Decoding Optimization for Chinese-English Machine Translation via a Dependent Syntax Language Model*

Ying Liu[1], Zhengtao Yu[1,2], Tao Zhang[1], and Xing Zhao[1,2]

[1] School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051, China
[2] Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming 650051, China
liuying_09@yeah.net, ztyu@hotmail.com, ztbpf@126.com, zhaoxing2011@163.com

Abstract. Decoding is a core process of the statistical machine translation, and determines the final results of it. In this paper, a decoding optimization for Chinese-English SMT with a dependent syntax language model was proposed, in order to improve the performance of the decoder in Chinese-English statistical machine translation. The data set was firstly trained in a dependent language model, and then calculated scores of NBEST list from decoding with the model. According to adding the original score of NBEST list from the decoder, the NBEST list of machine translation was reordered. The experimental results show that this approach can optimize the decoder results, and to some extent, improve the translation quality of the machine translation system.

Keywords: Language model, Statistical machine translation, Decoding, Dependent syntax.

1 Introduction

In recent years, more and more syntactic information is incorporated into the machine translation process. As for formal grammar aspect, the ITG model [1], such as Wu Dekai was firstly utilized in synchronization syntax for statistical machine translation. Chiang proposed a hierarchical phrase model, and had the similar phrases with ITG model based on formal grammar [2], but more powerful than ITG model in its rearranging aspect. In linguistics and syntax, Yamda et al [3,4]. Presented and developed the string-tree model, Liu Yang also developed a tree-string model. After then, Lin's and Quirk et al. proposed a statistical translation model [5] based on path-based conversion model and dependent syntax, respectively [6]. In recent two years, the emerging forest-string model, such as Mi Haiao et al.'s research, was also based on syntactic analysis [7].

* This paper is supported by National Nature Science Foundation (60863011), Yunnan Nature Science Foundation (2008CC023), Yunnan Young and Middle-aged Science and Technology Leaders Foundation (2007PY01-11).

H. Deng et al. (Eds.): AICI 2011, Part III, LNAI 7004, pp. 143–150, 2011.

Optimization in the decoding, a Luo Yi's beam search algorithm by Dynamic Programming Thought, Beam Search algorithm based on the design and implementation of a beam search decoder [8]. But did not make the optimization of the language model. In syntactic aspects of linguistic-based, syntactic information only plays a role in the template extraction process in both dependent and phrase syntax after source or target language syntactic analysis. Since the word's syntax relation information is not included in the template, which causes mechanical splicing template in the process of decoding. Consequently, there is no contribution to the decoding due to no effective use of syntactic relation information.

Therefore, in this paper, based on dependent tree-string translation method, we firstly constructed a dependent syntax language model, and then calculated scores of NBEST list from decoding with the model. According to adding the original score of NBEST list from the decoder, the NBEST list of machine translation was reordered to optimize the decoder results.

2 Process of Translation Decoding

In SMT system, the purpose of decoding is to find the most likely translation according to the input source language sentence from a given language model, translation model and other models based on the information. No matter how the model decoding, the process will be a discrete optimal solution for solving NP problems. To current statistical machine translation based on syntax of research shows that, whether it is based on dependency grammar or phrase structure grammar, In terms of whether the template phrase-phrase model, tree-string model, string-tree model, or forest-string model ,templates are taken in the training, including the template of lexicalization and non-lexicalization, the syntactic relation information in the training, such as dependency syntax of subj and phrase structure grammar of NP, the syntactic relation information plays an important role in the extraction template.

After template training is completed, most adopts direct translation model in the decoding, such as formula (1)

$$\Pr(s_1^I \mid t_1^J) = \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(s_1^I, t_1^J)]}{\sum_{\tilde{e}_1^I} \exp[\sum_{m=1}^{M} \lambda_m h_m(\tilde{e}_1^I, t_1^J)]} . \tag{1}$$

$h_m(s_1^I, t_1^J)$ is characteristic function ,Which $m = 1,...,M$, For each $h_m(s_1^I, t_1^J)$, model parameters corresponding λ_m. Our goal is to solve λ_m for making Chinese-English translation pair of the probability of each other's largest, namely $h_m(s_1^I, t_1^J)$ the largest. Where the denominator has no effect on the search results above formula, so our search model is(2):

$$\hat{\alpha} = \arg\max_{\alpha} \{ \sum_{m=1}^{M} \lambda_m h_m(s_1^I, t_1^J) \} . \tag{2}$$

When searching, the characteristic function as a constraint on the template stitching, in the normal decoding process produces NBEST candidate translation, the source language to dependency syntax parsing, from left to right, from bottom to top, changing NBEST candidate translation sequences. Because of containing variable in the template, the process of decoding change into derivation with probabilistic rules, the final selection with the maximum probability derivation. It can be seen for a same source language sentence, the design features of the same functions and decoding dependency syntax-based language model scores recalculated NBEST to try to get better translation results. Using the minimum Bayes risk decoding or re-scoring methods to compare the results of translation of multiple systems, after comparing the optimal sentences regard as the ultimate translation output.

3 Optimization of the Decoding Dependency Language Model

3.1 Statistical Language Model

In the field of statistical machine translation, dependency syntax for machine translation has some features, such as natural vocabulary, much closer to the semantic expression, and could better express the structural differences in languages, and because the dependency relationship is more than the phrase structure close to the significance of their expression between the words, so in this paper, dependency parsing in train the statistical language model.

Statistical language models are widely used to solve the problems of various natural language, such as speech recognition, machine translation, word segmentation, POS tagging, etc. It is the probability for the establishment of a language model, making the correct word sequence is more than the probability of word sequence error probability, $w_1,...,w_m$ sentence for the word sequence. The probability by the formula (3) calculation

$$p(w_1,...,w_m) = p(w_1)\prod_{i=2}^{m} p(w_i \mid w_1,...,w_{i-1}) . \quad (3)$$

Assuming that the conditional probability of the current word with the previous n-1 only words related to (4),

$$p(w_1,...,w_m) = p(w_1)\prod_{i=2}^{m} p(w_i \mid w_{i-n},...,w_{i-1}) . \quad (4)$$

Called Ngram model, this paper training is 3 per language model.

3.2 Dependency Language Model and Training

Statistical language model is used in various fields, such as natural language processing, speech recognition, intelligent input method, ORC recognition. The sequence is used to measure the compliance of a target level. To the English sentence. For example the following two English sentences:

S1: There is a cup on the table.
S2: There are a cup on the table.

From the perspective of English speakers, obviously, the first sentence is clearly more in line with the daily usage of English. But the computer does not know that knowledge, then the language model is to identify these knowledge. Language model is a Markov chain, Expression is a sequence of probability distribution, it can give the sample space of all possible sequences of class probability. In natural language, the language model that any word sequence is a sentence, but their probability are difference. In statistical machine translation, the decoding will match different templates, and ultimately the formation of many translations. Some of these translations the syntax used to meet the target language, while others do not match. Language model's job is to give the translation of sentences with different probabilities, and ultimately choose the best translation.

Dependency syntax language model based on dependency syntax module, including parts of speech and dependencies, dependency relationship which describes the relationship between two words, the potential of expressing a word binding with another word. In the field of statistical machine translation, dependency syntax for machine translation has some features, such as natural vocabulary, much closer to the semantic expression, and can better express the structural differences in languages, and because the dependency relationship is more than the phrase structure close to the significance of their expression between the words, so this paper to train the statistical dependency parsing language model, dependency language model training using the method of maximum likelihood estimation formula(5):

$$p(w_i \mid w_{i-n}, \ldots, w_{i-1}) = \frac{C(w_{i-n} \ldots w_{i-1}, w_i)}{C(w_{i-n} \ldots w_{i-1})}. \quad (5)$$

Count(X) Said X in the training corpus that the number of occurrences. This is not the word sequence, but the sentence sequence dependencies. Using medical field to 3636 in English as the training set and Stanford parser for training tools, to the training set parsing each sentence in English, the following is an example:

Cold is caused by a virus.
Syntactic analysis of the dependency tree is as follows:
nsubjpass(caused-3, Cold-1)
auxpass(caused-3, is-2)
det(virus-6, a-5)
agent(caused-3, virus-6)

The mark indicates that the left parentheses dependencies between the words, the numbers behind words indicate the position of words in the sentence. To remove the dependency tree of the words, in order to retain syntax tree of the syntactic relations between words, from the training set by 3636 through the dependency syntax analysis and the extracted sentence contains only dependency relations between the words "sentence". For the above sentence, namely:

Nsubjpass auxpass det agent

Sequence of this kind of dependency as a "sentence", then, language model training parameters usually SRILM tools to srilm tool training 3 per dependency syntax language model. Language model output is the probability of a word sequence, that is, the probability of ternary grammar. In the training text, given any passage of the medicine field for consecutive words text, call interface provided SRILM, phrases can be easily obtained given the probability on the training text.

3.3 Optimized for Decoder Figures

Re-calculate the candidate translation scores and change the N-BEST candidate translation sequences based on dependency syntax relationship language model. Relative to the decoding process in Fig.1, After increasing the dependency syntax relationship language model constraints in the decoding stage, the whole process of translation system showed in Fig.1. The translation process of dependency tree-string:

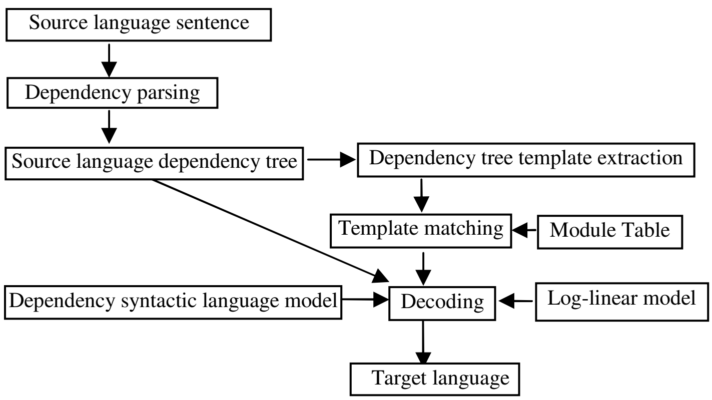

Fig. 1. Syntactic dependency relations role in the decoding stage

The task of the decoder is learned from the training text language model, translation model and the model of other additional information to determine the most the most similar translation of the source sentence. Based on dependency syntax relationship language model to re-calculate the candidate translation scores, and change NBEST candidate translation sequences. Relative to the decoding process of Fig.1, it's increasing the dependency syntax relationship language model constraints in the decoding stage, Fig.1 shows the translation process that whole process of translation system dependency tree-string.

From the struct tree of the source language to the string of the target language, the dependency tree-string, the translation model is used the role of randomization to describe the conversion of the relationship between the tree-tring, the decoding module framework showed in Fig.1. Input source language sentence, after dependent syntactic analysis, get the source language dependent tree. Then extracted template from the source language dependent tree, here we don't have to consider the target language, nor do word alignment constraints. The Process of extraction is bottom- up, first we get the child node's treelet, and then, the child node's assembled into parent node treelet. After getting the source language treelet, matching them with DTSC table which extracted from the modules table, if an extract of the source language side treelet is the same as the extracted treelet, we say that the template matching, and put it on the matching template table. After getting all the matching template table, we apply decoding algorithm to decode, and get the target language string.

Template table and template matching process played an important role in the whole translation process, and the translation efficiency mostly depends on whether the extracted templates is accurate. And for the medical domain, we use the template which defined in the artificial construct rules, it has many advantages such as translation accurate, adjustable sequence accurate and so on. This feature of the restricts areas avoid large-scale manual labor. Adding the template which defined in the artificial rules to the template extracted by generic domain, achieving template choice using the template matching algorithm in the decoding stage.

We selected 500 pairs of Chinese-English sentence pairs for test set, after producing N-BEST candidate translation during the normal decoding processes, record scores for each English sentence, analysis dependency syntax for each candidate translation, and get the dependency syntax sequence nodes, calculate the score of the node $DPni$ based on dependency syntax language model, so the i-sentence score of the NBEST candidate translation's final formula is (6):

$$scorei = Pni + DPni\ . \qquad (6)$$

To re-sort the translation of the NBEST candidate using a new $scorei$, than output one or N translation results. Decoding process is similar to the process of analyzing a single language, which is the source language side of "syntactic analysis", with the transition rules to match the right side of the source language and generate the structure tree of the target language with the left. Using bottom-up method to traverse the source language structure tree for each node when decoding, for the child tree which use the current node as root node, the decoder searches all matching TAT, to obtain the corresponding candidate translation of these nodes through TAT, and store them to the corresponding stack. The higher-level translation obtained assembly by the lower-level child node which identified by the TAT.

We want improve the translation speed to reduce the search space, the paper made the following restrictions: Limit the match TAT number <= 20 for each node.

For the same N-word translation of the left and the right end, decoder combine them, and cut out the translation with a low probability using window function, only keeping of the highest probability one.

The results of experiment show that we re-calculate the candidate translation scores based on dependency syntax relationship language model, change the N-BEST candidate translation sequences. Increasing the dependency syntax relationship language model constraints in the decoding stage, we have a certain limit on the search algorithm.

3.4 Experimental Results and Comparative Analysis

The model's training and decoding is a key step in SMT, training of the model's including extraction template and parameters of training that the method of template extraction also determines the way of decoding. It was adopted the dependency tree-string's translation method to extract dependency trees for string template formed template library, and then training the parameters of the characteristic function, ultimately a final recursive paste and replace formed decoding process. Training set is 3636 Chinese-English sentence pairs of medical domain, test set is randomly selected 500 Chinese-English sentence pairs. That the conditions of training set and test set shown in Table 1.

Table 1. Case Training Set and Test Set

Set	Sentences	Chinese	English
Training set	Sentences	3636	
	Sentences	35374	37581
Test set	Sentences	500	
	Sentences	4015	4681

Table 2. NBEST Sequence Changes and the Score

Name	Sentences
Training set	500
Total change	23
Positive change	17
Negative change	6
System	BLEU4
Increase the language model	0.2312
Increase the language model	0.2323

We adopt common BLEU evaluation method, the results of translation rate, into a domain system of rules and resources—Based on Dependent Tree-String Translation System to compare experimental system, for each sentence generated NBEST candidate translation output the best translation results, Table 2 shows adopted based on dependency syntax relationship after constraints the language model training set after the translation of the best candidate for change.

Table 2 shown after produced NBEST candidates in the translation of increase language model, we adopt based on dependency syntax relationship language model re-calculated scores, the output of the best candidates for translation, and BLEU2 evaluation score results.

It can be seen in 500 of the training set, about 4.6% of the candidate translation changes, which occurred about a positive change in 74% and 16% of the sentences appeared negative changes. On the whole, take the sentence for the unit, and evaluation standards for BLEU calculation of score, about 3.4% of the translation effect of the sentence be enhanced, higher than the open fields and improve the accuracy of machine translation.

4 Conclusion and Outlook

This paper presents a training the dependency language model in medical data sets based on dependent tree-string translation method, and this model calculates scores of NBEST list generated from decoding which changes sequence of machine translation of NBEST and restricts search algorithm, in order to optimize results of decoder. The experimental results show that this approach can optimize the decoder results and improve the translational quality of the machine translation system to some extent. In the next work, we need to incorporate more syntactic features for machine translation.

References

1. Wu, D.: Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics 23, 377–404 (1997)
2. Chiang, D.: A hierarchical phrase-based model for statistical machine translation. In: Proceedings of ACL, Ann Arbor, pp. 263–270 (2005)
3. Yamada, K., Knight, K.: A syntax-based statistical translation model. In: Proceedings of ACL 2001, pp. 523–530 (2001)
4. Galley, M., Hopkins, M., Knight, K., Marcu, D.: What's in a translation rule. In: Proceedings of HLT-NAACL 2004, pp. 273–280 (2004)
5. Lin, D., Cherry, C.: Word Alignment with Cohesion Constraint. In: Proceedings of HLT/NAACL 2003, Companion Volume, pp. 49–51. Edmonton, Canada (2003)
6. Quirk, C., Menezes, A., Cherry, C.: Dependency treelet translation: Syntactically informed phrasal SMT. In: Proceedings of ACL, Ann Arbor, Michigan, pp. 271–279 (June 2005)
7. Luo, Y.: Phrase-based statistical machine translation decoder beam search optimization and implementation. China University of Technology, Anhui (2007)
8. Deyi, X.: Research on Statistical Machine Translation Based on Bracketing Transduction Grammar and Dependency Grammar. Institute of Computing Technology Chinese Academy of Sciences, Beijing (2007)