DEPFIX: A System for Automatic Correction of Czech MT Outputs*

Rudolf Rosa, David Mareček and Ondřej Dušek
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague
{rosa,marecek,odusek}@ufal.mff.cuni.cz

Abstract

We present an improved version of DEPFIX (Mareček et al., 2011), a system for automatic rule-based post-processing of English-to-Czech MT outputs designed to increase their fluency. We enhanced the rule set used by the original DEPFIX system and measured the performance of the individual rules.

We also modified the dependency parser of McDonald et al. (2005) in two ways to adjust it for the parsing of MT outputs. We show that our system is able to improve the quality of the state-of-the-art MT systems.

1 Introduction

The today's outputs of Machine Translation (MT) often contain serious grammatical errors. This is particularly apparent in statistical MT systems (SMT), which do not employ structural linguistic rules. These systems have been dominating the area in the recent years (Callison-Burch et al., 2011). Such errors make the translated text less fluent and may even lead to unintelligibility or misleading statements. The problem is more evident in languages with rich morphology, such as Czech, where morphological agreement is of a relatively high importance for the interpretation of syntactic relations.

The DEPFIX system (Mareček et al., 2011) attempts to correct some of the frequent SMT systems' errors in English-to-Czech translations.[1] It analyzes the *target* sentence (the SMT output in Czech language) using a morphological tagger and a dependency parser and attempts to correct it by applying several rules which enforce consistency with the Czech grammar. Most of the rules use the *source* sentence (the SMT input in English language) as a source of information about the sentence structure. The *source* sentence is also tagged and parsed, and word-to-word alignment with the *target* sentence is determined.

In this paper, we present DEPFIX 2012, an improved version of the original DEPFIX 2011 system. It makes use of a new parser, described briefly in Section 3, which is adapted to handle the generally ungrammatical *target* sentences better. We have also enhanced the set of grammar correction rules, for which we give a detailed description in Section 4. Section 5 gives an account of the experiments performed to evaluate the DEPFIX 2012 system and compare it to DEPFIX 2011. Section 6 then concludes the paper.

2 Related Work

Our approach can be regarded as converse to the more common way of using an SMT system to automatically post-edit the output of a rule-based translation system, as described e.g. in (Simard et al., 2007) or (Lagarda et al., 2009).

The DEPFIX system is implemented in the

*This research has been supported by the European Union Seventh Framework Programme (FP7) under grant agreement n° 247762 (Faust), and by the grants GAUK116310, GA201/09/H057 (Res-Informatica), and LH12093.

[1] Although we apply the DEPFIX system just to SMT systems in this paper as it mainly targets the errors induced by this type of MT systems, it can be applied to virtually any MT system (Mareček et al., 2011).

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 362–368,
Montréal, Canada, June 7-8, 2012.

TectoMT/Treex NLP framework (Popel and Žabokrtský, 2010),[2] using the Morče tagger (Spoustová et al., 2007) and the MST parser (McDonald et al., 2005) trained on the CoNLL 2007 Shared Task English data (Nivre et al., 2007) to analyze the *source* sentences. The *source* and *target* sentences are aligned using GIZA++ (Och and Ney, 2003).

3 Parsing

The DEPFIX 2011 system used the MST parser (McDonald et al., 2005) with an improved feature set for Czech (Novák and Žabokrtský, 2007) trained on the Prague Dependency Treebank (PDT) 2.0 (Hajič and others, 2006) to analyze the *target* sentences. DEPFIX 2012 uses a reimplementation of the MST parser capable of utilizing parallel features from the *source* side in the parsing of the *target* sentence.

The *source* text is usually grammatical and therefore is likely to be analyzed more reliably. The *source* structure obtained in this way can then provide hints for the *target* parser. We use local features projected through the GIZA++ word alignment – i.e. for each *target* word, we add features computed over its aligned *source* word, if there is one.

To address the differences between the gold standard training data and SMT outputs, we "worsen" the treebank used to train the parser, i.e. introduce errors similar to those found in *target* sentences: The trees retain their correct structure, only the word forms are modified to resemble SMT output.

We have computed a "part-of-speech tag error model" on parallel sentences from the Prague Czech-English Dependency Treebank (PCEDT) 2.0 (Bojar et al., 2012), comparing the gold standard Czech translations to the output of an SMT system (Koehn et al., 2007) and estimating the Maximum Likelihood probabilities of errors for each part-of-speech tag. We then applied this error model to the Czech PCEDT 2.0 sentences and used the resulting "worsened" treebank to train the parser.

4 Rules

DEPFIX 2012 uses 20 hand-written rules, addressing various frequent errors in MT output. Each rule takes an analyzed *target* sentence as its input, often together with its analyzed *source* sentence, and attempts to correct any errors found – usually by changing morphosyntactic categories of a word (such as number, gender, case, person and dependency label) and regenerating the corresponding word form if necessary, more rarely by deleting superfluous particles or auxiliary words or changing the *target* dependency tree structure. However, neither word order problems nor bad lexical choices are corrected.

Many rules were already present in DEPFIX 2011. However, most were modified in DEPFIX 2012 to achieve better performance (denoted as ***modified***), and new rules were added (***new***). Rules not modified since DEPFIX 2011 are denoted as ***reused***.

The order of rule application is important as there are dependencies among the rules, e.g. FixPrepositionNounAgreement (enforcing noun-preposition congruency) depends on FixPrepositionalCase (fixing incorrectly tagged prepositional case). The rules are applied in the order listed in Table 2.

4.1 Analysis Fixing Rules

Analysis fixing rules try to detect and rectify tagger and parser errors. They do not change word forms and are therefore invisible on the output as such; however, rules of other types benefit from their corrections.

FixPrepositionalCase *(new)*

This rule corrects part-of-speech-tag errors in prepositional phrases. It looks for all words that depend on a preposition and do not match its part-of-speech tag case. It tries to find and assign a common morphological case fitting for both the word form and the preposition. Infrequent preposition-case combinations are not considered.

FixReflexiveTantum *(new)*

If the word form 'se' or 'si' is classified as reflexive tantum particle by the parser, but does not belong to an actual reflexive tantum verb (or a deverbative noun or an adjective), its dependency label is changed to a different value, based on the context.

FixNounNumber *(reused)*

If a noun is tagged as singular in *target* but as plural in *source*, the tag is likely to be incorrect. This rule tries to find a tag that would match both the

[2] http://ufal.mff.cuni.cz/treex

source number and the *target* word form, changing the *target* case if necessary.

FixPrepositionWithoutChildren *(reused)*

A *target* preposition with no child nodes is clearly an analysis error. This rule tries to find children for childless prepositions by projecting the children of the aligned *source* preposition to the *target* side.

FixAuxVChildren *(new)*

Since auxiliary verbs must not have child nodes, we rehang all their children to the governing full verb.

4.2 Agreement Fixing Rules

These rules relate to morphological agreement required by Czech grammar, which they try to enforce in case it is violated. Czech grammar requires agreement in morphological gender, number, case and person where applicable.

These rules typically use the *source* sentence only for confirmation.

FixRelativePronoun *(new)*

The Czech word relative pronoun 'který' is assigned gender and number identical to the closest preceding noun or pronoun, if the *source* analysis confirms that it depends on this noun/pronoun.

FixSubject *(modified)*

The subject (if the subject dependency label is confirmed by the *source* analysis) will have its case set to nominative; the number is changed if this leads to the word form staying unchanged.

FixVerbAuxBeAgreement *(modified)*

If an auxiliary verb is a child of an infinitive, the auxiliary verb receives the gender and number of the subject, which is a child of the infinitive (see also FixAuxVChildren).

FixSubjectPredicateAgreement *(modified)*

An active verb form receives the number and person from its subject (whose relation to the verb must be confirmed by the *source*).

FixSubjectPastParticipleAgreement *(modified)*

A past participle verb form receives the number and gender from its subject (confirmed by the *source* analysis).

FixPassiveAuxBeAgreement *(modified)*

An auxiliary verb 'být' ('to be') depending on a passive verb form receives its gender and number.

FixPrepositionNounAgreement *(modified)*

A noun or adjective depending on a preposition receives its case. The dependency must be confirmed in the *source*.

FixNounAdjectiveAgreement *(modified)*

An adjective (or an adjective-like pronoun or numeral) preceding its governing noun receives its gender, number and case.

4.3 Translation Fixing Rules

The following rules detect and correct structures often mistranslated by SMT systems. They usually depend heavily on the *source* sentence.

FixBy *(new)*

English preposition 'by' is translated to Czech using the instrumental case (if modifying a verb, e.g. 'built by David': 'postaveno Davidem') or using the genitive case (if modifying a noun, e.g. 'songs by David': 'písně Davida').

FixPresentContinuous *(modified)*

If the *source* sentence is in a continuous tense (e.g. 'Ondřej isn't experimenting.'), the auxiliary verb 'to be' must not appear on the output, which is often the case (e.g. *'Ondřej není experimentovat.'). This rule deletes the auxiliary verb in *target* and transfers its morphological categories to the main verb (e.g. 'Ondřej neexperimentuje.').

FixVerbByEnSubject *(new)*

If the subject of the *source* sentence is a personal pronoun, its following morphological categeries are propagated to the *target* predicate:

- person
- number (except for 'you', which does not exhibit number)
- gender (only in case of 'he' or 'she', which exhibit the natural gender)

FixOf *(new)*

English preposition 'of' modifying a noun is translated to Czech using the genitive case (e.g. 'pictures of Rudolf': 'obrázky Rudolfa').

FixAuxT *(reused)*

Reflexive tantum particles 'se' or 'si' not belonging to any verb or adjective are deleted. This situation usually occurs when the meaning of the *source* verb/adjective is lost in translation and only the particle is produced.

4.4 Other Rules

VocalizePrepos *(reused)*

Prepositions 'k', 's', 'v', 'z' are *vocalized* (i.e. changed to 'ke', 'se', 've', 'ze') where necessary. The vocalization rules in Czech are similar to 'a'/'an' distinction in English.

FixFirstWordCapitalization *(new)*

If the first word of *source* is capitalized and the first word of *target* is not, this rule capitalizes it.

5 Experiments and Results

For parameter tuning, we used datasets from the WMT10 translation task and translations by ONLINEB and CU-BOJAR systems.

5.1 Manual Evaluation

Manual evaluation of both DEPFIX 2011 and DEPFIX 2012 was performed on the WMT11[3] test set translated by ONLINEB. 500 sentences were randomly selected and blind-evaluated by two independent annotators, who were presented with outputs of ONLINEB, DEPFIX 2011 and DEPFIX 2012. (For 246 sentences, at least one of the DEPFIX setups modified the ONLINEB translation.) They provided us with a pairwise comparison of the three setups, with the possibility to mark the sentence as "indefinite" if translations were of equal quality. The results are given in Table 1.

In Table 2, we use the manual evaluation to measure the performance of the individual rules in DEPFIX 2012. For each rule, we ran DEPFIX 2012 with this rule disabled and compared the output to the output of the full DEPFIX 2012. The number of affected sentences on the whole WMT11 test set, given as "changed", represents the impact of the rule. The number of affected sentences selected for manual evaluation is listed as "evaluated". Finally, the annotators' ratings of the "evaluated" sentences (suggesting whether the rule improved or worsened the translation, or whether the result was indefinite) were counted and divided by the number of annotators to get the average performance of each rule. Please note that the lower the "evaluated" number, the lower the confidence of the results.

A / B	Setup 1 better	Setup 2 better	Indefinite
Setup 1 better	55%	1%	11%
Setup 2 better	1%	8%	4%
Indefinite	3%	2%	15%

Table 3: Inter-annotator agreement matrix for ONLINEB + DEPFIX 2012 as Setup 1 and ONLINEB as Setup 2.

The inter-annotator agreement matrix for comparison of ONLINEB + DEPFIX 2012 (denoted as Setup 1) with ONLINEB (Setup 2) is given in Table 3. The results for the other two setup pairs were similar, with the average inter-annotator agreement being 77%.

5.2 Automatic Evaluation

We also performed several experiments with automatic evaluation using the standard BLEU metric (Papineni et al., 2002). As the effect of DEPFIX in terms of BLEU is rather small, the results are not as confident as the results of manual evaluation.[4]

In Table 4, we compare the DEPFIX 2011 and DEPFIX 2012 systems and measure the contribution of parser adaptation (Section 3) and rule improvements (Section 4). It can be seen that the combined effect of applying both system modifications is greater than when they are applied alone. The improvement of DEPFIX 2012 over ONLINEB without DEPFIX is statistically significant at 95% confidence level.

The effect of DEPFIX 2012 on the outputs of some of the best-scoring SMT systems in the WMT12 Translation Task[5] is shown in Table 5. Although DEPFIX 2012 was tuned only on ONLINEB and CU-BOJAR system outputs, it improves the BLEU score of all the best-scoring systems, which suggests that

[3] http://www.statmt.org/wmt11

[4] As already noted by Mareček et al. (2011), BLEU seems not to be very suitable for evaluation of DEPFIX. See (Kos and Bojar, 2009) for a detailed study of BLEU performance when applied to evaluation of MT systems with Czech as the target language.

[5] http://www.statmt.org/wmt12

Setup 1	Setup 2	Differing sentences	Annotator	Setup 1 better	Setup 2 better	Indefinite
ONLINEB + DEPFIX 2011	ONLINEB	169	A	58%	13%	29%
			B	47%	11%	42%
ONLINEB + DEPFIX 2012	ONLINEB	234	A	65%	14%	21%
			B	59%	11%	30%
ONLINEB + DEPFIX 2012	ONLINEB + DEPFIX 2011	148	A	54%	24%	22%
			B	56%	22%	22%

Table 1: Manual pairwise comparison on 500 sentences from WMT11 test set processed by ONLINEB, ONLINEB + DEPFIX 2011 and ONLINEB + DEPFIX 2012. Evaluated by two independent annotators.

	Sentences							
Rule	changed	evaluated	impr.	%	wors.	%	indef.	%
FixPrepositionalCase	34	5	3	60	2	40	0	0
FixReflexiveTantum	1	0	–	–	–	–	–	–
FixNounNumber	80	11	5	45	5	45	1	9
FixPrepositionWithoutChildren	16	6	3	50	3	50	0	0
FixBy	75	13	10.5	81	1	8	1.5	12
FixAuxVChildren	26	6	4.5	75	0	0	1.5	25
FixRelativePronoun	56	8	6	75	2	25	0	0
FixSubject	142	18	13.5	75	3	17	1.5	8
FixVerbAuxBeAgreement	8	2	1	50	1	50	0	0
FixPresentContinuous	30	7	5.5	79	1	14	0.5	7
FixSubjectPredicateAgreement	87	10	5.5	55	1	10	3.5	35
FixSubjectPastParticipleAgreement	396	63	46.5	74	9.5	15	7	11
FixVerbByEnSubject	25	6	5	83	0	0	1	17
FixPassiveAuxBeAgreement	43	8	6	75	0.5	6	1.5	19
FixPrepositionNounAgreement	388	62	40	65	13	21	9	15
FixOf	84	13	11.5	88	0	0	1.5	12
FixNounAdjectiveAgreement	575	108	69.5	64	20	19	18.5	17
FixAuxT	38	7	4	57	1	14	2	29
VocalizePrepos	53	12	6	50	2.5	21	3.5	29
FixFirstWordCapitalization	0	0	–	–	–	–	–	–

Table 2: Impact and accuracy of individual DEPFIX 2012 rules using manual evaluation on 500 sentences from WMT11 test set translated by ONLINEB. The number of changed sentences is counted on the whole WMT11 test set, i.e. 3003 sentences. The numbers of improved, worsened and indefinite translations are averaged over the annotators.

DEPFIX setup	BLEU
without DEPFIX	19.37
DEPFIX 2011	19.41
DEPFIX 2011 + new parser	19.42
DEPFIX 2011 + new rules	19.48
DEPFIX 2012	19.56

Table 4: Performance of ONLINEB and various DEPFIX setups on the WMT11 test set.

System	BLEU
ONLINEB	16.25
ONLINEB + DEPFIX 2012	16.31
UEDIN	15.54
UEDIN + DEPFIX 2012	15.75
CU-BOJAR	15.41
CU-BOJAR + DEPFIX 2012	15.45
CU-TAMCH-BOJ	15.35
CU-TAMCH-BOJ + DEPFIX 2012	15.39

Table 5: Comparison of BLEU of baseline system output and corrected system output on WMT12 test set.

it is able to improve the quality of various SMT systems when applied to their outputs. (The improvement on UEDIN is statistically significant at 95% confidence level.) We submitted the ONLINEB + DEPFIX 2012 system to the WMT12 Translation Task as CU-DEPFIX.

6 Conclusion

We have presented two improvements to DEPFIX, a system of rule-based post-editing of English-to-Czech Machine Translation outputs proven by manual and automatic evaluation to improve the quality of the translations produced by state-of-the-art SMT systems. First, improvements in the existing rules and implementation of new ones, which can be regarded as an additive, evolutionary change. Second, a modified dependency parser, adjusted to parsing of SMT outputs by training it on a parallel treebank with worsened word forms on the Czech side. We showed that both changes led to a better performance of the new DEPFIX 2012, both individually and combined.

In future, we are planning to incorporate deeper analysis, devising rules that would operate on the deep-syntactic, or *tectogrammatical*, layer. The Czech and English tectogrammatical trees are more similar to each other, which should enable us to exploit more information from the *source* sentences. We also hope to be able to perform more complex corrections, such as changing the part of speech of a word when necessary.

Following the success of our modified parser, we would also like to modify the tagger in a similar way, since incorrect analyses produced by the tagger often hinder the correct function of our rules, sometimes leading to a rule worsening the translation instead of improving it.

As observed e.g. by Groves and Schmidtke (2009) for English-to-German and English-to-French translations, SMT systems for other language pairs also tend to produce reoccurring grammatical errors. We believe that these could be easily detected and corrected in a rule-based way, using an approach similar to ours.

References

Ondřej Bojar, Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of LREC 2012*, Istanbul, Turkey, May. ELRA, European Language Resources Association. In print.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July. Association for Computational Linguistics.

Declan Groves and Dag Schmidtke. 2009. Identification and analysis of post-editing patterns for MT. *Proceedings of MT Summit XII*, pages 429–436.

Jan Hajič et al. 2006. Prague Dependency Treebank 2.0. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T0 1, Philadelphia.

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 *ACL 2007, Proceedings of the 45th Annual Meeting of the*

Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Kamil Kos and Ondřej Bojar. 2009. Evaluation of machine translation metrics for czech as the target language. *The Prague Bulletin of Mathematical Linguistics*, 92(-1):135–148.

Antonio L. Lagarda, Vicent Alabau, Francisco Casacuberta, Roberto Silva, and Enrique Diaz-de Liano. 2009. Statistical post-editing of a rule-based machine translation system. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 217–220. Association for Computational Linguistics.

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. 2011. Two-step translation with grammatical post-processing. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 426–432. Association for Computational Linguistics.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL 2007 Shared Task. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, June.

Václav Novák and Zdeněk Žabokrtský. 2007. Feature engineering in maximum spanning tree dependency parser. In Václav Matoušek and Pavel Mautner, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 10th I nternational Conference on Text, Speech and Dialogue*, Lecture Notes in Computer Science, pages 92–98, Pilsen, Czech Republic. Springer Science+Business Media Deutschland GmbH.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: modular NLP framework. In *Proceedings of the 7th international conference on Advances in natural language processing*, IceTAL'10, pages 293–304, Berlin, Heidelberg. Springer-Verlag.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 508–515, Rochester, New York, April. Association for Computational Linguistics.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The best of two worlds: Cooperation of statistical and rule-based taggers for czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Praha.