The design for restructuring translation model based on multi-feature inference hypothesis

Fu Yan
Teaching and Research Institute of Foreign Languages,Bohai University,No19,Keji Road,Taihe District, Jinzhou City，liaoning Province, China
714097316@qq.com

Abstract

Syntactic reordering model is proposed on the basis of phrase-based statistical translation model in order to handle and count long-distance reordering in machine translation. In the method, various information obtained from monolingual and bilingual corpus is fully utilized under the maximum entropy mode framework. New collocation translation method is different from previous method with over-reliance on bilingual corpora in that monolingual corpora training translation model can be used. Context information is further introduced on the basis of matching internal information. EM algorithm is adopted to estimate context-based vocabulary translation probability. Meanwhile, syntax tree structure is segmented in the model according to the phrase segmentation, thereby avoiding inconsistency between phrases and syntactic structures. In the model, reordering sequence of some structures in syntax tree can be determined according to phrase alignment and word alignment in phrases. Sub-structure reordering probability is calculated according to reordering probability on each node, which is used as characteristic function of log-linear model. Experimental results of the model is significantly higher than score of classic phrase statistical translation model. The results show that syntactic reordering model is effective aiming at phrase-based statistical machine translation, syntactic knowledge and phrase translation process can be better combined. Experimental results show that the model is better than phrase-based statistical machine translation model in the aspect of translation knowledge generalization ability and translation results.

Keywords: translation model; multi-feature inference; EM algorithm; Blue-based evaluation

1 Introduction

High-quality collocation translation is very important for application in machine translation, cross-language information retrieval, second language learning, and many other aspects of natural language processing[1,3]. At present, research on translation knowledge acquisition is mostly based on parallel corpus. Previous scholars have realized a system of extracting collocation translated text from parallel English and French corpus aiming at collocation translation. Phrase and translation terms are also obtained from parallel corpus. However, because it is subject to complexity constraints of various word alignment assumption conditions, it is difficult to achieve satisfactory results in the aspects of dealing with word collocation, reordering, etc. in translation handling process[4,5]. Various phrase-based translation models are established in the forms of joint probability, etc. Continuous inter-translation word string alignment information is utilized on the basis of word alignment knowledge. After the phrase length exceeds some fixed value, the system performance can not be prominently improved due to inevitable sparse data problem of training corpus. In the paper, a new method of obtaining collocation translation knowledge from monolingual corpora is presented[6]. Experimental results show that BLEU scores of translated text can be prominently improved by the method in the paper. In the paper, log-linear model framework is eventually adopted for providing statistic machine translation model based on structural alignment[7]. Experiments show that model proposed in the paper not only has the advantages of good generalization ability in similar syntax -based models. Meanwhile, the test result is prominently better than previous result under the same condition.

2 Initial Establishment of Model

Node-to-node mapping relationship among syntax trees can not be established. For example, word alignment phenomena given in Figure 1 are universal in the actual research. If the syntax tree information is utilized to establish structure correspondence relationship, only original tree structure is decomposed and reconstructed, multi-level structure corresponding relationship can be established among syntactic structure combinations. In the paper, related concepts are introduced according to the research idea, thereby establishing structure mapping.

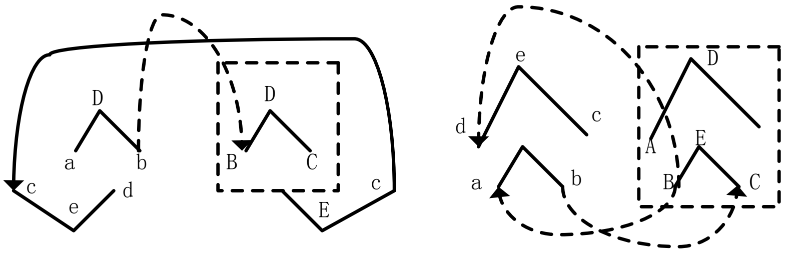

Figure 1 Isomerism of Tree Structure

3 Restructuring and Translation Model Design

Structure group R_1 belongs to non-syntactic structure and consolidation of non-sibling nodes. However, the structure is common in English-Chinese translation. According to our knowledge, current translation model based on synchronous generative grammar can not solve the problem. However, structure group can be utilized for marking phrases not meeting syntactic structure knowledge according to the above concept. For example, related syntax tree is decomposed for obtaining element structure groups of corresponding phrases: 1) Pr cause analysis; 2) V mechanism selection; 3) NP project analysis. Meanwhile, reconstructed element structure sequence can be obtained. Then, the same operation is applied to the target language syntactic tree. It is obvious that element structures can be recombined through restructuring. Bilingual structure mapping can be established with cell structure group as element.

4 Decoding Mechanism Design

Translation process: Input sentences are divided according to extracted source language phrases of phrase inter-translation alignment. Firstly, a pseudo-syntactic structure is randomly selected. Translated text of corresponding phrases in the pseudo-syntactic structure is given as the first phase in translated sentence. Corresponding scores are calculated. Then, a pseudo-syntactic structure can be randomly selected from the remaining pseudo-syntactic structures. Corresponding phrase translated text is provided as subsequent phrase in translated sentence. In addition, corresponding score after combination between the pseudo-syntactic structure and the former pseudo-syntactic structure is calculated. The above process is repeated until pseudo-syntactic structures of the entire syntactic tree are translated. Wherein, node reordering sequence is determined through inward word alignment in phrase inter-translation. Cluster search strategy based on multi-stack of similar reference [19] is adopted for decoding algorithm. All stacks are distinguished with covered word quantity in input sentences. Pruning strategies of combining same assumptions, threshold pruning and histogram pruning is adopted in each stack..

5 Translation Method Based on Assumption Reasoning Mechanism

As shown in Figure 2, Source language syntax tree is decomposed and covered to generate translation hypothesis by utilizing inter-translation source structure group by the translation model in the decoding process aiming at given source language sentence and its syntax tree. Figure 3 is combined. Inter-translation element structure group composed of structure groups R, R2 and Rn is utilized for respectively generating translation hypothesis. Finally, source language fragment covered by cell structure group completes conversion from source language to target language by utilizing phrase inter-translation pairs, namely lexicalization process, thereby completing the translation process.

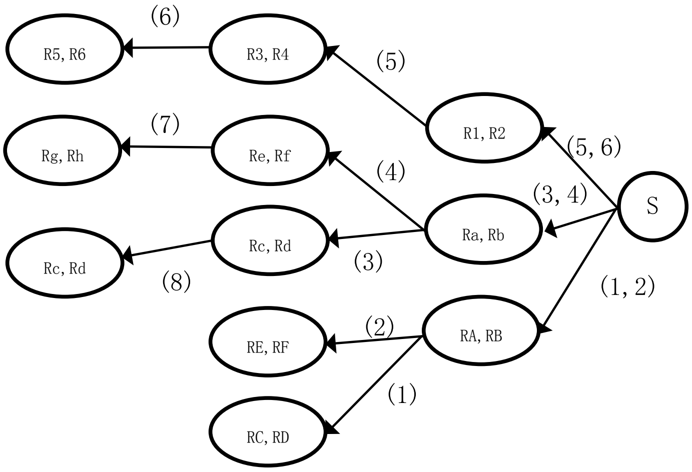

Figure 2 Assuming line of reasoning translation

6 Realization of Assumption Reasoning Model System

Comparison results between system based on structure alignment model realization and system based on phrase model are given as follows. In the paper, word alignment tools are used in experiment aiming at training part of corresponding model. Language modeling tools have translation task of Chinese-English translation. The same data set is used for training the features the same as previous part. Concrete training method is not described in the paper. Example of bilingual structure group automatically extracted in the experiment is also given. Meanwhile, conversion probability of converting from Chinese structure group into English structure group is also given.Only continuous vocabulary and vocabulary fragment knowledge are considered on the basis of generation based on assumption in phrase model search process. In the model, input syntactic tree is firstly decomposed as element structure sequence firstly, then sequence can be reordered according to structure inter-action group knowledge and target language syntactic knowledge. Translation of source language fragment is implemented in structure span. Therefore, the generalization ability of the model is better than the phrase-based model..

7 Experiment Result Analysis

The development set is composed of 1000 Chinese sentences. The test set is composed of 800 Chinese sentences. Each sentence in the development set and test set corresponds to 23 reference translated texts. Two-level Chinese parsing method based on headword-driven model is adopted for Chinese syntactic analysis. Firstly, influence of different reordering features on translation results is investigated. Label features, lexical features and vocabulary/label mixing features are respectively adopted. Labels and lexical features are adopted as reordering features at the same time. It can be seen that BLEU scores are improved compared with original system after introduction of syntactic reordering model. Wherein, translation results of adopting vocabulary features are significantly better than other translation results. However, the result is still lower than the results of only adopting lexical features. It is obvious that discrimination of reordering model is largely strengthened by introduction lexical features. The influence of label features is lower, and negative influence can be caused when it is acted with lexical features at the same time. Influence of word alignment reconstructing mode on final translation results is also examined in table 1. Two methods are respectively used for generating one-to-one word alignments, which is used for determining reordering sequence. Then words can be adopted as syntactic reordering features for reordering. It is obvious that the coverage rate of alignment point is more beneficial for estimating reordering than precision. It can be seen that the reordering error is slightly lowered after syntactic reordering model is adopted, thereby further demonstrating that linguistic syntactic information is helpful for global reordering.

Table 1 Translation evaluation data

System	Reordering results		
	wrong	Uncertain	Right
Before reordering	154	84	346
After reordering	79	96	357
Align restructuring	Translation evaluation		
Intersect	0.541		
Diag	0.589		

8 Conclusion

Bilingual structure alignment conversion knowledge can be integrated in the translation process according to statistical machine translation model proposed in the paper based on structure alignment. The proposed inter-translation structure group can be reconstruction of syntactic structures at different levels. Meanwhile, statistical machine translation reordering method with syntactic information as inspiration in the model can improve generalization ability of the model, which can effectively solve data sparseness problem. In the paper, main innovations are reflected as follows: 1) new method of obtaining collocation method by fully utilizing existing resources is proposed; 2) The proposed method based on monolingual corpus is utilized for effectively estimating context translation probability so that the method in the paper is superior to existing collocation translation method based on monolingual corpus. 3) Maximum entropy model provides system method for obtaining optimal collocation translation results by fully utilizing existing monolingual and bilingual resources. In addition, syntax-based phrase statistic translation reordering model is proposed in the paper. Pseudo-syntactic structure is defined on the basis of parsing tree. Reordering model is established through pseudo-syntactic structure. Translation quality is significantly improved by syntactic reordering model. It is obvious that introduction of linguistic syntactic structure is effective for statistical machine translation reordering. It should be observed that reordering method of sub-node full array is still greatly affected by sparse data. It is still difficult to effectively handle excessive sub-node quantity. We plan to introduce large-scale training data and jumping probability among sub-nodes for solving the problem in future work.

Acknowledgements

This work was financially supported by the Social Science Fund of Liaoning Province(L13DYY025)

References

[1] K.Knight, D.Marcu: *Machine translation in the year 2004*, Proceedings of the 2005 IEEE. International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, USA, (2005).

[2] Q.Liu : Journal of Chinese Information Processing Vol.17(2003),p.1-12.

[3] P.Koehn, F.Och, D.Marcu: *Statistical phrase-based translation*, Proceedings of the Conference on Human Language Technology, Edmonton, Canada, (2003).

[4] W.Cheng, J.Zhao, F.Liu: Journal of Chinese Language and Computing Vol.14(2004),p. 31-40.

[5] C.Tillman, T.Zhang: *A localized prediction model for statistical machine translation*, Proceedings of the 43rd Annual Meeting of the ACL, Ann Arbor,(2005).

[6] B.Xu, X.D.Shi,Q.Liu : Journal of Chinese Information Processing Vol.20(2006),p.1-9.

[7] S.B.Nie, H.Ney: Computational Linguistics Vol.30(2004),p.181-204.

[8]J.Torán: SIAM Journal on Computing Vol.33(2004),p.1093-1108.