Language Post Positioned Characteristic Based Chinese-Vietnamese Statistical Machine Translation Method

Jianyalin He[1,2],Zhengtao Yu[1,2],Changtao Lv[1,2], Hua Lai[1,2] , Shengxiang Gao[1,2] ,Yang Zhang[1,2]
[1](School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)
[2](Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)
Correspongding author:Zhengtao Yu, Email:ztyu@hotmail.com

***Abstract*—According to the typical characteristics of Vietnamese modifiers post positioned, a Chinese Vietnamese statistical machine translation method that fuses language post positioned characteristic function was proposed. In this method, firstly we analyzed the grammar differences between Chinese and Vietnamese, and extracted the difference of attribute position, adverbial position and qualifier order. Secondly we defined ordering block based on those difference, and added mapping method of ordering block to decoding algorithm in phrase based statistical machine translation model, then recorded N-best candidate translations and its score P produced by decoding. Thirdly reordered ordering block with language post positioned characteristic reordering algorithm and estimated score D by unconditional maximum likelihood probability distribution. Finally we chose the best translation based on score P and score D. We used lexicalized reordering model based phrase statistical machine translation as contrast experiment. The results of experiment show that our method effectively improves the quality of translation.**

Keywords-statistical machine translation; Chinese; Vietn amese; language post positioned

I. INTRODUCTION

Vietnam is one of the important neighbors in southwest China. The study of Chinese-Vietnamese statistical machine translation plays an important role in supporting Chinese and Vietnamese bilingual comprehension, public opinion analysis, information retrieval and cultural exchange. At present, the translation of Chinese -Vietnamese statistical machine is mainly focused on the construction of Chinese-Vietnamese bilingual resource library, the study of Chinese-Vietnamese word alignment, and translation research is still in its infancy. In the Chinese-Vietnamese word alignment, it significantly improved the accuracy of word alignment after incorporating the Vietnamese attributive post positioned into the log-linear model[1] . In the Vietnamese translation, the US Air Force during the Vietnam War funded the development of the first English-Vietnamese translation system LOGOS I, the system contains only 130 words and has the initial translation function; Yun Zhou et al.[2] developed the Chinese-Vietnamese machine translation system Han Viet 0.1 using the rule-based syntactic transformation method. The system can only convert local subtrees and can not convert the global subtree, and its rules can not cover all the linguistic phenomena. Phuoc Tran et al.[3] and Danhui Yan et al.[4] proposed the SVBUT model and the PVBUT model based on the Chinese-Vietnamese meaning relation, which mainly solves the problem of translation of unknown words. Google translation also integrates Chinese-Vietnamese translation system, but its translation quality is far from meeting the demand. The above-mentioned Chinese Vietnamese machine translation method does not consider the influence of the linguistic features of Chinese and Vietnamese on the translation effect, and the differences in different languages have a significant effect on the performance of the machine translation system. Therefore, how to integrate the language difference between Chinese and Vietnamese language into the machine translation system is a question worthy of study.

In terms of grammatical differences, Vietnamese and Chinese have both similarities and differences in grammar. The same point are the consistency of the main language of the two languages, they are SVO (the subject predicate object). For example, the Chinese is "我听音乐(I listen to music)", as Vietnamese is "Tôi (我) (me) nghe (听) (listen) nhạc (音乐) (music)"; The most striking difference is that the position of modifiers(attributive and adverbial) in the Vietnamese language is in the mirror-image relationship with the Chinese. That is, the adjectives in Vietnamese are located after their modified nouns, the adverbs are located after their modified adjectives and verbs. And the Chinese is just the opposite. For example, the Chinese is "她是一个漂亮的女孩(she is the most beautiful girl)", the Vietnamese is "Cô là một (她是) (she is) cô gái (女孩） (girl) xinh đẹp (美丽的) (beautiful) nhất (最)(the most)"[4] [5][6][7][8]. After the above analysis can be seen, Vietnamese and Chinese in the word order is significantly different, and these differences have a certain law: The position of modifier and modifier in Vietnamese language is opposite to that in Chinese; Modifiers and modifiers appear consecutively. By analyzing and summarizing these differences, some ordering rules can be constructed; Combine these rules into the phrase-based translation model[9][10][11][12] to explore the impact on the language characteristics of Chinese and Vietnamese on the performance of translation systems.

II. THE VIETNAMESE LANGUAGE POST POSITIONED CHARACTERISTIC

The difference between the language of the Vietnamese and the Chinese modifier (attributive, adverbial) and the central language is the most obvious difference between the two languages. For the problem of the position of the modifier, the attributive of the nouns in the Chinese language are always preceded by the noun, and the language in the Vietnamese language is usually located after the noun. The order for descriptive multi-tiered adverbs in Chinese is: 1) Predicate phrases; 2) Verb

978-1-5386-1981-0/17/$31.00 ©2017 IEEE