Reproducible Results in Parsing-Based Machine Translation: The JHU Shared Task Submission

Lane Schwartz *
University of Minnesota
Minneapolis, MN
lane@cs.umn.edu

Abstract

We present the Johns Hopkins University submission to the 2010 WMT shared translation task. We describe processing steps using open data and open source software used in our submission, and provide the scripts and configurations required to train, tune, and test our machine translation system.

1 Introduction

Research investigating natural language processing and computational linguistics can and should have an extremely low barrier to entry. The data with which we work is customarily available in common electronic formats. The computational techniques which we apply can typically be performed on commodity computing resources which are widely available. In short, there should be no reason why small research groups and even lone researchers should not be able to join and make substantive contributions furthering our field. The reality is less encouraging.

Many published articles describe novel techniques and provide interesting results, yet fail to describe technical details in sufficient detail to allow their results to be reproduced by other researchers. While there are notable and laudable exceptions, many publications fail to provide the source code and scripts necessary to reproduce results. The use of restricted data, not freely available for download by any interested researcher only compounds these problems. Pedersen (2008) rightly argues that the implementation details so often ignored in publications are in fact essential for our research to be reproducible science.

Reproducibility in machine translation is made more challenging by the complexity of experimental workflows. Results in machine translation tasks are dependent on a cascade of processing steps and configurations. While interesting subsets of these usually appear in experimental descriptions, many steps (preprocessing techniques, alignment parameters, translation rule extraction parameters, language model parameters, list of features used) are invariably omitted, even though these configurations are often critical to reproducing results.

This paper describes the Johns Hopkins University submission to the 2010 Workshop on Statistical Machine Translation shared translation task. Links to the software, scripts, and configurations used to run the experiments described herein are provided. The remainder of this paper is structured as follows. Section 2 lists the major examples of publicly available open source machine translation systems, parallel corpora, and machine translation workflow management systems. Section 3 describes the experimental workflow used to run the shared task translations, with the corresponding experimental design in section 4. Section 5 presents the shared task results.

2 Related Work

The last four years have witnessed the implementation and release of numerous open source machine translation systems. The widely used Moses system (Koehn et al., 2007) implements the standard phrase-based translation model. Parsing-based translation models are implemented by Joshua (Li et al., 2009), SAMT (Zollmann and Venugopal, 2006), and cdec (Dyer et al., 2010). Cunei (Phillips and Brown, 2009) implements statistical example-based translation. Olteanu et al. (2006) and Schwartz (2008) respectively provide additional open-source implementations of phrase-based and hierarchical decoders.

The SRILM (Stolcke, 2002), IRSTLM (Federico et al., 2008), and RandLM (Talbot and Osborne, 2007) toolkits enable efficient training and

*Research conducted as a visiting researcher at Johns Hopkins University

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 183–188,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

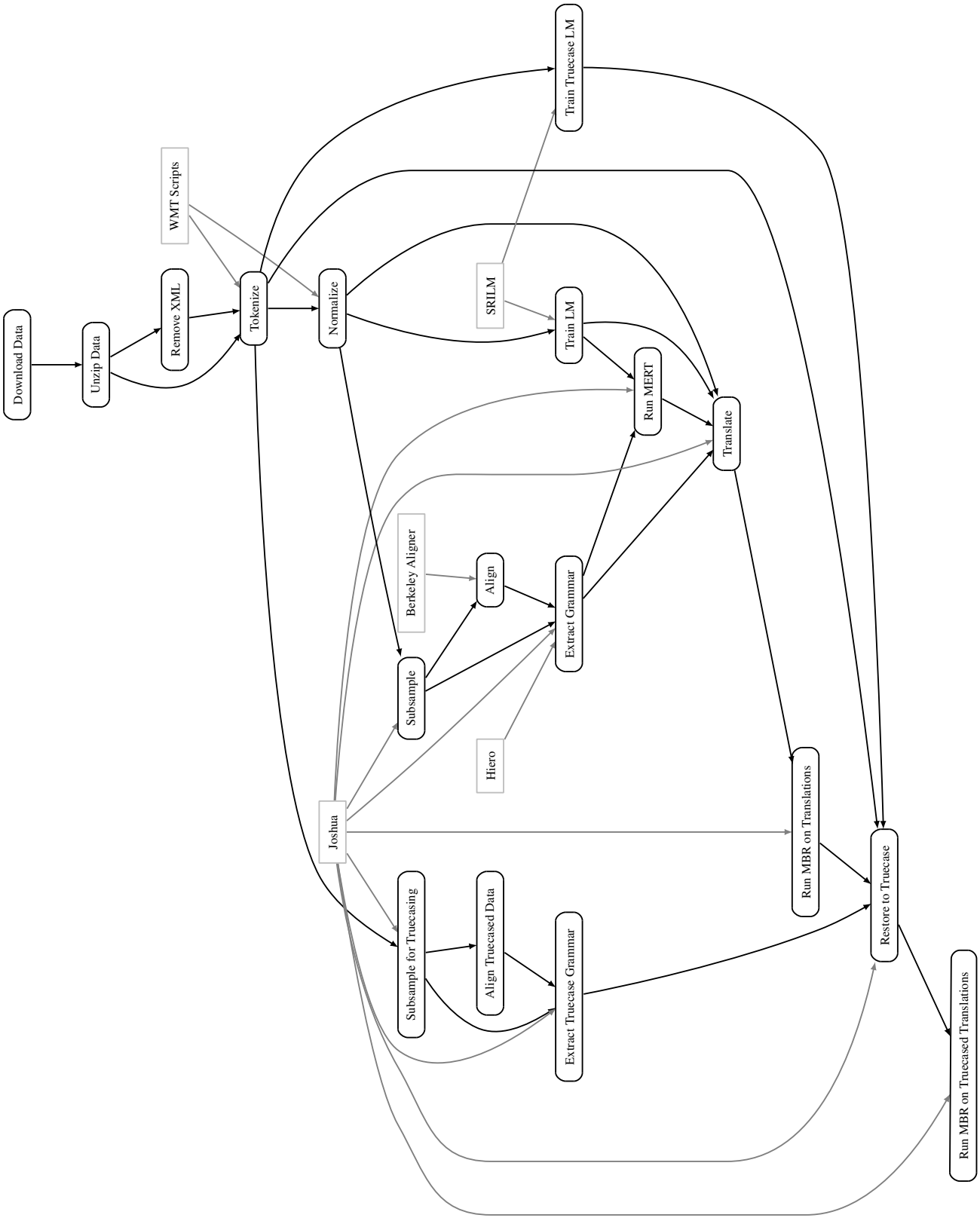

Figure 1: Machine translation workflow. Square nodes in grey indicate software and scripts. The scripts and configuration files used to implement and run this workflow are available for download at `http://sourceforge.net/projects/joshua/files/joshua/1.3/wmt2010-experiment.tgz/download`

querying of n-gram language models.

Freely available parallel corpora for numerous European languages have also been released in recent years. These include the Europarl (Koehn, 2005) and JRC-Acquis (Steinberger et al., 2006) legislative corpora, each of which includes data for most EU language pairs. The smaller News Commentary corpora (Callison-Burch et al., 2007; Callison-Burch et al., 2008) provide smaller amounts of parallel data in the news genre. The recent Fr-En 10^9 (Callison-Burch et al., 2009) corpus aggregates huge numbers of parallel French-English sentences from the web.

Open source systems to address the complex workflows required to run non-trivial machine translation experiments have also been developed. These include `experiment.perl` (Koehn et al., 2010), developed as a workflow management system at the University of Edinburgh, and LoonyBin (Clark et al., 2010), a general hyperworkflow management utility from Carnegie Melon University.

3 Managing Experiment Workflows

Running a statistical machine translation system to achieve state-of-the-art performance involves the configuration and execution of numerous interdependent intermediate tools. To manage task dependencies and tool configuration, our shared task workflow consists of a set of dependency scripts written for GNU Make (Stallman et al., 2006).

Figure 1 shows a graph depicting the steps in our experimental workflow, and the dependencies between steps. Each node in the graph represents a step in the workflow; each step is implemented as a Make script that defines how to run the tools required in that step. In each experiment, an additional configuration script is provided for each experimental step, defining the parameters to be used when running that step in the current experiment. Optional front-end wrapper scripts can also be provided, allowing for a complete experiment to be run - from downloading data and software through truecasing translated results - by executing a single make file.

This framework is also conducive to parallelization. Many tasks, such as preprocessing numerous training files, are not dependent on one another. In such cases `make` can be configured to execute multiple processes simultaneously on a single multi-processor machine. In cases where scheduled distributed computing environments such as the Sun Grid Engine are configured, make files can be processed by scheduler-aware `make` variants (`distmake`, SGE `qmake`, Sun Studio `dmake`) which distribute outstanding tasks to available distributed machines using the relevant distributed scheduler.

4 Experimental Configuration

Experimental workflows were configured[1] and run for six language pairs in the translation shared task: English-French, English-German, English-Spanish, French-English, German-English, and Spanish-English.

In all experiments, only data freely available for download was used. No restricted data from the LDC or other sources was used. Table 1 lists the parallel corpora used in training the translation model for each experiment. The monolingual corpora used in training each target language model are listed in table 2. In all experiments, newstest2008 was used as a development tuning corpus during minimum error rate training; newstest2009 was used as a development test set. The shared task data set newstest2010 was used as a final blind test set.

All data was automatically downloaded, unzipped, and preprocessed prior to use. Files provided in XML format were converted to plain text by selecting lines with `<seg>` tags, then removing the beginning and end tags for each segment; this processing was applied using GNU grep and sed. The `tokenize.perl` and `lowercase.perl` scripts provided for the shared task[2] were applied to all data.

Interpolated n-gram language models for the four target languages were built using the SRI Language Model Toolkit[3], with n-gram order set to 5. The Chen and Goodman (1998) technique for modified Kneser-Ney discounting (Kneser and Ney, 1995) was applied during language model training.

Following Li et al. (2009), a subset of the available training sentences was selected via subsam-

[1] http://sourceforge.net/projects/joshua/files/joshua/1.3/wmt2010-experiment.tgz/download

[2] http://www.statmt.org/wmt08/scripts.tgz with md5sum:
`tokenize.perl` 45cd1832827131013245eca76481441a
`lowercase.perl` a1958ab429b1e29d379063c3b9cd7062

[3] http://www-speech.sri.com/projects/srilm
SRILM version 1.5.7. Our experimental workflow requires that SRILM be compiled separately, with the $SRILM environment variable set to the install location.

Source	Target	Parallel Corpora
German	English	news-commentary10.de-en europarl-v5.de-en
English	German	news-commentary10.de-en europarl-v5.de-en
French	English	news-commentary10.fr-en europarl-v5.fr-en giga-fren.release2 undoc.2000.en-fr
English	French	news-commentary10.fr-en europarl-v5.fr-en giga-fren.release2 undoc.2000.en-fr
Spanish	English	news-commentary10.es-en europarl-v5.es-en undoc.2000.en-es
English	Spanish	news-commentary10.es-en europarl-v5.es-en undoc.2000.en-es

Table 1: Parallel training data used for training translation model, per language pair

Target	Monolingual Corpora
English	europarl-v5.en news-commentary10.en news.en.shuffled undoc.2000.en-fr.en giga-fren.release2.en
French	europarl-v5.fr news-commentary10.fr news.fr.shuffled undoc.2000.en-fr.fr giga-fren.release2.fr
German	europarl-v5.de news-commentary10.de news.de.shuffled
Spanish	europarl-v5.es news-commentary10.es news.es.shuffled undoc.2000.en-es.es

Table 2: Monolingual training data used for training language model, per target language

pling; training sentences are selected based on the estimated likelihood of each sentence being useful later for translating a particular test corpus.

Given a subsampled parallel training corpus, word alignment is performed using the Berkeley aligner[4] (Liang et al., 2006).

For each language pair, a synchronous context free translation grammar is extracted for a particular test set, following the methods of Lopez (2008) as implemented in (Schwartz and Callison-Burch, 2010). For the largest training sets (French-English and English-French) the original (Lopez, 2008) implementation included with Hiero was used to save time during training[5].

Because of the use of subsampling, the extracted translation grammars are targeted for use with a specific test set. Our experiments were begun prior to the release of the blind newstest2010 shared task test set. Subsampling was performed for the development tuning set, news-test2008, and the development test set, newstest2009. Once the newstest2010 test set was released, the process of subsampling, alignment, and grammar extraction was repeated to obtain translation grammars targeted for use with the shared task test set.

Our experiments used hierarchical phrase-based grammars containing exactly two nonterminals - the wildcard nonterminal X, and S, used to glue together neighboring constituents. Recent work has shown that parsing-based machine translation using SAMT (Zollmann and Venugopal, 2006) grammars with rich nonterminal sets can demonstrate substantial gains over hierarchical grammars for certain language pairs (Baker et al., 2009). Joshua supports such grammars; the experimental workflow presented here could easily be extended in future research to incorporate the use of SAMT grammars with additional language pairs.

The Z-MERT implementation (Zaidan, 2009) of minimum error rate training (Och, 2003) was used for parameter tuning. Tuned grammars were used by Joshua to translate all test sets. The Joshua decoder produces n-best lists of translations.

Rather than simply selecting the top candidate from each list, we take the preferred candidate after perform minimum Bayes risk rescoring (Kumar and Byrne, 2004).

Once a single translation has been extracted for each sentence in the test set, we repeat the procedures described above to train language and translation models for use in translating lowercased results into a more human-readable truecased form. A truecase language model is trained as above, but on the tokenized (but not normalized) monolingual target language corpus. Monotone word alignments are deterministically created, mapping normalized lowercase training text to the original truecase text. As in bilingual translation, subsampling is performed for the training set, and a translation grammar for lowercase-to-truecase is extracted. No tuning is

[4]http://berkeleyaligner.googlecode.com/files/berkeleyaligner_unsupervised-2.1.tar.gz — Berkeley aligner version 2.1

[5]It is expected that using the Joshua implementation should result in nearly identical results, albeit with somewhat more time required to extract the grammar.

performed. The Joshua decoder is used to translate the lowercased target language test results into truecase format. The `detokenize.perl` and `wrap-xml.perl` scripts provided for the shared task were manually applied to truecased translation results prior to final submission of results.

The code used for subsampling, grammar extraction, decoding, minimum error rate training, and minimum Bayes risk rescoring is provided with Joshua[6], with the exception of the original (Lopez, 2008) grammar extraction implementation.

5 Experimental Results

The experiments described in sections 3 and 4 above provided truecased translations for six language pairs in the translation shared task: English-French, English-German, English-Spanish, French-English, German-English, and Spanish-English. Table 3 lists the automatic metric scores for the newstest2010 test set, according to the BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) metrics.

Source	Target	BLEU	BLEU-cased	TER
German	English	21.3	19.5	0.660
English	German	15.2	14.6	0.738
French	English	27.7	26.4	0.614
English	French	23.8	22.8	0.681
Spanish	English	29.0	27.6	0.595
English	Spanish	28.1	26.5	0.596

Table 3: Automatic metric scores for the test set newstest2010

The submitted system ranked highest among shared task participants for the German-English task, according to TER.

In order to provide points of comparison with the 2009 Workshop on Statistical Machine Translation shared translation task participants, table 4 lists automatic metric scores for our systems' translations of the newstest2009 test set, which we used as a development test set.

6 Steps to Reproduce

The experiments in this paper can be reproduced by running the make scripts provided in the following file: `http://sourceforge.net/projects/joshua/files/joshua/1.3/wmt2010-experiment.tgz/download`. The README file details how to configure the workflow for your environment. Note that SRILM must be downloaded and compiled separately before running the experimental steps.

[6] http://sourceforge.net/projects/joshua/files/joshua/1.3/joshua-1.3.tgz/download — Joshua version 1.3

Source	Target	BLEU
German	English	18.19
English	German	13.57
French	English	26.41
English	French	25.28
Spanish	English	25.28
English	Spanish	24.02

Table 4: Automatic metric scores for the development test set newstest2009

Acknowledgements

This work was supported by the DARPA GALE program (Contract No HR0011-06-2-0001).

References

Kathy Baker, Steven Bethard, Michael Bloodgood, Ralf Brown, Chris Callison-Burch, Glen Coppersmith, Bonnie Dorr, Wes Filardo, Kendall Giles, Anni Irvine, Mike Kayser, Lori Levin, Justin Martineau, Jim Mayfield, Scott Miller, Aaron Phillips, Andrew Philpot, Christine Piatko, Lane Schwartz, and David Zajic. 2009. Semantically informed machine translation (SIMT). SCALE summer workshop final report, Human Language Technology Center Of Excellence.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT07)*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT08)*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT09)*, March.

Stanley Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University, Cambridge, MA, USA, August.

Jonathan Clark, Jonathan Weese, Byung Gyu Ahn, Andreas Zollman, Qin Gao, Kenneth Heafield, and Alon Lavie. 2010. The machine translation toolpack for LoonyBin: Automated management of experimental machine translation hyperworkflows. *The Prague Bulletin of Mathematical Linguistics*, 93:117–126, January.

C. Dyer, A. Lopez, J. Ganitkevitch, J. Weese, F. Ture, P. Blunsom, H. Setiawan, V. Eidelman, and P. Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proc. ACL (Demonstration Track)*, Uppsala, Sweden.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: An open source toolkit for handling large scale language models. In *Proc. Interspeech*, Brisbane, Australia.

Reinhard Kneser and Hermann Ney. 1995. Improved smoothing for mgram language modeling. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL-2007 Demo and Poster Sessions*.

Philipp Koehn, Anthony Rousseau, Ben Gottesmann, Aurora Marsye, Frédéric Blain, and Eun-Jin Park, 2010. *An Experiment Management System*. Fourth Machine Translation Marathon, Dublin, Ireland, January.

Philipp Koehn. 2005. A parallel corpus for statistical machine translation. In *Proceedings of MT-Summit*, Phuket, Thailand.

Shankar Kumar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of HLT/NAACL*.

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 *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece, March. Association for Computational Linguistics.

P. Liang, B. Taskar, and D. Klein. 2006. Alignment by agreement. In *North American Association for Computational Linguistics (NAACL)*, pages 104–111.

Adam Lopez. 2008. *Machine Translation by Pattern Matching*. Ph.D. thesis, University of Maryland.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of ACL*.

Marian Olteanu, Chris Davis, Ionut Volosen, and Dan Moldovan. 2006. Phramer an open source statistical pharse-based translator. In *HLT-NAACL 2006: Proceedings of the Workshop on Statistical Machine Translation*, New York, June.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Ted Pedersen. 2008. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470.

Aaron B. Phillips and Ralf D. Brown. 2009. Cunei machine translation platform: System description. In *3rd Workshop on Example-Based Machine Translation*, Dublin, Ireland.

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

Lane Schwartz. 2008. An open-source hierarchical phrase-based translation system. In *Proceedings of the 5th Midwest Computational Linguistics Colloquium (MCLC'08)*, May.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*.

Richard M. Stallman, Roland McGrath, and Paul D. Smith, 2006. *GNU Make*. Free Software Foundation, Boston, MA, 0.70 edition, April.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. 2006. The JRC-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colorado, September.

David Talbot and Miles Osborne. 2007. Randomised language modelling for statistical machine translation. In *Proceedings of ACL*.

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.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of the NAACL-2006 Workshop on Statistical Machine Translation (WMT-06)*, New York, New York.