Phrase Filtering for Content Words in Hierarchical Phrase-Based Model

Xing Wang[1], Jun Xie[2], Linfeng Song[2], Yajuan Lv [2], and Jianmin Yao[1]

[1] School of Computer Science &Technology, Soochow University, Suzhou, China
{20114227047,jyao}@suda.edu.cn
[2] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, China
{xiejun,songlinfeng,lvyajuan}@ict.ac.cn

Abstract. When hierarchical phrase-based statistical machine translation systems are used for language translation, sometimes the translations' content words were lost: source-side content words is empty when translated into target texts during decoding. Although the translations' BLEU score is very high, it is difficult to understand the translations because of the loss of the content words. In this paper, we propose a basic and efficient method for phrase filtering, with which the phrase' content words translation are checked to decide whether to use the phrase in decoding or not. The experimental results show that the proposed method alleviates the problem of the loss content words' and improves the BLEU scores.

Keywords: hierarchical phrase-based model, content words, phrase filtering.

1 Introduction

The Hierarchical phrase-based model (Chiang,2005)was proposed by David Chiang in 2005. The model is a synchronous context-free grammar translation model that learns context-free grammars from a bilingual word-aligned corpus. Learned grammar phrases are used during decoding. Compared to the phrase-based model(Koehn et al.,2003), the hierarchical phrase-based model using hierarchical phrases can capture the reordering between the phrases, the model has superior performance. Compared to syntax-based translation models(Liu et al.,2006; Xie et al.,2011), hierarchical phrase-based model learns context-free grammars from a bitext without any syntactic information. Therefore hierarchical phrase-based model becomes one of the most active models in study .

Hierarchical phrase-based model extracts huge numbers of synchronous context-free grammar phrases from a bilingual word-aligned corpus. However the model is prone to learn noisy phrases due to noise in the training corpus or wrong word alignment. Specifically, since some content words have no counterpart in translation, the model extracts noisy phrases omitting content words. Using phrases that omitting contents words leads to syntax and semantic errors in translation. Here are some samples that using some noisy phrases in our Baseline experiments:

P. Liu and Q. Su (Eds.): CLSW 2013, LNAI 8229, pp. 490–498, 2013.

Example1: 房间的灯不亮 —> the light does not work noisy phrase*：房间 的 X —> the X*

Example2: 您是在这儿用餐还是带走? —> here or to go ? noisy phrase*：您 是 在 X —>X*

The translation is fluent in example 1, but it uses a noisy phrase (房间 的 X -> the X). The noisy phrase leads to the lack of the content word "房间", and the translation remains ambiguous because of the use of noisy phrase. The translation in example 2 using the noisy phrases leads to the semantic error. The reason why the above phenomenon occurs is the use of noisy phrases during decoding. The alignment mistakes cause the model inevitably to extract the noisy phrases. Therefore, it is necessary to detect the phrases' content words translation. Filtering noisy phrases can reduce the model's phrases table and accelerate the decoding speed to improve the quality of the translation.

The remainder of the paper is organized as follows: Section 2 describes the proposed method of phrases filtering; Section 3 introduces the experimental setup; The experimental results and the discussions are presented in Section 4; Section 5 concludes the paper and suggests directions for future work.

2 Phrase Filtering

We can find using the noisy phrases causes the deterioration in the quality of the translation output through above examples. Alignment mistakes cause source side's content words have no counterpart in the target side, resulting in extracting noisy phrases. Using the noisy phrases during decoding, the translation's content words lost. However, BLEU metric, the most popular gain function for automated MT evaluation, treats all characters equally, so BLEU score can't describe the content words omitting in the translation.

Therefore it is necessary to filter noisy phrases. For each phrase, we recognize phrase's content words in source side, then check source side's content words have its counterpart or not. Filtering the phrases that source side's content words have no counterpart to avoid omitting content words during decoding.

2.1 Content Words Recognition

Before check source side's content words have counterpart or not, we need to recognize the content words in phrase's source side.

The ideal approach is tagging sentences in training corpus before word alignment step. then do word alignment with part-of-speech information. But this processing method is flawed: for example, the word "计划" in different contexts will be marked as the verb "计划/ v" or the noun "计划/ n". Word alignment toolkit GIZA++ treat " 计划/ v " and "计划/ n " as different words. Tagging the sentences means that we expand word alignment space, we will face a serious problem of data sparse. In this paper, Using above method with 260k sentence pairs of training corpus degrade the quality of translation.

Therefore, the reorganization of content words should consider both the data sparse problem and part-of-speech tagging accuracy. This paper proposes a compromise method shown in Figure 1:

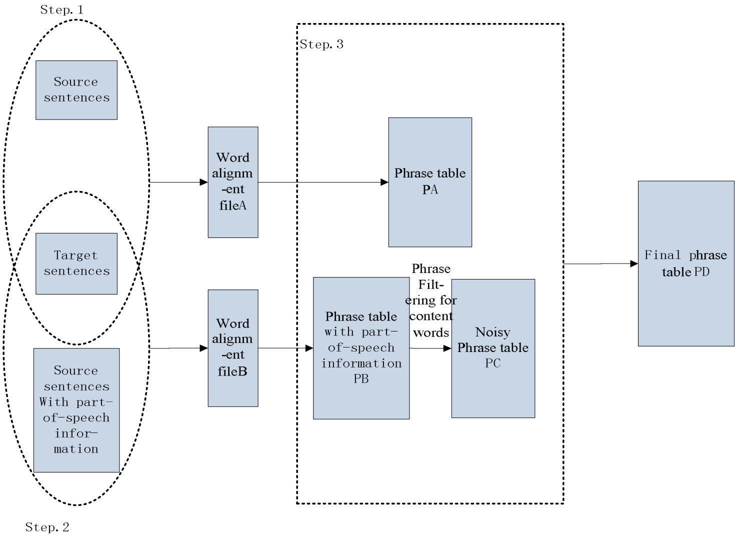

Fig. 1. The flowchart of content word recognition for phrase filtering

Step .1: Align words between source side and target side. Then we can obtain bilingual alignment information and bilingual word alignment file A.

Step 2: Tag the source sentence. Using source side with part-of-speech information, target side and the bilingual alignment information in Step.1 to generate bilingual word alignment file B.

Step.3: Bilingual word alignment file A can be used to generate a common phrase table PA. Bilingual word alignment file B can be used to generate a phrase table PB with source side's part-of-speech information. Filtering phrase table PB through checking source side's content words have counterpart or no, then the noisy phrase table PC can be obtained. The phrases in the noisy phrase table PC will be removed from the common phrase table PA. The remainder of the common phrase table PA is the final phrase table PD used during decoding.

2.2 Phrase Filtering

Alignment mistakes may appear in word align step. The alignment mistakes cause the model extract the noisy phrases, some noisy phrases' content words in source side have no counterpart in its target side, while a correct phrase's content words in source side should have the counterpart in the target side.

Definition 1: For a phrase X -> <γ, α>, if any content word in source side γ have no counterpart in the target side, we called it pseudo source side content words empty phrase.

Definition 2: If a phrase X -> <γ, α>, if any content word in target side α have no counterpart in the source side, we called it pseudo target side content words empty phrase.

Pseudo source side content words empty phrase and pseudo-target side content words empty phrase are called by a joint name pseudo content words empty phrase.

Why we call it pseudo content words empty phrase ? we can think about the situation: the content words' counterpart is appear at the other side in fact, but it has no alignment to its counterpart due to alignment mistakes. We will use a synonym expansion method to alleviate this problem in another section below.

Filtering the phrase by checking its' content words in source side have counterpart or not. pseudo content words empty phrases are filtered directly, the model won't use these phrases during decoding. For example, the phrase（*那样*$_1$ *的* *东西* 。$_2$ —> *like*$_1$ *that*$_1$.$_2$）(the subscripts indicate word alignments) , its source side words "的 "and "东西"have no counterpart in its target side. So this phrase is a pseudo content words empty phrase, it will be filtered out directly in the phrase filtering strep.

2.3 Hierarchical Phrase Filtering

Simply checking the phrase is pseudo content words empty phrase or not will filter a large number of qualified phrases, For example, the phrase （*做*$_1$ *了*$_1$ *很* *傻*$_2$ *的* *事* 。$_3$—> *acted*$_1$ *like* *a*$_4$ *fool*$_4$.$_3$）(source words "做""了"align to the target word"acted"). The phrase will be filtered because the source content word "事" has no counterpart in its target side. There are a lot of idiomatic expression that source side's content words can not find its counterpart in the word alignment step, then lots of qualified phrases are treated as content words empty phrases mistakenly, they will not appear during decoding.

According to the introduction of hierarchical phrase-based model, its phrases divided into the initial phrases and hierarchical phrases. The initial phrases mainly deal with the translation of the string. And the hierarchical phrases process sentence's structure, it is responsible for the sentence overall tone sequence. The above idiomatic expression misjudgment pseudo content words empty phrase generally does not involve the reordering of sentence, so we can only check hierarchical phrases in this step.

2.4 Synonym Expansion

In the above phrase filtering step, we only check phrases' content words in source side have counterpart or not. In this way, the phrase（*负*$_1$ *的* *力量* 。$_2$ —> *positive*$_1$ *force* .$_2$）,we filter the phrase because the source content word "力量" has no counterpart in its target side. It isn't a noisy rule in fact. We consider the use of alignment vocabulary translation probability to help filtering phrases, to make up for mistakenly filtered phrase.

Lexicalization probability is the probability of the word the A translated to the target language word B in a bilingual corpus. Now we consider the word"力量 "translation probability in the bilingual corpus in descending order of translation probability:

Pr(power | 力量) = 0.292
Pr(emphasized | 力量) = 0.138
Pr(strength | 力量) = 0.123
Pr(capability | 力量) = 0.076
Pr(of | 力量) = 0.046
…………

We use two pruning strategy to select the word A's basic translation in a bilingual corpus: First, the histogram pruning strategy. Set pruning value α, choose the top maximum probability α translation as lexicalized words translation. For example, let α = 5, basic translations of the word "力量" are {power, emphasized, strength, capability, of}. Second, the threshold pruning strategy. Lexicalization probability is greater than the pruning threshold β translation can be used as the basic translation of words. For example, let β = 0.1, the basic translations of the word "力量" are {power, emphasized, strength}.

As we know, concepts are represented by Wordnet (Miller 1995, Fellbaum 1998a) synonym sets and are expanded by following the typed links included in Wordnet. so we can use basic translation the WordNet synonym expansion. The word "power", the basic translation of the word "力量" , find its synonyms through WordNet. We can obtain the word "force" and other words, therefore the phrase （*负* $_1$ *的 力量 。* $_2$ —> *positive* $_1$ *force .* $_2$） is recalled. In this way we can make up for alignment mistakes, and recall the mistakenly filtered phrases.

Definition 3: If a phrase X -> <γ, α, ~>, if any content word in source side γ have no counterpart in the target side and the phrase can't be recalled through Wordnet Synonym expansion, called source side content words empty phrase.

Definition 3: If a phrase X -> <γ, α, ~>, if any content word in target side γ have no counterpart in the source side and the phrase can't be recalled through Wordnet Synonym expansion, called target side content words empty phrase.

Source side content words empty phrase and target side content words empty phrase are called by a joint name content words empty phrase.

3 Experimental Setup

We carried out MT experiments for oral translation from Chinese to English. The training data includes IWSLT10 test set and the indoor bilingual corpus that we obtained from the Web, consists of about 260k parallel pairs. The tuning set is IWSLT03 test set. The test set includes IWSLT07 test set and IWSLT08 test set. We uses Giza++ alignments (Och and Ney, 2000) symmetrized with the grow-diag-final-and heuristic. Chinese word segmentation and Chinese part-of-speech tagging were done by our indoor tool named PBCLAS. we used SRI Language Modeling Toolkit (Stolcke, 2002) to train a 5-gram model with modified Kneser-Ney smoothing on the 6,000k English sentences.

The experiment using the minimum error rate training (Och, 2003) for optimizing the feature weights. Our evaluation metric is IBM BLEU(Papineni et al.,2002),which performs case-insensitive matching of n-grams up to n = 4.

This paper set six systems to verify the function of filter phrase filtering, initial phrase filtration, filter hierarchical phrase, synonym expansion. They are divided into two groups: nouns group and nouns and pronouns group. Describe as follows:

Baseline: Using the common phrase table without any processing during decoding.

Filter-all: Using phrase table that all phrase checked during decoding.

Filter-nonX: Using phrase table that only initial phrase checked during decoding.

Filter-X: Using phrase table that only hierarchical phrase checked during decoding.

Filter-nonX + WordNet: Using phrase table that only initial phrase checked and recalled the mistakenly filtered through Wordnet synonym expansion during decoding.

Filter-X + WordNet: Using phrase table that only hierarchical phrase checked and recalled the mistakenly filtered through Wordnet synonym expansion during decoding.

4 Experimental Results and Discussions

The Baseline system use the phrase table without any processing during decoding. the phrase table consists of 7,885,749 phrases. Some phrases' source content words have no counterpart in the target side, Result in table 2 example : some system's output omit some content words. According to our statistics, 8 percent of all sentences in IWSLT07 test set and 6.4 percent of all sentences in IWSLT08 test set have the same problem in our Baseline's output. Therefore it is necessary to filter the phrase table.

Experiment divided into two groups: the first group is phrases' source nouns and pronouns recognize and check, and the second group is phrases' source nouns recognize and check.

Table 1. Experimental results on test set

		nouns and pronouns group			nouns group		
		IWSLT07	IWSLT08		IWSLT07	IWSLT08	
		BLEU(%)	BLEU(%)	number of filtered phrases	BLEU(%)	BLEU(%)	number of filtered phrases
Baseline		48.42	54.34		48.42	54.34	
Filter-all		48.68	54.05	333 ,613	49.82	53.08	143,114
Filter-nonX		48.49	54.33	192,455	48.66	54.38	87,947
Filter-X		48.73	53.67	141,158	**49.55**	**55.16**	55,167
Filter-nonX + WordNet	**α=5**	48.64	54.33	155,882	48.68	53.79	58,297
	β=0.1	48.49	54.33	178,514	48.66	54.38	68,195
Filter-X + WordNet	**α=5**	48.83	53.99	124,046	49.55	55.16	51,340
	β=0.1	48.73	53.67	132,057	49.55	55.16	52,473

In the first experiment group, compared to the Baseline, the system Filter-X's result shows that the filtering of initial phrases have an effect on the translation. From the example 1 in Table 2 we can see ,during decoding ,the translation's semantics is more accurate avoid using the noisy phrase. But the system Filter-nonX's BLEU score decrease in IWSLT08 test set. The reason is that some omitted content words expression are reasonable in the set references, such as Example 2 in Table 3.the phrase（X 什么 东西 —> what X）that used in the Baseline System is reasonable. System Filter-nonX+WordNet and systems the Filter-X+WordNet's BLEU score improve a little contrast to corresponding system Filter-nonX and system Filter-X, but they recall a number of mistakenly filtered phrases. the translation output changes not much because the recalled phrases' length is long and they are hardly used during decoding.

Table 2. Part of the quality improved obviously translation

example1	original text	房间 的 灯 不 亮 。
	Baseline's output	the light does not work .
	New output	the light in the room is not working .
example2	original text	原文：今天 有 好多 旅客 。
	Baseline's output	there are lots of passengers .
	New output	there are so many passengers today .

Table 3. Part of the quality improved not obviously translation

example1	original text	请问 这 是 什么 东西 ?
	Baseline's output	what is this ?
	New output	what is this item ?
example2	original text	请 给 我 炒 鸡蛋 。
	Baseline's output	scrambled eggs , please .
	New output	i 'd like scrambled eggs .

In the second experiment group. compared to the Baseline sytstem, System Filter-nonX's BLEU scores improve significant in two test sets: IWSLT07 test set, increased by 1.03 percent (0.4955 vs .4842) ,and IWSLT08 test set , increased by 0.82 percent (0.5516 vs 0.5434). Compared to the first experiment group, these systems are also greatly improved. This group retains part of idiomatic expression phrases. From the example3 in Figure 3, and the phrase（请 给 我 X —> X please）is filtered in the first experiment group, but it retain in the second experiment group. It is the same to the phrase（请 您 告诉 我 X —> please tell me the X）. System Filter-nonX+WordNet and system Filter-X+WordNet successfully recall a number of mistakenly filtered phrases but improve BLEU score a little. Overall, compared to the Baseline, other systems' BLEU score improves; it proved that our method is effective.

The system Filter-X's BLEU score improves more than the system Filter-nonX's BLEU score. Because system Filter-nonX filter more idiomatic expression phrases, removing these phrases deteriorate the translation quality. System Filter-nonX+WordNet and system Filter-X+WordNet 's results show that recalling the mistakenly filtered phrased though Wordnet can help system perform better.

5 Concludes the Paper and Future Work

This paper presents a phrase filtering method to filter phrase that source-side content words is empty when translated into target texts for hierarchical phrase-based model, which can accelerate the decoding speed as well as improve translation quality. Experimental results on IWSLT test sets show that the method can improve the BLEU score as well as solve content words omit problem in translation.

Future work will be focusing on the following two aspects. First, for the initial phrase processing. We learn to recognize the idiomatic expression to filter the initial phrase more effective, and we can adding more information to recognize the idiomatic expression phrases. Second, We can try to build the phrase filtering model, and use the soft constraints for the noisy phrases to improve the experimental performance.

Acknowledgement. The authors were supported by High-Technology R&D Program (863) Project No 2011AA01A207, National Natural Science Foundation of China (No. 61003152, 61272259, 61272260, 90920004), the Special fund project of the Ministry of Education Doctoral Program (No. 20103201110021), the Natural Science Foundation of Jiangsu Province (No. BK2011282), the Major Project of College Natural Science Foundation of Jiangsu Province (No. 11KJA520003) and the Natural Science Foundation of Jiangsu Province, Suzhou City (No. SYG201030， SH201212).

References

1. Chiang, D.: A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 263–270. Association for Computational Linguistics (2005)
2. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 48–54 (2003)
3. Liu, Y., Liu, Q., Lin, S.: 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, pp. 609–616 (2006)
4. Xie, J., Mi, H., Liu, Q.: A novel dependency-to-string model for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 216–226 (2011)

5. Och, F.J., Ney, H.: Improved Statistical Alignment Models. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics (2000)
6. Stolcke, A.: SRILM-an extensible language modeling toolkit. In: Proceedings of the 7th International Conference on Spoken Language Processing 2002, pp. 901–905 (2002)
7. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of 41st Annual Meeting on Association for Computational Linguistics, pp. 160–167. Association for Computational Linguistics
8. Papineni, K., Roukos, S., Ward, T., et al.: Bleu: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)
9. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM 38(11), 39–41 (1995)
10. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)