Phrase-based Parallel Fragments Extraction from Comparable Corpora

Xiaoyin Fu, Wei Wei, Shixiang Lu, Zhenbiao Chen and Bo Xu
Interactive Digital Media Technology Research Center, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
{xiaoyin.fu,wei.wei,shixiang.lu,zhenbiao.chen,xubo}@ia.ac.cn

Abstract

We present a phrase-based method to extract parallel fragments from the comparable corpora. We do this by introducing a force decoder based on the hierarchical phrase-based (HPB) translation model to detect the alignments in comparable sentence pairs. This method enables us to extract useful training data for statistical machine translation (SMT) system. We evaluate our method by fragment detection and large-scale translation tasks, which show that our method can effectively extract parallel fragments and improve the performance of the state-of-the-art SMT system.

1 Introduction

Parallel corpora are valuable resources for training a statistical translation system. In most cases, it has been an effective way to build state-of-the-art statistical models using a large scale of parallel corpora. However, the parallel corpora only exist in particular domains for a few number of language pairs, such as international conference recordings and legal texts. Since comparable corpora exist in large quantities with many languages, and the exploitation in them for extracting parallel data can be very useful for SMT system, the acquisition of parallel data from comparable corpora has caught much attention.

Various methods (Zhao and Vogel, 2002; Munteanu and Marcu, 2005; Abdul-Rauf and Schwenk, 2009; Smith et al., 2010) have been previously proposed to extract parallel data from comparable corpora at the sentence level. These methods share the same framework, which firstly identifies candidate document pairs and then extracts parallel sentences from the obtained documents. However, it is found that most of these sentences are comparable sentence pairs (Hong et al., 2010), which embed non-parallel fragments or even lack translations. Consider the comparable sentence pair from Chinese to English in Figure 1. Methods for extracting parallel sentences will bring in noise when these bilingual sentences are retained. But discarding them is also not a wise choice, as there are still some useful parallel fragments as the underlines shown in the figure.

中国 政府 开始 大力 发展 中国 的 经济 。

And developing the economy of China is the practical choice for the Chinese government .

Figure 1: Example of comparable sentence pairs. The parallel fragments are marked by underlines.

In order to deal with this problem, further efforts (Munteanu and Marcu, 2006; Quirk et al., 2007; Kumano et al., 2007; Lardilleux et al., 2012) were made to obtain parallel data at the fragment level. The work of (Riesa and Marcu, 2012) detected parallel fragments using the hierarchical alignment model. However, this approach obtains fragments from parallel sentence pairs, which limits its application in comparable corpora. (Hewavitharana and Vogel, 2011) have explored several alignment approaches to detect parallel fragments embedded in comparable sentences. However, these approaches extract fragments mainly using the lexical features and considering the words in parallel fragments are independent, which make it difficult to measure the alignments exactly.

In this paper, we present a phrase-based method, which considers both the lexical and phrasal features, to extract parallel fragments from comparable corpora. We introduce a force decoder based on the HPB translation model to detect parallel fragments for each sentence pair. The results show that our method can effectively extract parallel fragments from the comparable corpo-

International Joint Conference on Natural Language Processing, pages 972–976,
Nagoya, Japan, 14-18 October 2013.

ra and significantly improve the performance on Chinese-to-English translation tasks.

2 Parallel Fragments Extraction

2.1 HPB Translation Model

The HPB translation model (Chiang, 2005) has shown strong abilities in SMT for its capability in generalization. It is based on the weighted synchronous context-free grammar (SCFG). And the translation rule is represented as:

$$X \rightarrow \langle \alpha, \gamma, \sim \rangle \tag{1}$$

where X is a non-terminal, α and γ are source and target strings with terminals and non-terminals. $\sim$ describes a correspondence between the non-terminals in α and γ.

Two glue rules are added so that it prefers combining hierarchical phrases in a serial manner:

$$S \rightarrow \langle S_1 X_2, S_1 X_2 \rangle \tag{2}$$

$$S \rightarrow \langle X_1, X_1 \rangle \tag{3}$$

2.2 Force Decoding based on HPB model

The force decoding can be seen as a bilingual parsing process that generates derivation trees from both sides of the sentence pair with an existing H-PB model.

Let $\mathbf{e} = e_1^M$ and $\mathbf{f} = f_1^N$ be the source and target sentences in comparable corpora. For each of the sentence pair $\mathbf{e}$ and $\mathbf{f}$, the decoding process enumerates all of the possible bilingual derivation trees $\varPhi$ with HPB rules from bottom to up. At each node in these derivation trees, the decoder generates alignments by recursively combining phrases generated from the current node's children, and builds up larger and larger alignments. It should be noted that these nodes can be generated only if the alignments are exactly contained in both elements of the sentence pair. The derivation process works similarly to a CKY parser, moving bottom-up and generating larger constituents. However, the force decoder generates derivation trees for both of the bilingual sentences simultaneously and these trees do not have to span the entire sentences, especially in the non-parallel sentences, which is quite different with the CKY parser.

Still considering the comparable sentence pair in Figure1. Figure 2 gives an example of extracting one of the parallel fragments by force decoding with the following HPB rules:

$$X \rightarrow \langle \text{发展 } X_1, \text{developing } X_1 \rangle$$
$$X \rightarrow \langle X_1 \text{ 中国}, \text{China} \rangle$$
$$X \rightarrow \langle X_1 \text{ 经济}, \text{the economy} \rangle$$
$$X \rightarrow \langle X_1 \text{ 的 } X_2, X_2 \text{ of } X_1 \rangle$$

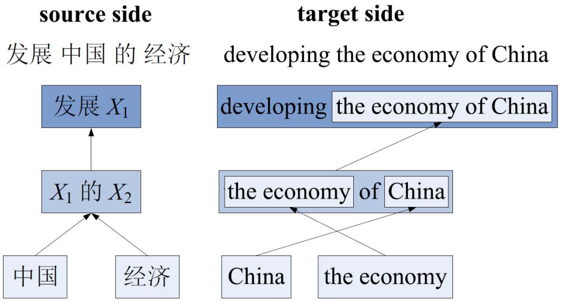

Figure 2: Example of derivation trees in force decoding. To give better illustration, the non-terminal rules from parent nodes are combined with the rules from child nodes on the target side.

It can be seen that the bilingual derivation trees, which represent the source and target fragments, are generated simultaneously. At the first derivation step, it is found that the Chinese words "中国" and "经济" from the source side can be translated into English as "China" and "the economy", which are exactly contained in the target sentence. Then we keep these words as the nodes in bilingual derivation trees, and continue to generate parent nodes by combining these child nodes bottom-up. The derivation process continues until there are no bilingual nodes that meet the words in both of bilingual sentences. At last, we will get the parallel fragments "发展 中国 的 经济" and "developing the economy of China" from the top of the derivation trees.

2.3 The Extension of HPB Rules

In our force decoding framework, there are some words that do not have translation rules, such as the out-of-vocabulary (OOV) words. This case could make up a large portion in the comparable corpora. To overcome this drawback, the HPB model has to be trained on a large scale of training data with a large vocabulary. Even so, there are still some of the words that may lack translations. We suppose these words can be translated into any of the sequential words in target sentences and add a special rule to our HPB model:

$$X \rightarrow \langle e_i, f_{(i',j')} \rangle, 1 \leq j' - i' \leq 2 \tag{4}$$

where i is the position of the word that do not have translation rules in source side, i' and j' are the start and end positions of the phrasal segment in target sentence. Here we restrict the length of the phrasal segment because larger segment tend to bring in noise in force decoding.

Moreover, in order to better evaluate the alignments between the parallel fragments, we extend the original HPB rules inspired by the work of (Čmejrek et al., 2009):

$$\langle X_1, X_1 f\rangle, \langle X_1, f X_1\rangle, \langle X_1 e, X_1\rangle, \langle e X_1, X_1\rangle \tag{5}$$

$$\langle X_1 X_2, X_2 X_1\rangle \tag{6}$$

in which rules (5) allow the HPB rules to insert and delete a single word, and rule (6) expands the standard glue rules and enables the aligning phrasal segments swap their constituents.

2.4 The Verification of Parallel Fragments

For each bilingual sentence pair, we can generate various alignment derivation trees. The derivation trees from source side are isomorphic to the target side because of the characteristic of SCFG.

In order to better evaluate the alignment for the derivation trees, each HPB rule in force decoding is associated with a score that is computed via the following log linear formula:

$$w(X \rightarrow \langle \alpha, \gamma, \sim\rangle) = \prod_i \phi_i(f, e)^{\lambda_i} \tag{7}$$

where $\phi_i(f, e)$ is a feature describing one particular aspect of the rule associated with the source and target phrases (f, e), and λ_i is the corresponding weight of the feature. Following the standard HPB model, features used in our force decoding are relative-frequency phrase translation probability $P(f|e)$ and its inverse $P(e|f)$, lexically weighted phrase translation probability $lex(f|e)$ and its inverse $lex(e|f)$.

Moreover, we consider the score of the special rule is:

$$w(X \rightarrow \langle e_i, f_{i',j'}\rangle) = \omega \times e^{-|j'-i'|} \tag{8}$$

in which, ω is the weight of the special rule.

After generating the derivation trees, we recursively traverse these trees at each node top-down, and extract parallel fragments from both sides with the following constraints:

1) The node in the derivation tree has a score greater than a threshold τ.

2) The node that represents the words from source side whose span is greater than 2.

The first constraint forces us to extract fragments with high alignment scores, as there are some alignment errors in HPB rules. And the second constraint makes us be more confident in the alignment scores over the larger fragments. The recursive traversal from derivation trees stops, once a fragment pair has been extracted.

For each sentence pair, different parallel fragments are extracted from derivation trees. Then we combine these fragments if there are overlaps in both source and target side. Otherwise, we keep these fragments as independent pairs.

3 Experiments

In our experiments, we compared our fragments extraction method with the PESA method explored by (Hewavitharana and Vogel, 2011), which is based on the lexical features.

3.1 Data and Evaluation Setup

We used the parallel corpora from **LDC**[1] to train our HPB model in force decoding. The HPB model was trained following (Chiang, 2007) with word alignment by running GIZA++ (Och and Ney, 2003). We downloaded comparable data from the online news sites: the BBC, and Xinhua News. The candidate sentence pairs (**Raw**) had been extracted following the approach of (Munteanu and Marcu, 2005) as we only focused on the performance of parallel fragments extraction. The sizes of these corpora are listed in Table 1.

Data Sets	#Sentences	#Chinese	#English
LDC	3.4M	64M	70M
Raw	2.6M	42M	49M

Table 1: Numbers of sentences and words for the parallel and comparable corpora.

We evaluated the quality of the extracted parallel fragments in two different ways:

Fragments Evaluation We obtained manual alignments for 600 sentence pairs and extracted parallel segments up to 10 words that are consistent with the annotated word alignment. We also removed the segments less than 3 words for the

[1] LDC2002E18, LDC2002T01, LDC2003E07, LDC2003E14, LDC2003T17, LDC2004T07, LDC2004T08, LDC2005T06, LDC2005T10, LDC2005T34, LDC2006T04, LDC2007T09.

constraint as described in Section 2. Then we tested the performance with the manual annotation.

Translation Evaluation We evaluated the fragments on Chinese-to-English translation tasks. We used a HPB translation system with a 4-gram language model trained on about 4 billion words of English using SRI Language Toolkit (Stolcke, 2002). We tuned parameters of the SMT system using minimum error-rate training (Och, 2003) to maximize the BLEU-4 (Papineni et al., 2002) on NIST 2005, and evaluated on the standard test sets, NIST 2006 and NIST 2008.

3.2 Experimental results

3.2.1 Performance on Fragments Extraction

We first compared the our method (HPB-FD) with PESA by fragments extraction. To give credit to our fragments extraction, we used partial matches to evaluate the performance of our extract method, following the way of (Hewavitharana and Vogel, 2011). The precision and recall were defined based on the tokens in the extracted target fragments that were also exists in the reference. And the F1 score was calculated in the standard way.

	Exact	**P**	**R**	**F1**
PESA	60.36	88.42	84.74	86.54
HPB-FD	74.12	94.36	88.90	91.55

Table 2: The results for fragments extraction with PESA and HPB-FD.

Table 2 gives the performance of PESA and our HPB-FD method. The results are presented as percentages of: exact matches found (**Exact**), precision (**P**), recall (**R**) and **F1**. It can be seen that our method can effectively extract parallel fragments from the comparable corpora. Comparing to PESA, our extraction method has higher scores in both Exact and F1 measure. This demonstrates that extracting fragments by our force decoding method can be more effectively to evaluate parallel fragments in comparable corpora.

3.2.2 Performance on Machine Translation

We then evaluated the extracted parallel fragments with the HPB translation system. In the baseline system, translation model (LDC) was trained on the LDC corpora that had been cleaned and thought to be less noisy. In the contrast experiments, we trained three translation models. The first model (LDC+Raw) was trained on the LDC with the extracted comparable sentences. The second model (LDC+PESA) was trained on the LDC and fragments that were extracted by PESA. And the third (LDC+HPB-FD) was trained on LDC and fragments that were extracted by HPB-FD. Table 3 lists the BLEU scores obtained by different training data.

	NIST 2006	**NIST 2008**
LDC	28.07	26.12
LDC+Raw	28.20(+0.13)	26.05(-0.07)
LDC+PESA	28.65(+0.58)	26.62(+0.50)
LDC+HPB-FD	**29.01(+0.94)**	**26.93(+0.81)**

Table 3: The translation performance with different training data. BLEU score gains are significant with $p < 0.01$.

Comparing to the baseline system, all the adding training data get stable improvements in translation performance except for the comparable sentences. It suggests that the simple increment in training data does not always lead to better performance. The superiority of parallel corpora confirms that, the quality is more important than quantity in collecting training data. Moreover, comparing to the parallel fragments extracted by PESA, our method get better translation results in both translation tasks, which also suggests our method can effectively extract parallel fragments from comparable corpora for the SMT system.

4 Conclusions

Parallel data in the real world is increasing continually. However, we cannot always get the translation performance improved by simply enlarging our training data. The collection of parallel data is expensive, and to our best knowledge, there is not a unified method to detect parallel fragments automatically.

We have presented an effective phrase-based method, which combines the lexical and phrasal features, for extracting parallel fragments from comparable corpora. The similarity between the source and target fragments is measured by the force decoding based on the existing HPB model. Experimental results show that our method can effectively detect the parallel fragments and achieve significant improvements over the baseline HPB translation system on the large scale Chinese-to-English translation tasks.

Acknowledgments

We thank the anonymous reviewers for their valuable comments and suggestions. This work was supported by the National High Technology Research and Development Program ("863" Program) of China under Grant No.2011AA01A207 and also supported by the National Program on Key Basic Research Project ("973" Program) under Grant No.2013CB329302.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the Use of Comparable Corpora to Improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 16–23.

David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 263–270.

David Chiang. 2007. Hierarchical Phrase-based Translation. *Computational Linguistics*, 33(2):201–228.

Martin Čmejrek, Bowen Zhou, Bing Xiang. 2009. Enriching SCFG Rules Directly from Efficient Bilingual Chart Parsing. In *Proceedings of the International Workshop on Spoken language Transaltion*, pages 136–143.

Sanjika Hewavitharana and Stephan Vogel. 2011. Extracting parallel phrases from comparable data. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 61–68.

Gumwon Hong, Chi-Ho Li, Ming Zhou and, Hae-Chang Rim. 2010. An Empirical Study on Web Mining of Parallel Data. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 474–482.

Tadashi Kumano, Hideki Tanaka and Takenobu Tokunaga. 2007. Extracting Phrasal Alignments from Comparable Corpora by Using Joint Probability SMT Model. In *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 95–103.

Adrien Lardilleux, François Yvon and Yves Lepage. 2012. Hierarchical Sub-sentential Alignment with Anymalign. In *Proceedings of the 16th annual meeting of the European Association for Machine Translation*, pages 279–286.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 81–88.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 160–167.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the ACL*, pages 311–318.

Chris Quirk, Raghavendra U. Udupa, and Arul Menezes. 2007. Generative Models of Noisy Translations with Applications to Parallel Fragment Extraction. In *Proceedings of the Machine Translation Summit XI*, pages 377–384.

Jason Riesa and Daniel Marcu. 2012. Automatic Parallel Fragment Extraction from Noisy Data. In *Proceedings of the 2012 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 538–542.

Jason R. Smith and Chris Quirk and Kristina Toutanova. 2010. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 403–411.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving Machine Transaltion Performance by Exploiting Non-parallel Corpora. *Computational Linguistics*, 31(4).

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proceedings of International Conference on Spoken Language Processing*, pages 901–904.

Bing Zhao and Setphan Vogel. 2002. Adaptive Parallel Sentences Mining from Web Bilingual News Collection. In *2002 IEEE Int. Conf. on Data Mining*, pages 745–748.