CMU Syntax-Based Machine Translation at WMT 2011

Greg Hanneman and **Alon Lavie**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
{ghannema, alavie}@cs.cmu.edu

Abstract

We present the Carnegie Mellon University Stat-XFER group submission to the WMT 2011 shared translation task. We built a hybrid syntactic MT system for French–English using the Joshua decoder and an automatically acquired SCFG. New work for this year includes training data selection and grammar filtering. Expanded training data selection significantly increased translation scores and lowered OOV rates, while results on grammar filtering were mixed.

1 Introduction

During the past year, the statistical transfer machine translation group at Carnegie Mellon University has continued its work on large-scale syntactic MT systems based on automatically acquired synchronous context-free grammars (SCFGs). For the 2011 Workshop on Machine Translation, we built a hybrid MT system, including both syntactic and non-syntactic rules, and submitted it as a constrained entry to the French–English translation task. This is our fourth yearly submission to the WMT shared translation task.

In design and construction, the system is similar to our submission from last year's workshop (Hanneman et al., 2010), with changes in the methods we employed for training data selection and SCFG filtering. Continuing WMT's general trend, we worked with more data than in previous years, basing our 2011 system on 13.9 million sentences of parallel French–English training data and an English language model of 1.8 billion words. Decoding was carried out in Joshua (Li et al., 2009), an open-source framework for parsing-based MT. We managed our experiments with LoonyBin (Clark and Lavie, 2010), an open-source tool for defining, modifying, and running complex experimental pipelines.

We describe our system-building process in more detail in Section 2. In Section 3, we evaluate the system's performance on WMT development sets and examine the aftermath of training data selection and grammar filtering. Section 4 concludes with possible directions for future work.

2 System Construction

2.1 Training Data Selection

WMT 2011's provided French–English training data consisted of 36.8 million sentence pairs from the Europarl, news commentary, UN documents, and Giga-FrEn corpora (Table 1). The first three of these are, for the most part, clean data resources that have been successfully employed as MT corpora for a number of years. The Giga-FrEn corpus, though the largest, is also the least precise, as its Web-crawled data sources are less homogeneous and less structured than the other corpora. Nevertheless, Pino et al. (2010) found significant improvements in French–English MT output quality by including it. Our goal for this year was to strike a middle ground: to avoid computational difficulties in using the entire 36.8 million sentence pairs of training data, but to mine the Giga-FrEn corpus for sentences to increase our system's vocabulary coverage.

Our method of training data selection proceeded as follows. We first tokenized all the parallel training

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 365–371,
Edinburgh, Scotland, UK, July 30–31, 2011. ©2011 Association for Computational Linguistics

Corpus	Released	Used
Europarl	1,825,077	1,614,111
News commentary	115,562	95,138
UN documents	12,317,600	9,352,232
Giga-FrEn	22,520,400	2,839,466
Total	**36,778,639**	**13,900,947**

Table 1: Total number of training sentence pairs released, by corpus, and the number used in building our system.

data using the Stanford parser's tokenizer (Klein and Manning, 2003) for English and our own in-house script for French. We then passed the Europarl, news commentary, and UN data through a filtering script that removed lines longer than 95 tokens in either language, empty lines, lines with excessively imbalanced length ratios, and lines containing tokens of more than 25 characters in either language. From the filtered data, we computed a list of the source-side vocabulary words along with their frequency counts. Next, we searched the Giga-FrEn corpus for relatively short lines on the source side (up to 50 tokens long) that contained either a new vocabulary word or a word that had been previously seen fewer than 20 times. Such lines were added to the filtered training data to make up our system's final parallel training corpus.

The number of sentences retained from each data source is listed in Table 1; in the end, we trained our system from 13.9 million parallel sentences. With the Giga-FrEn data included, the source side of our parallel corpus had a vocabulary of just over 1.9 million unique words, compared with a coverage of 545,000 words without using Giga-FrEn.

We made the decision to leave the training data in mixed case for our entire system-building process. At the cost of slightly sparser estimates for word alignments and translation probabilities, a mixed-case system avoids the extra step of building a statistical recaser to treat our system's output.

2.2 Grammar Extraction and Scoring

Once we had assembled the final training corpus, we annotated it with statistical word alignments and constituent parse trees on both sides. Unidirectional word alignments were provided by MGIZA++ (Gao and Vogel, 2008), then symmetrized with the grow-diag-final-and heuristic (Koehn et al., 2005). For generating parse trees, we used the French and English grammars of the Berkeley statistical parser (Petrov and Klein, 2007).

Except for minor bug fixes, our method for extracting and scoring a translation grammar remains the same as in our WMT 2010 submission. We extracted both syntactic and non-syntactic portions of the translation grammar. The non-syntactic grammar was extracted from the parallel corpus and word alignments following the standard heuristics of phrase-based SMT (Koehn et al., 2003). The syntactic grammar was produced using the method of Lavie et al. (2008), which decomposes each pair of word-aligned parse trees into a series of minimal SCFG rules. The word alignments are first generalized to node alignments, where nodes s and t are aligned between the source and target parse trees if all word alignments in the yield of s land within the yield of t and vice versa. Minimal SCFG rules are derived from adjacent levels of node alignments: the labels from each pair of aligned nodes forms a rule's left-hand side, and the right-hand side is made up of the labels from the frontier of aligned nodes encountered when walking the left-hand side's subtrees. Within a phrase length limit, each aligned node pair generate an all-terminal phrase pair rule as well.

Since both grammars are extracted from the same Viterbi word alignments using similar alignment consistency constraints, the phrase pair rules from the syntactic grammar make up a subset of the rules extracted according to phrase-based SMT heuristics. We thus share instance counts between identical phrases extracted in both grammars, then delete the non-syntactic versions. Remaining non-syntactic phrase pairs are converted to SCFG rules, with the phrase pair forming the right-hand side and the dummy label PHR::PHR as the left-hand side. Except for the dummy label, all nonterminals in the final SCFG are made up of a syntactic category label from French joined with a syntactic category label from English, as extracted in the syntactic grammar. A sampling of extracted SCFG rules is shown in Figure 1.

The combined grammar was scored according to the 22 translation model features we used last year. For a generic SCFG rule of the form $\ell_s :: \ell_t \rightarrow$

$$\text{PHR} :: \text{PHR} \rightarrow [,\ \text{ainsi qu'}] :: [\text{as well as}]$$

$$\text{V} :: \text{VBN} \rightarrow [\text{modifiées}] :: [\text{modified}]$$

$$\text{NP} :: \text{NP} \rightarrow [\text{les conflits armés}] :: [\text{armed conflict}]$$

$$\text{AP} :: \text{SBAR} \rightarrow [\text{tel qu' VPpart}^1] :: [\text{as VP}^1]$$

$$\text{NP} :: \text{NP} \rightarrow [\text{D}^1\ \text{N}^2\ \text{A}^3] :: [\text{CD}^1\ \text{JJ}^3\ \text{NNS}^2]$$

Figure 1: Sample extracted SCFG rules. They include non-syntactic phrase pairs, single-word and multi-word syntactic phrase pairs, partially lexicalized hierarchical rules, and fully abstract hierarchical rules.

$[r_s] :: [r_t]$, we computed 11 maximum-likelihood features as follows:

- Phrase translation scores $P(r_s \mid r_t)$ and $P(r_t \mid r_s)$ for phrase pair rules, using the larger non-syntactic instance counts for rules that were also extracted syntactically.
- Hierarchical translation scores $P(r_s \mid r_t)$ and $P(r_t \mid r_s)$ for syntactic rules with nonterminals on the right-hand side.
- Labeling scores $P(\ell_s :: \ell_t \mid r_s)$, $P(\ell_s :: \ell_t \mid r_t)$, and $P(\ell_s :: \ell_t \mid r_s, r_t)$ for syntactic rules.
- "Not syntactically labelable" scores $P(\ell_s :: \ell_t = \text{PHR} :: \text{PHR} \mid r_s)$ and $P(\ell_s :: \ell_t = \text{PHR} :: \text{PHR} \mid r_t)$, with additive smoothing ($n = 1$), for all rules.
- Bidirectional lexical scores for all rules with lexical items, calculated from a unigram lexicon over Viterbi-aligned word pairs as in the Moses decoder (Koehn et al., 2007).

We also included the following 10 binary indicator features using statistics local to each rule:

- Three low-count features that equal 1 when the extracted frequency of the rule is exactly equal to 1, 2, or 3.
- A syntactic feature that equals 1 when the rule's label is syntactic, and a corresponding non-syntactic feature that equals 1 when the rule's label is PHR::PHR.
- Five rule format features that equal 1 when the rule's right-hand side has a certain composition. If a_s and a_t are true when the source and target sides contain only nonterminals, respectively, our rule format features are equal to a_s, a_t, $a_s \wedge \bar{a}_t$, $\bar{a}_s \wedge a_t$, and $\bar{a}_s \wedge \bar{a}_t$.

Finally, our model includes a glue rule indicator feature that equals 1 when the rule is a generic glue rule. In the Joshua decoder, glue rules monotonically stitch together adjacent parsed translation fragments at no model cost.

2.3 Language Modeling

This year, our constrained-track system made use of part of the English Gigaword data, along with other provided text, in its target-side language model. From among the data released directly for WMT 2011, we used the English side of the Europarl, news commentary, French–English UN document, and English monolingual news corpora. From the English Gigaword corpus, we included the entire Xinhua portion and the most recent 13 million sentences of the AP Wire portion. Some of these corpora contain many lines that are repeated a disproportionate number of times — the monolingual news corpus in particular, when filtered to only one occurrence of each sentence, reaches only 27% of its original line count. As part of preparing our language modeling data, we deduplicated both the English news and the UN documents, the corpora with the highest percentages of repeated sentences. We also removed lines containing more than 750 characters (about 125 average English words) before tokenization.

The final prepared corpus was made up of approximately 1.8 billion words of running text. We built a 5-gram language model from it with the SRI language modeling toolkit (Stolcke, 2002). To match the treatment given to the training data, the language model was also built in mixed case.

2.4 Grammar Filtering for Decoding

As is to be expected from a training corpus of 13.9 million sentence pairs, the grammars we extract according to the procedure of Section 2.2 are quite large: approximately 2.53 billion non-syntactic and 440 million syntactic rule instances, for a combined grammar of 1.26 billion unique rules. In preparation for tuning or decoding, we are faced with the engineering challenge of selecting a subset of the grammar that contains useful rules and fits in a reasonable amount of memory.

Before even extracting a syntactic grammar, we passed the automatically generated parse trees on the training corpus through a small tag-correction script as a pre-step. In previous experimentation, we noticed that a surprising proportion of cardinal numbers in English had been tagged with labels other than CD, their correct tag. We also found errors in labeling marks of punctuation in both English and French, when again the canonical labels are unambiguous. To fix these errors, we forcibly overwrote the labels of English tokens made up of only digits with CD, and we overwrote the labels of 25 English and 24 French marks of punctuation or other symbols with the appropriate tag as defined by the relevant treebank tagging guidelines.

After grammar extraction and combination of syntactic and non-syntactic rules, we ran an additional filtering step to reduce derivational ambiguity in the case where the same SCFG right-hand side appeared with more than one left-hand-side label. For each right-hand side, we sorted its possible labels by extracted frequency, then threw out the labels in the bottom 10% of the left-hand-side distribution.

Finally, we ran a main grammar filtering step prior to tuning or decoding, experimenting with two different filtering methods. In both cases, the phrase pair rules in the grammar were split off and filtered so that only those whose source sides completely matched the tuning or test set were retained.

The first, more naive grammar filtering method sorted all hierarchical rules by extracted frequency, then retained the most frequent 10,000 rules to join all matching phrase pair rules in the final translation grammar. This is similar to the basic grammar filtering we performed for our WMT 2010 submission. It is based on the rationale that the most frequently extracted rules in the parallel training data are likely to be the most reliably estimated and also frequently used in translating a new data set. However, it also passes through a disproportionate number of fully abstract rules — that is, rules whose right-hand sides are made up entirely of nonterminals — which can apply more recklessly on the test set because they are not lexically grounded.

Our second, more advanced method of filtering made two improvements over the naive approach. First, it controlled for the imbalance of hierarchical rules by splitting the grammar's partially lexicalized rules into a separate group that can be filtered independently. Second, it applied a lexical-match filter such that a partially lexicalized rule was retained only if all its lexicalized source phrases up to bigrams matched the intended tuning or testing set. The final translation grammar in this case was made up of three parts: all phrase pair rules matching the test set (as before), the 100,000 most frequently extracted partially lexicalized rules whose bigrams match the test set, and the 2000 most frequently extracted fully abstract rules.

3 Experimental Results and Analysis

We tuned each system variant on the newstest2008 data set, using the Z-MERT package (Zaidan, 2009) for minimum error-rate training to the BLEU metric. We ran development tests on the newstest2009 and newstest2010 data sets; Table 2 reports the results obtained according to various automatic metrics. The evaluation consists of case-insensitive scoring according to METEOR 1.0 (Lavie and Denkowski, 2009) tuned to HTER with the exact, stemming, and synonymy modules enabled, case-insensitive BLEU (Papineni et al., 2002) as implemented by the NIST `mteval-v13` script, and case-insensitive TER 0.7.25 (Snover et al., 2006).

Table 2 gives comparative results for two major systems: one based on our WMT 2011 data selection as outlined in Section 2.1, and one based on the smaller WMT 2010 training data that we used last year (8.6 million sentence pairs). Each system was run with the two grammar filtering variants described in Section 2.4: the 10,000 most frequently extracted hierarchical rules of any type ("10k"), and a combination of the 2000 most frequently extracted abstract rules and the 100,000 most frequently extracted partially lexicalized rules that matched the test set ("2k+100k"). Our primary submission to the WMT 2011 shared task was the fourth line of Table 2 ("WMT 2011 2k+100k"); we also made a constrastive submission with the system from the second line ("WMT 2010 2k+100k").

Using part of the Giga-FrEn data — along with the additions to the Europarl, news commentary, and UN document courses released since last year

	newstest2009			newstest2010		
System	**METEOR**	**BLEU**	**TER**	**METEOR**	**BLEU**	**TER**
WMT 2010 10k	54.94	24.77	56.53	56.66	25.78	55.06
WMT 2010 2k+100k	55.16	24.88	56.19	56.89	26.05	54.66
WMT 2011 10k	55.82	26.02	54.77	58.13	27.71	52.96
WMT 2011 2k+100k	55.77	26.01	54.70	57.88	27.38	53.04

Table 2: Development test results for systems based on WMT 2010 data (without the Giga-FrEn corpus) and WMT 2011 data (with some Giga-FrEn). The fourth line is our primary shared-task submission.

Applications	**10k**	**2k+100k**
Unique rules	1,305	1,994
Rule instances	14,539	12,130

Table 3: Summary of 2011 system syntactic rule applications on both test sets.

— is beneficial to translation quality, as there is a clear improvement in metric scores between the 2010 and 2011 systems. Our BLEU score improvements of 1.2 to 1.9 points are statistically significant according to the paired bootstrap resampling method (Koehn, 2004) with $n = 1000$ and $p < 0.01$. They are also larger than the 0.7- to 1.1-point gains reported by Pino et al. (2010) when the full Giga-FrEn was added. The 2011 system also shows a significant reduction in the out-of-vocabulary (OOV) rate on both test sets: 38% and 47% fewer OOV types, and 44% and 45% fewer OOV tokens, when compared to the 2010 system.

Differences between grammar filtering techniques, on the other hand, are much less significant according to all three metrics. Under paired bootstrap resampling on the newstest2009 set, the grammar variants in both the 2010 and 2011 systems are statistically equivalent according to BLEU score. On newstest2010, the 2k+100k grammar improves over the 10k version ($p < 0.01$) in the 2010 system, but the situation is reversed in the 2011 system.

We investigated differences in grammar use with an analysis of rule applications in the two variants of the 2011 system, the results of which are summarized in Table 3. Though the configuration with the 2k+100k grammar does apply syntactic rules 20% more frequently than its 10k counterpart, the 10k system uses overall 53% more unique rules. One contributing factor to this situation could be that the fully abtract rule cutoff is set too low compared to the increase in partially lexicalized rules. The effect of the 2k+100k filtering is to reduce the number of abstract rules from 4000 to 2000 while increasing the number of partially lexicalized rules from 6000 to 100,000. However, we find that the 10k system makes heavy use of some short, meaningful abstract rules that were excluded from the 2k+100k system. The 2k+100k grammar, by contrast, includes a long tail of less frequently used partially lexicalized grammar rules.

In practice, there is a balance between the use of syntactic and non-syntactic grammar rules during decoding. We highlight an example of how both types of rules work together in Figure 2, which shows our primary system's translation of part of newstest2009 sentence 2271. The French source text is given in italics and segmented into phrases. The SCFG rules used in translation are shown above each phrase, where numerical superscripts on the nonterminal labels indicate those constituents' relative ordering in the original French sentence. (Monotonic glue rules are not shown.) While non-syntactic rules can be used for short-distance reordering and fixed phrases, such as *téléphones mobiles* ↔ *mobile phones*, the model prefers syntactic translations for more complicated patterns, such as the head–children reversal in *appareils musicaux portables* ↔ *portable music devices*.

4 Conclusions and Future Work

Compared to last year, the two main differences in our current WMT submission are: (1) a new training data selection strategy aimed at increasing system vocabulary without hugely increasing corpus size, and (2) a new method of grammar filtering that emphasizes partially lexicalized rules over fully ab-

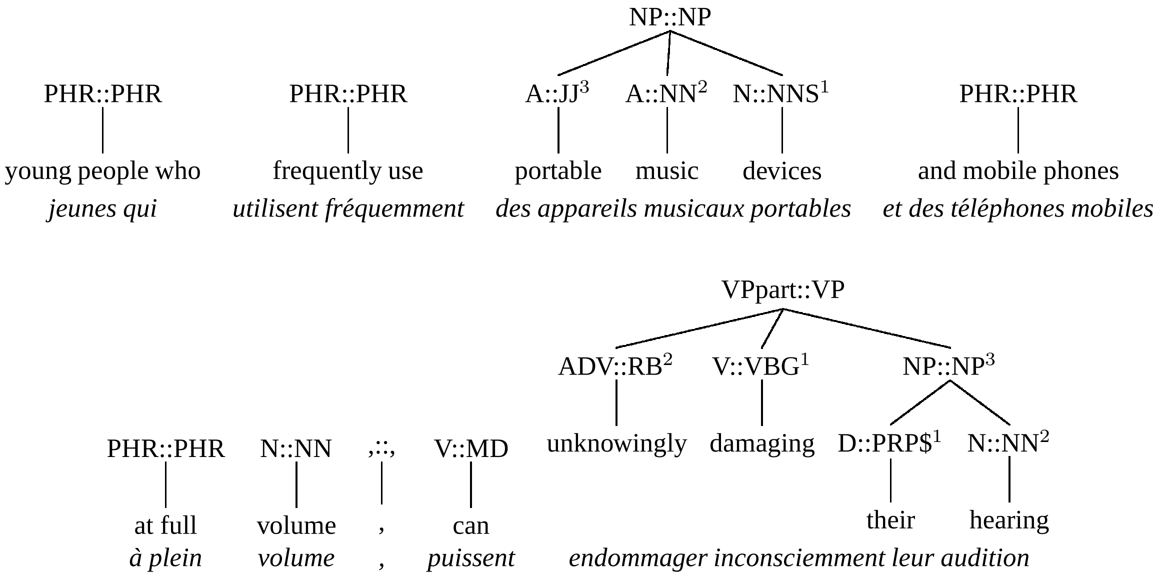

Figure 2: Our primary submission's translation of a partial sentence from the newstest2009 set, showing a combination of syntactic and non-syntactic rules.

stract ones.

Based on the results presented in Section 3, we feel confident in declaring vocabulary-based filtering of the Giga-FrEn corpus a success. By increasing the size of our parallel corpus by 26%, we more than tripled the number of unique words appearing in the source text. In conjunction with supplements to the Europarl, news commentary, and UN document corpora, this improvement led to 44% fewer OOV tokens at decoding time on two different test sets, as well as a boost in automatic metric scores of 0.6 METEOR, 1.2 BLEU, and 1.5 TER points compared to last year's system. We expect to employ similar data selection techniques when building future systems, especially as the amount of parallel data available continues to increase.

We did not, however, find significant improvements in translation quality by changing the grammar filtering method. As discussed in Section 3, limiting the grammar to only 2000 fully abstract rules may not have been enough, since additional abstract rules applied fairly frequently in test data if they were available. We plan to experiment with larger filtering cutoffs in future work. A complementary solution could be to increase the number of partially lexicalized rules. Although we found mixed results in their application within our current system, the success of Hiero-derived MT systems (Chiang, 2005; Chiang, 2010) shows that high translation quality can be achieved with rules that are only partially abstract. A major difference between such systems and our current implementation is that ours, at 102,000 rules, has a much smaller grammar.

Acknowledgments

This research was supported in part by U.S. National Science Foundation grants IIS-0713402 and IIS-0915327, as well as by the DARPA GALE program. Thanks to Kenneth Heafield for processing the English monolingual data and building the language model file, and to Jonathan Clark for LoonyBin support and bug fixes. We also thank Yahoo! for the use of the M45 research computing cluster, where we ran many steps of our experimental pipeline.

References

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 263–270, Ann Arbor, MI, June.

David Chiang. 2010. Learning to translate with source and target syntax. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1443–1452, Uppsala, Sweden, July.

Jonathan Clark and Alon Lavie. 2010. LoonyBin: Keeping language technologists sane through automated management of experimental (hyper)workflows. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 1301–1308, Valletta, Malta, May.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, OH, June.

Greg Hanneman, Jonathan Clark, and Alon Lavie. 2010. Improved features and grammar selection for syntax-based MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 82–87, Uppsala, Sweden, July.

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems 15*, pages 3–10. MIT Press, Cambridge, MA.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 48–54, Edmonton, Alberta, May–June.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of IWSLT 2005*, Pittsburgh, PA, October.

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 *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July.

Alon Lavie and Michael J. Denkowski. 2009. The METEOR metric for automatic evaluation of machine translation. *Machine Translation*, 23(2–3):105–115.

Alon Lavie, Alok Parlikar, and Vamshi Ambati. 2008. Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora. In *Proceedings of the Second ACL Workshop on Syntax and Structure in Statistical Translation*, pages 87–95, Columbus, OH, June.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren N.G. Thornton, Jonathan Weese, and Omar F. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evalution of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, July.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of NAACL HLT 2007*, pages 404–411, Rochester, NY, April.

Juan Pino, Gonzalo Iglesias, Adrià de Gispert, Graeme Blackwood, Jaime Brunning, and William Byrne. 2010. The CUED HiFST system for the WMT10 translation shared task. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 155–160, Uppsala, Sweden, July.

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 the Seventh Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA, August.

Andreas Stolcke. 2002. SRILM — an extensible language modeling toolkit. In *Proceedings of the Seventh International Conference on Spoken Language Processing*, pages 901–904, Denver, CO, September.

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.