Context-extended Phrase Reordering Model for Pivot-based Statistical Machine Translation

Xiaoning Zhu, Tiejun Zhao, Yiming Cui and Conghui Zhu
School of Computer Science and Technology
Harbin Institute of Technology
Harbin, China
{xnzhu, tjzhao, ymcui, chzhu}@mtlab.hit.edu.cn

***Abstract*—For translation between language pairs which is lack of bilingual data, pivot-based SMT uses a pivot language as a "bridge" to generate source-target translation, inducing from source-pivot and pivot-target translation. However, due to the missing of the context information, the reordering model was hard to obtain with the conventional methods. In this paper, we present a context-extended phrase reordering model for pivot-based statistical machine translation by extending the context information in source, pivot and target language. Experimental results show that our method leads to significant improvements over the baseline system on European Parliament data.**

Keywords-pivot; reordering; context; machine translation

I. INTRODUCTION

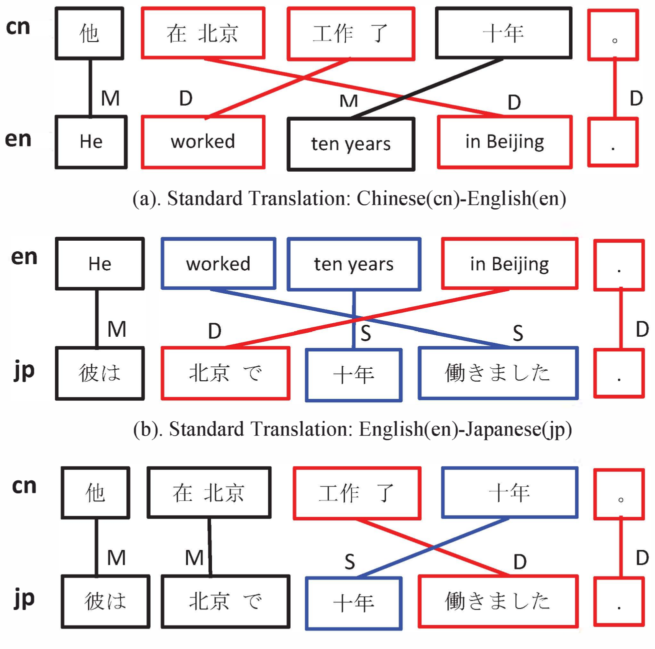
Statistical machine translation (SMT) uses bilingual corpora to build translation models. Thus the quantity and quality of the bilingual data heavily affect the performance of SMT systems. For frequently used language pairs, such as Chinese-English, it is easy to collect large amount of bilingual corpus. However, for most of language pairs, only a limited amount of bilingual data is available. Thus, it is difficult to build a high-performance SMT system with small scale bilingual data.

To overcome the data sparsely of SMT, a conventional solution is to introduce a pivot language as a "bridge" to connect the source and target language [1-4], where there exist large amounts parallel corpora in source-pivot and pivot-target languages. Among various pivot-based SMT approach, the triangulation method is a representative work. Given a source-pivot and a pivot-target phrase translation model, the triangulation method proposes to build a source-target phrase translation model by multiplying the posterior probabilities of source-pivot and pivot-target phrase pairs. However, due to the missing of the context information, the reordering model is hard to obtain with the triangulation method.

In phrase-based statistical machine translation, phrase reordering is a very important issue. Among the various reordering models, the lexicalized reordering model [5,6] is a commonly used method in current SMT systems. For each phrase pair, the lexicalized reordering model defines three types of orientations: directly follows a previous phrase (monotone), swapped with a previous phrase (swap), or not connected to the previous phrase (discontinuous). See Figure 1.(a) and Figure 1.(b) for illustrations.

In this paper, our work is based on the lexicalized reordering model. When applying the lexicalized reordering model into the triangulation method, a key problem is that the context information is missing in the phrase table. See Figure 1. for an illustration. In Figure 1 we can see that the reordering model of Chinese-Japanese translation (Figure 1.(c)) can be induced from Chinese-English (Figure 1.(a)) and English-Japanese translation (Figure 1.(b)). However, if we do not obtain adequate context information, we cannot induce the lexical orientation of the phrase. (e.g. if we do not know the previous phrase and following phrase of the Japanese phrase "十年", we cannot obtain the lexical orientation of the phrase.). Thus, we proposed a novel approach to extend the context information in source, pivot and target language to learn the lexicalized reordering model. Note that, this work can also be easily applied to other reordering models, such as the maximum entropy based phrase reordering model [7].

The remainder of this paper is organized as follows. In Section 2, we describe the related work. We introduce the context-extended phrase reordering model in Section 3. In Section 4, we describe the experiments and analyze the performance of our model, respectively. And Section 5 gives a brief conclusion of the paper.



(a). Standard Translation: Chinese(cn)-English(en)

(b). Standard Translation: English(en)-Japanese(jp)

(c). Pivot Translation: Chinese(cn)-Japanese(jp)

Figure 1. An example of lexicalized reordering model in standard SMT and pivot-based SMT.

978-1-4673-9596-0 /15/$31.00 ©2015 IEEE

II. Related Work

To the best of our knowledge, only two papers were done for phrase reordering model in pivot-based SMT.

Wu and Wang [1] use a small standard reorder model which was trained with standard bilingual source-target data for pivot-based SMT. The problem is that the source-target data are not always available.

Henríquez [7] proposed a phrase table combination method to learn phrase reordering model. Conforms to the lexicalized reordering model, the approach also defines three types of orientations: monotone, swap, and discontinuous. Given the orientations of source-pivot and pivot-target phrase pairs:

A swap move on the source-pivot system is dissolved if the same phrase is swapped again on the pivot-target system. Therefore it is a monotonous move.

A monotonous move followed by a swap means a swap from source-target. It is the same if the swap is performed first and then the monotonous move.

A discontinuous move always generates a final discontinuous move, regardless of which move is performed before it.

Figure 2.(a) to Figure 2.(d) shows an example of the rules explained above.

The problem of this method is that, on the one hand, the reordering probabilities are not accurate enough due to the non-uniformity of the probability space [8]; on the other hand, due to the missing of the context information, the orientation may not correct when combining the reordering tables. Figure 2.(e) and Figure 2.(f), a discontinuous move followed by a discontinuous move may not always be a discontinuous move.

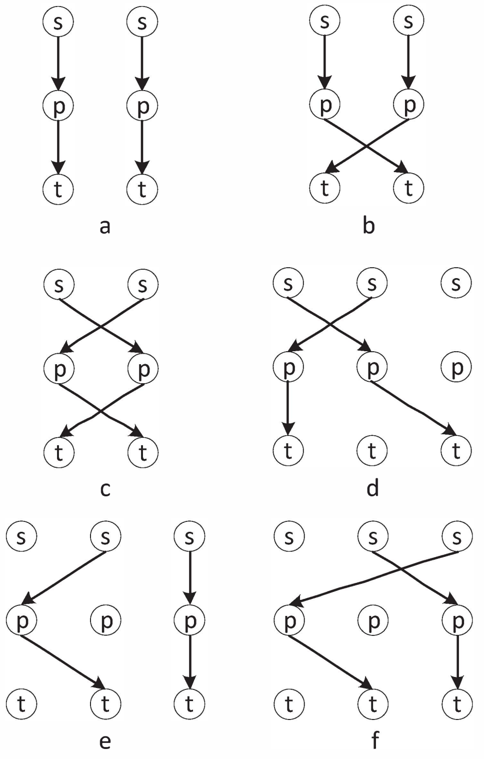


Figure 2: a) Two monotonous moves. b) A monotonous with the previous phrase followed by a swap with the next phrase. c) Two consecutive swap moves. d) A swap move followed by a discontinuous move. e) Two discontinuous move lead to a monotonous move. f) Two discontinuous move lead to a swap move.

III. Context-extended Approach

The key issue of our approach is to search the context information and define whether a source-target phrase pair is matched in context. In this paper, we define that a source-target phrase pair is context-match if its context is matched via the pivot language. See Figure 1 for an illustration. If we want to know the orientation of the Japanese phrase “十年”, we need to obtain its previous phrase “北京 で” and following phrase “働きました”. Besides, we also need to obtain the corresponding Chinese phrases of these three Japanese phrases via English phrases. If all these phrases are matched, we named the phrase pair is context-match. Because the source-pivot and pivot-target bilingual data are independent, these corresponding phrases are not always available. To overcome this problem, we propose a mixed string including the fragments of the phrase and part of speech (POS), which can search the context maximum. The rationality of the mixed string relies on the phrase with the similar structure (depend on POS), which often share the same orientation.

We use an example for illustration. Considering the following phrases (subscripts indicate word alignments):

P1:
进行$_1$ 综合$_2$ 改革$_3$
carry$_1$ out$_1$ comprehensive$_2$ reforms$_3$

P2:
进行$_1$ 深度$_2$ 改革$_3$
carry$_1$ out$_1$ deep$_2$ reforms$_3$

P3:
进行$_1$ X_2 改革$_3$
carry$_1$ out$_1$ X_2 reforms$_3$

P4:
进行$_1$ 综合$_2$ 治理$_3$
carry$_1$ out$_1$ comprehensive$_2$ management$_3$

P5:
进行$_1$ 综合$_2$ X_3
carry$_1$ out$_1$ comprehensive$_2$ X_3

From the example we can see that: phrase *P1* to *P5* share the same orientation. The reason is that these phrases either share the same structure or share the same fragments. For e.g., phrase *P1* and phrase *P2* share the same structure; phrase P1 and phrase *P5* share the same fragments.

Given two parallel corpus (source-pivot and pivot-target), the context-match of each phrase can be extracted efficiently according to Algorithm 1.

where:

1. p(*t*) / p(*s*), the corresponding pivot phrase of target phrase *t* and source phrase *s*

2. pre(*t*) / pre(*s*), the previous phrase of *t* or *s*

3. fol(*t*) / fol(*s*), the following phrase of *t* or *s*

4. t-pos / s-pos, the part of speech of *t* or *s*

5. t-frag / s-frag, the fragment of *t* or *s*

6. t-posfrag / s-posfrag, the part of speech and fragment of *t* or *s*

7.CM / CM fol / CM pre , the context-match value of the phrase/ the previous phrase and the following phrase

```
Algorithm 1 Context Match
Input: Source-Pivot and Pivot-Target Corpus
Output: Context Match of Each Phrase
for all phrase t in target language do
    if p($t$) = p($s$) then
        if p(pre($t$)) = p(pre($s$)) then
            CMpre=1
        else if p(pre($t$-$pos$)) = p(pre($s$-$pos$)) then
            CM pre=1/2
        else if p(pre($t$-$frag$)) = p(pre($s$-$frag$)) then
            CM pre=matched word/length of phrase
        else if p(pre($t$-$posfrag$))=p(pre($s$-$posfrag$)) then
            CM pre=matched word/2length of phrase
        end if
        if p(fol($t$)) = p(fol($s$)) then
            CMfol=1
        else if p(fol($t$-$pos$)) = p(fol($s$-$pos$)) then
            CM fol=1/2
        else if p(fol($t$-$frag$)) = p(fol($s$-$frag$)) then
            CM fol=matched word/length of phrase
        else if p(fol($t$-$posfrag$))=p(fol($s$-$posfrag$)) then
            CM fol=matched word/2length of phrase
        end if
            CM= CM $pre$+ CM $fol$
    end if
end for
```

With the context-match of each phrase, we can get the orientations following method proposed by Tillman [5]. Given three types of orientations: monotone(m), swap(s) and discontinuous(d), we can calculate the phrase reordering model following:

$$p_o(\text{orientation}|\bar{s},\bar{t}) = \frac{\text{count}(\text{orientation},\bar{s},\bar{t})}{\sum_o \text{count}(o,\bar{s},\bar{t})} \qquad (1)$$

$$\text{orientation} \in \{\text{m},\text{s},\text{d}\} \qquad (2)$$

where $\bar{s}$ represents the phrase in source language and $\bar{t}$ represents the phrase in target language.

For the phrases which our context-extended method cannot cover, we use the reordering table combination method to generate its reordering [7].

IV. Experiments

In our experiments, the word alignment was obtained by GIZA++ and the heuristics grow-diag-final is adopted as refinement rule. Our translation system is Moses, and the parameters were tuned by MERT.

In order to make a comprehensive comparison, we built three baseline systems to compare with our system. The basic baseline system is the triangulation method based pivot approach without reordering model. The triangulation method with a reordering model based on movement distance and the reordering table combination method are also built as our baseline system.

To evaluate the translation quality, we used case-insensitive BLEU-4 as our evaluation metric. The statistical signifi-cance using 95% confidence intervals were measured with paired bootstrap resampling.

We mainly test our approach on Europarl corpus, which is a multi-lingual corpus including 21 European languages. Due to the size of the data, we only select 11 languages which were added to Europarl from 04/1996 or 01/1997, including Danish (da), German (de), Greek (el), English (en), Spanish (es), Finnish (fi), French (fr), Italian (it) Dutch (nl) Portuguese (pt) and Swedish (sv). In order to avoid a trilingual scenario, we split the training corpus into 2 parts by the year of the data: the data released in odd years were used for training source-pivot translation model and the data released in even years were used for training pivot-target translation model.

We perform our experiments on different translation directions. As a most widely used language in the world, English was used as the pivot language for granted when carrying out experiments on different translation directions.

Several test sets have been released for the Europarl corpus. In our experiments, we used WMT07 as our development data and WMT08 as our test data. We compare our context-extended reordering model with three baseline systems in various translation directions. The results were shown in Table 4. The parameters were tuned by MERT on WMT07 and were tested on WMT08. Limited by the length of the paper, we only show the test set results on WMT08. The trend of the development set results on WMT07 is similar to WMT08.

According to the results, several observations can be made from the table.

1. Generally, the three reordering models are better than the triangulation method without reordering model, and the context-extended reordering model is better than the other reordering models and the triangulation method without reordering model. The performance of these methods in descending order: the context-extended method, the reordering table composition method, the distance-based method and the triangulation method without reordering.

2. In all 90 language directions, our context-extended method achieves notable improvements over the baseline system without reordering and the distance-based method. Furthermore, our context-extended method is better than the reordering table composition method in 83 language directions, and the context-extended method is significantly better than the reordering table composition method in 31 language directions.

3. The improvements of our context-extended method are not consistent in different translation directions. The improvement ranges from 0.06 (sv-el) to 1.6 (sv-da) compared with the no-ordering method, and ranges from -0.01 (fr-de) to 0.74 (es-pt) compared with the reordering table composition method. One possible reason is that the performance highly is related with choice of pivot language, and characteristics of source, pivot and target languages.

4. In some language pairs, the reordering table composition method is comparable with our context-extended method. However, in most situations, our approach is better than the reordering table composition method in general.

TABLE I. EXPERIMENTAL RESULTS ON EUROPARL WITH DIFFERENT TRANSLATION DIRECTIONS

	TGT SRC	da	de	el	es	fi	fr	it	nl	pt	sv
No reorder	**da**	-	19.23	20.15	27.19	14.55	24.10	20.35	22.11	24.15	28.18
Distance			19.54	20.31	27.30	14.67	24.11	20.50	22.19	24.29	28.65
Henríquez			19.83	20.38	27.41	14.88	24.19	20.70	22.24	24.37	28.98
CE-based			**20.39***	20.39	27.41	14.99	24.21	20.71	22.30	24.41	**29.54***
No reorder	**de**	23.12	-	19.70	26.11	12.71	22.27	18.70	23.63	22.95	21.09
Distance		23.56		19.88	26.45	12.87	22.39	18.90	23.87	23.09	21.35
Henríquez		23.87		20.03	26.76	12.99	22.65	19.01	24.01	23.32	21.78
CE-based		24.04		**20.37***	26.81	13.02	22.81	19.21	24.13	23.33	21.84
No reorder	**el**	23.21	18.01	-	32.07	13.17	27.33	23.04	20.60	27.52	22.61
Distance		23.33	18.20		32.32	13.25	27.49	23.17	20.69	27.69	22.89
Henríquez		23.47	18.34		32.65	13.33	27.66	23.29	20.81	28.01	23.04
CE-based		23.59	**18.77***		**33.05***	13.37	**27.97***	**23.74***	20.81	28.29	**23.41***
No reorder	**es**	25.24	19.61	27.17	-	13.81	32.80	27.56	22.32	34.57	24.73
Distance		25.48	19.89	27.22		13.98	33.14	27.78	22.35	34.87	24.98
Henríquez		25.76	20.01	27.31		14.03	33.31	27.99	22.40	35.25	25.31
CE-based		**26.25***	20.23	27.47		14.26	**33.88***	28.26	22.42	**35.99***	25.43
No reorder	**fi**	18.18	13.19	14.71	20.05	-	17.51	14.55	15.49	17.20	16.53
Distance		18.29	13.27	14.96	20.11		17.72	14.76	15.58	17.38	16.75
Henríquez		18.43	13.33	15.11	20.37		17.94	14.93	15.71	17.51	16.89
CE-based		18.55	13.39	15.25	**20.83***		**18.35***	**15.32***	15.89	**17.91***	**17.23***
No reorder	**fr**	25.54	19.99	26.44	37.37	13.76	-	28.41	22.61	33.73	24.53
Distance		25.67	20.11	26.61	37.74	13.89		28.67	22.69	33.91	24.66
Henríquez		25.69	20.20	26.87	37.89	13.99		28.89	22.76	34.10	24.71
CE-based		**26.04***	20.19	26.95	38.09	14.07		29.15	22.85	**34.73***	24.73
No reorder	**it**	22.33	17.71	24.13	34.33	13.20	30.13	-	21.34	30.71	22.01
Distance		22.61	17.89	24.19	34.59	13.31	30.39		21.45	30.91	22.04
Henríquez		22.69	18.01	24.23	34.81	13.49	30.61		21.59	31.22	22.12
CE-based		**23.00***	**18.31***	24.41	**35.49***	13.56	**31.09***		21.64	**31.62***	22.19
No reorder	**nl**	22.39	19.81	18.51	24.57	11.94	21.27	18.27	-	21.62	19.75
Distance		22.54	20.01	18.76	24.87	12.02	21.29	18.33		21.69	19.98
Henríquez		22.79	20.31	19.00	25.01	12.17	21.39	18.34		21.69	20.11
CE-based		**23.12***	**20.75***	19.11	**25.72***	12.29	21.39	18.36		21.69	20.27
No reorder	**pt**	24.05	19.10	25.19	36.44	13.31	32.31	28.01	21.49	-	22.80
Distance		24.12	19.31	25.33	36.65	13.51	32.51	28.22	21.63		23.01
Henríquez		24.37	19.61	25.45	36.98	13.89	32.98	28.49	21.89		23.16
CE-based		24.59	19.70	25.59	**37.63***	14.07	**33.32***	**28.97***	22.10		23.21
No reorder	**sv**	31.13	20.18	22.05	29.13	15.37	25.53	21.01	22.28	25.54	-
Distance		31.65	20.49	22.09	29.24	15.47	25.65	21.06	22.45	25.78	
Henríquez		32.03	20.74	22.11	29.36	15.51	25.71	21.18	22.76	25.89	
CE-based		**32.73***	**21.20***	22.11	29.43	15.55	25.72	21.18	22.89	26.02	

V. CONCLUSION

This paper proposes a novel approach to learn phrase reordering model for pivot-based SMT by extending the context of phrase in source, pivot and target language. To extend the context information, we introduced a mixed string to cover more phrases, including the fragments of the phrase and POS. The experimental results on Europarl show significant improvements over the baseline systems.

ACKNOWLEDGMENT

The work of this paper is funded by the project of National Natural Science Foundation of China (No.61173073 & No.61272384) and the project of National High Technology Research and Development Program of China (863 Program) (No.2015AA015405).

REFERENCES

[1] H. Wu and H. Wang, “Pivot Language Approach for Phrase-Based Statistical Machine Translation,” In Proceedings of 45th Annual Meeting of the Association for Computational Linguistics, 2007, pp. 856-863.

[2] M. Utiyama and H. Isahara, “A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation,” In Proceedings of Human Language Technology: the Conference of the North American Chapter of the Association for Computational Linguistics, 2007, pp 484-491.

[3] N. Bertoldi, M. Barbaiani, M. Federico, and R. Cattoni, “Phrase-Based statistical machine translation with Pivot Languages,” In Proceedings of the 5th International Workshop on Spoken Language Translation (IWSLT),2008, pp 143-149.

[4] T. Cohn and M. Lapata, “Machine Translation by Triangulation: Make Effective Use of Multi-Parallel Corpora,” In Proceedings of 45th Annual Meeting of the Association for Computational Linguistics, 2007, pp 828-735.

[5] C. Tillmann, ”A Unigram Orientation Model for Statistical Machine Translation,” In Proceedings of HLT-NAACL 2004

[6] P. Koehn, A. Axelrod, A. Mayne, C. Callison-Burch, M. Osborne and D. Talbot, “Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation,” In International Workshop on Spoken Language Translation.

[7] Carlos A. Henríquez Q., Rafael E. Banchs, José B. Mariño, “Learning Reordering Models for Statistical Machine Translation with a Pivot Language. Internal Report,” http://nlp.lsi.upc.edu/publications/papers/henriquez2010b.pdf

[8] X Zhu, Z. He, H. Wu, C. Zhu, H. Wang, T. Zhao, “Improving Pivot-Based Statistical Machine Translation by Pivoting the Co-occurrence Count of Phrase Pairs,” In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1665–1675.