The CMU Machine Translation Systems at WMT 2013: Syntax, Synthetic Translation Options, and Pseudo-References

Waleed Ammar Victor Chahuneau Michael Denkowski Greg Hanneman
Wang Ling Austin Matthews Kenton Murray Nicola Segall Yulia Tsvetkov
Alon Lavie Chris Dyer*
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
*Corresponding author: cdyer@cs.cmu.edu

Abstract

We describe the CMU systems submitted to the 2013 WMT shared task in machine translation. We participated in three language pairs, French–English, Russian–English, and English–Russian. Our particular innovations include: a label-coarsening scheme for syntactic tree-to-tree translation and the use of specialized modules to create "synthetic translation options" that can both generalize beyond what is directly observed in the parallel training data and use rich source language context to decide how a phrase should translate in context.

1 Introduction

The MT research group at Carnegie Mellon University's Language Technologies Institute participated in three language pairs for the 2013 Workshop on Machine Translation shared translation task: French–English, Russian–English, and English–Russian. Our French–English system (§3) showcased our group's syntactic system with coarsened nonterminal types (Hanneman and Lavie, 2011). Our Russian–English and English–Russian system demonstrate a new multi-phase approach to translation that our group is using, in which **synthetic translation options** (§4) to supplement the default translation rule inventory that is extracted from word-aligned training data. In the Russian-English system (§5), we used a CRF-based transliterator (Ammar et al., 2012) to propose transliteration candidates for out-of-vocabulary words, and used a language model to insert or remove common function words in phrases according to an n-gram English language model probability. In the English–Russian system (§6), we used a conditional logit model to predict the most likely inflectional morphology of Russian lemmas, conditioning on rich source syntactic features (§6.1). In addition to being able to generate inflected forms that were otherwise unobserved in the parallel training data, the translations options generated in this matter had features reflecting their appropriateness given much broader source language context than usually would have been incorporated in current statistical MT systems.

For our Russian–English system, we additionally used a secondary "pseudo-reference" translation when tuning the parameters of our Russian–English system. This was created by automatically translating the Spanish translation of the provided development data into English. While the output of an MT system is not always perfectly grammatical, previous work has shown that secondary machine-generated references improve translation quality when only a single human reference is available when BLEU is used as an optimization criterion (Madnani, 2010; Dyer et al., 2011).

2 Common System Components

The decoder infrastructure we used was cdec (Dyer et al., 2010). Only the constrained data resources provided for the shared task were used for training both the translation and language models. Word alignments were generated using the Model 2 variant described in Dyer et al. (2013). Language models used modified Kneser-Ney smoothing estimated using KenLM (Heafield, 2011). Translation model parameters were discriminatively set to optimize BLEU on a held-out development set using an online passive aggressive algorithm (Eidelman, 2012) or, in the case of

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 70–77,
Sofia, Bulgaria, August 8-9, 2013

the French–English system, using the hypergraph MERT algorithm and optimizing towards BLEU (Kumar et al., 2009). The remainder of the paper will focus on our primary innovations in the various system pairs.

3 French-English Syntax System

Our submission for French–English is a tree-to-tree translation system that demonstrates several innovations from group's research on SCFG-based translation.

3.1 Data Selection

We divided the French–English training data into two categories: clean data (Europarl, News Commentary, UN Documents) totaling 14.8 million sentence pairs, and web data (Common Crawl, Giga-FrEn) totaling 25.2 million sentence pairs. To reduce the volume of data used, we filtered non-parallel and other unhelpful segments according to the technique described by Denkowski et al. (2012). This procedure uses a lexical translation model learned from just the clean data, as well as source and target n-gram language models to compute the following feature scores:

- French and English 4-gram log likelihood (normalized by length);
- French–English and English–French lexical translation log likelihood (normalized by length); and,
- Fractions of aligned words under the French–English and English–French models.

We pooled previous years' WMT news test sets to form a reference data set. We computed the same features. To filter the web data, we retained only sentence for which each feature score was no lower than two standard deviations below the mean on the reference data. This reduced the web data from 25.2 million to 16.6 million sentence pairs. Parallel segments from all parts of the data that were blank on either side, were longer than 99 tokens, contained a token of more than 30 characters, or had particularly unbalanced length ratios were also removed. After filtering, 30.9 million sentence pairs remained for rule extraction: 14.4 million from the clean data, and 16.5 million from the web data.

3.2 Preprocessing and Grammar Extraction

Our French–English system uses parse trees in both the source and target languages, so tokenization in this language pair was carried out to match the tokenizations expected by the parsers we used (English data was tokenized with the Stanford tokenizer for English and an in-house tokenizer for French that targets the tokenization used by the Berkeley French parser). Both sides of the parallel training data were parsed using the Berkeley latent variable parser.

Synchronous context-free grammar rules were extracted from the corpus following the method of Hanneman et al. (2011). This decomposes each tree pair into a collection of SCFG rules by exhaustively identifying aligned subtrees to serve as rule left-hand sides and smaller aligned subtrees to be abstracted as right-hand-side nonterminals. Basic subtree alignment heuristics are similar to those by Galley et al. (2006), and composed rules are allowed. The computational complexity is held in check by a limit on the number of RHS elements (nodes and terminals), rather than a GHKM-style maximum composition depth or Hiero-style maximum rule span. Our rule extractor also allows "virtual nodes," or the insertion of new nodes in the parse tree to subdivide regions of flat structure. Virtual nodes are similar to the A+B extended categories of SAMT (Zollmann and Venugopal, 2006), but with the added constraint that they may not conflict with the surrounding tree structure.

Because the SCFG rules are labeled with nonterminals composed from both the source and target trees, the nonterminal inventory is quite large, leading to estimation difficulties. To deal with this, we automatically coarsening the nonterminal labels (Hanneman and Lavie, 2011). Labels are agglomeratively clustered based on a histogram-based similarity function that looks at what target labels correspond to a particular source label and vice versa. The number of clusters used is determined based on spikes in the distance between successive clustering iterations, or by the number of source, target, or joint labels remaining. Starting from a default grammar of 877 French, 2580 English, and 131,331 joint labels, we collapsed the label space for our WMT system down to 50 French, 54 English, and 1814 joint categories.[1]

[1]Selecting the stopping point still requires a measure of intuition. The label set size of 1814 chosen here roughly corresponds to the number of joint labels that would exist in the grammar if virtual nodes were not included. This equivalence has worked well in practice in both internal and published experiments on other data sets (Hanneman and Lavie, 2013).

Extracted rules each have 10 features associated with them. For an SCFG rule with source left-hand side ℓ_s, target left-hand side ℓ_t, source right-hand side r_s, and target right-hand side r_t, they are:

- phrasal translation log relative frequencies $\log f(r_s \mid r_t)$ and $\log f(r_t \mid r_s)$;
- labeling relative frequency $\log f(\ell_s, \ell_t \mid r_s, r_t)$ and generation relative frequency $\log f(r_s, r_t \mid \ell_s, \ell_t)$;
- lexical translation log probabilities $\log p_{lex}(r_s \mid r_t)$ and $\log p_{lex}(r_t \mid r_s)$, defined similarly to Moses's definition;
- a rarity score $\frac{\exp(\frac{1}{c})-1}{\exp(1)-1}$ for a rule with frequency c (this score is monotonically decreasing in the rule frequency); and,
- three binary indicator features that mark whether a rule is fully lexicalized, fully abstract, or a glue rule.

Grammar filtering. Even after collapsing labels, the extracted SCFGs contain an enormous number of rules — 660 million rule types from just under 4 billion extracted instances. To reduce the size of the grammar, we employ a combination of lossless filtering and lossy pruning. We first prune all rules to select no more than the 60 most frequent target-side alternatives for any source RHS, then do further filtering to produce grammars for each test sentence:

- Lexical rules are filtered to the sentence level. Only phrase pairs whose source sides match the test sentence are retained.
- Abstract rules (whose RHS are all nonterminals) are globally pruned. Only the 4000 most frequently observed rules are retained.
- Mixed rules (whose RHS are a mix of terminals and nonterminals) must match the test sentence, and there is an additional frequency cutoff.

After this filtering, the number of completely lexical rules that match a given sentence is typically low, up to a few thousand rules. Each fully abstract rule can potentially apply to every sentence; the strict pruning cutoff in use for these rules is meant to focus the grammar to the most important general syntactic divergences between French and English. Most of the latitude in grammar pruning comes from adjusting the frequency cutoff on the mixed rules since this category of rule is by far the most common type. We conducted experiments with three different frequency cutoffs: 100, 200, and 500, with each increase decreasing the grammar size by 70–80 percent.

3.3 French–English Experiments

We tuned our system to the newstest2008 set of 2051 segments. Aside from the official newstest2013 test set (3000 segments), we also collected test-set scores from last year's newstest2012 set (3003 segments). Automatic metric scores are computed according to BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011), and TER (Snover et al., 2006), all computed according to MultEval v. 0.5 (Clark et al., 2011). Each system variant is run with two independent MERT steps in order to control for optimizer instability.

Table 1 presents the results, with the metric scores averaged over both MERT runs. Quite interestingly, we find only minor differences in both tune and test scores despite the large differences in filtered/pruned grammar size as the cutoff for partially abstract rules increases. No system is fully statistically separable (at $p < 0.05$) from the others according to MultEval's approximate randomization algorithm. The closest is the variant with cutoff 200, which is generally judged to be slightly worse than the other two. METEOR claims full distinction on the 2013 test set, ranking the system with the strictest grammar cutoff (500) best. This is the version that we ultimately submitted to the shared translation task.

4 Synthetic Translation Options

Before discussing our Russian–English and English–Russian systems, we introduce the concept of **synthetic translation options**, which we use in these systems. We provide a brief overview here; for more detail, we refer the reader to Tsvetkov et al. (2013).

In language pairs that are typologically similar, words and phrases map relatively directly from source to target languages, and the standard approach to learning phrase pairs by extraction from parallel data can be very effective. However, in language pairs in which individual source language words have many different possible translations (e.g., when the target language word could have many different inflections or could be surrounded by different function words that have no

System	Dev (2008) BLEU	METR	TER	Test (2012) BLEU	METR	TER	Test (2013) BLEU	METR	TER
Cutoff 100	22.52	31.44	59.22	27.73	33.30	53.25	28.34	* 33.19	53.07
Cutoff 200	22.34	31.40	59.21	* 27.33	33.26	53.23	* 28.05	* 33.07	53.16
Cutoff 500	22.80	31.64	59.10	27.88	* 33.58	53.09	28.27	* 33.31	53.13

Table 1: French–English automatic metric scores for three grammar pruning cutoffs, averaged over two MERT runs each. Scores that are statistically separable ($p < 0.05$) from both others in the same column are marked with an asterisk (*).

direct correspondence in the source language), we can expect the standard phrasal inventory to be incomplete, except when very large quantities of parallel data are available or for very frequent words. There simply will not be enough examples from which to learn the ideal set of translation options. Therefore, since phrase based translation can only generate input/output word pairs that were directly observed in the training corpus, the decoder's only hope for producing a good output is to find a fluent, meaning-preserving translation using incomplete translation lexicons. Synthetic translation option generation seeks to fill these gaps using secondary generation processes that produce possible phrase translation alternatives that are not directly extractable from the training data. By filling in gaps in the translation options used to construct the sentential translation search space, global discriminative translation models and language models can be more effective than they would otherwise be.

From a practical perspective, synthetic translation options are attractive relative to trying to build more powerful models of translation since they enable focus on more targeted translation problems (for example, transliteration, or generating proper inflectional morphology for a single word or phrase). Since they are translation options and not complete translations, many of them may be generated.

In the following system pairs, we use synthetic translation options to augment hiero grammar rules learned in the usual way. The synthetic phrases we include augment draw from several sources:

- transliterations of OOV Russian words (§5.3);
- English target sides with varied function words (for example, given a phrase that translates into *cat* we procedure variants like *the cat*, *a cat* and *of the cat*); and,
- when translating *into* Russian, we generate phrases by first predicting the most likely Russian lemma for a source word or phrase, and then, conditioned on the English source context (including syntactic and lexical features), we predict the most likely inflection of the lemma (§6.1).

5 Russian–English System

5.1 Data

We used the same parallel data for both the Russian–English and English Russian systems. Except for filtering to remove sentence pairs whose log length ratios were statistical outliers, we only filtered the Common Crawl corpus to remove sentence pairs with less than 50% concentration of Cyrillic characters on the Russian side. The remaining data was tokenized and lower-cased. For language models, we trained 4-gram Markov models using the target side of the bitext and any available monolingual data (including Gigaword for English). Additionally, we trained 7-gram language models using 600-class Brown clusters with Witten-Bell smoothing.[2]

5.2 Baseline System

Our baseline Russian–English system is a hierarchical phrase-based translation model as implemented in cdec (Chiang, 2007; Dyer et al., 2010). SCFG translation rules that plausibly match each sentence in the development and deftest sets were extracted from the aligned parallel data using the suffix array indexing technique of Lopez (2008). A Russian morphological analyzer was used to lemmatize the training, development, and test data, and the "noisier channel" translation approach of Dyer (2007) was used in the Russian–English system to let unusually inflected surface forms back off to per-lemma translations.

[2] http://www.ark.cs.cmu.edu/cdyer/ru-600/.

5.3 Synthetic Translations: Transliteration

Analysis revealed that about one third of the unseen Russian tokens in the development set consisted of named entities which should be transliterated. We used individual Russian-English word pairs in Wikipedia parallel headlines [3] to train a linear-chained CRF tagger which labels each character in the Russian token with a sequence of zero or more English characters (Ammar et al., 2012). Since Russian names in the training set were in nominative case, we used a simple rule-based morphological generator to produce possible inflections and filtered out the ones not present in the Russian monolingual corpus. At decoding, unseen Russian tokens are fed to the transliterator which produces the most probable 20 transliterations. We add a synthetic translation option for each of the transliterations with four features: an indicator feature for transliterations, the CRF unnormalized score, the trigram character-LM log-probability, and the divergence from the average length-ratio between an English name and its Russian transliteration.

5.4 Synthetic Translations: Function Words

Slavic languages like Russian have a large number of different inflected forms for each lemma, representing different cases, tenses, and aspects. Since our training data is rather limited relative to the number of inflected forms that are possible, we use an English language model to generate a variety of common function word contexts for each content word phrase. These are added to the phrase table with a feature indicating that they were not actually observed in the training data, but rather hallucinated using SRILM's `disambig` tool.

5.5 Summary

Table 5.5 summarizes our Russian-English translation results. In the submitted system, we additionally used MBR reranking to combine the 500-best outputs of our system, with the 500-best outputs of a syntactic system constructed similarly to the French–English system.

Table 2: Russian-English summary.

Condition	BLEU
Baseline	30.8
Function words	30.9
Transliterations	31.1

6 English–Russian System

The bilingual training data was identical to the filtered data used in the previous section. Word alignments was performed after lemmatizing the Russian side of the training corpus. An unpruned, modified Kneser-Ney smoothed 4-gram language model (Chen and Goodman, 1996) was estimated from all available Russian text (410 million words) using the KenLM toolkit (Heafield et al., 2013).

A standard hierarchical phrase-based system was trained with rule shape indicator features, obtained by replacing terminals in translation rules by a generic symbol. MIRA training was performed to learn feature weights.

Additionally, word clusters (Brown et al., 1992) were obtained for the complete monolingual Russian data. Then, an unsmoothed 7-gram language model was trained on these clusters and added as a feature to the translation system. Indicator features were also added for each cluster and bigram cluster occurence. These changes resulted in an improvement of more than a BLEU point on our held-out development set.

6.1 Predicting Target Morphology

We train a classifier to predict the inflection of each Russian word independently given the corresponding English sentence and its word alignment. To do this, we first process the Russian side of the parallel training data using a statistical morphological tagger (Sharoff et al., 2008) to obtain lemmas and inflection tags for each word in context. Then, we obtain part-of-speech tags and dependency parses of the English side of the parallel data (Martins et al., 2010), as well as Brown clusters (Brown et al., 1992). We extract features capturing lexical and syntactical relationships in the source sentence and train structured linear logistic regression models to predict the tag of each English word independently given its part-of-speech.[4] In practice, due to the large size of the corpora and of the feature space dimension, we were only able to use about 10% of the available bilingual data, sampled randomly from the Common Crawl corpus. We also restricted the

[3] We contributed the data set to the shared task participants at http://www.statmt.org/wmt13/wiki-titles.ru-en.tar.gz

[4] We restrict ourselves to verbs, nouns, adjectives, adverbs and cardinals since these open-class words carry most inflection in Russian.

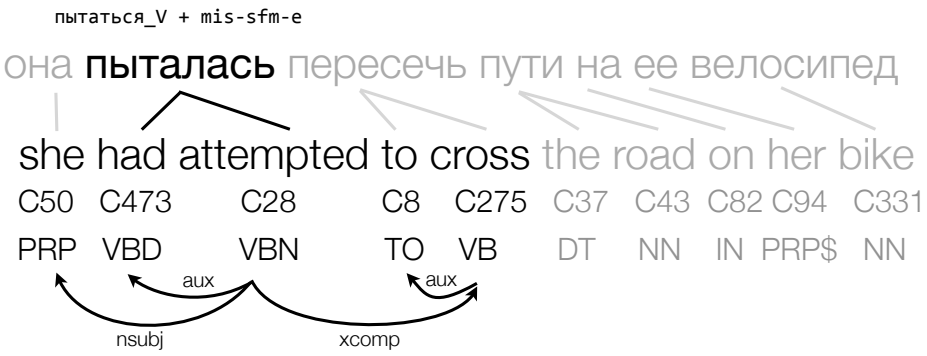

Figure 1: The classifier is trained to predict the verbal inflection `mis-sfm-e` based on the linear and syntactic context of the words aligned to the Russian word; given the stem пытаться (*pytat'sya*), this inflection paradigm produces the observed surface form пыталась (*pytalas'*).

set of possible inflections for each word to the set of tags that were observed with its lemma in the full monolingual training data. This was necessary because of our choice to use a tagger, which is not able to synthesize surface forms for a given lemma-tag pair.

We then augment the standard hierarchical phrase-base grammars extracted for the baseline systems with new rules containing inflections not necessarily observed in the parallel training data. We start by training a non-gappy phrase translation model on the bilingual data where the Russian has been lemmatized.[5] Then, before translating an English sentence, we extract translation phrases corresponding to this specific sentence and re-inflect each word in the target side of these phrases using the classifier with features extracted from the source sentence words and annotations. We keep the original phrase-based translation features and add the inflection score predicted by the classifier as well as indicator features for the part-of-speech categories of the re-inflected words.

On a held-out development set, these synthetic phrases produce a 0.3 BLEU point improvement. Interestingly, the feature weight learned for using these phrases is positive, indicating that useful inflections might be produced by this process.

7 Conclusion

The CMU systems draws on a large number of different research directions. Techniques such as MBR reranking and synthetic phrases allow different contributors to focus on different translation problems that are ultimately recombined into a single system. Our performance, in particular, on English–Russian machine translation was quite satisfying, we attribute our biggest gains in this language pair to the following:

- Our inflection model that predicted how an English word ought best be translated, given its context. This enabled us to generate forms that were not observed in the parallel data or would have been rare *independent of context* with precision.
- Brown cluster language models seem to be quite effective at modeling long-range morphological agreement patterns quite reliably.

[5] We keep intact words belonging to non-predicted categories.

Acknowledgments

We sincerely thank the organizers of the workshop for their hard work, year after year, and the reviewers for their careful reading of the submitted draft of this paper. This research work was supported in part by the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF-10-1-0533, by the National Science Foundation under grant IIS-0915327, by a NPRP grant (NPRP 09-1140-1-177) from the Qatar National Research Fund (a member of the Qatar Foundation), and by computing resources provided by the NSF-sponsored XSEDE program under grant TG-CCR110017. The statements made herein are solely the responsibility of the authors.

References

Waleed Ammar, Chris Dyer, and Noah A. Smith. 2012. Transliteration by sequence labeling with lattice encodings and reranking. In *NEWS workshop at ACL*.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computional Linguistics*, 18(4):467–479.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318, Santa Cruz, California, USA, June. Association for Computational Linguistics.

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

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Crontrolling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 176–181, Portland, Oregon, USA, June.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland, UK, July.

Michael Denkowski, Greg Hanneman, and Alon Lavie. 2012. The cmu-avenue french-english translation system. In *Proceedings of the NAACL 2012 Workshop on Statistical Machine Translation*.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan 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. of ACL*.

Chris Dyer, Kevin Gimpel, Jonathan H. Clark, and Noah A. Smith. 2011. The CMU-ARK German-English translation system. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proc. of NAACL*.

Chris Dyer. 2007. The 'noiser channel': Translation from morphologically complex languages. In *Proceedings of WMT*.

Vladimir Eidelman. 2012. Optimization strategies for online large-margin learning in machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 961–968, Sydney, Australia, July.

Greg Hanneman and Alon Lavie. 2011. Automatic category label coarsening for syntax-based machine translation. In *Proceedings of SSST-5: Fifth Workshop on Syntax, Semantics, and Structure in Statistical Translation*, pages 98–106, Portland, Oregon, USA, June.

Greg Hanneman and Alon Lavie. 2013. Improving syntax-augmented machine translation by coarsening the label set. In *Proceedings of NAACL-HLT 2013*, pages 288–297, Atlanta, Georgia, USA, June.

Greg Hanneman, Michelle Burroughs, and Alon Lavie. 2011. A general-purpose rule extractor for SCFG-based machine translation. In *Proceedings of SSST-5: Fifth Workshop on Syntax, Semantics, and Structure in Statistical Translation*, pages 135–144, Portland, Oregon, USA, June.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, UK, July.

Shankar Kumar, Wolfgang Macherey, Chris Dyer, and Franz Och. 2009. Efficient minimum error rate training and minimum Bayes-risk decoding for translation hypergraphs and lattices. In *Proc. of ACL-IJCNLP*.

Adam Lopez. 2008. Tera-scale translation models via pattern matching. In *Proc. of COLING*.

Nitin Madnani. 2010. *The Circle of Meaning: From Translation to Paraphrasing and Back*. Ph.D. thesis, Department of Computer Science, University of Maryland College Park.

André F. T. Martins, Noah A. Smith, Eric P. Xing, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. 2010. Turbo parsers: Dependency parsing by approximate variational inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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, Pennsylvania, USA, July.

Serge Sharoff, Mikhail Kopotev, Tomaz Erjavec, Anna Feldman, and Dagmar Divjak. 2008. Designing and evaluating a russian tagset. In *Proc. of LREC*.

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, Massachusetts, USA, August.

Yulia Tsvetkov, Chris Dyer, Lori Levin, and Archna Batia. 2013. Generating English determiners in phrase-based translation with synthetic translation options. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 138–141, New York, New York, USA, June.