Research Article

A Chinese-Naxi Tree-to-Tree Machine Translation Method Based on Subtree Alignment

Shengxiang Gao, Zhengtao Yu, Chao Liu, Lin Chen, and Xudong Hong

School of Information Engineering and Automation, Kunming University of Science and Technology, No. 727 South Jingming Road, Chenggong District, Kunming 650500, China

Correspondence should be addressed to Zhengtao Yu; ztyu@hotmail.com

Received 2 May 2015; Revised 27 August 2015; Accepted 27 September 2015

Academic Editor: Dane Quinn

Copyright © 2015 Shengxiang Gao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In allusion to the syntactic differences between Chinese and Naxi language, the thesis presents a tree-to-tree method of Chinese-Naxi machine translation based on subtree alignment. In this method, we define a subtree alignment model, providing its inference probability, and solve the alignment missing problem of Chinese-Naxi alignment by updating nodes (insert or delete). And then we train the subtree alignment model by EM algorithm, merging subtree alignment to the translation model. Finally we extract the template of Chinese-Naxi translation, adopting the extraction algorithm based on matrix, and implement the Chinese-Naxi machine translation. Result of the contrast experiment shows that, compared to the Chinese-Naxi machine translation method based on tree-to-tree translation, the translation accuracy increased after importing subtree alignment.

1. Introduction

Naxi characters are the only existing and active pictograph over the world. Work of Chinese-Naxi machine translation can contribute to the inheriting of Naxi characters effectively, as well as bilingual learning of Chinese-Naxi language. However, there are great differences between Naxi language and Chinese [1–4]. There are unique features in Naxi language such as “verbs being in the end of sentences,” “attribute being postpositional,” and “no auxiliary word,” which cause the alignment missing in Chinese-Naxi bilingual alignment. For example, the Chinese sentence “我喜欢吃黑雪山的松果” (I like to eat pine cones from black snow mountain) is translated to the Naxi sentence “ ☞ (I) ☞ (pine cones) ☞ (black snow mountain) ☞ (like) ☞ (to eat).” Due to the features of pictograph, Naxi language has no auxiliary words like “过” and “的” in Chinese, which causes the alignment missing when executing alignment procedure, as well as the confusion of the original sentence. Some solutions were presented to solve this problem. Xiao et al. presented a word realignment method for statistical machine translation [5], which contributes to the accuracy of machine translation by using the inconsistency of bidirectional word alignment. Xiao also presented a subtree alignment based on unsupervised

learning [6, 7], which can extract more rules compared to word alignment and works well in solving alignment missing problem. As to the Chinese-Naxi word alignment, Li and Yu et al. did some researches on dependency-tree-to-string translation model [8, 9], which improved the translation accuracy in dependency-tree-to-string translation model. Li’s Chinese-Naxi machine translation method based on dependency-tree-to-string translation model made full use of the Chinese syntactic information, but no consideration to Naxi syntactic features. In fact those great syntactic differences between Chinese and Naxi language make the syntax analysis an important role in bilingual machine translation. Therefore, on basis of Li’s dependency-tree-to-string translation method, Gao et al. proposed a Chinese-Naxi machine translation method based on Naxi dependency language model, used dependency parsing syntax information of Naxi side to construct a Naxi dependency language model and fused it into decoding process; to a certain degree, this improved the accuracy of Chinese-Naxi machine translation [10]; However, about the alignment missing, caused by the other characteristic of Naxi language, no auxiliary word, this method did not refer to.

In allusion to aforementioned problem, if we adopt the statistical machine translation method based on

TABLE 1: Expression of the four conditional probabilities.

Rules	Probabilities
$NP(NN(\text{妈妈})) \rightarrow NP(NN(\text{妈妈}))$	$P_{nt}(NP NP)P_{lex}(\text{妈妈} \text{妈妈})$ $P_{tree}("NP(NN)" NP)$
$PP(P(\text{用}) NP(NN(\text{芦苇}))) \rightarrow NN(\text{芦苇})$	$P_{nt}(NN PP)P_{lex}(\text{芦苇} \text{用芦苇})$
$VP(MSP(\text{来}) VP(VV(\text{织布}))) \rightarrow VV(\text{织布})$	$P_{nt}(VV VP)P_{lex}(\text{织布} \text{来织布})$
$VP(PP VP) \rightarrow VP(NN VV)$	$P_{nt}(VP VP)P_{tree}("VP(NN VV)" VP)$ $P_{reorder}("1 - 1, 2 - 2" (PP, VP), (NN, VV))$
$IP(NP VP) \rightarrow IP(NP VP)$	$P_{nt}(IP IP)P_{tree}("IP(NP VP)" IP)$ $P_{reorder}("1 - 1, 2 - 2" (NP, VP), (NP, VP))$

dependency-tree-to-string model, there would be a large quantity of alignment missing due to the lack of consideration to Naxi syntactic features, which would cause the information loss of misaligned word in translation result. In the thesis, we proposed a tree-to-tree translation template of Chinese-Naxi translation based on subtree alignment, and in the procedure of extraction, we synthesize the syntactic features of each language, extend the translation template with misaligned word, and implement the translation of misaligned word in bilingual alignment finally.

2. A Chinese-Naxi Tree-to-Tree Translation Template Extraction Method Based on Subtree Alignment

2.1. Definition of the Subtree Alignment Model. Given a syntax tree pair (S, T) , we define the best inference d^* to realize subtree alignment, and its formal definition is as follows:

$$d^* = \underset{d \in D(S, T)}{\operatorname{argmax}} P_\theta(T, d | S). \quad (1)$$

Here, $D(S, T)$ is the inferential space.

2.2. Inference Probability of the Subtree Alignment Model. The inference probability of the subtree alignment model is as follows:

$$\begin{aligned}
 p(T, d | S) \equiv & \prod_{r \in d} p_{nt}(\text{root}(t_r) | \text{root}(s_r)) \\
 & \cdot p_{tree}(\text{tree}(t_r) | \text{root}(t_r)) \\
 & \cdot p_{lex}(\text{lex}(t_r) | \text{lex}(s_r)) \\
 & \cdot p_{reorder}(\emptyset_r | \text{vnt}(s_r), \text{vnt}(t_r)).
 \end{aligned} \quad (2)$$

Here, there are four factor probabilities: $p_{nt}(\cdot)$ is the mapping probability of nonterminal symbols between Chinese and Naxi. $p_{tree}(\cdot)$ is the generating probability of subtree T which is extended from a node in Naxi side. $p_{lex}(\cdot)$ is the mapping probability of terminal symbols between Chinese and Naxi. $p_{reorder}(\cdot)$ is the coding rule probability of reordering frontier nonterminal. The four conditional probabilities are expressed by bilingual aligned subtree instance, as shown in Figure 1, respectively.

FIGURE 1: Bilingual syntax tree with subtree alignment.

According to the aligning rules of syntax tree with Chinese-Naxi subtree alignment in Figure 1, the four conditional probabilities may be expressed as shown in Table 1.

2.3. Node Deletion and Insertion. In the processing of Chinese-Naxi subtree alignment, if the Naxi language side is empty subtree, the inference probability is changed into

$$\begin{aligned}
 P(T, d | S) \equiv & \prod_{r \in d} P_{nt}(\text{root}(\varepsilon) | \text{root}(s_r)) \\
 & \cdot P_{tree}(\text{tree}(\varepsilon) | \text{root}(\varepsilon)) \\
 & \cdot P_{lex}(\text{lex}(\varepsilon) | \text{lex}(s_r)) \\
 & \cdot P_{reorder}(\emptyset_r | \text{vnt}(s_r), \text{vnt}(\varepsilon)).
 \end{aligned} \quad (3)$$

Here, ε is a special symbol, an expression of empty subtree.

```

(1) Function TrainModelWithEM( $\{(S_1, T_1), \dots, (S_n, T_n)\}$ )
(2)   Set  $\{\theta_{t_{nt}|s_{nt}}^{(0)}\}$  = an initial model
(3)   For  $k = 0$  to  $K - 1$  do
(4)     Foreach non-terminal symbol pair  $(s_{nt}, t_{nt})$  do // for each corresponding non-terminal symbols
(5)        $EC(\theta_{t_{nt}|s_{nt}}) = 0$ 
       E-step: //Compute Expected Count
(6)       Foreach tree pair  $(S, T)$  in sequence  $\{(S_1, T_1), \dots, (S_n, T_n)\}$  do //for each sub-tree pair
(7)         Foreach node pair  $(u, v)$  with symbol pair  $(t_{nt}, s_{nt})$  in  $(S, T)$  do // for frontier node pair
(8)           Foreach rule  $r$  rooted at  $(u, v)$  do //Extract rule  $r$  from the sample
(9)              $EC(\theta_{t_{nt}|s_{nt}}) += \frac{\sum_{d:r \in d \wedge r \text{ is rooted at } (u,v)} P_{\theta^{(k)}}(T, d | S)}{\sum_{d'} P_{\theta^{(k)}}(T, d' | S)}$ 
       M-step: //Compute Maximum Value
(10)      Foreach non-terminal symbol pair  $(s_{nt}, t_{nt})$  do
(11)         $\theta_{t_{nt}|s_{nt}}^{(k+1)} = \frac{EC(\theta_{t_{nt}|s_{nt}})}{\sum_{t'_{nt}} EC(\theta_{t'_{nt}|s_{nt}})}$ 
(12)      return  $\{\theta_{t_{nt}|s_{nt}}^{(K)}\}$ 

```

ALGORITHM 1

If the Chinese subtree is empty, the inference probability is changed into

$$\begin{aligned}
P(T, d | S) \equiv & \prod_{r \in d} P_{nt}(\text{root}(t_r) | \text{root}(\varepsilon)) \\
& \cdot P_{\text{tree}}(\text{tree}(t_r) | \text{root}(t_r)) \\
& \cdot P_{\text{lex}}(\text{lex}(t_r) | \text{lex}(\varepsilon)) \\
& \cdot P_{\text{reorder}}(\emptyset_r | \text{vnt}(\varepsilon), \text{vnt}(t_r)).
\end{aligned} \tag{4}$$

Here, ε is also a special symbol, an expression of empty subtree.

2.4. Training Subtree Alignment Model. EM algorithm [11] is used to train the model parameters and the four conditional probabilities; the training process to $p_{nt}(\cdot)$ is shown in Algorithm 1.

Here, s_{nt} and t_{nt} indicate the nonterminal of source language side and the one of target language side, respectively. $EC(\theta_{t_{nt}|s_{nt}})$ is the Expected Count of $\theta_{t_{nt}|s_{nt}}$. K is the iteration times. u and v express the tree node of source ending and the one of target ending, respectively.

3. A Chinese-Naxi Tree-to-Tree Translation Template Extracting Algorithm Based on Subtree Alignment

Before extracting translation template, we must carry on syntactic analysis at both source language side (Chinese) and target language side (Naxi) in bilingual training corpus to figure out the subtree alignment relation. And then, we can do the extraction by adopting the extracting algorithm based on matrix [12]. According to the subtree alignment relation we can obtain the phrase tree as shown in Figure 1.

The extracting algorithm is shown in Algorithm 2.

TABLE 2: The experimental corpus.

Corpus type	Sentence pairs (k)	Naxi words (k)	Chinese words (k)
Bilingual aligned parallel corpus	35	548	822
Translation model corpus	24	295	547
Development set	16	146	235
Test set	7	70	109

Here, M represents a subtree alignment matrix for the pair of trees (S, T) . $P_{\min}$ is an empirical threshold to control how often rules are pruned. In this work, it is 10^{-7} by default.

4. Experiment and Result Analysis

4.1. Experimental Design. For the reason that research on machine translation of Naxi language is not as well-developed as that of English, public corpus is short of support. To our study, we collected 35,000 Chinese-Naxi parallel sentences from bilingual textbooks of elementary education and dialogues. After word segmentation, dependence syntax parsing, and bilingual word alignment annotation, we developed a Chinese-Naxi corpus for statistical machine translation, from which we selected 16,000 bilingual aligned parallel sentence pairs as development set and 7,000 sentence pairs as test set, and the experimental corpus is shown in Table 2.

In order to validate the accuracy of tree-to-tree Chinese-Naxi translation method based on subtree alignment, we design some contrast experiments to the translation methods, respectively, based on tree-to-string, string-to-tree, tree-to-tree, and tree-to-tree model with subtree alignment [13, 14]. In the experiments, we take “NiuTrans” (1.3.0 version) as the basic prototype system [7]. From the sentence on source side (Chinese), the corresponding Chinese syntax trees are

```

(1) Function MatrixExtract( $S, T, M$ ) // Extract alignment matrix rules
(2)   ForEach node  $u$  in  $S$  do
(3)     ForEach node  $v$  in  $T$  do
(4)       If IsEXTRACTABLE( $\{(u, v)\}, M$ ) do // If the fuction returns true
(5)         next loop
(6)       ForEach tree-fragment pair  $(s_r, t_r)$  rooted at  $(u, v)$  do
(7)         ForEach frontier alignment  $a$  between  $(s_r, t_r)$  do
(8)           If IsEXTRACTABLE( $a, M$ ) then
(9)              $r = \text{CREATERULE}(s_r, t_r, a)$  // Create rule  $r$ 
(10)             $\text{rules.ADD}(r)$ 
(11)   return rules
(12) Function IsEXTRACTABLE( $a, M$ )
(13)   ForEach alignment  $(p, q)$  in  $a$  do
(14)     If probability of  $(p, q)$  in  $M < P_{\min}$  then // When probability less than threshold  $P_{\min}$ 
(15)       return false
(16)   return true

```

ALGORITHM 2

TABLE 3: The results of comparative experiments.

Chinese-Naxi syntax translation system	Experimental set	
	Development set (15187)	Test set (7327)
	Measure	
	BLEU4 [%]	BLEU4 [%]
Chinese-Naxi syntax translation system based on tree-to-string model	22.41	22.53
Chinese-Naxi syntax translation system based on string-to-tree model	22.45	22.58
Chinese-Naxi syntax translation system based on tree-to-tree model	23.59	24.74
Chinese-Naxi syntax translation system based on tree-to-tree model with subtree alignment	25.05	26.16

generated by Berkeley Parser, a syntactic analyzer based on Penn Treebank [15], and from the sentence on target side (Naxi), the corresponding Naxi syntax trees are also generated by Naxi syntactic analyzer which is developed by us. A 3-gram language model was trained by SRILM toolkit. EM algorithm trained a subtree alignment model. Matrix algorithm based subtree alignment extracted tree-to-tree translation template and got 512 template rules. Minimum Error Rate Training (MERT) trained our logarithm linear model and turned the weight parameters. The BLEU4 [16] is selected as the experimental measure.

4.2. Experimental Results and Analysis. The results of comparative experiments are shown in Table 3.

As we can see from Table 3, the Chinese-Naxi syntax translation system based on tree-to-tree model with subtree alignment is +1.5% of BLEU4 higher than that of the Chinese-Naxi syntax translation system based on tree-to-tree model with word alignment and +2.6% higher than that of the Chinese-Naxi syntax translation system based on tree-to-string model. The improvement is attributed to the fact that the Chinese-Naxi translation model based on tree-to-tree translation with subtree alignment solves alignment missing between Chinese and Naxi by deleting or inserting node. For example, when translating the Chinese sentence “我喜欢吃黑雪山的松果” (I like to eat pine cones from

black snow mountain), the proposed method can delete the null alignment word “的” and thus got better result “ꞑꞑꞑ (I) ꞑꞑ (pine cones) ꞑꞑ (black snow mountain) ꞑꞑ (like) ꞑꞑ (to eat).” The experimental results show that the proposed translation model based on tree-to-tree translation with subtree alignment improves the accuracy of Chinese-Naxi machine translation. However, there are also some errors in our translation. After detailed analysis, we found that the main reasons are the characteristics of Naxi language, such as verbs being in the end of sentence, attribute being postpositional, polysemy for one word, and synonyms for multiple words, and the data sparseness that the small size of corpus brings about.

5. Conclusions

This thesis puts forward a Chinese-Naxi syntactic statistical machine translation method based on tree-to-tree model with subtree alignment. The experimental results show that when compared with the tree-to-tree translation model with word alignment, the tree-to-tree machine translation model with subtree alignment wins +1.5% of BLEU4. In the next step, we will expand the scale of corpus. At the same time, aiming at serious structural differences between Chinese syntax and Naxi syntax, we will continue to work on the Chinese-Naxi tree-to-tree machine translation model and

attempt to integrate semantic information into the translation model to improve quality and performance of Chinese-Naxi syntactic statistical machine translation.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work is supported by National Natural Science Foundation of China (Grant nos. 61163022 and 61163004).

References

- [1] S. Mu, "Naxi Dongba hieroglyphic dictionary outline," *Lexicographical Studies*, no. 4, pp. 117–121, 1997.
- [2] S. Yu, "Naxi Dongba language study on applied literature," in *Proceedings of the 99th International Academic Seminar on Dongba Culture: Exploration of Yu Zhenjin's Dongba*, pp. 608–636, Lijiang, China, 1999.
- [3] L. He, "Discussion on inheritance of Dongba culture," *Social Sciences in Yunnan*, vol. 6, no. 1, pp. 83–87, 2004.
- [4] H. Guo, J. Zhao, and K. Shuai, "Research on Naxi pictograph recognition," *Chinese Journal of Scientific Instrument*, vol. 28, no. 4, pp. 749–750, 2007.
- [5] T. Xiao, T. Li, R. Chen, J. Zhu, and H. Wang, "Word realignment for statistical machine translation," *Journal of Chinese Information Processing*, vol. 24, no. 1, pp. 100–116, 2010.
- [6] T. Xiao and J. Zhu, "Unsupervised sub-tree alignment for tree-to-tree translation," *Journal of Artificial Intelligence Research*, vol. 48, pp. 733–782, 2013.
- [7] T. Xiao, J. Zhu, H. Zhang, and Q. Li, "NiuTrans: an open source toolkit for phrase-based and syntax-based machine translation," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL '12)*, pp. 19–24, Association for Computational Linguistics, Jeju Island, Republic of Korea, July 2012.
- [8] L. Li, Z. Yu, C. Mao, and J. Guo, "The extracting method of Chinese-Naxi translation template based on improved dependency tree-to-string," in *Chinese Lexical Semantics: 14th Workshop, CLSW 2013, Zhengzhou, China, May 10–12, 2013. Revised Selected Papers*, vol. 8229 of *Lecture Notes in Computer Science*, pp. 350–358, Springer, Berlin, Germany, 2013.
- [9] Z. Yu, T. Zhang, J. Guo, C. Mao, and J. Li, "Naxi-English bilingual word alignment based on language feature model of Naxi," in *Proceedings of the 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE '11)*, pp. 100–105, Tokushima, Japan, November 2011.
- [10] S. Gao, X. Yang, Z. Yu, X. Pan, and J. Guo, "Chinese-Naxi machine translation method based on Naxi dependency language model," *International Journal of Machine Learning and Cybernetics*, 2015.
- [11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series: B*, vol. 39, no. 1, pp. 1–38, 1977.
- [12] Y. Liu, T. Xia, X. Xiao, and Q. Liu, "Weighted alignment matrices for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, pp. 1017–1026, Association for Computational Linguistics, East Stroudsburg, Pa, USA, August 2009.
- [13] Y. Liu, Q. Liu, and S. Lin, "Tree-to-string alignment template for statistical machine translation," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL-44 '06)*, pp. 609–616, Association for Computational Linguistics, Sydney, Australia, July 2006.
- [14] D. Xiong, Q. Liu, and S. Lin, "Maximum entropy based phrase reordering model for statistical machine translation," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL '06)*, pp. 521–528, Association for Computational Linguistics, Sydney, Australia, August 2006.
- [15] H. Cao, T. Zhao, and S. Li, "Parsing Penn Chinese treebank (CTB) with head-driven model," *Chinese High Technology Letters*, vol. 17, no. 1, pp. 15–20, 2007.
- [16] K. Papineni, S. Roukos, and T. Ward, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*, pp. 311–318, Association for Computational Linguistics, Philadelphia, Pa, USA, July 2002.