PhraseFix: Statistical Post-Editing of TectoMT

Petra Galuščáková, Martin Popel, and Ondřej Bojar
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czech Republic
{galuscakova,popel,bojar}@ufal.mff.cuni.cz

Abstract

We present two English-to-Czech systems that took part in the WMT 2013 shared task: TectoMT and PhraseFix. The former is a deep-syntactic transfer-based system, the latter is a more-or-less standard statistical post-editing (SPE) applied on top of TectoMT. In a brief survey, we put SPE in context with other system combination techniques and evaluate SPE vs. another simple system combination technique: using synthetic parallel data from TectoMT to train a statistical MT system (SMT). We confirm that PhraseFix (SPE) improves the output of TectoMT, and we use this to analyze errors in TectoMT. However, we also show that extending data for SMT is more effective.

1 Introduction

This paper describes two submissions to the WMT 2013 shared task:[1] TectoMT – a deep-syntactic tree-to-tree system and PhraseFix – statistical post-editing of TectoMT using Moses (Koehn et al., 2007). We also report on experiments with another hybrid method where TectoMT is used to produce additional (so-called *synthetic*) parallel training data for Moses. This method was used in CU-Bojar and CU-Depfix submissions, see Bojar et al. (2013).

2 Overview of Related Work

The number of approaches to system combination is enormous. We very briefly survey those that form the basis of our work reported in this paper.

2.1 Statistical Post-Editing

Statistical post-editing (SPE, see e.g. Simard et al. (2007), Dugast et al. (2009)) is a popular method for improving outputs of a rule-based MT system. In principle, SPE could be applied to any type of *first-stage system* including a statistical one (Oflazer and El-Kahlout, 2007; Béchara et al., 2011), but most benefit could be expected from post-editing rule-based MT because of the complementary nature of weaknesses and advantages of rule-based and statistical approaches.

SPE is usually done with an off-the-shelf SMT system (e.g. Moses) which is trained on output of the first-stage system aligned with reference translations of the original source text. The goal of SPE is to produce translations that are better than both the first-stage system alone and the second-stage SMT trained on the original training data.

Most SPE approaches use the reference translations from the original training parallel corpus to train the second-stage system. In contrast, Simard et al. (2007) use human-post-edited first-stage system outputs instead. Intuitively, the latter approach achieves better results because the human-post-edited translations are closer to the first-stage output than the original reference translations. Therefore, SPE learns to perform the changes which are needed the most. However, creating human-post-edited translations is laborious and must be done again for each new (version of the) first-stage system in order to preserve its full advantage over using the original references.[2]

Rosa et al. (2013) have applied SPE on English→Czech SMT outputs. They have used the approach introduced by Béchara et al. (2011), but no improvement was achieved. However, their rule-based post-editing were found helpful.

Our SPE setting (called PhraseFix) uses TectoMT as the first-stage system and Moses as the second-stage system. Ideally, TectoMT pre-

[1] http://www.statmt.org/wmt13

[2] If more reference translations are available, it would be beneficial to choose such references for training SPE which are most similar to the first-stage outputs. However, in our experiments only one reference is available.

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 141–147,
Sofia, Bulgaria, August 8-9, 2013

serves well-formed syntactic sentence structures, and the SPE (Moses) fixes low fluency wordings.

2.2 MT Output Combination

An SPE system is trained to improve the output of a single first-stage system. Sometimes, more (first-stage) systems are available, and we would like to combine them. In *MT output selection*, for each sentence one system's translation is selected as the final output. In *MT output combination*, the final translation of each sentence is a combination of phrases from several systems. In both approaches, the systems are treated as black boxes, so only their outputs are needed. In the simplest setting, all systems are supposed to be equally good/reliable, and the final output is selected by voting, based on the number of shared n-grams or language model scores. The number and the identity of the systems to be combined therefore do not need to be known in advance. More sophisticated methods learn parameters/weights specific for the individual systems. These methods are based e.g. on confusion networks (Rosti et al., 2007; Matusov et al., 2008) and joint optimization of word alignment, word order and lexical choice (He and Toutanova, 2009).

2.3 Synthetic Data Combination

Another way to combine several first-stage systems is to employ a standard SMT toolkit, e.g. Moses. The core of the idea is to use the n first-stage systems to prepare synthetic parallel data and include them in the training data for the SMT.

Corpus Combination (CComb) The easiest method is to use these n newly created parallel corpora as additional training data, i.e. train Moses on a concatenation of the original parallel sentences (with human-translated references) and the new parallel sentences (with machine-translated pseudo-references).

Phrase Table Combination (PTComb) Another method is to extract n phrase tables in addition to the original phrase table and exploit the Moses option of multiple phrase tables (Koehn and Schroeder, 2007). This means that given the usual five features (forward/backward phrase/lexical log probability and phrase penalty), we need to tune $5 \cdot (n+1)$ features. Because such MERT (Och, 2003) tuning may be unstable for higher n, several methods were proposed where the $n+1$ phrase tables are merged into a single one (Eisele et al., 2008; Chen et al., 2009). Another issue of phrase table combination is that the same output can be achieved with phrases from several phrase tables, leading to spurious ambiguity and thus less diversity in n-best lists of a given size (see Chen et al. (2009) for one possible solution). CComb does not suffer from the spurious ambiguity issue, but it does not allow to tune special features for the individual first-stage systems.

In our experiments, we use both CComb and PTComb approaches. In PTComb, we use TECTOMT as the only first-stage system and Moses as the second-stage system. We use the two phrase tables separately (the merging is not needed; $5 \cdot 2$ is still a reasonable number of features in MERT). In CComb, we concatenate English↔Czech parallel corpus with English↔"synthetic Czech" corpus translated from English using TECTOMT. A single phrase table is created from the concatenated corpus.

3 TECTOMT

TECTOMT is a **linguistically-motivated** tree-to-tree deep-syntactic translation system with transfer based on Maximum Entropy context-sensitive translation models (Mareček et al., 2010) and Hidden Tree Markov Models (Žabokrtský and Popel, 2009). It employs some rule-based components, but the most important tasks in the analysis-transfer-synthesis pipeline are based on statistics and machine learning. There are three main reasons why it is a suitable candidate for SPE and other hybrid methods.

- TECTOMT has quite **different distribution and characteristics of errors** compared to standard SMT (Bojar et al., 2011).
- TECTOMT is **not tuned for BLEU** using MERT (its development is rather driven by human inspection of the errors although different setups are regularly evaluated with BLEU as an additional guidance).
- TECTOMT uses deep-syntactic dependency language models in the transfer phase, but it does **not use standard n-gram language models** on the surface forms because the current synthesis phase supports only 1-best output.

The version of TECTOMT submitted to WMT 2013 is almost identical to the WMT 2012 version. Only a few rule-based components (e.g. detection of surface tense of English verbs) were refined.

Corpus	Sents	Tokens	
		Czech	English
CzEng	15M	205M	236M
^{tmt}(CzEng)	15M	197M	236M
Czech Web Corpus	37M	627M	–
WMT News Crawl	25M	445M	–

Table 1: Statistics of used data.

	BLEU	1-TER
TECTOMT	14.71±0.53	35.61±0.60
PHRASEFIX	17.73±0.54	35.63±0.65
Filtering	14.68±0.50	35.47±0.57
Mark Reliable Phr.	**17.87±0.55**	35.57±0.66
Mark Identities	**17.87±0.57**	**35.85±0.68**

Table 2: Comparison of several strategies of SPE. Best results are in bold.

4 Common Experimental Setup

All our systems (including TECTOMT) were trained on the CzEng (Bojar et al., 2012) parallel corpus (development and evaluation subsets were omitted), see Table 1 for statistics. We translated the English side of CzEng with TECTOMT to obtain "synthetic Czech". This way we obtained a new parallel corpus, denoted ^{tmt}(CzEng), with English ↔ synthetic Czech sentences. Analogically, we translated the WMT 2013 test set (newstest2013) with TECTOMT and obtained ^{tmt}(newstest2013). Our baseline SMT system (Moses) trained on CzEng corpus only was then also used for WMT 2013 test set translation, and we obtained ^{smt}(newstest2013). For all MERT tuning, newstest2011 was used.

4.1 Alignment

All our parallel data were aligned with GIZA++ (Och and Ney, 2003) and symmetrized with the "grow-diag-final-and" heuristics. This applies also to the synthetic corpora ^{tmt}(CzEng), ^{tmt}(newstest2013),[3] and ^{smt}(newstest2013).

For the SPE experiments, we decided to base alignment on (genuine and synthetic Czech) lemmas, which could be acquired directly from the TECTOMT output. For the rest of the experiments, we approximated lemmas with just the first four lowercase characters of each (English and Czech) token.

4.2 Language Models

In all our experiments, we used three language models on truecased forms: News Crawl as provided by WMT organizers,[4] the Czech side of CzEng and the Articles section of the Czech Web Corpus (Spoustová and Spousta, 2012).

We used SRILM (Stolcke, 2002) with modified Kneser-Ney smoothing. We trained 5-grams on CzEng; on the other two corpora, we trained 7-grams and pruned them if the (training set) perplexity increased by less than 10^{-14} relative. The domain of the pruned corpora is similar to the test set domain, therefore we trained 7-grams on these corpora. Adding CzEng corpus can then increase the results only very slightly – training 5-grams on CzEng is therefore sufficient and more efficient.

Each of the three LMs got its weight assigned by MERT. Across the experiments, Czech Web Corpus usually gained the largest portion of weights (40±17% of the total weight assigned to language models), WMT News Crawl was the second (32±15%), and CzEng was the least useful (15±7%), perhaps due to its wide domain mixture.

5 SPE Experiments

We trained a base SPE system as described in Section 2.1 and dubbed it PHRASEFIX.

First two rows of Table 2 show that the first-stage TECTOMT system (serving here as the baseline) was significantly improved in terms of BLEU (Papineni et al., 2002) by PHRASEFIX ($p < 0.001$ according to the paired bootstrap test (Koehn, 2004)), but the difference in TER (Snover et al., 2006) is not significant.[5] The preliminary results of WMT 2013 manual evaluation show only a minor improvement: TECTOMT=0.476 vs. PHRASEFIX=0.484 (higher means better, for details on the ranking see Callison-Burch et al. (2012)).

[3] Another possibility was to adapt TECTOMT to output source-to-target word alignment, but GIZA++ was simpler to use also due to different internal tokenization in TECTOMT and our Moses pipeline.

[4] The deep-syntactic LM of TECTOMT was trained only on this News Crawl data – `http://www.statmt.org/wmt13/translation-task.html` (sets 2007–2012).

[5] The BLEU and TER results reported here slightly differ from the results shown at `http://matrix.statmt.org/matrix/systems_list/1720` because of different tokenization and normalization. It seems that statmt.org disables the `--international-tokenization` switch, so e.g. the correct Czech quotes (*„word“*) are not tokenized, hence the neighboring tokens are never counted as matching the reference (which is tokenized as " *word* ").

Despite of the improvement, PHRASEFIX's phrase table (synthetic Czech ↔ genuine Czech) still contains many wrong phrase pairs that worsen the TECTOMT output instead of improving it. They naturally arise in cases where the genuine Czech is a too loose translation (or when the English-Czech sentence pair is simply misaligned in CzEng), and the word alignment between genuine and synthetic Czech struggles.

Apart from removing such garbage phrase pairs, it would also be beneficial to have some control over the SPE. For instance, we would like to generally prefer the original output of TECTOMT except for clear errors, so only reliable phrase pairs should be used. We examine several strategies:

Phrase table filtering. We filter out all phrase pairs with forward probability ≤ 0.7 and all singleton phrase pairs. These thresholds were set based on our early experiments. Similar filtering was used by Dugast et al. (2009).

Marking of reliable phrases. This strategy is similar to the previous one, but the low-frequency phrase pairs are not filtered-out. Instead, a special feature marking these pairs is added. The subsequent MERT of the SPE system selects the best weight for this indicator feature. The frequency and probability thresholds for marking a phrase pair are the same as in the previous case.

Marking of identities A special feature indicating the equality of the source and target phrase in a phrase pair is added. In general, if the output of TECTOMT matched the reference, then such output was probably good and does not need any post-editing. These phrase pairs should be perhaps slightly preferred by the SPE.

As apparent from Table 2, marking either reliable phrases or identities is useful in our SPE setting in terms of BLEU score. In terms of TER measure, marking the identities slightly improves PHRASEFIX. However, none of the improvements is statistically significant.

6 Data Combination Experiments

We now describe experiments with phrase table and corpus combination. In the training step, the source-language monolingual corpus that serves as the basis of the synthetic parallel data can be:

- the source side of the original parallel training corpus (resulting in ^{tmt}(CzEng)),
- a huge source-language monolingual corpus for which no human translations are available (we have not finished this experiment yet),
- the source side of the test set (resulting in ^{tmt}(newstest2013) if translated by TECTOMT or ^{smt}(newstest2013) if translated by baseline configuration of Moses trained on CzEng), or
- a combination of the above.

There is a trade-off in the choice: the source side of the test set is obviously most useful for the given input, but it restricts the applicability (all systems must be installed or available online in the testing time) and speed (we must wait for the slowest system and the combination).

So far, in PTComb we tried adding the full synthetic CzEng ("CzEng + ^{tmt}(CzEng)"), adding the test set ("CzEng + ^{tmt}(newstest2013)" and "CzEng + ^{smt}(newstest2013)"), and adding both ("CzEng + ^{tmt}(CzEng) + ^{tmt}(newstest2013)"). In CComb, we concatenated CzEng and full synthetic CzEng ("CzEng + ^{tmt}(CzEng)").

There are two flavors of PTComb: either the two phrase tables are used both at once as alternative decoding paths ("Alternative"), where each source span is equipped with translation options from any of the tables, or the synthetic Czech phrase table is used only as a back-off method if a source phrase is not available in the primary table ("Back-off"). The back-off model was applied to source phrases of up to 5 tokens.

Table 3 summarizes our results with phrase table and corpus combination. We see that adding synthetic data unrelated to the test set does bring only a small benefit in terms of BLEU in the case of CComb, and we see a small improvement in TER in two cases. Adding the (synthetic) translation of the test set helps. However, adding translated source side of the test set is helpful only if it is translated by the TECTOMT system. If our baseline system is used for this translation, the results even slightly drop.

Somewhat related experiments for pivot languages by Galuščáková and Bojar (2012) showed a significant gain when the outputs of a rule-based system were added to the training data of Moses. In their case however, the genuine parallel corpus was much smaller than the synthetic data. The benefit of unrelated synthetic data seems to vanish with larger parallel data available.

Training Data for Moses	Decoding Type	BLEU	1-TER
baseline: CzEng	—	18.52±0.57	36.41±0.66
^{tmt}(CzEng)	—	15.96±0.53	33.67±0.63
CzEng + ^{tmt}(CzEng)	CComb	18.57±0.57	36.47±0.64
CzEng + ^{tmt}(CzEng)	PTComb Alternative	18.42±0.58	36.47±0.65
CzEng + ^{tmt}(CzEng)	PTComb Back-off	18.38±0.57	36.25±0.65
CzEng + ^{tmt}(newstest2013)	PTComb Alternative	18.68±0.57	37.00±0.65
CzEng + ^{smt}(newstest2013)	PTComb Alternative	18.46±0.54	36.59±0.65
CzEng + ^{tmt}(CzEng) + ^{tmt}(newstest2013)	PTComb Alternative	**18.85±0.58**	**37.03±0.66**

Table 3: Comparison of several strategies used for Synthetic Data Combination (PTComb – phrase table combination and CComb – corpus combination).

	BLEU	Judged better
SPE	17.73±0.54	123
PTComb	**18.68±0.57**	**152**

Table 4: Automatic (BLEU) and manual (number of sentences judged better than the other system) evaluation of SPE vs. PTComb.

7 Discussion

7.1 Comparison of SPE and PTComb

Assuming that our first-stage system, TECTOMT, guarantees the grammaticality of the output (sadly often not quite true), we see SPE and PTComb as two complementary methods that bring in the goods of SMT but risk breaking the grammaticality. Intuitively, SPE feels less risky, because one would hope that the post-edits affect short sequences of words and not e.g. the clause structure. With PTComb, one relies purely on the phrase-based model and its well-known limitations with respect to grammatical constraints.

Table 4 compares the two approaches empirically. For SPE, we use the default PHRASEFIX; for PTComb, we use the option "CzEng + ^{tmt}(newstest2013)". The BLEU scores are repeated.

We ran a small manual evaluation where three annotators judged which of the two outputs was better. The identity of the systems was hidden, but the annotators had access to both the source and the reference translation. Overall, we collected 333 judgments over 120 source sentences. Of the 333 judgments, 17 marked the two systems as equally correct, and 44 marked the systems as incomparably wrong. Across the remaining 275 non-tying comparisons, PTComb won – 152 vs. 123.

We attribute the better performance of PTComb to the fact that, unlike SPE, it has direct access to the source text. Also, the risk of flawed sentence structure in PTComb is probably not too bad, but this can very much depend on the language pair. English→Czech translation does not need much reordering in general.

Based on the analysis of the better marked results of the PTComb system, the biggest problem is the wrong selection of the word and word form, especially for verbs. PTComb also outperforms SPE in processing of frequent phrases and subordinate clauses. This problem could be solved by enhancing fluency in SPE or by incorporating more training data. Another possibility would be to modify TECTOMT system to produce more than one-best translation as the correct word or word form may be preserved in sequel translations.

7.2 Error Analysis of TECTOMT

While SPE seems to perform worse, it has a unique advantage: it can be used as a feedback for improving the first stage system. We can either inspect the filtered SPE phrase table or differences in translated sentences.

After submitting our WMT 2013 systems, this comparison allowed us to spot a systematic error in TECTOMT tagging of latin-origin words:

source	*pancreas*
TECTOMT	*slinivek* [plural]
PHRASEFIX	*slinivky* [singular] *břišní*

The part-of-speech tagger used in TECTOMT incorrectly detects *pancreas* as plural, and the wrong morphological number is used in the synthesis. PHRASEFIX correctly learns that the plural form *slinivek* should be changed to singular *slinivky*, which has also a higher language model score. Moreover, PHRASEFIX also learns that the trans-

lation of *pancreas* should be two words (*břišní* means *abdominal*). TECTOMT currently uses a simplifying assumption of 1-to-1 correspondence between content words, so it is not able to produce the correct translation in this case.

Another example shows where PHRASEFIX recovered from a lexical gap in TECTOMT:

source	*people who are strong-willed*
TECTOMT	*lidé , kteří jsou silná willed*
PHRASEFIX	*lidí , kteří mají silnou vůli*

TECTOMT's primary translation model considers *strong-willed* an OOV word, so a back-off dictionary specialized for hyphen compounds is used. However, this dictionary is not able to translate *willed*. PHRASEFIX corrects this and also the verb *jsou = are* (the correct Czech translation is *mají silnou vůli = have a strong will*).

Finally, PHRASEFIX can also break things:

source	*You won't be happy here*
TECTOMT	*Nebudete šťastní tady*
PHRASEFIX	*Vy tady šťastní* [you here happy]

Here, PHRASEFIX damaged the translation by omitting the negative verb *nebudete = you won't*.

8 Conclusion

Statistical post-editing (SPE) and phrase table combination (PTComb) can be seen as two complementary approaches to exploiting the mutual benefits of our deep-transfer system TECTOMT and SMT.

We have shown that SPE improves the results of TECTOMT. Several variations of SPE have been examined, and we have further improved SPE results by marking identical and reliable phrases using a special feature. However, SMT still outperforms SPE according to BLEU and TER measures. Finally, employing PTComb, we have improved the baseline SMT system by utilizing additional data translated by the TECTOMT system. A small manual evaluation suggests that PTComb is on average better than SPE, though in about one third of sentences SPE was judged better. In our future experiments, we plan to improve SPE by applying techniques suited for monolingual alignment, e.g. feature-based aligner considering word similarity (Rosa et al., 2012) or extending the parallel data with vocabulary identities to promote alignment of the same word form (Dugast et al., 2009). Marking and filtering methods for SPE also deserve a deeper study. As for PTComb, we plan to combine several sources of synthetic data (including a huge source-language monolingual corpus).

Acknowledgements

This research is supported by the grants GAUK 9209/2013, FP7-ICT-2011-7-288487 (MosesCore) of the European Union and SVV project number 267 314. We thank the two anonymous reviewers for their comments.

References

Hanna Béchara, Yanjun Ma, and Josef van Genabith. 2011. Statistical post-editing for a statistical MT system. *MT Summit XIII*, pages 308–315.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A Grain of Salt for the WMT Manual Evaluation. In *Proc. of WMT*, pages 1–11, Edinburgh, Scotland. ACL.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The Joy of Parallelism with CzEng 1.0. In *Proc. of LREC*, pages 3921–3928, Istanbul, Turkey. ELRA.

Ondřej Bojar, Rudolf Rosa, and Aleš Tamchyna. 2013. Chimera – Three Heads for English-to-Czech Translation. In *Proc. of WMT*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proc. of WMT*, Montreal, Canada. ACL.

Yu Chen, Michael Jellinghaus, Andreas Eisele, Yi Zhang, Sabine Hunsicker, Silke Theison, Christian Federmann, and Hans Uszkoreit. 2009. Combining Multi-Engine Translations with Moses. In *Proc. of WMT*, pages 42–46, Athens, Greece. ACL.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2009. Statistical Post Editing and Dictionary Extraction: Systran/Edinburgh Submissions for ACL-WMT2009. In *Proc. of WMT*, pages 110–114, Athens, Greece. ACL.

Andreas Eisele, Christian Federmann, Hervé Saint-Amand, Michael Jellinghaus, Teresa Herrmann, and Yu Chen. 2008. Using Moses to Integrate Multiple Rule-Based Machine Translation Engines into a Hybrid System. In *Proc. of WMT*, pages 179–182, Columbus, Ohio. ACL.

Petra Galuščáková and Ondřej Bojar. 2012. Improving SMT by Using Parallel Data of a Closely Related Language. In *Proc. of HLT*, pages 58–65, Amsterdam, Netherlands. IOS Press.

Xiaodong He and Kristina Toutanova. 2009. Joint Optimization for Machine Translation System Combination. In *Proc. of EMNLP*, pages 1202–1211, Singapore. ACL.

Philipp Koehn and Josh Schroeder. 2007. Experiments in Domain Adaptation for Statistical Machine Translation. In *Proc. of WMT*, pages 224–227, Prague, Czech Republic. ACL.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of ACL*, pages 177–180, Prague, Czech Republic. ACL.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proc. of EMNLP*, Barcelona, Spain.

David Mareček, Martin Popel, and Zdeněk Žabokrtský. 2010. Maximum entropy translation model in dependency-based MT framework. In *Proc. of MATR*, pages 201–206. ACL.

Evgeny Matusov, Gregor Leusch, Rafael E. Banchs, Nicola Bertoldi, Daniel Dechelotte, Marcello Federico, Muntsin Kolss, Young-Suk Lee, Jose B. Marino, Matthias Paulik, Salim Roukos, Holger Schwenk, and Hermann Ney. 2008. System Combination for Machine Translation of Spoken and Written Language. *IEEE*, 16(7):1222–1237.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of ACL*, Sapporo, Japan.

Kemal Oflazer and Ilknur Durgar El-Kahlout. 2007. Exploring different representational units in English-to-Turkish statistical machine translation. In *Proc. of WMT*, pages 25–32. ACL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318, Stroudsburg, PA, USA. ACL.

Rudolf Rosa, Ondřej Dušek, David Mareček, and Martin Popel. 2012. Using Parallel Features in Parsing of Machine-Translated Sentences for Correction of Grammatical Errors. In *Proc. of SSST*, pages 39–48, Jeju, Republic of Korea. ACL.

Rudolf Rosa, David Mareček, and Aleš Tamchyna. 2013. Deepfix: Statistical Post-editing of Statistical Machine Translation Using Deep Syntactic Analysis. Sofia, Bulgaria. ACL.

Antti-Veikko Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie Dorr. 2007. Combining Outputs from Multiple Machine Translation Systems. In *Proc. of NAACL*, pages 228–235, Rochester, New York. ACL.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical Phrase-Based Post-Editing. In *Proc. of NAACL*, pages 508–515, Rochester, New York. ACL.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of Association for Machine Translation in the Americas*, pages 223–231.

Johanka Spoustová and Miroslav Spousta. 2012. A High-Quality Web Corpus of Czech. In *Proc. of LREC*, Istanbul, Turkey. ELRA.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. of ICSLP*, pages 257–286.

Zdeněk Žabokrtský and Martin Popel. 2009. Hidden Markov Tree Model in Dependency-based Machine Translation. In *Proc. of IJCNLP*, pages 145–148, Suntec, Singapore.