Rule-Based Preordering on Multiple Syntactic Levels in Statistical Machine Translation

Ge Wu, Yuqi Zhang

Institute for Anthropomatics
Karlsruhe Institute of Technoloy, Germany
utcur@student.kit.edu, yuqi.zhang@kit.edu

Abstract

We propose a novel data-driven rule-based preordering approach, which uses the information of syntax tree and word alignment to reorder the words in source sentences before decoding in a phrase-based SMT system between English and Chinese. The preorering algorithm extracts reordering patterns from multiple levels of the syntax trees of the training data and applies the rules on the source sentences in a similar manner. We've conducted experiments in English-to-Chinese and Chinese-to-English translation directions. Our results show that the approach has led to improved translation quality both when it was applied separately on the baseline or when it was combined with some other reordering approaches. We report an improvement of 0.43 in BLEU score when our preordering approach was used in addition to the short rule [1], long rule[2] and tree rule [3] based preordering approaches in the English-to-Chinese translation direction. There's also an improvement of 0.3 in BLEU score when our preordering approach was used in addition to the aforementioned preordering approaches in the Chinese-to-English translation direction. Through the translation examples, we've also found improvements in syntactic structure with our preordering approach.

1. Introduction

Word order is a general issue when we want to translate text from one language to the other. Different languages normally have different word orders and the difference could be very huge. Among all the languages, Chinese is one language which is very different from English, because they belong to different language families and have long period of separately development. Both languages have a Subject-Verb-Object order, but they also have a lot of differences in word order. Especially sentences in both languages can sometimes have completely different syntactic structures. The differences may involve long-distance or unstructured position changes.

Most state-of-the-art phrase-based SMT systems use language model, phrase table or decoder to adjust the word order. Or an additional reordering model is used in the log-linear model for word reordering. However, these methods may have some disadvantages, such as some don't handle long-distance reordering, some don't handle unstructured reordering and some are rather time consuming.

Encouraged by the results from the paper [1], [2] and [3], we further propose a new data-driven, rule-based preordering method, which extracts and applies reordering rules based on syntax tree. The method is called Multi-Level-Tree (MLT) reordering, which orders the constituents on multiple levels of the syntax tree all together. This preordering method rearranges the words in source language into a similar order as they are supposed to be in the target language before translation. With the appropriate word order, better translation quality can be achieved. Especially, our preordering method is very suitable for translation between language pairs like English and Chinese, which have very different word orders. Besides, the method can also be combined with the above mentioned rule-based reordering methods to achieve better translation quality.

2. Related Work

Word reordering is an important problem for statistical machine translation, which has long been addressed.

In a phrase-based SMT system, there are several possibilities to change the word orders. Words can be reordered during the decoding phase by setting a window, which allows the decoder to choose the next word for translation. Reordering could also be influenced by the language model, because the language model give probability of how a certain word is likely to follow. Different language model may give different probability, which further influences the decision made by log-linear model. Other ways to change the word orders include using distance based reordering models or lexicalized reordering models [4, 5]. The lexicalized reordering model reorders the phrases by using information of how the neighboring phrases change orientations.

Another way to achieve word reordering is to detach it from decoding phase and do it separately in a pre-process before decoding, in order to reduce the time for translation. This kind of preordering approaches use linguistic information to modify the word orders. Preordering can also be rule-based, which extracts different types of reordering rules by observing reordering patterns from the training data and apply the rules to the sentences to be translated. Depends on how the rules are defined, different information may be used such as word alignments, POS tags, syntax trees, etc.

Some early approaches use manually defined reordering rules based on the linguistic information for particular languages [6, 7, 8, 9]. Later come the data-driven methods [10, 11], which learn the reordering rules automatically.

Rottman and Vogel 2007 [1] introduced the idea of extracting reordering rules from the POS tag sequences of training data and use them for reordering. Niehues and Kolss 2009 [2] went further, and developed a method for long-distance word reordering, which works good on German-English translation task due to the long-distance shift of verbs. The method extracts discontinuous reordering rules in addition to the continuous ones, which contains a placeholder to match several words and enables the word to shift cross long distance.

Afterwards, Herrmann et al. 2013 [3] introduced a new approach to reorder the words based on syntax tree, which leads to further improvements on translation quality. The algorithm takes syntactic structure of the sentences into account and extract the rules from the syntax tree by detecting the reordering of child sequences. It also has the variant based only on part of the child sequences which is suitable for language with flat syntactic structure such as German.

However, these approaches which are bases on POS tag sequences or syntax trees are mostly designed for languages like German and are not especially adapted for languages like Chinese. As Chinese has very different word orders, a reordering approach, which can further explore the syntactic structure of Chinese and utilize this information for reordering, is desirable.

The hierarchical phrase-based translation model [12] is especially suitable for Chinese translation, and provide very good translation results. It extracts hierarchical rules by using information of the syntactic structure. Phrases from different hierarchies, or so-called phrases of phrases, are reordered during the decoding.

The idea of phrases on different hierarchies has inspired us to create this preordering method based on multiple levels of the syntax tree. Besides, we also hope to detach the reordering from decoding phase and do it separately in a preprocess before decoding, in order to reduce the time for translation. This kind of preordering approaches use linguistic information to modify the word orders.

Oracle reordering has also shown values for evaluating the potential of preordering. [13] introduced the permutation distance metrics which can be used to measure reordering quality. And [14] described how we can construct permutations from the word alignment as oracle reordering.

3. Motivation

The word order between English and Chinese differs very significantly. For one, the words in Chinese have generally different origins as those in English, which leads to very different vocabulary and word construction. Sometimes it is very hard to find corresponding words in the other language. For example, some prepositions in Chinese have very different usage than those prepositions in English. Also the continuous writing of Chinese without spaces makes this problem more severe, since word boundaries are not always so clear in Chinese. The text needs to be segmented first before translation. A word segmentation process is used to separate the words, but the results may not always be ideal.

For the other, both languages have sometimes very different sentence structures. Thus, a word-for-word translation between English and Chinese is often unnatural or difficult to understand. Each of them has some sentence patterns that don't exist or rarely used in the other. In Chinese, a modifier is often put before the part that it modifies. While in English, it is very common that the modifier is put after the part that it modifies. Besides, English sentences with a lot of long clauses may be more suitable to translate into several Chinese sentences, because in Chinese people don't tend to use long clauses in general.

Some typical problems of word orders between English and Chinese that we've found in the data are as follows:

- *Pre-modifier instead of post-modifier.* In Chinese people tend to use pre-modifier rather than post-modifier. This involves the position change of adverbials, relative clauses and preposition phrases during translation.
- *Constuction of questions.*
- *Special sentence constructions.* For example, *Bâ*-construction in Chinese and sentence constructions such as *there be* and inverted negative sentences in English don't have correspondence in the other language in general.
- *Long distance word position change.*

Not all these reordering problem can be solved well by using reordering rules based on POS tag sequence or single level of syntax tree. In order to improve the reordering between English and Chinese, we need reordering method that can handle more complicated, unstructural word order change. Inspired by the ideas of reordering on syntax tree and hierarchical phrases, we created the MLT reordering, which reorders words based on multiple syntactic levels and can handle long distance word position change and complicated word position change very well.

4. Multi-Level-Tree Reordering

4.1. Reordering on Multiple Syntactic Levels

Reordering patterns are based on multiple levels of the syntax tree. Figure 1

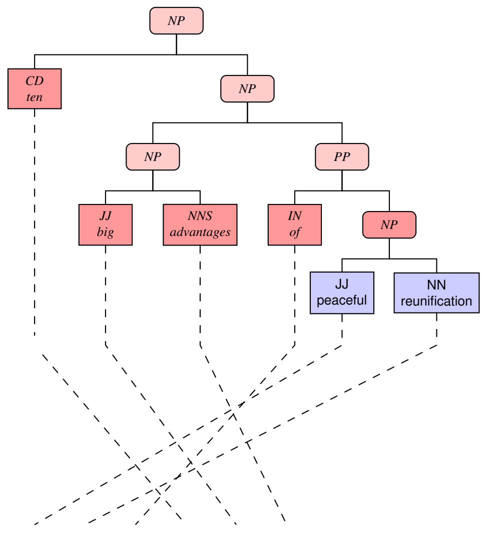

Figure 1: Reordering pattern from multiple syntactic levels

4.2. Rule Extraction

4.3. Rule Application

5. Results

6. Conclusion

6.1. First page

The first page should have the paper title, author(s), and affiliation(s) centered on the page across both columns. The remainder of the text must be in the two-column format, staying within the indicated image area.

6.1.1. Paper Title

The paper title must be in boldface. All non-function words must be capitalized, and all other words in the title must be lower case. The paper title is centered across the top of the two columns on the first page as indicated above.

6.1.2. Authors' Name(s)

The authors' name(s) and affiliation(s) appear centered below the paper title. If space permits, include a mailing address here. The templates indicate the area where the title and author information should go. These items need not be confined to the number of lines indicated; papers with multiple authors and affiliations may require two or more lines. Note that the submission version of technical papers *should be anonymized for review*.

6.1.3. Abstract

Each paper must contain an abstract that appears at the beginning of the paper.

6.2. Basic layout features

- Proceedings will be printed in A4 format. The layout is designed so that files, when printed in US Letter format, include all material but margins are not symmetric. Although this is not an absolute requirement, if at all possible, **PLEASE TRY TO MAKE YOUR SUBMISSION IN A4 FORMAT.**
- Two columns are used except for the title part and possibly for large figures that need a full page width.
- Left margin is 20 mm.
- Column width is 80 mm.
- Spacing between columns is 10 mm.
- Top margin 25 mm (except first page 30 mm to title top).
- Text height (without headers and footers) is maximum 235 mm.
- Headers and footers must be left empty (they will be added for printing).
- Check indentations and spacings by comparing to this example file (in pdf format).

6.2.1. Headings

Section headings are centered in boldface with the first word capitalized and the rest of the heading in lower case. Subheadings appear like major headings, except they start at the left margin in the column. Sub-sub-headings appear like subheadings, except they are in italics and not boldface. See the examples given in this file. No more than 3 levels of headings should be used.

6.3. Text font

Times or Times Roman font is used for the main text. Recommended font size is 9 points which is also the minimum allowed size. Other font types may be used if needed for special purposes. While making the final PostScript file, remember to include all fonts!

LaTeX users: DO NOT USE Computer Modern FONT FOR TEXT (Times is specified in the style file). If possible, make the final document using POSTSCRIPT FONTS. This is necessary given that, for example, equations with non-ps Computer Modern are very hard to read on screen.

6.4. Figures

All figures must be centered on the column (or page, if the figure spans both columns). Figure captions should follow each figure and have the format given in Fig. 2.

Figures should preferably be line drawings. If they contain gray levels or colors, they should be checked to print well on a high-quality non-color laser printer.

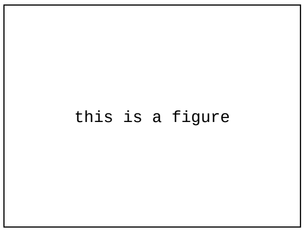

Figure 2: *Schematic diagram of speech production.*

6.5. Tables

An example of a table is shown as Table 1. Somewhat different styles are allowed according to the type and purpose of the table. The caption text may be above or below the table.

Table 1: *This is an example of a table.*

ratio	decibels
1/1	0
2/1	≈ 6
3.16	10
10/1	20
1/10	-20

6.6. Equations

Equations should be placed on separate lines and numbered. Examples of equations are given below. Particularly,

$$x(t) = s(f_\omega(t)) \tag{1}$$

where $f_\omega(t)$ is a special warping function

$$f_\omega(t) = \frac{1}{2\pi j} \oint_C \frac{\nu^{-1k} d\nu}{(1 - \beta\nu^{-1})(\nu^{-1} - \beta)} \tag{2}$$

A residue theorem states that

$$\oint_C F(z)dz = 2\pi j \sum_k Res[F(z), p_k] \tag{3}$$

Applying (3) to (1), it is straightforward to see that

$$1 + 1 = \pi \tag{4}$$

Make sure to use `\eqref` when refering to equation numbers. Finally we have proven the secret theorem of all speech sciences (see equation (3) above). No more math is needed to show how useful the result is!

6.7. Hyperlinks

Hyperlinks can be included in your paper. Moreover, be aware that the paper submission procedure includes the option of specifying a hyperlink for additional information. This hyperlink will be included in the CD-ROM. Particularly pay attention to the possibility, from this single hyperlink, to have further links to information such as other related documents, sound or multimedia.

If you choose to use active hyperlinks in your paper, please make sure that they present no problems in printing to paper.

6.8. Page numbering

Final page numbers will be added later to the document electronically. *Please don't make any headers or footers!*.

6.9. References

The reference format is the standard for IEEE publications. References should be numbered in order of appearance, for example

7. Experiments

Please make sure to give all the necessary details regarding your experimental setting so as to ensure that your results could be reproduced by other teams.

8. Conclusions

This paper has described a novel approach for doing wonderful stuff such as ...

9. Acknowledgements

The IWSLT 2014 organizing committee would like to thank the organizing committees of INTERSPEECH 2004 for their help and for kindly providing the template files.

10. References

[1] Rottmann, K. and Vogel, S., "Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model," 2007.

[2] Niehues, J. and Kolss, M., "A POS-Based Model for Long-Range Reorderings in SMT," in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, (Athens, Greece), pp. 206–214, Association for Computational Linguistics, 2009.

[3] Herrmann, T., Weiner, J., Niehues, J., and Waibel, A., "Analyzing the Potential of Source Sentence Reordering in Statistical Machine Translation," 2013.

[4] Tillmann, C., "A Unigram Orientation Model for Statistical Machine Translation," in *Proceedings of HLT-NAACL 2004: Short Papers*, pp. 101–104, Association for Computational Linguistics, 2004.

[5] Koehn, P., Axelrod, A., Birch, A., Callison-Burch, C., Osborne, M., Talbot, D., and White, M., "Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation," in *IWSLT*, pp. 68–75, 2005.

[6] Collins, M., Koehn, P., and Kučerová, I., "Clause Restructuring for Statistical Machine Translation," in *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 531–540, Association for Computational Linguistics, 2005.

[7] Popovic, M. and Ney, H., "POS-Based Word Reorderings for Statistical Machine Translation," in *International Conference on Language Resources and Evaluation*, pp. 1278–1283, 2006.

[8] Habash, N., "Syntactic Preprocessing for Statistical Machine Translation," *MT Summit XI*, pp. 215–222, 2007.

[9] Wang, C., Collins, M., and Koehn, P., "Chinese Syntactic Reordering for Statistical Machine Translation," in *EMNLP-CoNLL*, pp. 737–745, Citeseer, 2007.

[10] Zhang, Y., Zens, R., and Ney, H., "Chunk-Level Reordering of Source Language Sentences with Automatically Learned Rules for Statistical Machine Translation," in *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, pp. 1–8, Association for Computational Linguistics, 2007.

[11] Crego, J. M. and Habash, N., "Using Shallow Syntax Information to Improve Word Alignment and Reordering for SMT," in *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 53–61, Association for Computational Linguistics, 2008.

[12] Chiang, D., "Hierarchical Phrase-Based Translation," *computational linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[13] Birch, A., Osborne, M., and Blunsom, P., "Metrics for MT Evaluation: Evaluating Reordering," *Machine Translation*, vol. 24, Mar. 2010.

[14] Birch, A., "Reordering Metrics for Statistical Machine Translation," 2011.