3) the word “đẹp” has two meanings whereas the machine translation still lacks word-sense to select a proper meaning of the word to suit the context of the sentence. And sentence (4) the word “đắt” has two meanings, “This house is expensive” is correct, or “This house is high” is incorrect.

In our experiments, the characteristic of this translation is high quality translation results if the input sentence matches the translation pattern and this translation pattern is correct. However, translation patterns and translation word dictionaries are usually made manually. Therefore, there are many costs in making a pattern-based machine translation system. Anyhow, this machine translation for Vietnamese Cham has been important implication for studying, teaching, translating as well as preservation of Cham script and Cham language.

7. Conclusions

In this paper, we proposed a technique pattern-based model to translate Vietnamese into Cham language. Initially we tested program based on a bilingual corpus, bilingual dictionaries and translation sample with limited resources, through observations the results achieved were relatively satisfying. Therefore, to develop a machine translation application for Vietnamese-Cham is necessary.

For future work, in order to make the application well translated, we are interested in building large enough resources, good quality for bilingual corpus, bilingual dictionaries and translation sample. Besides, we propose a new model for a MT system that will combine rule-based and example-based approach and will be applied to the Vietnamese - Cham translation.

REFERENCES

[1] Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. *In A. Elithorn. and R. Bannerji (eds.) Artificial and Human Intelligence* . Nato Publications. pp. 181-207.

[2] Sato, S., & Nagao M. (1990). Toward memory-based translation. *Proceedings of COLING*, Helsinki, Finland, pp. 247-252.

[3] Mamta (2015). A Review of Various Approaches Used for Machine Translation. *International Journal of Advance Research in Computer Science and Management Studies*. Vol. 3, 2321-7782.

[4] H.Maruyama (1993). Pattern-based translation:Context-free Transducer and Its Applications to Practical NLP. *In Proc. of Natural Language Pacific Rim Symposium*, 232-237

[5] Murakami, Isamu., & Masato (2013). Pattern-Based Statistical Machine Translation for NTCIR-10 PatentMT. *Proceedings of the 10th NTCIR Conference*, Tokyo, Japan, 18-21.

[6] Russel S.Tomlin (1986). *Basic word order: Functional principles*. Croom Helm, London,UK.

[7] Dinh Dien (2006). *Natural language processing*. Publishing House: National University ,TP.HCM .