Using Parallel Features in Parsing of Machine-Translated Sentences for Correction of Grammatical Errors *

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

Abstract

In this paper, we present two dependency parser training methods appropriate for parsing outputs of statistical machine translation (SMT), which pose problems to standard parsers due to their frequent ungrammaticality. We adapt the MST parser by exploiting additional features from the source language, and by introducing artificial grammatical errors in the parser training data, so that the training sentences resemble SMT output.

We evaluate the modified parser on DEPFIX, a system that improves English-Czech SMT outputs using automatic rule-based corrections of grammatical mistakes which requires parsed SMT output sentences as its input. Both parser modifications led to improvements in BLEU score; their combination was evaluated manually, showing a statistically significant improvement of the translation quality.

1 Introduction

The machine translation (MT) quality is on a steady rise, with mostly statistical systems (SMT) dominating the area (Callison-Burch et al., 2010; Callison-Burch et al., 2011). Most MT systems do not employ structural linguistic knowledge and even the state-of-the-art MT solutions are unable to avoid making serious grammatical errors in the output, which often leads to unintelligibility or to a risk of misinterpretations of the text by a reader.

*This research has been supported by the EU Seventh Framework Programme under grant agreement n° 247762 (Faust), and by the grants GAUK116310 and GA201/09/H057.

This problem is particularly apparent in target languages with rich morphological inflection, such as Czech. As Czech often conveys the relations between individual words using morphological agreement instead of word order, together with the word order itself being relatively free, choosing the correct inflection becomes crucial.

Since the output of phrase-based SMT shows frequent inflection errors (even in adjacent words) due to each word belonging to a different phrase, a possible way to address the grammaticality problem is a combination of statistical and structural approach, such as SMT output post-editing (Stymne and Ahrenberg, 2010; Mareček et al., 2011).

In this paper, we focus on improving SMT output parsing quality, as rule-based post-editing systems rely heavily on the quality of SMT output analysis. Parsers trained on gold standard parse trees often fail to produce the expected result when applied to SMT output with grammatical errors. This is partly caused by the fact that when parsing highly inflected free word-order languages the parsers have to rely on morphological agreement, which, as stated above, is often erroneous in SMT output.

Training a parser specifically by creating a manually annotated treebank of MT systems' outputs would be very expensive, and the application of such treebank to other MT systems than the ones used for its generation would be problematic. We address this issue by two methods of increasing the quality of SMT output parsing:

- a different application of previous works on bitext parsing – exploiting additional features from the source language (Section 3), and

Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 39–48,
Jeju, Republic of Korea, 12 July 2012.

- introducing artificial grammatical errors in the target language parser training data, so that the sentences resemble the SMT output in some ways (Section 4). This technique is, to our knowledge, novel with regards to its application to SMT and the statistical error model.

We test these two techniques on English-Czech MT outputs using our own reimplementation of the MST parser (McDonald et al., 2005) named RUR[1] parser. and evaluate their contribution to the SMT post-editing quality of the DEPFIX system (Mareček et al., 2011), which we outline in Section 5. We describe the experiments carried out and present the most important results in Section 6. Section 7 then concludes the paper and indicates more possibilities of further improvements.

2 Related Work

Our approach to parsing with parallel features is similar to various works which seek to improve the parsing accuracy on parallel texts ("bitexts") by using information from both languages. Huang et al. (2009) employ "bilingual constraints" in shift-reduce parsing to disambiguate difficult syntactic constructions and resolve shift-reduce conflicts. Chen et al. (2010) use similar subtree constraints to improve parser accuracy in a dependency scenario. Chen et al. (2011) then improve the method by obtaining a training parallel treebank via SMT. In recent work, Haulrich (2012) experiments with a setup very similar to ours: adding alignment-projected features to an originally monolingual parser.

However, the main aim of all these works is to improve the parsing accuracy on correct parallel texts, i.e. human-translated. This paper applies similar methods, but with a different objective in mind – increasing the ability of the parser to process ungrammatical SMT output sentences and, ultimately, improve rule-based SMT post-editing.

Xiong et al. (2010) use SMT parsing in translation quality assessment, providing syntactic features to a classifier detecting erroneous words in SMT output, yet they do not concentrate on improving parsing accuracy – they employ a link grammar parser, which is robust, but not tuned specifically to process ungrammatical input.

There is also another related direction of research in parsing of parallel texts, which is targeted on parsing under-resourced languages, e.g. the works by Hwa et al. (2005), Zeman and Resnik (2008), and McDonald et al. (2011). They address the fact that parsers for the language of interest are of low quality or even non-existent, whereas there are high-quality parsers for the other language. They exploit common properties of both languages and delexicalization. Zhao et al. (2009) uses information from word-by-word translated treebank to obtain additional training data and boost parser accuracy.

This is different from our situation, as there exist high performance parsers for Czech (Buchholz and Marsi, 2006; Nivre et al., 2007; Hajič et al., 2009). Boosting accuracy on correct sentences is not our primary goal and we do not intend to *replace* the Czech parser by an English parser; instead, we aim to increase the robustness of an already *existing* Czech parser by adding knowledge from the corresponding English source, parsed by an English parser.

Other works in bilingual parsing aim to parse the parallel sentences directly using a grammar formalism fit for this purpose, such as Inversion Transduction Grammars (ITG) (Wu, 1997). Burkett et al. (2010) further include ITG parsing with word-alignment in a joint scenario. We concentrate here on using dependency parsers because of tools and training data availability for the examined language pair.

Regarding treebank adaptation for parser robustness, Foster et al. (2008) introduce various kinds of artificial errors into the training data to make the final parser less sensitive to grammar errors. However, their approach concentrates on mistakes made by humans (such as misspellings, word repetition or omission etc.) and the error models used are handcrafted. Our work focuses on morphology errors often encountered in SMT output and introduces statistical error modelling.

3 Parsing with Parallel Features

This section describes our SMT output parsing setup with features from analyzed *source* sentences. We

[1]The abbreviation "RUR" parser stands for "Rudolph's Universal Robust" parser.

explain our motivation for the inclusion of parallel features in Section 3.1, then provide an account of the parsers used (including our RUR parser) in Section 3.2, and finally list all the monolingual and parallel features included in the parser training (in Sections 3.3 and 3.4, respectively).

3.1 Motivation

An advantage of SMT output parsing over general dependency parsing is that one can also make use of *source* – English sentences in our case. Moreover, although SMT output is often in many ways ungrammatical, *source* is usually grammatical and therefore easier to process (in our case especially to tag and parse). This was already noticed in Mareček et al. (2011), who use the analysis of *source* sentence to provide additional information for the DEPFIX rules, claiming it to be more reliable than the analysis of SMT output sentence.

We have carried this idea further by having devised a simple way of making use of this information in parsing of the SMT output sentences: We parse the *source* sentence first and include features computed over the parsed *source* sentence in the set of features used for parsing SMT output. We first align the *source* and SMT output sentences on the word level and then use alignment-wise local features – i.e. for each SMT output word, we add features computed over its aligned *source* word, if applicable (cf. Section 3.4 for a listing).

3.2 Parsers Used

We have reimplemented the MST parser (McDonald et al., 2005) in order to provide for a simple insertion of the parallel features into the models.

We also used the original implementation of the MST parser by McDonald et al. (2006) for comparison in our experiments. To distinguish the two variants used, we denote the original MST parser as MCD parser,[2] and the new reimplementation as RUR parser.

We trained RUR parser in a first-order non-projective setting with single-best MIRA. Dependency labels are assigned in a second stage by a MIRA-based labeler, which has been implemented according to McDonald (2006) and Gimpel and Cohen (2007).

We used the Prague Czech-English Dependency Treebank[3] (PCEDT) 2.0 (Bojar et al., 2012) as the training data for RUR parser – a parallel treebank created from the Penn Treebank (Marcus et al., 1993) and its translation into Czech by human translators. The dependency trees on the English side were converted from the manually annotated phrase-structure trees in Penn Treebank, the Czech trees were created automatically using MCD. Words of the Czech and English sentences were aligned by GIZA++ (Och and Ney, 2003).

We apply RUR parser only for SMT output parsing; for *source* parsing, we use MCD parser trained on the English CoNLL 2007 data (Nivre et al., 2007), as the performance of this parser is sufficient for this task.

3.3 Monolingual Features

The set of monolingual features used in RUR parser follows those described by McDonald et al. (2005). For parsing, we use the features described below. The individual features are computed for both the parent node and the child node of an edge and conjoined in various ways. The *coarse morphological tag* and *lemma* are provided by the Morče tagger (Spoustová et al., 2007).

- *coarse morphological tag* – Czech two-letter coarse morphological tag, as described in (Collins et al., 1999),[4]
- *lemma* – morphological lemma,
- context features: *preceding coarse morphological tag*, *following coarse morphological tag* – coarse morphological tag of a neighboring node,
- *coarse morphological tags in between* – bag of coarse morphological tags of nodes positioned between the parent node and the child node,

[2] MCD uses k-best MIRA, does first- and second-order parsing, both projectively and non-projectively, and can be obtained from `http://sourceforge.net/projects/mstparser`.

[3] `http://ufal.mff.cuni.cz/pcedt`

[4] The first letter is the main POS (12 possible values), the second letter is either the morphological case field if the main POS displays case (i.e. for nouns, adjectives, pronouns, numerals and prepositions; 7 possible values), or the detailed POS if it does not (22 possible values).

- *distance* – signed bucketed distance of the parent and the child node in the sentence (in # of words), using buckets 1, 2, 3, 4, 5 and 11.

To assign dependency labels, we use the same set as described above, plus the following features (called "non-local" by McDonald (2006)), which make use of the knowledge of the tree structure.

- *is first child*, *is last child* – a boolean indicating whether the node appears in the sentence as the first/last one among all the child nodes of its parent node,

- *child number* – the number of syntactic children of the current node.

3.4 Parallel Features

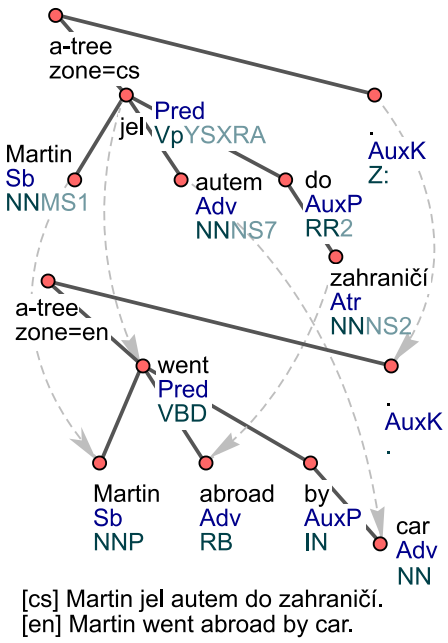

Figure 1: Example sentence for parallel features illustration (see Table 1).

In RUR parser we use three types of parallel features, computed for the parent and child node of an edge, which make use of the *source* English nodes aligned to the parent and child node.

- *aligned tag*: morphological tag following the Penn Treebank Tagset (Marcus et al., 1993) of the English node aligned to the Czech node

Feature	Feature value on parent node	Feature value on child node
word form	jel	Martin
aligned tag	VBD	NNP
aligned dep. label	Pred	Sb
aligned edge existence	true	
word form	jel	autem
aligned tag	VBD	NN
aligned dep. label	Pred	Adv
aligned edge existence	false	
word form	do	zahraničí
aligned tag	—	RB
aligned dep. label	—	Adv
aligned edge existence	—	
word form	#root#	.
aligned tag	#root#	.
aligned dep. label	AuxS	AuxK
aligned edge existence	true	

Table 1: Parallel features for several edges in Figure 1.

- *aligned dependency label*: dependency label of the English node aligned to the Czech node in question, according to the PCEDT 2.0 label set (Bojar et al., 2012)

- *aligned edge existence*: a boolean indicating whether the English node aligned to the Czech parent node is also the parent of the English node aligned to the Czech child node

The parallel features are conjoined with the monolingual *coarse morphological tag* and *lemma* features in various ways.

If there is no *source* node aligned to the parent or child node, the respective feature cannot be computed and is skipped.

An example of a pair of parallel sentences is given in Figure 1 with the corresponding values of parallel features for several edges in Table 1.

4 Worsening Treebanks to Simulate Some of the SMT Frequent Errors

Addressing the issue of great differences between the gold standard parser training data and the actual analysis input (SMT output), we introduced artificial inconsistencies into the training treebanks, in order to make the parsers more robust in the face of grammar errors made by SMT systems. We have concen-

trated solely on modeling incorrect word flection, i.e. the dependency trees retained their original correct structures and word lemmas remained fixed, but the individual inflected word forms have been modified according to an error model trained on real SMT output. We simulate thus, with respect to morphology, a treebank of parsed MT output sentences.

In Section 4.1 we describe the steps we take to prepare the worsened parser training data. Section 4.2 contains a description of our monolingual greedy alignment tool which is needed during the process to map SMT output to reference translations.

4.1 Creating the Worsened Parser Training Data

The whole process of treebank worsening consists of five steps:

1. We translated the English side of PCEDT[5] to Czech using SMT (we chose the Moses system (Koehn et al., 2007) for our experiments) and tagged the resulting translations using the Morče tagger (Spoustová et al., 2007).

2. We aligned the Czech side of PCEDT, now serving as a reference translation, to the SMT output using our Monolingual Greedy Aligner (see Section 4.2).

3. Collecting the counts of individual errors, we estimated the Maximum Likelihood probabilities of changing a correct fine-grained morphological tag (of a word from the reference) into a possibly incorrect fine-grained morphological tag of the aligned word (from the SMT output).

4. The tags on the Czech side of PCEDT were randomly sampled according to the estimated "fine-grained morphological tag error model". In those positions where fine-grained morphological tags were changed, new word forms were generated using the Czech morphological generator by Hajič (2004).[6]

We use the resulting "worsened" treebank to train our parser described in Section 3.2.

4.2 The Monolingual Greedy Aligner

Our monolingual alignment tool, used in treebank worsening to tie reference translations to MT output (see Section 4.1), scores all possible alignment links and then greedily chooses the currently highest scoring one, creating the respective alignment link from word A (in the reference) to word B (in the SMT output) and deleting all scores of links from A or to B, so that one-to-one alignments are enforced. The process is terminated when no links with a score higher than a given threshold are available; some words may thus remain unaligned.

The score is computed as a linear combination of the following four features:

- word form (or lemma if available) similarity based on Jaro-Winkler distance (Winkler, 1990),
- fine-grained morphological tag similarity,
- similarity of the relative position in the sentence,
- and an indication whether the word following (or preceding) A was already aligned to the word following (or preceding) B.

Unlike bilingual word aligners, this tool needs no training except for setting weights of the four features and the threshold.[7]

5 The DEPFIX System

The DEPFIX system (Mareček et al., 2011) applies various rule-based corrections to Czech-English SMT output sentences, especially of morphological agreement. It also employs the parsed *source* sentences, which must be provided on the input together with the SMT output sentences.

The corrections follow the rules of Czech grammar, e.g. requiring that the clause subject be in the

[5]This approach is not conditioned by availability of parallel treebanks. Alternatively, we might translate any text for which reference translations are at hand. The model learned in the third step would then be applied (in the fourth step) to a different text for which parse trees are available.

[6]According to the "fine-grained morphological tag error model", about 20% of fine-grained morphological tags were changed. In 4% of cases, no word form existed for the new fine-grained morphological tag and thus it was not changed.

[7]The threshold and weights were set manually using just ten sentence pairs. The resulting alignment quality was found sufficient, so no additional weights tuning was performed.

nominative case or enforcing subject-predicate and noun-attribute agreements in morphological gender, number and case, where applicable. Morphological properties found violating the rules are corrected and the corresponding word forms regenerated.

The *source* sentence parse, word-aligned to the SMT output using GIZA++ (Och and Ney, 2003), is used as a source of morpho-syntactic information for the correction rules. An example of a correction rule application is given in Figure 2.

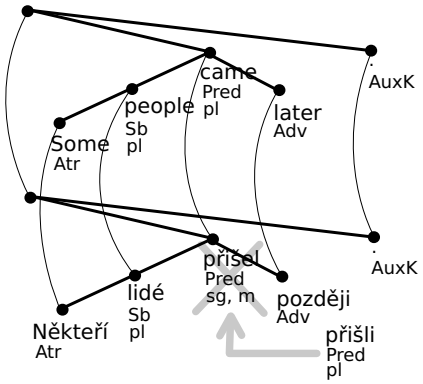

Figure 2: Example of fixing subject-predicate agreement. The Czech word *přišel [he came]* has a wrong morphological number and gender. Adapted from Mareček et al. (2011).

The system is implemented within the TectoMT/Treex NLP framework (Popel and Žabokrtský, 2010). Mareček et al. (2011) feed the DEPFIX system with analyses by the MCD parser trained on gold-standard treebanks for parsing of English *source* sentences as well as Czech SMT output.

6 Experiments and Results

We evaluate RUR parser indirectly by using it in the DEPFIX system and measuring the performance of the whole system. This approach has been chosen instead of direct evaluation of the SMT output parse trees, as the task of finding a correct parse tree of a possibly grammatically incorrect sentence is not well defined and considerably difficult to do.

We used WMT10, WMT11 and WMT12 English to Czech translation test sets, `newssyscombtest2010`, `newssyscombtest2011` and `newstest2012`,[8] (denoted as WMT10, WMT11 and WMT12) for the automatic evaluation. The data sets include the source (English) text, its reference translation and translations produced by several MT systems. We used the outputs of three SMT systems: GOOGLE,[9] UEDIN (Koehn et al., 2007) and BOJAR (Bojar and Kos, 2010).

For the manual evaluation, two sets of 1000 randomly selected sentences from WMT11 and from WMT12 translated by GOOGLE were used.

6.1 Automatic Evaluation

Table 2 shows BLEU scores (Papineni et al., 2002) for the following setups of DEPFIX:

- SMT output: output of an SMT system without applying DEPFIX
- MCD: parsing with MCD
- RUR: parsing with RUR (Section 3.2)
- RUR+PARA: parsing with RUR using parallel features (Section 3.4)
- RUR+WORS: parsing with RUR trained on worsened treebank (Section 4)
- RUR+WORS+PARA: parsing with RUR trained on worsened treebank and using parallel features

It can be seen that both of the proposed ways of adapting the parser to parsing of SMT output often lead to higher BLEU scores of translations post-processed by DEPFIX, which suggests that they both improve the parsing accuracy.

We have computed 95% confidence intervals on 1000 bootstrap samples, which showed that the BLEU score of RUR+WORS+PARA was significantly higher than that of MCD and RUR parser in 4 and 3 cases, respectively (results where RUR+WORS+PARA achieved a significantly higher score are marked with '*'). On the other hand, the score of neither RUR+WORS+PARA nor RUR+WORS and RUR+PARA was ever significantly lower than the score of MCD or RUR parser. This leads us to believe that the two proposed methods are able to produce slightly better SMT output parsing results.

[8] `http://www.statmt.org/wmt10`, `http://www.statmt.org/wmt11`, `http://www.statmt.org/wmt12`

[9] `http://translate.google.com`

Test set	WMT10			WMT11			WMT12		
SMT system	BOJAR	GOOGLE	UEDIN	BOJAR	GOOGLE	UEDIN	BOJAR	GOOGLE	UEDIN
SMT output	*15.85	*16.57	*15.91	*16.88	*20.26	*17.80	14.36	16.25	*15.54
MCD	16.09	16.95	*16.35	*17.02	20.45	*18.12	14.35	**16.32**	*15.65
RUR	16.08	*16.85	*16.29	17.03	20.42	*18.09	14.37	16.31	15.66
RUR+PARA	**16.13**	*16.90	*16.35	17.05	20.47	18.19	14.35	16.31	15.72
RUR+WORS	16.12	16.96	*16.45	17.06	**20.53**	18.21	**14.40**	16.31	15.71
RUR+WORS+PARA	**16.13**	**17.03**	**16.54**	**17.12**	**20.53**	**18.25**	14.39	16.30	**15.74**

Table 2: Automatic evaluation using BLEU scores for the unmodified SMT output (output of BOJAR, GOOGLE and UEDIN systems on WMT10, WMT11 and WMT12 test sets), and for SMT output parsed by various parser setups and processed by DEPFIX. The score of RUR+WORS+PARA is significantly higher at 95% confidence level than the scores marked with '*' on the same data.

6.2 Manual Evaluation

Performance of RUR+WORS+PARA setup was manually evaluated by doing a pairwise comparison with other setups – SMT output, MCD and RUR parser. The evaluation was performed on both the WMT11 (Table 4) and WMT12 (Table 5) test set. 1000 sentences from the output of the GOOGLE system were randomly selected and processed by DEPFIX, using the aforementioned SMT output parsers. The annotators then compared the translation quality of the individual variants in differing sentences, selecting the better variant from a pair or declaring two variants "same quality" (indefinite). They were also provided with the *source* sentence and a reference translation. The evaluation was done as a blind test, with the sentences randomly shuffled.

The WMT11 test set was evaluated by two independent annotators. (The WMT12 test set was evaluated by one annotator only.) The inter-annotator agreement and Cohen's kappa coefficient (Cohen and others, 1960), shown in Table 3, were computed both including all annotations ("with indefs"), and disregarding sentences where at least one of the annotators marked the difference as indefinite ("without indefs") – we believe a disagreement in choosing the better translation to be more severe than a disagreement in deciding whether the difference in quality of the translations allows to mark one as being better.

For both of the test sets, RUR+WORS+PARA significantly outperforms both MCD and RUR baseline, confirming that a combination of the proposed modifications of the parser lead to its better performance. Statistical significance of the results was confirmed by a one-sided pairwise t-test, with the following differences ranking: RUR+WORS+PARA better = 1, baseline better = -1, indefinite = 0.

RUR+WORS+PARA	with indefs		without indefs	
compared to	IAA	Kappa	IAA	Kappa
SMT output	77%	0.54	92%	0.74
MCD	79%	0.66	95%	0.90
RUR	75%	0.60	94%	0.85

Table 3: Inter-annotator agreement on WMT11 data set translated by GOOGLE

6.3 Inspection of Parser Modification Benefits

For a better understanding of the benefits of using our modified parser, we inspected a small number of parse trees, produced by RUR+WORS+PARA, and compared them to those produced by RUR. In many cases, the changes introduced by RUR+WORS+PARA were clearly positive. We provide two representative examples below.

Subject Identification

Czech grammar requires the subject to be in nominative case, but this constraint is often violated in SMT output and a parser typically fails to identify the subject correctly in such situations. By worsening the training data, we make the parser more robust in this respect, as the worsening often switches the case of the subject; by including parallel features, especially the *aligned dependency label* feature, RUR+WORS+PARA parser can often identify the subject as the node aligned to the *source* subject.

Annotator	Baseline	Differing sentences	Out of the differing sentences					
			RUR+WORS+PARA better		baseline better		indefinite	
			count	percent	count	percent	count	percent
A	SMT output	422	301	71%	79	19%	42	10%
	MCD	211	120	57%	65	31%	26	12%
	RUR	217	123	57%	64	29%	30	14%
B	SMT output	422	284	67%	69	16%	69	16%
	MCD	211	107	51%	56	26%	48	23%
	RUR	217	118	54%	53	24%	46	21%

Table 4: Manual comparison of RUR+WORS+PARA with various baselines, on 1000 sentences from WMT11 data set translated by GOOGLE, evaluated by two independent annotators.

Annotator	Baseline	Differing sentences	Out of the differing sentences					
			RUR+WORS+PARA better		baseline better		indefinite	
			count	percent	count	percent	count	percent
A	SMT output	420	270	64%	88	21%	62	15%
	MCD	188	86	45%	64	34%	38	20%
	RUR	187	96	51%	57	30%	34	18%

Table 5: Manual comparison of RUR+WORS+PARA with various baselines, on 1000 sentences from WMT12 data set translated by GOOGLE.

Governing Noun Identification

A parser for Czech typically relies on morphological agreement between an adjective and its governing noun (in morphological number, gender and case), which is often violated in SMT output. Again, RUR+WORS+PARA is more robust in this respect, *aligned edge existence* now being the crucial feature for the correct identification of this relation.

7 Conclusions and Future Work

We have studied two methods of improving the parsing quality of Machine Translation outputs by providing additional information to the parser.

In Section 3, we propose a method of integrating additional information known *at runtime*, i.e. the knowledge of the source sentence (*source*), from which the sentence being parsed (SMT output) has been translated. This knowledge is provided by extending the parser feature set with new features from the source sentence, projected through word-alignment.

In Section 4, we introduce a method of utilizing additional information known *in the training phase*, namely the knowledge of the ways in which SMT output differs from correct sentences. We provide this knowledge to the parser by adjusting its training data to model some of the errors frequently encountered in SMT output, i.e. incorrect inflection forms.

We have evaluated the usefulness of these two methods by integrating them into the DEPFIX rule-based MT output post-processing system (Mareček et al., 2011), as MT output parsing is crucial for the operation of this system. When used with our improved parsing, the DEPFIX system showed better performance both in automatic and manual evaluation on outputs of several, including state-of-the-art, MT systems.

We believe that the proposed methods of improving MT output parsing can be extended beyond their current state. The parallel features used in our setup are very few and very simple; it thus remains to be examined whether more elaborate features could help utilize the additional information contained in the source sentence to a greater extent. Modeling other types of SMT output inconsistencies in parser training data is another possible step.

We also believe that the methods could be adapted for use in other applications, e.g. automatic classification of translation errors, confidence estimation or multilingual question answering.

References

Ondřej 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, 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.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics.

David Burkett, John Blitzer, and Dan Klein. 2010. Joint parsing and alignment with weakly synchronized grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–135. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July. Association for Computational Linguistics.

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.

Wenliang Chen, Jun'ichi Kazama, and Kentaro Torisawa. 2010. Bitext dependency parsing with bilingual subtree constraints. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 21–29. Association for Computational Linguistics.

Wenliang Chen, Jun'ichi Kazama, Min Zhang, Yoshimasa Tsuruoka, Yujie Zhang, Yiou Wang, Kentaro Torisawa, and Haizhou Li. 2011. SMT helps bitext dependency parsing. In *EMNLP*, pages 73–83. ACL.

Jacob Cohen et al. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Michael Collins, Lance Ramshaw, Jan Hajič, and Christoph Tillmann. 1999. A statistical parser for Czech. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 505–512, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jennifer Foster, Joachim Wagner, and Josef Van Genabith. 2008. Adapting a WSJ-trained parser to grammatically noisy text. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 221–224. Association for Computational Linguistics.

Kevin Gimpel and Shay Cohen. 2007. Discriminative online algorithms for sequence labeling- a comparative study.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18. Association for Computational Linguistics.

Jan Hajič. 2004. *Disambiguation of rich inflection: computational morphology of Czech*. Karolinum.

Martin Haulrich. 2012. *Data-Driven Bitext Dependency Parsing and Alignment*. Ph.D. thesis.

Liang Huang, Wenbin Jiang, and Qun Liu. 2009. Bilingually-constrained (monolingual) shift-reduce parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1222–1231. Association for Computational Linguistics.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.*, 11:311–325, September.

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.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Comput. Linguist.*, 19:313–330, June.

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. 2011. Two-step translation with grammatical post-processing. In Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan, editors, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 426–432, Edinburgh, UK. 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.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 216–220, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 62–72. Association for Computational Linguistics.

Ryan McDonald. 2006. *Discriminative learning and spanning tree algorithms for dependency parsing*. Ph.D. thesis, Philadelphia, PA, USA. AAI3225503.

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.

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.

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.

Sara Stymne and Lars Ahrenberg. 2010. Using a grammar checker for evaluation and postprocessing of statistical machine translation. In *Proceedings of LREC*, pages 2175–2181.

William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods (American Statistical Association)*, pages 354–359.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Deyi Xiong, Min Zhang, and Haizhou Li. 2010. Error detection for statistical machine translation using linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 604–611. Association for Computational Linguistics.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. *NLP for Less Privileged Languages*, page 35.

Hai Zhao, Yan Song, Chunyu Kit, and Guodong Zhou. 2009. Cross language dependency parsing using a bilingual lexicon. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 55–63. Association for Computational Linguistics.