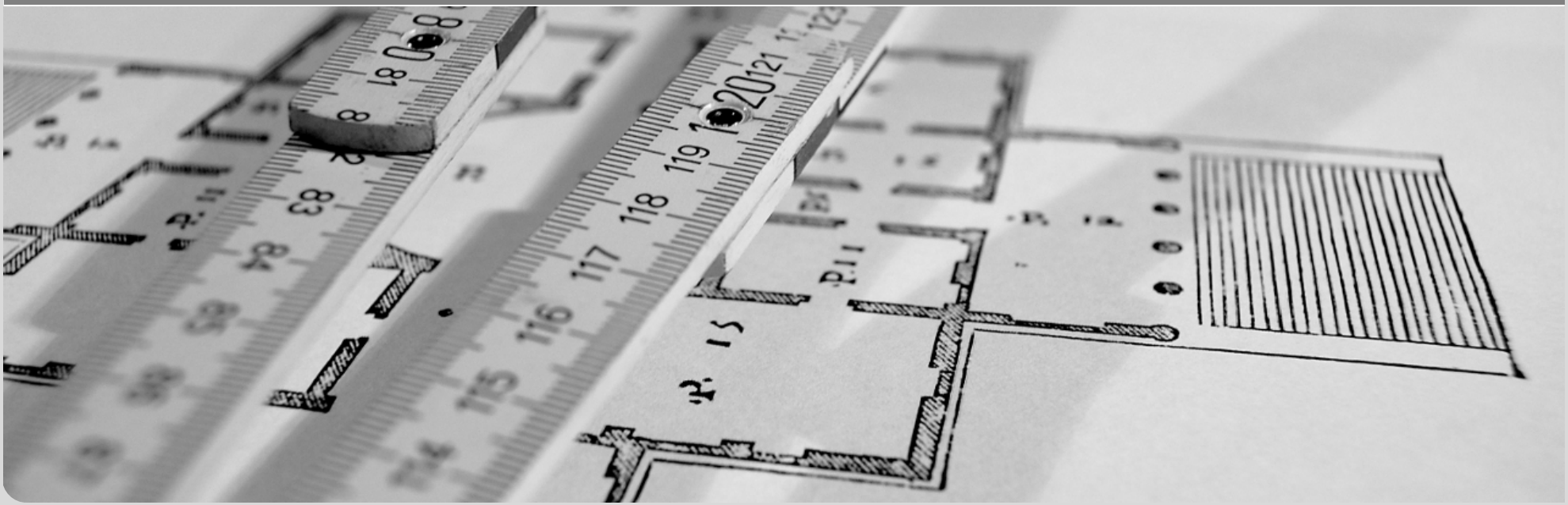
Karlsruhe Institute of Technology

Chinese Preordering on Multiple Syntactic Levels

Ge Wu | September 2, 2014

INSTITUTE FOR ANTHROPOMATICS AND ROBOTICS (IAR)



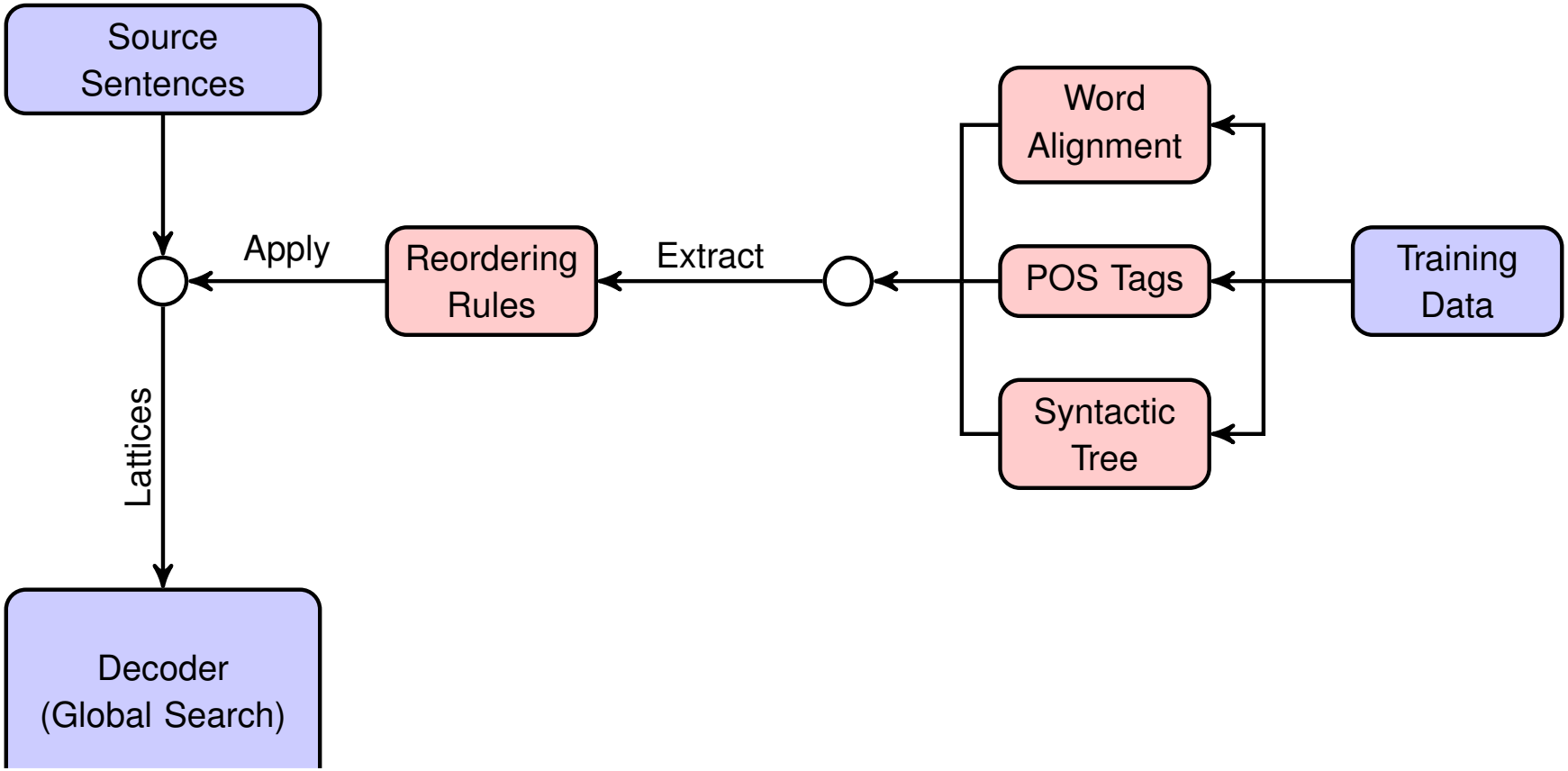
Outline

Introduction

Goal & Motivation & Reason, etc.

- X

Preordering System

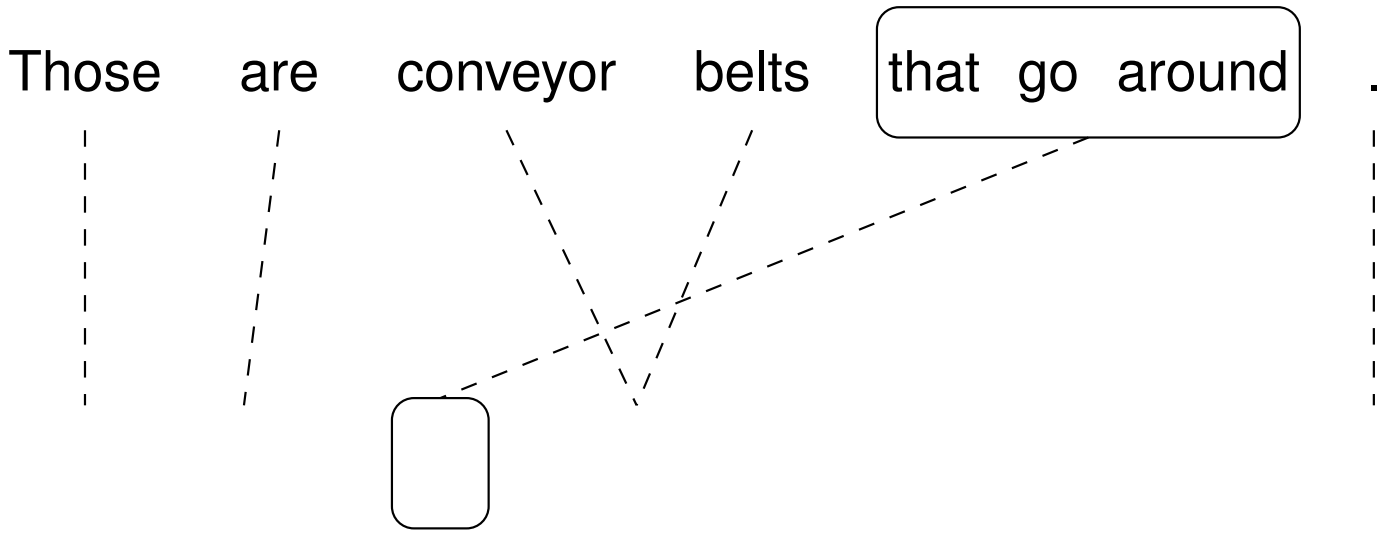


Rreordering Rules

- Short rules
- Long rules
- Tree rules

Chinese Word Orders

- Premodifier instead of postmodifier
  - Adverbials
  - Relative clauses
  - Preposition phrases



Chinese Word Orders

- Questions
- Special sentence constructions
  - *There aren't many people around that are really involved with architecture as clients.*
  - *Never would India have thought on this scale before.*

Reordering Problems

- Long distance position change
- Reordering on multiple syntactic levels

Reordering Approach

Rule Extraction & Application

- X

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: BLEU score overview of English-to-Chinese system

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	1.38 %
Tree Rules	23.35	7.11 %
MLT Rules	23.96	9.91 %

Table: BLEU score overview of Chinese to English systems

Conclusion

- Better translation quality
- Better syntactic structure
- Space for further improvement

Outlook

- Other rule types
- Better reordering approaches
- Vector presentation as feature
- Reordering with less information

Thank you for your attention

References I

- Alexandra Birch. „Reordering Metrics for Statistical Machine Translation“. In: (2011).
- Alexandra Birch, Miles Osborne, and Phil Blunsom. „Metrics for MT Evaluation: Evaluating Reordering“. In: *Machine Translation* 24.1 (Mar. 2010). ISSN: 0922-6567. DOI: 10.1007/s10590-009-9066-5. URL: http://dx.doi.org/10.1007/s10590-009-9066-5.
- Phil Blunsom, Edward Grefenstette, Nal Kalchbrenner, et al. „A Convolutional Neural Network for Modelling Sentences“. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014.

References II

David Chiang. „Hierarchical Phrase-Based Translation“. In: *computational linguistics* 33.2 (2007), pp. 201–228.

Michael Collins, Philipp Koehn, and Ivona Kučerová. „Clause Restructuring for Statistical Machine Translation“. In: *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2005, pp. 531–540.

Josep M Crego and Nizar Habash. „Using Shallow Syntax Information to Improve Word Alignment and Reordering for SMT“. In: *Proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics. 2008, pp. 53–61.

References III

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. „Generating Typed Dependency Parses from Phrase Structure Parses“. In: *Proceedings of LREC*. Vol. 6. 2006, pp. 449–454.

Nizar Habash. „Syntactic Preprocessing for Statistical Machine Translation“. In: *MT Summit XI* (2007), pp. 215–222.

Teresa Herrmann, Jan Niehues, and Alex Waibel. „Combining Word Reordering Methods on Different Linguistic Abstraction Levels for Statistical Machine Translation“. In: *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 39–47. URL: http://www.aclweb.org/anthology/W13-0805.

References IV

- Teresa Herrmann et al. *Analyzing the Potential of Source Sentence Reordering in Statistical Machine Translation*. 2013.
- Philipp Koehn. *Statistical Machine Translation*. 1st. New York, NY, USA: Cambridge University Press, 2010. ISBN: 0521874157, 9780521874151.
- Philipp Koehn et al. „Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation“. In: *IWSLT*. 2005, pp. 68–75.
- Uri Lerner and Slav Petrov. „Source-Side Classifier Preordering for Machine Translation“. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP '13)*. 2013.
- Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. „Building a Large Annotated Corpus of English: The Penn Treebank“. In: *Computational linguistics* 19.2 (1993), pp. 313–330.

References V

- Tomas Mikolov et al. „Efficient Estimation of Word Representations in Vector Space“. In: *arXiv preprint arXiv:1301.3781* (2013).
- Jan Niehues and Muntsin Kolss. „A POS-Based Model for Long-Range Reorderings in SMT“. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Association for Computational Linguistics. Athens, Greece, 2009, pp. 206–214.
- Kishore Papineni et al. „BLEU: a Method for Automatic Evaluation of Machine Translation“. In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2002, pp. 311–318.
- Maja Popovic and Hermann Ney. „POS-Based Word Reorderings for Statistical Machine Translation“. In: *International Conference on Language Resources and Evaluation*. 2006, pp. 1278–1283.

References VI

- Kay Rottmann and Stephan Vogel. *Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model*. 2007.
- Beatrice Santorini. „Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd revision)“. In: (1990).
- Christoph Tillmann. „A Unigram Orientation Model for Statistical Machine Translation“. In: *Proceedings of HLT-NAACL 2004: Short Papers*. Association for Computational Linguistics. 2004, pp. 101–104.
- Chao Wang, Michael Collins, and Philipp Koehn. „Chinese Syntactic Reordering for Statistical Machine Translation“. In: *EMNLP-CoNLL*. Citeseer. 2007, pp. 737–745.

References VII

Yuqi Zhang, Richard Zens, and Hermann Ney. „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*. Association for Computational Linguistics. 2007, pp. 1–8.