Joshua 3.0: Syntax-based Machine Translation with the Thrax Grammar Extractor

Juri Ganitkevitch[1], **Yuan Cao**[1], **Chris Callison-Burch**[1], **Matt Post**[2], and **Jonathan Weese**[1]
[1]Center for Language and Speech Processing
[2]Human Language Technology Center of Excellence
Johns Hopkins University

Abstract

Pro, compact grammars, paraphrase pivoting
TODO Juri: write this

1 Introduction

TODO Juri: clean this up and flesh it out.

Joshua is an open-source[1] toolkit for hierarchical machine translation of human languages. The original version of Joshua (Li et al., 2009) was a reimplementation of the Python-based Hiero machine-translation system (Chiang, 2007); it was later extended (Li et al., 2010) to support richer formalisms, such as SAMT (Zollmann and Venugopal, 2006).

2 Compact Grammar Representation

TODO Juri: intro into this part.

2.1 Packed Synchronous Tries

Memory usage is a limitation of both the Joshua and cdec extractors. Translation models can be very large, and many feature scores require accumulation of statistical data from the entire set of extracted rules. Since it is impractical to keep the entire grammar in memory, rules are usually sorted on disk and then read sequentially.

2.1.1 Source-Side Trie

TODO Juri: describe source-side format

2.1.2 Target-Side Trie

TODO Juri: describe target-side format

Figure 1: TODO Juri: Decoding versus load time plot.

2.1.3 Attached Data

TODO Juri: discuss attached data idea, describe feature format, alignments

2.2 Quantization

TODO Juri: discuss features taking the most spaces, quantization in the spirit of KenLM and BerkeleyLM.

2.3 Optimizations

TODO Juri: what did we do to improve decoding speed?

2.4 Experiments

TODO Juri: brief rundown of experiments

[1] http://github.com/joshua-decoder/joshua

Language pair	sentences (K)	words (M)
cs–en	332	4.7
de–en	279	5.5
en–cs	487	6.9
en–de	359	7.2
en–fr	682	12.5
fr–en	792	14.4

Table 1: TODO Juri: some BLEU scores for quantized versus not quantized.

Figure 2: TODO Yuan: Plot of iterations/score for various classifiers, pointing out that our built-in perceptron is doing well.

3 Y-PRO: Pairwise Ranking Optimization in Joshua

TODO Yuan: give a brief description of PRO, highlight the compatibility with Z-MERT's easily plugged in metrics. Also highlight the supported classifiers (which we should fix in the main repository)

3.1 Experiments

TODO Yuan: Describe the experiments you did for convergence/speed/translation quality

4 Thrax: Paraphrase Extraction at Scale

TODO Juri: describe paraphrase stage and integration with Thrax features

5 Future work

TODO All: Ideas? Sparse features?

Language pair	sentences (K)	words (M)
cs–en	332	4.7
de–en	279	5.5
en–cs	487	6.9
en–de	359	7.2
en–fr	682	12.5
fr–en	792	14.4

Table 2: TODO Yuan: Table of MERT versus PRO (with various classifiers) showing numer of iterations, time needed and scores on dev and test.

Language pair	sentences (K)	words (M)
cs–en	332	4.7
de–en	279	5.5
en–cs	487	6.9
en–de	359	7.2
en–fr	682	12.5
fr–en	792	14.4

Table 3: TODO Juri: Table of large grammars we extracted.

References

Ondřej Bojar and Zdeněk Žabokrtský. 2009. CzEng0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92. in print.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

David Chiang. 2010. Learning to translate with source and target syntax. In *Proc. ACL*, Uppsala, Sweden, July.

Jonathan H. Clark and Alon Lavie. 2010. Loonybin: Keeping language technologists sane through automated management of experimental (hyper) workflows. In *Proc. LREC*.

Jeffrey Dean and Sanjay Ghemawat. 2004. Mapreduce: Simplified data processing on large clusters. In *OSDI*.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proc. ACL 2010 System Demonstrations*, pages 7–12.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proc. NAACL*, Boston, Massachusetts, USA, May.

Philipp Koehn, Franz Josef Och, and Daniel Marcu.

2003. Statistical phrase-based translation. In *Proc. NAACL*, Morristown, NJ, USA.

Shankar Kumar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proc. NAACL*, Boston, Massachusetts, USA, May.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proc. WMT*, Athens, Greece, March.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Ann Irvine, Sanjeev Khudanpur, Lane Schwartz, Wren N.G. Thornton, Ziyuan Wang, Jonathan Weese, and Omar F. Zaidan. 2010. Joshua 2.0: a toolkit for parsing-based machine translation with syntax, semirings, discriminative training and other goodies. In *Proc. WMT*.

Yang Liu, Yajuan Lü, and Qun Liu. 2009. Improving tree-to-tree translation with packed forests. In *Proc. ACL*, Suntec, Singapore, August.

Adam Lopez. 2008. Tera-scale translation models via pattern matching. In *Proc. COLING*, Manchester, UK, August.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proc. ACL*, Sydney, Australia, July.

Lane Schwartz and Chris Callison-Burch. 2010. Hierarchical phrase-based grammar extraction in joshua: Suffix arrays and prefix trees. *The Prague Bulletin of Mathematical Linguistics*, 93:157–166, January.

Ashish Venugopal and Andreas Zollmann. 2009. Grammar based statistical MT on Hadoop: An end-to-end toolkit for large scale PSCFG based MT. *The Prague Bulletin of Mathematical Linguistics*, 91:67–78.

Jonathan Weese. 2011. A systematic comparison of synchronous context-free grammars for machine translation. Master's thesis, Johns Hopkins University, May.

Omar F. Zaidan. 2009. Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88.

Min Zhang, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan, and Sheng Li. 2008. A tree sequence alignment-based tree-to-tree translation model. In *Proc. ACL*, Columbus, Ohio, June.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proc. NAACL Workshop on Statistcal Machine Translation*, New York, New York.

Andreas Zollmann, Ashish Venugopal, Franz Josef Och, and Jay Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical mt. In *Proc. COLING*.