- Introduction & Background
- Objective
- Methodology
- Result
- Discussion
- Conclusion
- Thanks

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

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

Mohamad Budi

Ph.D Student. Mohamad Budi
Prof.Dr. Mohamad Bin Bilal Ali
Dr. Noor Dayana Abd Halim

1. Introduction & Background

- 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, because Cham grammar structure is unclear, hence, we need to create the bilingual corpus, bilingual dictionary and translation sample to support for this method.

Next

2. Background

- In the pioneering work of Nagao, machine translation (MT) is the translation of text by a computer with no human involvement. (Nagao, 1984; Sato & Nagao, 1990).
- Today 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 are being used. (Mamta, 2015).
- Pattern-based MT is a traditional method that was proposed in the 1960s (H.Maruyama, 1993).

3. Objective

The objectives of the research to be investigated are as follows:

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

4.1Vietnamese and Cham Simple Sentence Structure

- Many linguists classify that Cham language is a member of the Malayo-Polynesian and Cham language is polysyllabic word forms.
- Vietnamese language is a part of Austroasiatic language. Used Latin alphabet with additional diacritics for tones and certain letters. Vietnamese language is Isolating language with single syllabic.

Next

Vietnamese and Cham Simple Sentence Structure

The similarities both Vietnamese and Cham language were relatively orderly. Grammatical structure: 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 (Russel, 1986). However, the structure of Vietnamese and Cham language is similar: SOV.

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

Next

4.2 Vietnamese Cham Pattern-Based Machine Translation

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

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.

Next

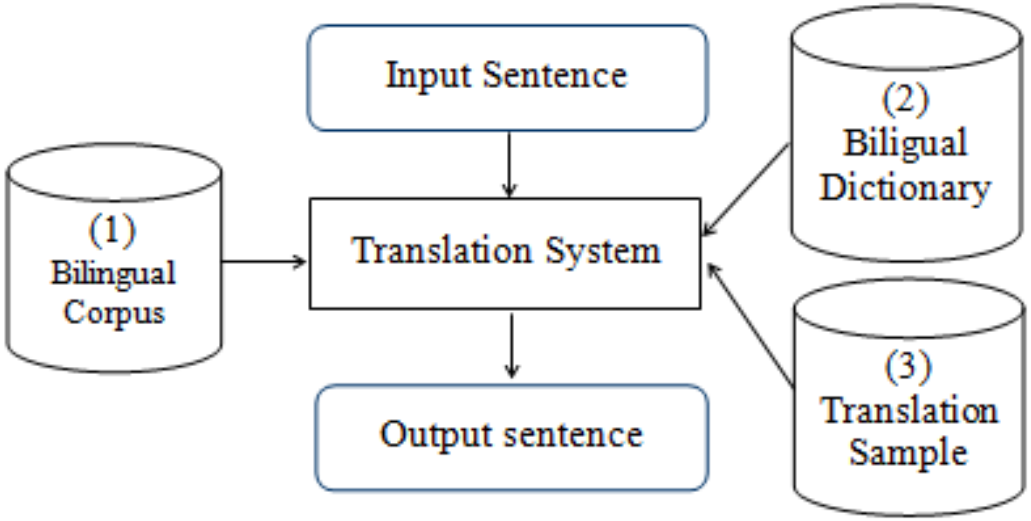

Figure 2: Pattern-based Vietnamese Cham Machine Translation Model

Next

UTM UNIVERSITI TEKNOLOGI MALAYSIA

4.2.1 Bilingual Corpus (cont.)

Bilingual corpus is a pairs of Vietnamese-Cham sentences.
Each pair of sentences are separated by a blank line.
Bilingual corpus are simple and standard sentences.
Corpus does not contain the text in confused sentences. (Dinh Dien, 2006)

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

Figure 3. Vietnamese- Cham Sample Sentence Construction

Where are you?
Abang daok pak halei?

Next

4.2.2 Translation Sample Construction

+ Case sample sentence with a variable

No.	Sentence one variable
1.	Anh học chữ X khi nào ?
	

When did you learn **X** language?

Abang bac akhar **Y** habier?

"Anh học chữ Cham khi nào?" (when did you learn **Cham** language?).

Translate 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: 

Next

4.2.2 Translation Sample Construction

+ Case sample sentence with a variable

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

When did you learn **X** language?

Abang bac akhar **Y** habier?

Figure 4. Sample Sentence With a Variable

Sentence (1), translate sentence: "Anh học chữ Cham khi nào?" then the word "Cham" is replaced by "X" and its meaning "[illegible]" is also replaced by "Y" and the translation result is: [illegible]? (when did you learn **Cham** language?).

Next

Case sample sentence with many variable

Assuming exist sentence “X like Y”. In Vietnamese “**X1** thích **X2**”. and the corresponding translation “Y1 [illegible] Y2”

To translate the sentence “**I** like **Cat**”, in Vietnamese “**Tôi** thích **mèo**” then “**Tôi**” is replaced by “**X1**” and its meaning “[illegible]” was replaced by "Y1".

Step 1: “[illegible] Y2”

Similarly, "**mèo**" was replaced by "**X2**" and its meaning “[illegible]” was replaced by "Y2". Meaning of complete sentences was “[illegible]” (I like cat).

Sample sentence (V-C) is the general case of sample sentences (V1-C1) and sample sentences (V2-C2).

(V) Tôi muốn X1 một X2 *(I want X1 a X2)*

(C) ကို ကျော်ဟု Y1 နှာ Y2

(V1) Tôi muốn mua một X *(I want buy a X)*

(C1) ကို ကျော်ဟု ဝယ် နှာ Y

(V2) Tôi muốn X một con mèo *(I want X a Cat)*

(C2) ကို ကျော်ဟု Y နှာ ကြောင် တကောင်

Figure 5. Sample Sentence With Many Variable

Next

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.

For turning a word (a) = “SPAKE” into (b)= “PARK”

		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}	c_{21} subst	c_{22} delete	c_{23}	c_{24}
A	c_{30}	c_{31} insert	???		
K					
E					

First delete “S”, second insert “R” and third delete “E” in sentence (a).

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

Next

4.3 String Edit Distance (cont.)

		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

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.

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

UTM UNIVERSITI TEKNOLOGI MALAYSIA

5. Result (cont.)

Table 2. Sample Sentence of Multiple Meaning

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

Back

6. Discussion

Based on Table 2 the results showed that, sentence (1) the sample sentence one variabe "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: "**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**".

6. Discussion (cont.)

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

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.

Thanks for listening

Terima kasih

ขอบคุณครับ

Next



Prof.Dr. Mohamad Bin Bilal Ali
Dr. Noor Dayana Abd Halim
PhD. Student. Mohamad Budi

University Technology Malaysia (UTM)- Malaysia

Back