Using a Grammar Checker for Evaluation and Postprocessing of Statistical Machine Translation

Sara Stymne, Lars Ahrenberg

Department of Computer and Information Science
Linköping University, Sweden
{sarst,lah}@ida.liu.se

Abstract

One problem in statistical machine translation (SMT) is that the output often is ungrammatical. To address this issue, we have investigated the use of a grammar checker for two purposes in connection with SMT: as an evaluation tool and as a postprocessing tool. As an evaluation tool the grammar checker gives a complementary picture to standard metrics such as Bleu, which do not account for grammaticality. We use the grammar checker as a postprocessing tool by applying the error correction suggestions it gives. There are only small overall improvements of the postprocessing on automatic metrics, but the sentences that are affected by the changes are improved, as shown both by automatic metrics and by a human error analysis. These results indicate that grammar checker techniques are a useful complement to SMT.

1. Introduction

One problem with standard statistical machine translation systems is that their output tends to be ungrammatical, since there generally is no linguistic knowledge used in the systems. We investigate how a grammar checker can be used to address this issue. Grammar checkers are developed to find errors in texts produced by humans, but in this study we investigate if they can also be used to find errors made by machines. We identify two novel usages of the grammar checker for machine translation (MT): as an evaluation tool and as a postprocessing tool.

We have performed experiments for English-Swedish translation using a factored phrase-based statistical machine translation (PBSMT) system based on Moses (Koehn et al., 2007) and the mainly rule-based Swedish grammar checker Granska (Domeij et al., 2000; Knutsson, 2001). The combination of a grammar checker and a MT system could be used for other architectures and language pairs as well, however. We have performed experiments on six translation systems that differ on two dimensions: the amount of training data, and the amount of linguistic knowledge used in the system.

To be able to use the grammar checker as an evaluation tool, we performed an error analysis of the grammar checker on SMT output. We then defined three crude measures based on the error identification by the grammar checker. All three measures are error rates based on the grammar checker error categories. The difference between them is that they use different subsets of the categories. All three measures give a complementary picture to two standard MT metrics, since they are better at accounting for the fluency and grammaticality of the machine translation output.

We used the grammar checker as an automatic postprocessing tool on the SMT output, by using the correction suggestions given for many errors. We applied the suggestions only for categories that had a high precision on the error analysis on SMT output. A human error analysis showed that the corrections were successful in most cases. There were only small improvements on automatic metrics on the full test sets, however, but this was mainly due to the fact that the postprocessing only affected a small proportion of the sentences.

The remainder of the paper is organized as follows: Section 2 describes the translation systems used in the study. In Section 3 the grammar checker Granska is introduced, and an error analysis of Granska on SMT output is described. The experiments and results are presented in Section 4 for evaluation and in Section 5 for postprocessing. Section 6 contains a discussion of related work, and Section 7 contains the conclusion and suggestions for future work.

2. SMT System

The translation system used is a standard PBSMT setup using the Moses decoder (Koehn et al., 2007) and the SRILM toolkit for sequence models (Stolcke, 2002). We take advantage of the factored translation framework in Moses (Koehn and Hoang, 2007), where factors other than surface form can be used to represent words, which allows the inclusion of linguistic knowledge such as lemmas and part-of-speech tags. To tune feature weights minimum error rate training is used (Och, 2003).

The system is trained and tested on the Europarl corpus (Koehn, 2005). The Swedish target side of the training corpus is part-of-speech tagged using the Granska tagger (Carlberger and Kann, 1999). The training corpus is filtered to remove sentences longer than 40 words and with a length ratio of more than 1 to 7. 500 sentences are used for tuning and 2000 sentences for testing.

In order to evaluate the use of the grammar checker we trained six different systems that are varied on two dimensions: the amount of training data and the amount of linguistic knowledge in the systems, in the form of extra factors on the output side. For the corpus size we have used two sizes of the training corpus for the translation model: a large training corpus with 701,157 sentences, and a small training corpus with 100,000 sentences. In both cases the sequence models were trained on the large corpus.

Text: Averaging vore med tre timmar per dag , det är den mest omfattande mänskliga aktivitet efter sover och - för vuxna - arbete .

Rule: stav1@stavning Span: 1-1 Words: Averaging

Rule: kong10E@kong Span: 14-15 Words: mänskliga aktivitet
mänskliga aktiviteten
mänsklig aktivitet

Figure 1: Example of filtered Granska output for a sentence with two errors

The linguistic knowledge used in the system are varied by the use of different factors on the output side. There are three system setups for each corpus size: the *none* system with only a standard language model on surface form, and two systems with an additional sequence model. The *POS* system use standard part-of-speech tags and the *morph* system use morphologically enriched part-of-speech tags. An example of the annotation is shown in (1).

(1) *EN:* my question is important .
none: min fråga är viktig .
POS: min|ps fråga|nn är|vb viktig|jj .|mad
morph: min|ps.utr.sin.def fråga|nn.utr.sin.ind.nom är|vb.prs.akt.kop viktig|jj.pos.utr.sin.ind.nom .|mad

Using part-of-speech tags can be expected to improve word order. The use of morphological tags can also improve agreement (Stymne et al., 2008).

3. Grammar Checker

A grammar checker is a tool that can identify grammar errors in texts. Often they also include other errors such as spelling errors and stylistic errors. Grammar checkers tend to be authoring tools or writing aids, that is, they are designed to be used by a human who assess the alarms and suggestions given by the tools, rather than as software that can be applied automatically.

For MT output, both types of grammar checkers could be considered useful. A grammar checker system that is an authoring tool could be used to highlight suspicious problems and to suggest changes in translations that are sent to a human posteditor. An automatic grammar checking system, on the other hand, could be used to automatically improve MT output, regardless of whether the translations are being used directly, e.g. for gisting, or if they are being sent for further postediting by humans. If the purpose of using the grammar checker is evaluation it would clearly be preferable with an automatic grammar checker.

3.1. Granska

We use the Swedish grammar checker Granska (Domeij et al., 2000; Knutsson, 2001), which is a hybrid, mainly rule-based grammar checker. The main modules in Granska are: a tokenizer; a probabilistic Hidden Markov model-based tagger, which tags texts both with part-of-speech and morphology (Carlberger and Kann, 1999); the spell checker Stava (Kann et al., 2001); and a rule matcher, which identifies errors and generates error descriptions and correction suggestions. The rule matcher contains hand-written rules in an object-oriented rule language developed for Granska. Granska finds both grammar and spelling errors, and in many cases it also gives correction suggestions. The grammar errors are divided into thirteen main categories, of which many are in turn further divided into a number of subcategories.

Granska can output XML, from which we extracted the necessary information for our purposes, exemplified in Figure 1. In the output we see the whole sentence (*Text:*), and a list of errors that were found in each sentence. For each error we know the main rule type and rule subcategory that applied (*Rule:*), the position in the sentence where it occurs (*Span:*), the words that are wrong (*Words:*), and possibly one or several correction suggestions. For the sentence in Figure 1, Granska detected a spelling error, the unknown English word *Averaging*, for which it has no correction suggestions, and a noun phrase agreement error, for which it has two suggestions, of which the first one is correct in this context. There are also other errors in the sentence which Granska does not find, mainly because it is not designed for this type of malformed output.

3.2. Error Analysis of Granska on SMT Output

In order to use the grammar checker for SMT it is useful to know on which categories the grammar checker produces good results and good correction suggestions for SMT output. Granska has been evaluated on human output in previous studies, but with different results on different text types (Knutsson, 2001). Applying Granska on a different type of text than it was designed for can also affect its performance; for instance, its precision for a subset of its error categories degrades from 92% on texts written by adults to 35% on texts written by primary school children (Sofkova Hashemi, 2007). Machine translation output is another very different text type, on which we cannot expect the behavior of Granska to be the same as for human texts. Thus we performed an error analysis of the performance of the grammar checker on the translation output from the tuning process, a total of 11,574 words, from Europarl. The evaluation was performed by one native Swedish speaker.

The grammar errors found in the sample only belonged to five of the thirteen error categories in Granska:

- **Agreement NP** – Errors in noun phrase agreement of determiners, nouns and adjectives for number, gender and definiteness
- **Agreement predicatives** – Errors in agreement of predicatives with the subject or object on number, gender and definiteness

Type	Error identification		Correction suggestions			
	Correct	False	Correct1	Correct2+	Wrong	None
Agreement NP	64	10	48	10	4+10	2+0
Agreement Predicatives	21	1	20	–	1+1	–
Split compounds	12	14	8	–	3+13	1+1
Verb	31	18	11	2	–	18+18
Word order	9	0	8	–	1+0	–

Table 1: Analysis of grammar errors that are identified by Granska on tuning data. Correct1 means that the top suggestion is correct, and Correct2+ that some other suggestion is correct. For cases where suggestions are wrong or not given, the numbers are divided between errors that are correctly identified and errors that are erroneously identified.

Type	Granska evaluation on human text			Granska evaluation on SMT output
	Recall	Precision	Precision range	Precision
Agreement NP	0.83	0.44	0.11–0.72	0.86
Agreement Predicative	0.69	0.32	0.00–0.44	0.95
Split compounds	0.46	0.39	0.00–0.67	0.46
Verb	0.97	0.83	0.71–0.91	0.63
Word order	0.75	0.38	0.00–1.00	1.00

Table 2: Comparison of an evaluation of Granska on human text (Knutsson, 2001) with the error analysis on SMT output. Precision range is the extreme precision values on the five different text types

- **Split compounds** – Compounds that are written as two separate words, instead of as one word
- **Verb** – Errors that are related to verbs, such as missing verbs, wrong form of verbs, or two finite verbs
- **Word order** – Wrong word order

Table 1 summarizes the results for these five categories. The performance varies a lot between the error categories, with good performance of error identification on agreement and word order, but worse on split compounds and verbs. Looking into the different subcategories shows that the false alarms for verbs mostly belong to three categories: *missing verb*, *missing finite verb*, and *infinitive marker without a verb*. When we excluded these categories there are 17 verb errors left, of which only 1 is wrong.

The quality of the correction suggestions also varies between the categories. In the verb category, suggestions are never wrong, but they are given in a minority of the cases where they are correctly identified, and never for the false alarms. For split compounds, on the other hand, the majority of the correction suggestions are incorrect, mainly since suggestions are given to nearly all false alarms. For NP agreement, all false alarms have correction suggestions, but the majority of the correction suggestions are still correct. Predicative agreement and verb errors have correct suggestions for nearly all identified errors.

There are very few pure spelling errors in the translation output, since the words all come from the corpus the SMT system is trained on. The grammar checker still identifies 161 spelling errors, of which the majority are untranslated foreign words (49.0%) and proper names (32.9%). The only correction suggestions that are useful for spelling errors is the capitalization of lower-cased proper names, which occur in 9 cases.

The error analysis on SMT output can be contrasted to an earlier evaluation of Granska on human texts performed by Knutsson (2001). That evaluation was performed on 201,019 words from five different text types: sport news, foreign news, government texts, popular science and student essays. The results from that evaluation on the error categories found in the SMT sample are contrasted with the SMT error analysis in Table 2. The performance of Granska varies a lot between the five text types, as shown by the precision range. The precision on SMT output is better than the average precision on human texts on all error categories except verb errors. This is promising for the use of Granska on SMT output. The human texts were annotated with all present errors, which meant that recall could be calculated. The recall is rather high on all categories except split compounds. No annotation of all errors were done in our SMT error analysis, and thus we cannot calculate recall. It could be expected to be a lot lower than on human texts, however, since Granska was not developed with SMT errors in mind.

4. Grammar Checker as an Evaluation Tool

A disadvantage of most current evaluation metrics for MT, is that they do not take grammar into account. Thus, the grammar checker could complement existing metrics by indicating the grammaticality of a text, simply by counting the number of errors. This only accounts for the fluency of the output, however, so it needs to be used in addition to other metrics that can account for adequacy. In this study we compare the translation results on three crude measures based on the grammar checker with the results on two standard metrics: Bleu (Papineni et al., 2002), which mainly measures n-gram precision, and TER (Snover et al., 2006), which is an error rate based on a modified Levenshtein distance.[1] The scores for all metrics are calculated based on a single reference translation.

Based on the error analysis in Section 3.2. we define three grammar checker based metrics. Grammar error ratio

[1]Percent notation is used for all Bleu and TER scores.

Size	Factors	Bleu	TER	GER_1	GER_2	SGER
Large	none	22.18	66.42	0.196	0.293	0.496
	POS	21.63	66.88	0.228	0.304	0.559
	morph	22.04	66.63	0.125	0.195	0.446
Small	none	21.16	67.49	0.244	0.359	0.664
	POS	20.79	67.79	0.282	0.375	0.718
	morph	19.45	69.52	0.121	0.245	0.600

Table 3: Results for the six basic systems. Bleu have higher values for better systems, whereas TER and the grammar checker metrics are error rates, and have lower values for better systems.

1, GER_1, is the average number of the grammar errors that had a high precision on error categorization on the development set (all errors except split compounds and three verb categories), and grammar error ratio 2, GER_2, is the average number of all grammar errors per sentence. Spelling and grammar error ratio, SGER, is the average total number of identified errors per sentence. All three metrics are error rates, with the value of 0 for systems with no identified errors, and a higher value when more errors are identified. There is no upper bound on the metrics.

Table 3 shows the results of the evaluation. As expected the systems trained on the large corpus are all better than the small systems on the standard metrics. On the new grammar checker metrics the large systems are all better on SGER but not always on the GER metrics. The ranking between the six systems are the same on the two standard metrics, but differ for the Granska metrics. The *morph* system is best for the small corpus on the Granska metrics, but markedly worse on the standard metrics. We believe that this is because the small corpus is too small for the morphology to be useful, due to sparsity issues, so that the grammar is improved, but the word selection is worse. Also on the large corpus, the morph sequence model does improve the grammaticality of the output, but here it performs nearly equal to the *none* system, and better than the *POS* system on the standard metrics. Using a POS sequence model gives the worst performance on all metrics for both corpus sizes.

There are some interesting differences between the three Granska metrics. SGER gives markedly worse performance on the small corpus, which is not the case for the two GER metrics. This is mainly due to the fact that SGER is the only metric that includes spelling errors, and many of the spelling errors are English out-of-vocabulary words, which are just passed through the translation system. With a smaller training corpus, the number of out-of-vocabulary words increases. If we want the metric to give a picture of the coverage of the system, this information is clearly useful. GER_1 gives the best value for the *morph* system with the small corpus. This could be an indication that using only these few error categories makes this metric less robust to bad translations, since the grammar checker performance most likely degrades when the SMT output is too bad for it to analyze in a meaningful way. On the other hand GER_1 gives a larger difference between the *none* and *POS* systems, than GER_2, which also uses error categories with a bad performance on SMT output.

Size	Factors	Bleu	TER	Changes
Large	none	22.34	66.37	382
	POS	21.81	66.84	429
	morph	22.17	66.54	259
Small	none	21.30	67.47	456
	POS	20.95	67.75	514
	morph	19.52	69.48	249

Table 4: Results and number of changes for the six systems when Granska is used for postprocesing

5. Grammar Checker for Postprocessing

We wanted to see if the suggestions of the grammar checker could be used to improve MT by automatic postprocessing, where we apply the suggestions from the grammar checker. We have chosen to accept the correction suggestions for the categories where a large majority of the suggestion were correct in the error analysis in Section 3.2. In the test systems there were between 8 and 16 errors in categories that did not appear in the error analysis. These errors were ignored. When there were several correction suggestions for an error, we always chose the first suggestion. For most of the errors, such as agreement errors, the corrections were performed by changing one or several word forms. For other categories, the word order was changed, or words were deleted or inserted.

Table 4 shows the results for the systems with postprocessing. There are consistent but very small improvements for all systems on both metrics, compared to the unprocessed scores in Table 3. The Bleu scores had an absolute improvement by at most 0.18, and the TER scores by just 0.09 at most. As could be expected, the systems with bad scores on the Granska metrics have more corrections. More interestingly, the improvements as measured by the metrics were not generally larger on the *none* and *POS* systems than on the *morph* systems with much fewer corrections.

One of the likely reasons for the small improvements on the two metrics was that only a relatively small proportion of the sentences were affected by the postprocessing. To investigate this, we calculated scores on only the subset of sentences that were affected by the changes, shown in Table 5. The subsets of sentences are all different, so the scores can not be directly compared between the systems, or to the scores on the full test sets. Not surprisingly, the improvements are much larger on the subsets than on the full test set for all systems. The difference in the change is much larger on Bleu, with an absolute improvement of around 0.70 for the large systems and of around 0.50 for the small systems.

Size	Factors	Bleu		TER		No. sentences
		Basic	Postproc.	Basic	Postproc.	
Large	none	19.44	20.12	68.99	68.77	335
	POS	18.87	19.61	69.42	69.26	373
	morph	18.47	19.29	70.28	69.69	238
Small	none	18.72	19.26	69.96	69.88	395
	POS	17.74	18.27	71.00	70.88	452
	morph	16.79	17.24	72.01	71.77	241

Table 5: Results on the subsets of sentences where Granska postprocessing led to a change. Note that these subsets are different for the six systems, so these scores are not comparable between the systems.

Size	Factors	Good	Neutral	Bad
Large	none	73	19	8
	POS	77	17	6
	morph	68	19	13
Small	none	74	19	7
	POS	73	17	10
	morph	68	20	12

Table 6: Error analysis of the 100 first Granska-based changes for each system

The difference on TER is still relatively small, except for the *morph* systems, especially for the large system, which has a TER improvement of 0.59.

To further investigate the quality of the error corrections, an error analysis was performed by one native Swedish speaker on the first 100 error corrections for each system. The corrections were classified into three categories: good, bad, and neutral. The good changes were improvements compared to not changing anything, and the bad changes resulted in a worse translation than before the change. For the neutral category the translations before and after the change were of equal quality.

The results of the analysis are summarized in Table 6. A large majority of the changes were improvements, which indicates that the two standard metrics used are not that good in capturing this type of improvement. The corrections for the *morph* systems, are slightly worse than for the other systems. This is to a large degree due to the fact that these systems have fewer agreement errors than the other systems, and agreement errors are generally the easiest errors to correct.

Table 7 shows some examples of the different types of errors. In the good example a NP agreement error is fixed by switching the indefinite article so it gets the correct gender, and a verb with the wrong tense, present, is changed to perfect tense. In the first neutral example an NP agreement error which is mixed between definite and indefinite, is changed and becomes syntactically correct, but indefinite instead of the preferred definite. The second neutral example concerns agreement of an adjectival predicative with a collective noun, where both the original plural adjective, and the changed singular adjective are acceptable. In the first bad example an attributive adjective has been mistaken for a head noun, resulting in a change from a correct NP to an incorrect NP with two noun forms. The second bad example contains an untranslated English plural genitive noun, which are given Swedish plural inflection by Granska.

Even though the performed corrections are generally good, the number of errors with useful suggestions is low, which makes the overall effect of the corrections small. There are many more actual errors in the output, which are not found by Granska. For the postprocessing technique to be even more useful we need to be able to identify and correct more of the errors, either by modifying the grammar checker or by developing a custom SMT output checker.

6. Related Work

Automatic metrics are usually based on the matching of words in the translation hypothesis to words in one or several human reference translations in some way, as is done by Bleu (Papineni et al., 2002) and TER (Snover et al., 2006). These types of metrics do not generalize any linguistic knowledge; they only rely on the matching of surface strings. There are metrics with extended matching, e.g. Meteor (Lavie and Agarwal, 2007), which uses stemming and synonyms from WordNet, and TERp (Snover et al., 2009), which uses paraphrases. There are also some metrics that incorporate other linguistic levels, such as part-of-speech (Popović and Ney, 2009), dependency structures (Owczarzak et al., 2007), or deeper linguistic knowledge such as semantic roles and discourse representation structure (Giménez and Márquez, 2008).

Controlled language checkers, which can be viewed as a type of grammar checker, have been suggested in connection to MT, but for preprocessing of the source language, see Nyberg et al. (2003) for an overview. The controlled language checkers tend to be authoring tools, as in Mitamura (1999) and de Koning (1996) for English and Sågvall Hein (1997) for Swedish, which are used by humans before feeding a text to a usually rule-based MT system.

Automatic postprocessing has been suggested before for MT, but not by using a grammar checker. Often postprocessing has targeted specific phenomena, such as correcting English determiners (Knight and Chander, 1994), merging German compounds (Stymne, 2009), or applying word substitution (Elming, 2006). The combination of a statistical MT system and the mainly rule-based grammar checker can also be viewed as a hybrid MT system, on which there has been much research, see e.g., Thurmair (2009) for an overview. Carter and Monz (2009) discuss the issue of applying tools that are developed for human texts, in their case

Good	Original	Att ge nya befogenheter till en kommitté av ministrar främjas genom *en oansvarigt sekretariat* skulle inte *utgör* någon typ av framsteg ...
	Changed	Att ge nya befogenheter till en kommitté av ministrar främjas genom *ett oansvarigt sekretariat* skulle inte *ha utgjort* någon typ av framsteg ...
Neutral	Original	Det är viktigt att fylla *den kulturella vakuum* mellan våra två regioner
	Changed	Det är viktigt att fylla *ett kulturellt vakuum* mellan våra två regioner
	Correct	Det är viktigt att fylla *det kulturella vakuumet* mellan våra två regioner
	Original	Jag hör ibland sägas att rådet är så *engagerade* i Berlin ...
	Changed	Jag hör ibland sägas att rådet är så *engagerat* i Berlin ...
Bad	Original	Skulle det inte vara värt att ansvar på alla nivåer i *den beslutsfattande* processen tydligare, snarare än att försöka gå framåt ...
	Changed	Skulle det inte vara värt att ansvar på alla nivåer i *det beslutsfattandet* processen tydligare, snarare än att försöka gå framåt ...
	Original	Dokumentet kommer att överlämnas till europeiska rådet i Biarritz i några *days'* tid .
	Changed	Dokumentet kommer att överlämnas till europeiska rådet i Biarritz i några *daysar* tid .
	Correct	Dokumentet kommer att överlämnas till europeiska rådet i Biarritz i några *dagars* tid .

Table 7: Some examples of error corrections from the different categories

statistical parsers, on SMT output.

7. Conclusion and Future Work

We have explored the use of a grammar checker for MT, and shown that it can be useful both for evaluation and for postprocessing. A more large scale investigation with a more thorough error analysis would be useful, both for Swedish with the Granska grammar checker and for other languages and tools. In particular we want to see if the results are as useful for other architectures of the MT system and of the grammar checker, as for the combination of a statistical MT system and a mainly rule-based grammar checker. We also plan to apply grammar checkers for other languages on standard datasets such as that of the WMT shared task[2] (Callison-Burch et al., 2009), where we could correlate the grammar checker performance with human judgments and several automatic metrics for many different MT systems.

Using the grammar checker for evaluation gives a complimentary picture to the Bleu and TER metrics, since Granska accounts for fluency to a higher extent. Granska needs to be combined with some other metric to account for adequacy, however. A possibility for future research would be to combine a grammar checker and some measure of adequacy into one metric.

In the postprocessing scenario we used the grammar checker as a black box with good results. The grammar checker was, however, only able to find a small proportion of all errors present in the SMT output, since it was not developed with SMT in mind. One possibility to find more errors is to extend the rules in the grammar checker. Another possibility would be a tighter integration between the SMT system and the grammar checker. As an example, the grammar checker tags the translation output, which is error-prone. Instead part-of-speech tags from factored translation systems could be used directly by a grammar checker, without re-tagging. A third option would be to develop a new grammar checker targeted at MT errors rather than human errors, which could be either a stand-alone tool or a module in a MT system.

[2]The shared translation task at the Workshop of Statistical Machine Translation, `http://www.statmt.org/wmt09/translation-task.html`

8. References

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece.

Johan Carlberger and Viggo Kann. 1999. Implementing an efficient part-of-speech tagger. *Software Practice and Experience*, 29:815–832.

Simon Carter and Christof Monz. 2009. Parsing statistical machine translation output. In *Proceedings of the 4th Language & Technology Conference*, Poznań , Poland.

Michiel de Koning. 1996. Bringing controlled language support to the desktop. In *Proceedings of the EAMT Machine Translation Workshop*, pages 11–20, Vienna, Austria.

Rickard Domeij, Ola Knutsson, Johan Carlberger, and Viggo Kann. 2000. Granska – an efficient hybrid system for Swedish grammar checking. In *Proceedings of the 12th Nordic Conference on Computational Linguistics (Nodalida'99)*, pages 49–56, Trondheim, Norway.

Jakob Elming. 2006. Transformation-based correction of rule-based MT. In *Proceedings of the 11th Annual Conference of the European Association for Machine Translation*, pages 219–226, Oslo, Norway.

Jesús Giménez and Lluìs Márquez. 2008. A smorgasbord of features for automatic MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198, Columbus, Ohio, USA.

Viggo Kann, Rickard Domeij, Joachim Hollman, and Mikael Tillenius. 2001. Implementation aspects and applications of a spelling correction algorithm. *Quantitative Linguistics*, 60:108–123.

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *12th National conference*

of the American Association for Artificial Intelligence, Seattle, Washington, USA.

Ola Knutsson. 2001. *Automatisk språkgranskning av svensk text*. Licentiate thesis, Royal Institute of Technology (KTH), Stockholm, Sweden.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural language Processing and Computational Natural Language Learning*, pages 868–876, Prague, Czech Republic.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL, demonstration session*, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic.

Teruko Mitamura. 1999. Controlled language for multilingual machine translation. In *Proceedings of MT Summit VII*, pages 46–52, Singapore, Singapore.

Eric Nyberg, Teruko Mitamura, and Willem-Olaf Huijsen. 2003. Controlled language for authoring and translation. In Harold Somers, editor, *Computers and Translation. A translator's guide*, pages 245–282. John Benjamins, Amsterdam.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 42nd Annual Meeting of the ACL*, pages 160–167, Sapporo, Japan.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Dependency-based automatic evaluation for machine translation. In *Proceedings of the Workshop on Syntax and Structure in Statistical Translation*, pages 80–87, Rochester, New York, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, Pennsylvania, USA.

Maja Popović and Hermann Ney. 2009. Syntax-oriented evaluation measures for machine translation output. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 29–32, Athens, Greece.

Anna Sågvall Hein. 1997. Language control and machine translation. In *Proceedings of the 7th International Conference of Theoretical and Methodological Issues in Machine Translation*, pages 103–110, Santa Fe, New Mexico, USA.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human notation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece.

Sylvana Sofkova Hashemi. 2007. Ambiguity resolution by reordering rules in text containing errors. In *Proceedings of 10th International Conference on Parsing Technologies*, pages 69–79, Prague, Czech Republic.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the Seventh International Conference on Spoken Language Processing*, pages 901–904, Denver, Colorado, USA.

Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2008. Effects of morphological analysis in translation between German and English. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 135–138, Columbus, Ohio, USA.

Sara Stymne. 2009. A comparison of merging strategies for translation of German compounds. In *Proceedings of the EACL 2009 Student Research Workshop*, pages 61–69, Athens, Greece.

Gregor Thurmair. 2009. Comparing different architectures of hybrid machine translation systems. In *Proceedings of MT Summit XII*, pages 340–347, Ottawa, Ontario, Canada.