Karlsruhe Institute of Technology

Analyzing the Potential of Source Sentence Reordering in Statistical Machine Translation for Chinese

Master Thesis of

Ge Wu

At the Department of Informatics
Institute for Anthropomatics and Robotics (IAR)

Advisor: Alex Waibel
Second Advisor: Yuqi Zhang

Duration: 1st Febrary 2014 – 22nd June 2014

Abstract

todo

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

22nd June 2014

Contents

1. Introduction

1.1. Motivation

Word reordering is a general issue when we want to translate text from one language to the other. Different languages normally have different word reordering and the difference could be huge, when two languages are isolated from each other. Depend on the language itself, the word reordering could have very distinguish features. For example, 45% of the languages in the world has a subject-object-verb(SOV) order. Unlike in English, verbs are put after object in these languages. Japanese is a popular language among them. Instead of saying "The black cat climbed to the tree top.", people would say "The black cat the tree top to climbed." in Japanese. Another example is Spanish, in which people often put the adjective after the modified nouns. An example from the paper [LP13] shows how people would order the words differently:

English	The black cat climbed to the tree top.
Japanese	The black cat the tree top to climbed.
Spanish	The cat black climbed to the top tree.

Table 1.1.: Word orders of three different languages

Since different word orders are a common issue among languages, we propose several pre-reordering methods and evaluate them in this thesis. Before translation, the words in source language are rearranged into a similar word order as the target language's through these methods. With the appropriate word order, better translation quality will be achieved.

1.2. Objective and Contribution

The ground of this thesis are three papers about data driven, rule based pre-reordering: [RV], [NK] and [HWNW]. In this thesis, we tried to

asset is data driven

original (mltilayer)

try to extend to other language

hiarchical [Chi07]

conclusion goal is

1.3. Structure

In this chapter we mainly describe the background and objective of this thesis, including the related research in the next section of this chapter. In the chapter 2 we shows the fundamental knowledge, which is related and relevant to our research. In chapter 3 we introduce our reordering methods in detail. The experiment setup and results are present in chapter 4, together with the evaluation of the methods we use. In the last chapter we conclude this work with an overall discussion of our methods. We also point out some possible directions for future research.

1.4. Related Work

todo

2. Foundations

2.1. Pre-Reodering system

2.2. Alignment

2.3. Part-of-Speech (POS) Tag

2.4. Parse Tree

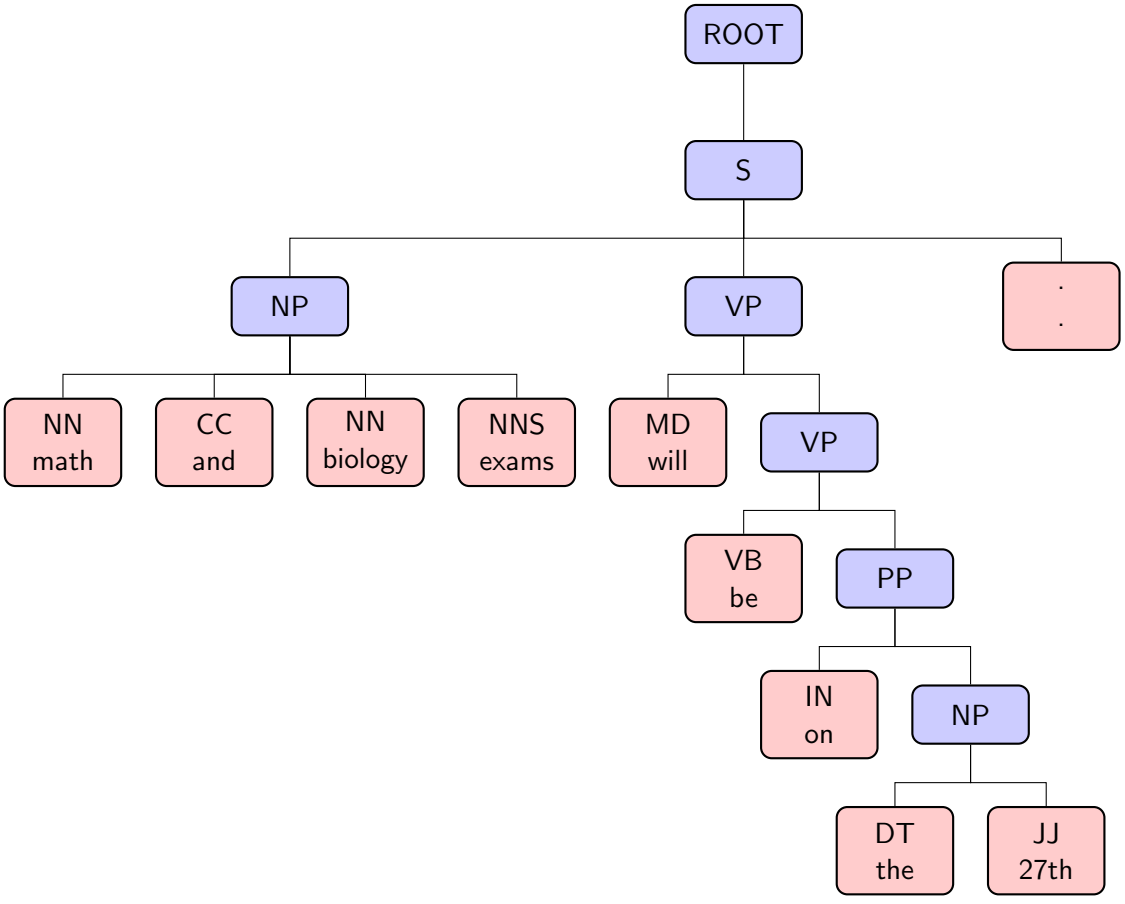

Figure 2.1.: Example of a parse tree

2.5. Lattices

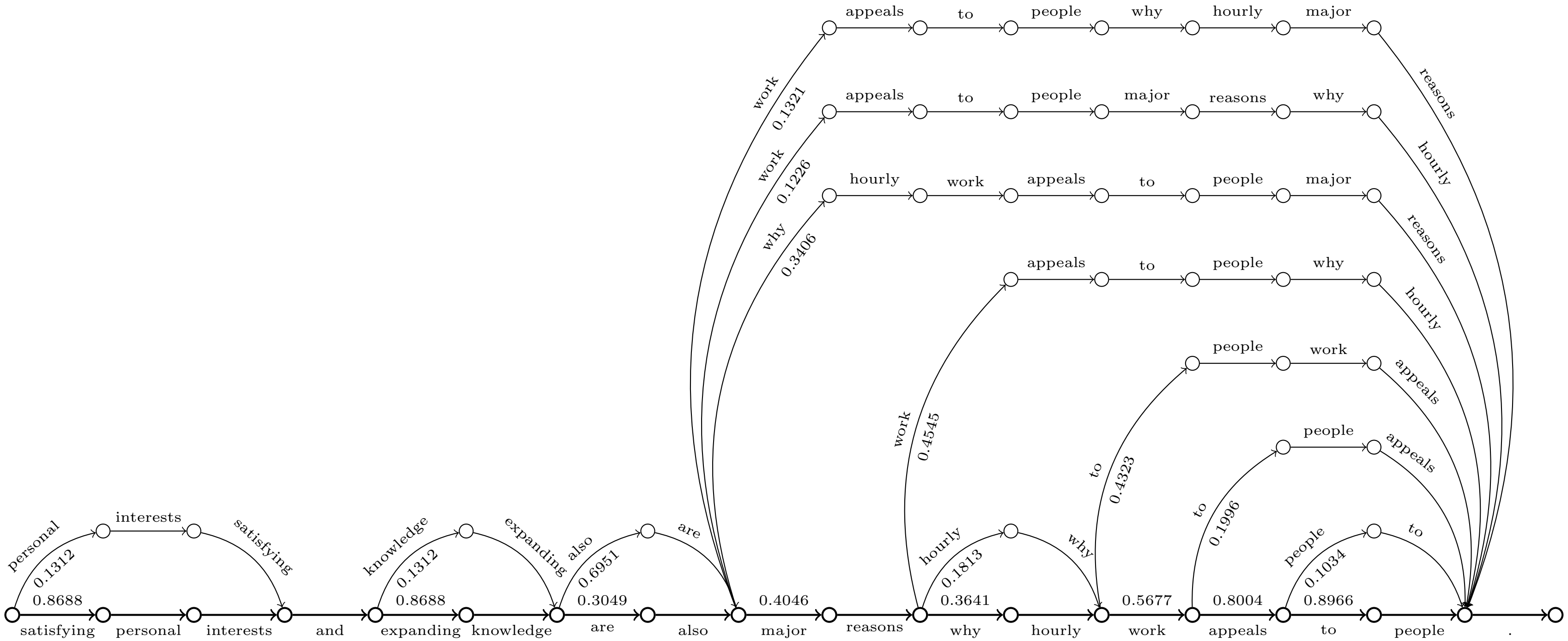

Figure 2.2.: Example of a word lattice

3. Reordering Approach

3.1. Reordering Problem in Chinese-English Translation

4. Evaluation

...

4.1. Experiement Setup

... Criterien [BOB10]

4.2. Overall Result

	BLEU Score	Improvement
Baseline	21.80	
+Short Rules	22.90	5.05 %
+Long Rules	23.13	6.10 %
+Tree Rules	23.84	9.36 %
+MLT Rules	23.96	9.91 %
Oracle Reordering	26.80	22.94 %

Table 4.1.: Result of Chinese to English translation, case-insensative

	BLEU Score	Improvement
Baseline	12.07	
+Short Rules	12.50	3.56 %
+Long Rules	12.99	7.62 %
+Tree Rules	13.38	10.85 %
+MLT Rules	13.68	13.34 %
Oracle Reordering	18.58	53.94 %

Table 4.2.: Results of English to Chinese translation

4.3. Effect dependency tree?

4.4. Alternative Scanning?

4.5. Effect of Different Left Side

4.6. Effect of Different Threshold

4.7. Research on other language pair

	BLEU Score	Improvement
Baseline	18.45	
+Short Rules	19.09	3.47 %
+Long Rules	19.16	3.85 %
+Tree Rules	19.34	4.82 %
+MLT Rules	1.00	-94.58 %
Oracle Reordering	1.00	-94.58 %

Table 4.3.: Results of English to German translation

	BLEU Score	Improvement
Baseline	18.45	
+Short Rules	19.09	3.47 %
+Long Rules	19.16	3.85 %
+Tree Rules	19.34	4.82 %
+MLT Rules	1.00	-94.58 %
Oracle Reordering	1.00	-94.58 %

Table 4.4.: Results of German to English translation

4.8. Experiement Result

4.9. Evaluation

5. Conclusion

5.1. Discussion

5.2. Conclusion

5.3. Outlook

Appendix

A. First Appendix Section

ein Bild

Figure A.1.: A figure

...

List of Tables

List of Figures

Bibliography

[BOB10] A. Birch, M. Osborne, and P. Blunsom, "Metrics for mt evaluation: Evaluating reordering," *Machine Translation*, vol. 24, no. 1, pp. 15–26, Mar. 2010. [Online]. Available: http://dx.doi.org/10.1007/s10590-009-9066-5

[Chi07] D. Chiang, "Hierarchical phrase-based translation," *computational linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[HWNW] T. Herrmann, J. Weiner, J. Niehues, and A. Waibel, "Analyzing the potential of source sentence reordering in statistical machine translation."

[LP13] U. Lerner and S. Petrov, "Source-side classifier preordering for machine translation," in *Proc. of EMNLP '13*, 2013.

[NK] J. Niehues and M. Kolss, "A pos-based model for long-range reorderings in smt."

[RV] K. Rottmann and S. Vogel, "Word reordering in statistical machine translation with a pos-based distortion model."