Two-step translation with grammatical post-processing*

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

Abstract

This paper describes an experiment in which we try to automatically correct mistakes in grammatical agreement in English to Czech MT outputs. We perform several rule-based corrections on sentences parsed to dependency trees. We prove that it is possible to improve the MT quality of majority of the systems participating in WMT shared task. We made both automatic (BLEU) and manual evaluations.

1 Introduction

This paper is a joint report on two English-to-Czech submissions to the WMT11 shared translation task. The main contribution is however the proposal and evaluation of a rule-based post-processing system DEPFIX aimed at correcting errors in Czech grammar applicable to any MT system. This is somewhat the converse of other approaches (e.g. Simard et al. (2007)) where a statistical system was applied for the post-processing of a rule-based one.

2 Our phrase-based systems

This section briefly describes our underlying phrase-based systems. One of them (CU-BOJAR) was submitted directly to the WMT11 manual evaluation, the other one (CU-TWOSTEP) was first corrected by the proposed method (Section 3 below) and then submitted under the name CU-MARECEK.

*This research has been supported by the European Union Seventh Framework Programme (FP7) under grant agreement n° 247762 (Faust), n° 231720 (EuroMatrix Plus), and by the grants GAUK 116310 and GA201/09/H057.

2.1 Data for statistical systems

Our training parallel data consists of CzEng 0.9 (Bojar and Žabokrtský, 2009), the News Commentary corpus v.6 as released by the WMT11 organizers, the EMEA corpus, a corpus collected from the transcripts of TED talks (http://www.ted.com), the parallel news and separately some of the parallel web pages of the European Commission (http://ec.europa.eu), and the Official Journal of the European Union as released by the Apertium consortium (http://apertium.eu/data).

A custom web crawler was used for the European Commission website. English and Czech websites were matched according to their URLs. Unfortunately, Czech websites very often contain untranslated parts of English texts. Because of this, we aimed especially at the news articles, which are very often translated correctly and also more relevant for the shared task. Texts were segmented using trainable tokenizer (Klyueva and Bojar, 2008) and deduplicated. Processed texts were automatically aligned by Hunalign (Varga and others, 2005).

The data from the Official Journal were first converted from XML to plain text. The documents were paired according to their filenames. To better handle the nature of these data, we decided to divide the documents into two classes based on the average number of words per sentence: "lists" are documents with less than 2.8 words per sentence, other documents are called "texts". The corresponding "lists" were aligned line by line. The corresponding "texts" were automatically segmented by trainable tokenizer and aligned automatically by Hunalign.

We use the following two Czech language mod-

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 426–432,
Edinburgh, Scotland, UK, July 30–31, 2011.

els, their weights are optimized in MERT:

- 5-gram LM from the Czech side of CzEng (excluding the Navajo section). The LM was constructed by interpolating LMs of the individual domains (news, EU legislation, technical documentation, etc.) to achieve the lowest perplexity on the WMT08 news test set.
- 6-gram LM from the monolingual data supplied by WMT11 organizers (news of the individual years and News Commentary), the Czech National Corpus and a web collection of Czech texts. Again, the final LM is constructed by interpolating the smaller LMs[1] for the WMT08 news test set.

2.2 Baseline Moses (CU-BOJAR)

The system denoted CU-BOJAR for English-to-Czech is simple phrase-based translation, i.e. Moses without factors. We tokenized, lemmatized and tagged all texts using the tools wrapped in TectoMT (Popel and Žabokrtský, 2010). We further tokenize e.g. dashed words ("23-year") after all the processing is finished. Phrase-based MT is then able to handle such expressions both at once, or decompose them as needed to cover unseen variations. We use lexicalized reordering (orientation-bidirectional-fe). The translation runs in "supervised truecase", which means that we use the output of our lemmatizers to decide whether the word should be lowercased or should preserve uppercasing. After the translation, the first letter in the output is simply uppercased. The model is optimized using Moses' standard MERT on the WMT09 test set.

The organizers of WMT11 encouraged participants to apply simple normalization to their data (both for training and testing).[2] The main purpose of the normalization is to improve the consistency of typographical rules. Unfortunately, some of the automatic changes may accidentally damage the meaning of the expression.[3] We therefore opted to submit the output based on *non-normalized* test sets as our primary English-to-Czech submission.

We invested much less effort into the submission called CU-BOJAR for Czech-to-English. The only interesting feature there is the use of alternative decoding paths to translate either from the Czech form or from the Czech lemma equipped with meaning-bearing morphological properties, e.g. the number of nouns. Bojar and Kos (2010) used the same setup with simple lemmas in the fallback decoding path. The enriched lemmas perform marginally better.

2.3 Two-step translation

Our two-step translation is essentially the same setup as detailed by Bojar and Kos (2010): (1) the English source is translated to simplified Czech, and (2) the simplified Czech is monotonically translated to fully inflected Czech. Both steps are simple phrase-based models. Instead of word forms, the simplified Czech uses lemmas enriched by a subset of morphological features selected manually to encode only properties overt both in English and Czech such as the tense of verbs or number of nouns. Czech-specific morphological properties indicating various agreements (e.g. number and gender of adjectives, gender of verbs) are imposed in the second step solely on the basis of the language model.

The first step uses the same parallel and monolingual corpora as CU-BOJAR, except the LMs being trained on the enriched lemmas, not on word forms. The second step uses exactly the same LM as CU-BOJAR but the phrase-table is extracted from all our Czech monolingual data (phrase length limit of 1.)

3 Grammatical post-processing

Phrase-based machine translation systems often have problems with grammatical agreement, especially on longer dependencies. Sometimes, there is a mistake in agreement even between adjacent words because each one belongs to a different phrase. The goal of our post-processing is to correct forms of some words so that they do not violate grammatical rules (eg. grammatical agreement).

The problem is how to find the correct syntactic relations in the output of an MT system. Parsers trained on correct sentences can rely on grammatical agreement, according to which they determine

[1] The interpolated LM file (gzipped ARPA format) is 5.1 GB so we applied LM pruning as implemented in SRI toolkit with the threshold 10^{-14} to reduce the file size to 2.3 GB.

[2] http://www.statmt.org/wmt11/normalize-punctuation.perl

[3] Fixing the ordering of the full stop and the quote is wrong because the order (at least in Czech typesetting) depends on whether it is the full sentence or a final phrase that is captured in the quotes. Even riskier are rules handling decimal and thousand separators in numbers. While there are language-specific conventions, they are not always followed and the normalization can in such cases confuse the order of magnitude by 3.

the dependencies between words. Unfortunately, the agreement in MT outputs is often wrong and the parser fails to produce a correct parse tree. Therefore, we would need a parser trained on a manually annotated treebank consisting of specific outputs of machine translation systems. Such a treebank does not exist and we do not even want to create one, because the MT systems are changing constantly and also because manual annotation of texts that are often not even understandable would be almost a superhuman task.

The DEPFIX system was implemented in TectoMT framework (Popel and Žabokrtský, 2010). MT outputs were tagged by Morče tagger (Spoustová et al., 2007) and then parsed with MST parser (McDonald et al., 2005) that was trained on the Prague Dependency Treebank (Hajič and others, 2006), i.e. on correct Czech sentences. We used an improved implementation with some additional features especially tuned for Czech (Novák and Žabokrtský, 2007). The parser accuracy is much lower on the "noisy" MT output sentences, but a lot of dependencies in which we are to correct grammatical agreement are determined correctly. Adapting the parser for outputs of MT systems will be addressed in the coming months.

A typical example of a correction is the agreement between the subject and the predicate: they should share the morphological number and gender. If they do not, we simply change the number and gender of the predicate in agreement with the subject.[4] An example of such a changed predicate is in Figure 1.

Apart from the dependency tree of the target sentence, we can also use the dependency tree of the source sentence. Source sentences are grammatically correct and the accuracy of the tagger and the parser is accordingly higher there. Words in the source and target sentences are aligned using GIZA++[5] (Och and Ney, 2003) but verbose outputs of the original MT systems would be possibly a better option. The rules for fixing grammatical agreement between words can thus consider also the dependency relations and morphological caregories of their English counterparts in the input sentence.

[4]In this case, we suppose that the number of the subject has a much higher chance to be correct.

[5]GIZA++ was run on lemmatized texts in both directions and intersection symmetrization was used.

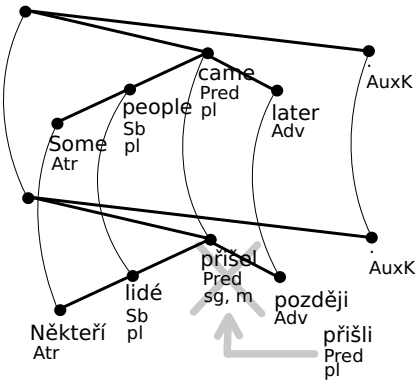

Figure 1: Example of fixing subject-predicate agreement. The Czech word *přišel [he came]* has a wrong morphological number and gender.

3.1 Grammatical rules

We have manually devised a set of the following rules. Their input is the dependency tree of a Czech sentence (MT output) and its English source sentence (MT input) with the nodes aligned where possible. Each of the rules fires if the specified conditions ("IF") are matched, executes the command ("DO") , usually changing one or more morphological categories of the word, and generates a new word form for any word which was changed.

The rules make use of several morphological categories of the word (`node:number`, `node:gender`...), its syntactic relation to its parent in the dependency tree (`node:afun`) and the same information for its English counterpart (`node:en`) and other nodes in the dependency trees.

The order of the rules in this paper follows the order in which they are applied; this is important, as often a rule changes a morphological category of a word which is then used by a subsequent rule.

3.1.1 Noun number (NounNum)

In Czech, a word in singular sometimes has the same form as in plural. Because the tagger often fails to tag the word correctly, we try to correct the tag of a noun tagged as singular if its English counterpart is in plural, so that the subsequent rules can work correctly.

We trust the form of the word but changing the number may also require to change the morphological case (i.e. the tagger was wrong with both number and case). In such cases we choose the first (linearly

from nominative to instrumentative) case matching the form. The rule is:

IF: `node:pos = noun &`
`node:number = singular &`
`node:en:number = plural`
DO: `node:number := plural;`
`node:case := find_case(node:form, plural);`

3.1.2 Subject case (SubjCase)

The subject of a Czech sentence must be in the nominative case. Since the parser often fails in marking the correct word as a subject, we use the English source sentence and presuppose that the Czech counterpart of the English subject is also a subject in the Czech sentence.

IF: `node:en:afun = subject`
DO: `node:case := nominative;`

3.1.3 Subject-predicate agreement (SubjPred)

Subject and predicate in Czech agree in their morphological number. To identify a Czech Subject, we trust the subject in the English sentence. Then we copy the number from the (Czech) Subject to the Czech Predicate.

IF: `node:en:afun = subject &`
`parent:afun = predicate`
DO: `parent:number := node:number;`

3.1.4 Subject-past participle agreement (SubjPP)

Czech past participles agree with subject in morphological gender.

IF: `node:pos = noun|pronoun &`
`node:en:afun = subject &`
`parent:pos = verb_past_participle`
DO: `parent:number := node:number;`
`parent:gender := node:gender;`

3.1.5 Preposition without children (PrepNoCh)

In our dependency trees, the preposition is the parent of the words it belongs to (usually a noun). A preposition without children is incorrect so we find nodes aligned to its English counterpart's children and rehang them under the preposition.

IF: `node:afun = preposition &`
`!node:has_children &`
`node:en:has_children`
DO: `foreach node:en:child;`
`node:en:child:cs:parent := node;`

3.1.6 Preposition-noun agreement (PrepNoun)

Every prepositions gets a morphological case assigned to it by the tagger, with which the dependent noun should agree.

IF: `parent:pos = preposition &`
`node:pos = noun`
DO: `node:case := parent:case;`

3.1.7 Noun-adjective agreement (NounAdj)

Czech adjectives and nouns agree in morphological gender, number and case. We assume that the noun is correct and change the adjective accordingly.

IF: `node:pos = adjective &`
`parent:pos = noun`
DO: `node:gender := parent:gender;`
`node:number := parent:number;`
`node:case := parent:case;`

3.1.8 Reflexive particle deletion (ReflTant)

Czech reflexive verbs are accompanied by reflexive particles ('se' and 'si'). We delete particles not beloning to any verb (or adjective derived from a verb).

IF: `node:form = 'se'|'si' &`
`node:pos = pronoun &`
`parent:pos != verb|verbal_adjective`
DO: `remove node;`

4 Experiments and results

We tested our CU-TWOSTEP system with DEPFIX post-processing on both WMT10 and WMT11 testing data. This combined system was submitted to shared translation task as CU-MARECEK. We also ran the DEPFIX post-processing on all other participating systems.

4.1 Automatic evaluation

The achieved BLEU scores are shown in Tables 1 and 2. They show the scores before and after the DEPFIX post-processing. It is interesting that the improvements are quite different between the years 2010 and 2011 in terms of their BLEU score. While the average improvement on WMT10 test set was 0.21 BLEU points, it was only 0.05 BLEU points on the WMT11 test set. Even the results of the same TWOSTEP system differ in a similar way, so it must have been caused by the different data.

system	before	after	improvement
cu-twostep	*15.98*	*16.13*	*0.15 (0.05 - 0.26)*
cmu-heaf.	16.95	17.04	0.09 (-0.01 - 0.20)
cu-bojar	15.85	16.09	0.24 (0.14 - 0.36)
cu-zeman	12.33	12.55	0.22 (0.12 - 0.32)
dcu	13.36	13.59	0.23 (0.13 - 0.37)
dcu-combo	18.79	18.90	0.11 (0.02 - 0.23)
eurotrans	10.10	10.11	0.01 (-0.04 - 0.07)
koc	11.74	11.91	0.17 (0.08 - 0.26)
koc-combo	16.60	16.86	0.26 (0.16 - 0.37)
onlineA	11.81	12.08	0.27 (0.17 - 0.38)
onlineB	16.57	16.79	0.22 (0.11 - 0.33)
potsdam	12.34	12.57	0.23 (0.14 - 0.35)
rwth-combo	17.54	17.79	0.25 (0.15 - 0.35)
sfu	11.43	11.83	0.40 (0.29 - 0.52)
uedin	15.91	16.19	0.28 (0.18 - 0.40)
upv-combo	17.51	17.73	0.22 (0.10 - 0.34)

Table 1: Depfix improvements on the WMT10 systems in BLEU score. Confidence intervals, which were computed on 1000 bootstrap samples, are in brackets.

system	before	after	improvement
cu-twostep	16.57	16.60	0.03 (-0.07 - 0.13)
cmu-heaf.	20.24	20.32	0.08 (-0.03 - 0.19)
commerc2	09.32	09.32	0.00 (-0.04 - 0.04)
cu-bojar	16.88	16.85	-0.03 (-0.12 - 0.07)
cu-popel	14.12	14.11	-0.01 (-0.06 - 0.03)
cu-tamch.	16.32	16.28	-0.04 (-0.14 - 0.06)
cu-zeman	14.61	14.80	0.19 (0.09 - 0.29)
jhu	17.36	17.42	0.06 (-0.03 - 0.16)
online-B	20.26	20.31	0.05 (-0.06 - 0.16)
udein	17.80	17.88	0.08 (-0.02 - 0.17)
upv-prhlt.	20.68	20.69	0.01 (-0.08 - 0.11)

Table 2: Depfix improvements on the WMT11 systems in BLEU score. Confidence intervals are in brackets.

4.2 Manual evaluation

Two independent annotators evaluated DEPFIX manually on the outputs of CU-TWOSTEP and ONLINE-B. We randomly selected 1000 sentences from the `newssyscombtest2011` data set and the appropriate translations made by these two systems. The annotators got the outputs before and after DEPFIX post-processing and their task was to decide which translation[6] from these two is better and label it by the letter *'a'*. If it was not possible to determine which is better, they labeled both by *'n'*.

[6]They were also provided with the source English sentence and the reference translation. The options were shuffled and indentical candidate sentences were collapsed.

A / B	improved	worsened	indefinite	total
improved	273	20	15	308
worsened	12	59	7	78
indefinite	53	35	42	130
total	338	114	64	516

Table 5: Matrix of the inter-annotator agreement

rule	fired	impr.	wors.	% impr.
SubjCase	51	46	5	90.2
SubjPP	193	165	28	85.5
NounAdj	434	354	80	81.6
NounNum	156	122	34	78.2
PrepNoun	135	99	36	73.3
SubjPred	68	48	20	70.6
ReflTant	15	10	5	66.7
PrepNoCh	45	29	16	64.4

Table 6: Rules and their utility.

Table 3 below shows that about 60% of sentences fixed by DEPFIX were improved and only about 20% were worsened. DEPFIX worked a little better on the ONLINE-B, making fewer changes but also fewer wrong changes. It is probably connected with the fact that overall better translations by ONLINE-B are easier to parse.

The matrix of inter-annotator agreement is in Table 5. Our two annotators agreed in 374 sentences (out of 516), that is 72.5%. On the other hand, if we consider only cases where both annotators chose different translation as better (no indefinite marks), we get only 8.8% disagreement (32 out of 364).

Using the manual evaluation, we can also measure performance of the individual rules. Table 6 shows the number of all, improved or worsened sentences where a particular rule was applied. Definitely, the most useful rule (used often and quite reliable) was the one correcting noun-adjective agreement, followed by the subject-pastparticiple agreement rule.

In each changed sentence, two rules (not necessarily related ones) were applied on average.

4.3 Manual evaluation across data sets

The fact that the improvements in BLEU scores on WMT10 test set are much higher has led us to one more experiment: we compare manual annotations of 330 sentences from each of the WMT10 and

system	annotator	changed	improved		worsened		indefinite	
			count	%	count	%	count	%
cu-bojar-twostep	A	269	152	56.5	39	14.5	78	29.0
cu-bojar-twostep	B	269	173	64.3	50	18.6	46	17.1
online-B	A	247	156	63.1	39	15.9	52	21.1
online-B	B	247	165	66.8	64	25.9	18	7.3

Table 3: Manual evaluation of the DEPFIX post-processing on 1000 randomly chosen sentences from WMT11 test set.

test set	changed	improved		worsened		indefinite		BLEU		
		count	%	count	%	count	%	before	after	diff
newssyscombtest2010	104	52	50.0	20	19.2	32	30.8	16.99	17.38	0.39
newssyscombtest2011	101	66	65.3	19	18.8	16	15.8	13.99	13.87	-0.12

Table 4: Manual and automatic evaluation of the DEPFIX post-processing on CU-TWOSTEP system across different datasets. 330 sentences were randomly selected from each of the WMT10 and WMT11 test sets. Both manual scores and BLEU are computed only on the sentences that were changed by the DEPFIX post-processing.

WMT11 sets as translated by CU-TWOSTEP and corrected by DEPFIX. Table 4 shows that WMT10 and WMT11 are comparable in manually estimated improvement (50–65%). BLEU does not indicate that and even estimates a drop in quality on this subset WMT11. (The absolute BLEU scores differ from BLEUs on the whole test sets but we are interested only in the change of the scores.) BLEU is thus not very suitable for the evaluation of DEPFIX.

5 Conclusions and future work

Manual evaluation shows that our DEPFIX approach to improving MT output quality is sensible. Although it is unable to correct many serious MT errors, such as wrong lexical choices, it can improve the grammaticality of the output in a way that the language model often cannot, which leads to output that is considered to be better by humans. We also suggest that BLEU is not appropriate metric for measuring changes in grammatical correctness of sentences, especially with inflective languages.

An advantage of our method is that it is possible to apply it on output of any MT system (although it works better for phrase-based MT systems). While DEPFIX has been developed using the output of CU-BOJAR, the rules we devised are not specific to any MT system. They simply describe several grammatical rules of Czech language that can be machine-checked and if errors are found, the output can be corrected. Moreover, our method only requires the source sentence and the translation output for its operation – i.e. it is not necessary to modify the MT system itself.

We are now considering modifications of the parser so that it is able to parse the incorrect sentences produced by MT. Theoretically it would be possible to train the parser on annotated ungrammatical sentences, but we do not want to invest such annotation labour. Instead, when parsing the Czech sentence we will make the parser utilize the information contained in the parse tree of the English sentence, which is usually correct. We will probably also have to make the parser put less weight to the often incorrect tagger output. An alternative is to avoid parsing of the target and project the source parse to the target side using word alignments, if provided by the MT system.

Because some of our rules are able to work using only the tagger output, we will also try to apply them before the parsing as they might help the parser by correcting some of the tags.

We will also try several modifications of the tagger, but the English sentence does not help us so much here, because it does not contain any information regarding the most common errors – incorrect assignment of morphological gender and case. However, it could help with part of speech and morphological number disambiguation. Moreover, it would be probably helpful for us if the tagger included several most probable hypotheses, as the single-output-only disambiguation is often erroneous on ungrammatical sentences.

References

Ondrej Bojar and Kamil Kos. 2010. 2010 Failures in English-Czech Phrase-Based MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 60–66, Uppsala, Sweden, July. Association for Computational Linguistics.

Ondřej Bojar and Zdeněk Žabokrtský. 2009. CzEng0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92.

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

Natalia Klyueva and Ondřej Bojar. 2008. UMC 0.1: Czech-Russian-English Multilingual Corpus. In *Proceedings of International Conference Corpus Linguistics*, pages 188–195, Saint-Petersburg.

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.

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.

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.

Dániel Varga et al. 2005. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing*, pages 590–596, Borovets, Bulgaria.