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 – 10th August 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.

10th August 2014

Contents

1. Introduction **1**
- 1.1. Motivation . . . 1
- 1.2. Objective and Contribution . . . 1
- 1.3. Structure . . . 2
- 1.4. Related Work . . . 2

2. Foundations **3**
- 2.1. SMT System . . . 3
- 2.2. Rules Based Pre-Reodering . . . 3
- 2.3. Word Alignment . . . 4
- 2.4. Part-of-Speech Tag . . . 5
- 2.5. Syntactic Tree . . . 5
- 2.6. Reordering Rule Types . . . 6
- 2.7. Oracle Reordering . . . 6
- 2.8. Lattices . . . 6
- 2.9. BLEU Score . . . 8

3. Reordering Approach **9**
- 3.1. Reordering Problem in Chinese-English Translation . . . 9
- 3.2. Motivation of our Pre-reordering system . . . 9
- 3.3. The reordering algorithm . . . 9

4. Evaluation **11**
- 4.1. English to Chinese Systems . . . 11
  - 4.1.1. Experimental Setup . . . 11
  - 4.1.2. Results . . . 12
  - 4.1.3. Evaluation . . . 12
- 4.2. Chinese to English Systems . . . 13
  - 4.2.1. Experimental Setup . . . 13
  - 4.2.2. Results . . . 13
  - 4.2.3. Evaluation . . . 13
- 4.3. Conclusion . . . 13

5. Conclusion **15**
- 5.1. Discussion . . . 15
- 5.2. Conclusion . . . 15
- 5.3. Outlook . . . 15

Appendix **17**
- A. Score list of systems . . . 17
- B. Documentation of Pre-Reordering System . . . 17

List of Tables **19**

List of Figures **21**

Bibliography **23**

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

special problem of chinese: segmentation

2. Foundations

2.1. SMT System

Statistical machine translation (SMT) is the state of art machine translation paradigm. It uses a typical log-linear model which is composed of a decoder and different statistical models including phrase table, reordering model and language model. All the models are weighted with parameters which are tuned from the development data. Besides development data, training data are used for training the alignment, phrase table and other models. And test data are used for evaluation purpose. The architecture of a SMT system could be illustrated as figure 2.1.

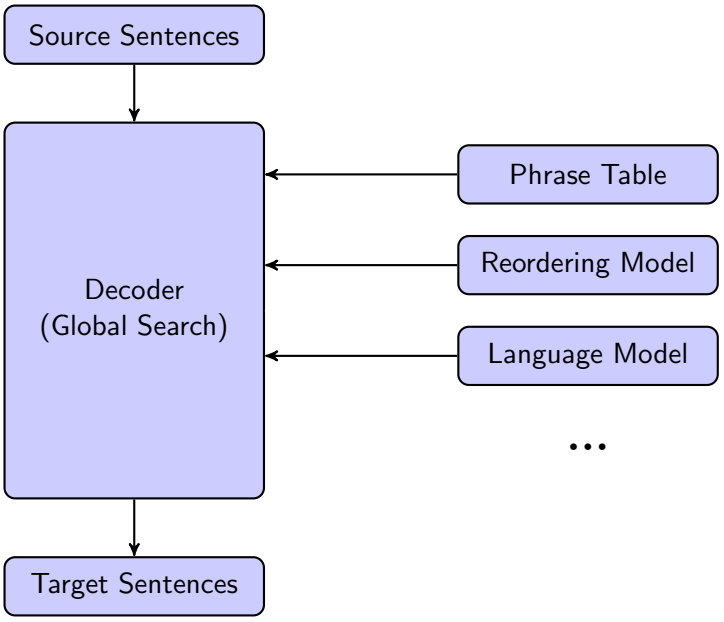

Figure 2.1.: Architecture of SMT system

2.2. Rules Based Pre-Reodering

Our pre-reordering method is based on reordering rules. Reordering rules are rules that tell us how we should reordering the sentences in source language before translating them.

In our system, the rules are generated by using the word alignment, part-of-speech (POS) tag and syntactic tree, all of which are calculated based on the training data. After we apply the reordering rules to the source sentences, word lattices are generated. The word lattices contains all the reordering possibilities of the source sentences and are further passed to the decoder for translating. The pre-reodering system could be illustrated as figure 2.2.

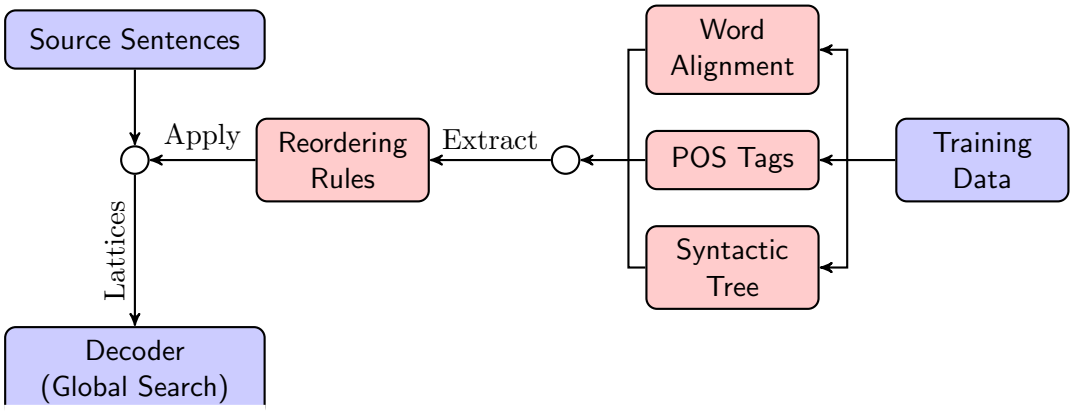

Figure 2.2.: Pre-reordering system

A more detailed description of word alignment, POS tag, syntactic tree, reordering rules and word lattices is also clarified in the following sections.

The reordering approach we used for extracting and applying the rules is introduced in the next chapter.

2.3. Word Alignment

Word alignment indicate the possible alignment between words in the source sentence and words in the target sentence. For example, figure 2.3 shows an alignment between an English sentence and a Chinese sentence.

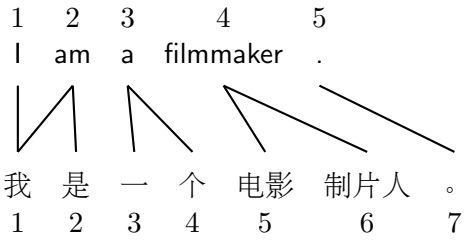

Figure 2.3.: Example of word alignment

Or it may be simply presented as index pair in the file system.

$$1-1 \quad 2-1 \quad 2-2 \quad 3-3 \quad 3-4 \quad 4-5 \quad 4-6 \quad 5-7$$

The word alignment could be trained with the GIZA++ tool by using Expectation Maximization (EM) algorithm. From the word alignment of training data we can see the patterns how the words are reordered before and after translation. Therefore, we could extract these reordering rules and apply them on the text, which is to be translated.

2.4. Part-of-Speech Tag

Part-of-speech (POS) tags are markups of words in the text, which corresponds their linguistic role in the text. Depends on the definition of the roles, the set of POS tags could be different. Besides, different languages may have different POS tag set, since they may have different linguistic features, which are relevant to translation.

The	domestic	consumption	market	for	animal	products	is	very	great	.
DT	JJ	NN	NN	IN	NN	NNS	VBZ	RB	JJ	SENT

Figure 2.4.: Example of POS tags

Figure 2.4 shows an tagged English sentence.

Tagset? English & Chinese how to tag?

2.5. Syntactic Tree

TODO

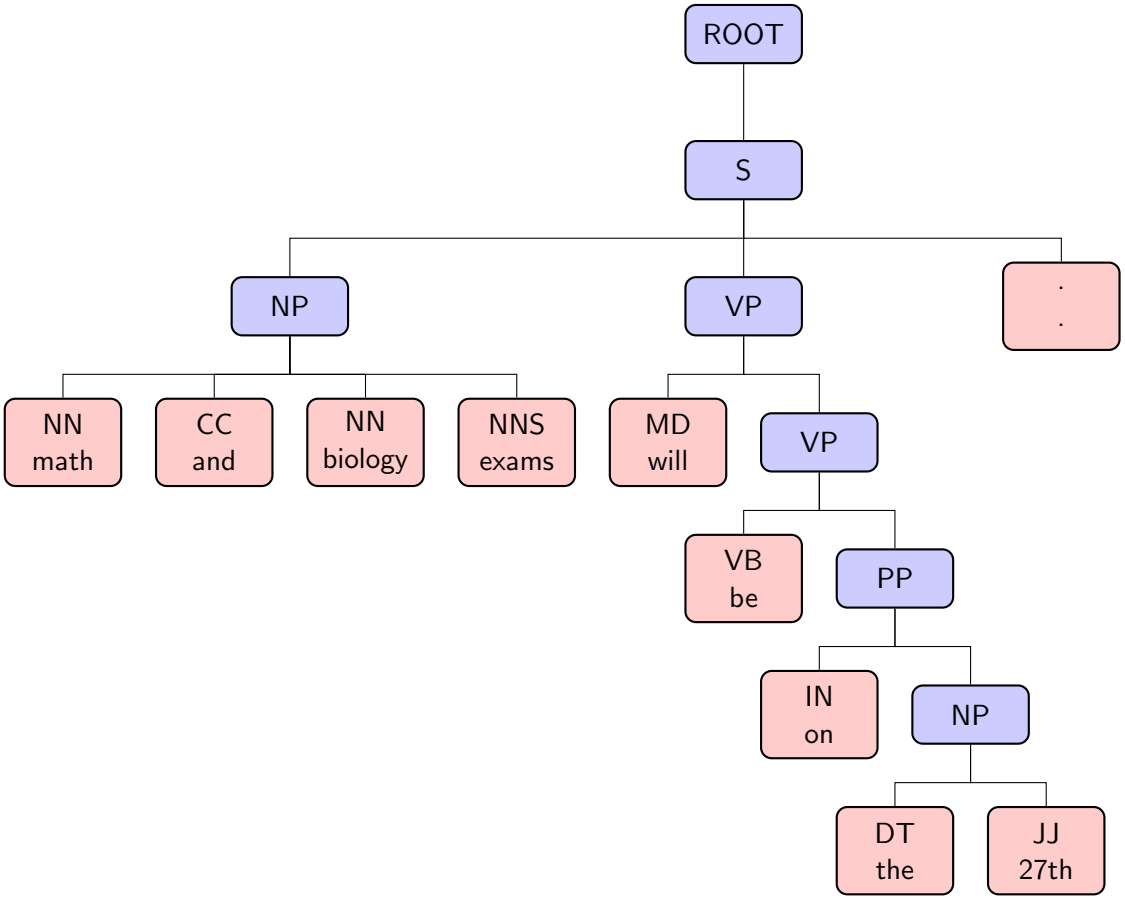

Figure 2.5.: Example of a parse tree

2.6. Reordering Rule Types

2.7. Oracle Reordering

2.8. Lattices

A word lattice could be presented with a directed acyclic graph. The graph contains nodes and transitions, with each transition labeled with a word. One example is showed in the next page. The word lattice in the example groups different word reorderings of the same English sentence together, with each reordering corresponding a path from the beginning node to the end node. The word lattice provides different p pass to decoder, probability on transitions, probability of reordering.

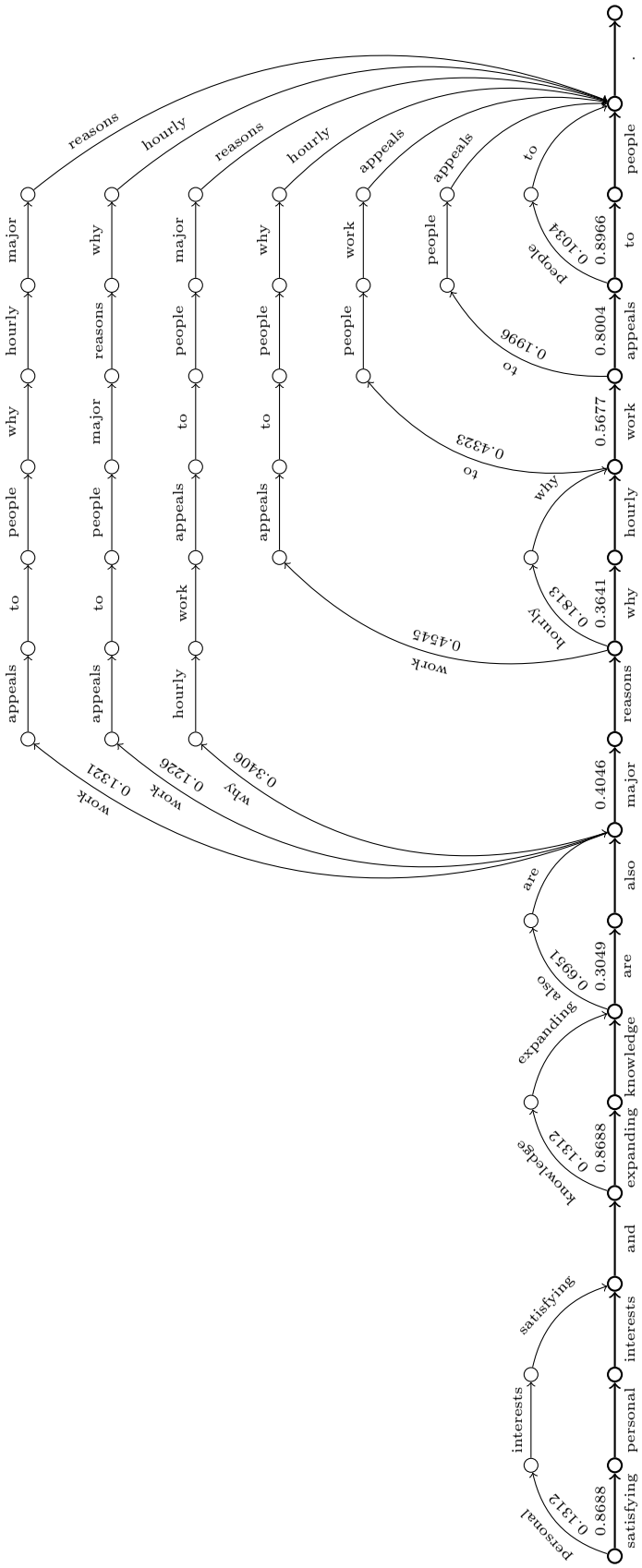

Figure 2.6.: Example of a word lattice

2.9. BLEU Score

BLEU is the de facto standard in machine translation[BOB10]. We use BLEU score to evaluate the SMT system throughout this paper.

3. Reordering Approach

3.1. Reordering Problem in Chinese-English Translation

3.2. Motivation of our Pre-reordering system

3.3. The reordering algorithm

4. Evaluation

We've conducted two sets of experiments to test our reordering methods. The first set of experiments is designed for testing the English-to-Chinese translation, which is described in the first section of this chapter. The second set of experiments is designed for the other translating direction, which is described in the second section. Through the experiments on both translating direction, we could get a better overview of our methods' effect.

Both sections are composed of three parts: experimental setup, result and evaluation. The first part describes details of the system configurations and experimental data. The second part shows the BLEU scores of different systems for comparison. And the last part evaluates the improvement with translation examples from the experiments.

4.1. English to Chinese Systems

4.1.1. Experimental Setup

We performed experiments with or without different reordering methods covering the English to Chines translation direction. The reordering methods included the reordering with short rules, long rules, tree rules and our MLT rules. The system was trained on news text from the LDC corpus and subtitles from TED talks. The development data and test data were both news text from the LDC corpus. The system was a phrase based SMT system, which used a 6-gram language model with Knersey-Ney smoothing. Besides the pre-reordering, no lexical reordering or other reordering method was used. The text was translated through a monotone decoder. The Chinese text were first segmented into words before use. The reordering rules were extracted based on the alignment, POS tags and syntactic trees from the training data. One reference of the test data was used for evaluating the BLEU score. Table 4.1 shows the size of data used in the system.

Data Set		Sentence Count	Word Count		Size (Byte)	
			English	Chinese	English	Chinese
Training Data	LDC	303K	10.96M	8.56M	60.88M	47.27M
	TED	151K	2.58M	2.86M	14.24M	15.63K
Development Data		919	30K	25K	164K	142K
Test Data		1663	47K	38K	263K	220K

Table 4.1.: BLEU score overview of English to Chinese systems

4.1.2. Results

	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	13.81	14.42 %
Oracle Reordering	18.58	53.94 %
Long Rules	12.31	1.99 %
Tree Rules	13.30	10.19 %
MLT Rules	13.68	13.34 %

Table 4.2.: BLEU score overview of English to Chinese systems

Table 4.2 shows the BLEU scores for configurations with different reordering methods. The table consist of 2 sections. the first row of the top section shows results of the baseline, which involves no reordering. In the following rows of the top section, different types of reordering rules are combined gradually, each type per row, and the improvements are showed. For example, the row with "+MLT Rules" presents the configuration with all the rule types including MLT rules and all the other rules in the rows above. All the improvements are calculated comparing to the baseline in percentage. Each row with a certain reordering type presents all the different variations of the type and the best score under these configurations are shown. For example, long rules also presents the left rules and right rules type. In the lower section of the table, rules types are not combined and the effect of each rule type is shown.

4.1.3. Evaluation

The results shows increasing scores when we used reordering methods from short rules, long rules, tree rule to MLT rules. And better BLEU scores were achieved when we combined the different reordering rules. The MLT rules improved the BLEU score, not only when we used it alone, but also when we added to the other existing reordering rules. But taking a close look at the gap between BLEU score of oracle reordering and the best BLEU score we've achieved, we can also tell, there's still much potential for improvement.

We also found improvement in the translated text by analyzing it manually. Some examples are listed in table 4.3.

Source	Hu Jintao also extended deep condolences on the death of the Chinese victims and expressed sincere sympathy to the bereaved families.
No MLT	胡锦涛 还 表示 深切 哀悼 的 受害者 家属 的 死亡 , 向 迁难者 家属 表示 诚挚 的 慰问 。
MLT	胡锦涛 还 对 中国 迁难者 表示 哀悼 , 向 迁难者 家属 表示 诚挚 的 慰问 。
Source	The Dalai Lama will go to visit Washington this month.
No MLT	达赖 喇嘛 将 访问 华盛顿 的 这 一 个 月 。
MLT	达赖 喇嘛 将 本 月 访问 华盛顿 。

Table 4.3.: Examples of improvements in translated text

Each section of table 4.3 shows translation of a sentence in its source language and target language. The translation in the row with "No MLT" comes from the configuration without using MLT reordering, and the translation in the row with "MLT" comes from the configuration with using MLT reordering. From the examples, we can see that the MLT reordering improved the sentence structure significantly. In the last example, the part of the sentence "visit Washington this month" was parsed as the structur "(visit (Washington) ((this) (month)))" in the syntactic tree, but the correct Chinese translation has word order corresponds to "this month visit Washington". This reordering inserts the word "visit" between the words "this month" and "Washington", which are on a lower level of the syntactic tree. This behavior could not be done by the tree-rule-based reordering, which only change order of constituents on the same tree level.

From this experiments we can draw the conclusion that our reordering method indeed improves the English-to-Chinese translation quality obviously, no matter when we apply it alone or when we combine it with other reordering methods mentioned in this paper. And we could further justify our claim that the MLT reordering method changes the word order more significantly to improve the translation quality.

4.2. Chinese to English Systems

4.2.1. Experimental Setup

The experiments for Chinese-to-English systems had a similar setup as described in the last section. The parallel data used in the English-to-Chinese system were also used in this experiment by changing the role of the source language and target language. We only used the LDC data set for training in this system, and the TED data were not used. And the test data had three English references for evaluating the results instead of one as in the previous system.

4.2.2. Results

	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	24.14	10.73 %
Oracle Reordering	26.80	22.94 %
Long Rules	22.10	6.10 %
Tree Rules	23.35	9.36 %
MLT Rules	23.96	10.73 %

Table 4.4.: BLEU score overview of Chinese to English systems

4.2.3. Evaluation

4.3. Conclusion

5. Conclusion

And taking a close look at the BLEU score generated with oracle reordering, we can tell there's still potential for improvement.

5.1. Discussion

5.2. Conclusion

5.3. Outlook

better algorithm for reordering between chinese english

distributive representation

Appendix

A. Score list of systems

ein Bild

Figure A.1.: A figure

...

B. Documentation of Pre-Reordering System

1. Integration of the pre-reordering system

In this section, we explain the details of our pre-reordering system and how to integrate it into the SMT system of our faculty at KIT.

2. other script

List of Tables

1.1. Word orders of three different languages 1

4.1. BLEU score overview of English to Chinese systems 11
4.2. BLEU score overview of English to Chinese systems 12
4.3. Examples of improvements in translated text 12
4.4. BLEU score overview of Chinese to English systems 13

List of Figures

2.1. Architecture of SMT system 3
2.2. Pre-reordering system 4
2.3. Example of word alignment 4
2.4. Example of POS tags 5
2.5. Example of a parse tree 5
2.6. Example of a word lattice 7

A.1. A figure 17

Bibliography

[BOB10] A. Birch, M. Osborne, and P. Blunsom, "Metrics for mt evaluation: Evaluating reordering," *Machine Translation*, vol. 24, no. 1, 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."