An Evaluation of Post-processing Google Translations with Microsoft® Word

Håkan Burden, Jones Belhaj, Magnus Bergqvist, Joakim Gross, Kristofer Hansson Aspman, Ali Issa, Kristoffer Morsing, Quishi Wang

Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden
burden@chalmers.se

1. Introduction

A common problem with statistical machine translation (SMT) is that the candidate translations are often extra-grammatical. Recent research has tried to improve on the quality of the translations made by SMT systems by post-processing the candidate translations using a grammar checker (Huet et al., 2010; Stymne and Ahrenberg, 2010). The SMT systems in these studies require training on parallel corpora, while most end-users want a general-purpose translator and neither have the necessary knowledge nor a representative corpora for training an SMT system. Thus the popularity of systems such as Google Translate.

2. Evaluation through Replication

We decided to evaluate the performance of Google Translate and the possible improvements on grammatical fluency through post-processing the candidate translations by Microsoft® Word, replicating previous research done by Stymne and Ahrenberg (2010).

2.1 Replicated Study

Stymne and Ahrenberg (2010) evaluate the impact of post-processing SMT candidate translations by first training Moses (Koehn et al., 2007) on 701 157 English-Swedish sentence pairs taken from the EuroParl corpus (Koehn, 2005). The resulting SMT system was then evaluated on 2 000 sentences of EuroParl. The translations were post-processed by Granska (Carlberger et al., 2004), a grammar checker for Swedish. If there were more than one possible correction according to Granska the first was always chosen. The impact of the grammar checker was evaluated by both an automatic analysis using BLEU (Papineni et al., 2002) as well as a manual inspection of the first 100 suggested corrections made by Granska.

2.2 Replication Setup

In our replication we chose to use Google Translate instead of Moses since we wanted an SMT system that did not require any training before usage. As grammar checker we chose Microsoft® Word 2010 (MS Word) since this is a widely used word processor. If there were more than one possible grammatical correction for a candidate translation the first was always chosen. In the case a sentence was high-lighted as extra-grammatical but there were no suggestions on how to correct the translation it was left unchanged. We used the same 2 000 sentences for evaluation as Stymne and Ahrenberg. The BLEU score was calculated by using iBLEU (Madnani, 2011). In our evaluation we went one step further than the replicated study by asking a human translator to analyse the candidate translations as well as the suggested grammatical corrections.

3. Results

3.1 BLEU Scores

In Table 1 the BLEU scores from the original study are given together with the scores from our replication. The last two rows show the BLEU scores for the subset of candidate translations that were corrected by a grammar checker. In both cases Google Translate outperforms the results from Stymne and Ahrenberg with a 35%-increase in BLEU-score. The lower BLEU-scores for the corrected translations might be explained by the fact that these sets only contain a 100 sentences each while BLEU is more reliable for larger evaluation sets (Owczarzak et al., 2007).

3.2 Manual Inspection

The first 100 candidate translations that had a possible correction according to MS Word were manually inspected by the authors and graded. The three possible grades were 'Good', 'Neutral' and 'Bad', the same as in the replicated study. An example of how each grade was used to evaluate the corrections is given below.

MS Word made the suggestion to change the definite noun *utmaningen*, meaning *the challenge*, to the indefinite form *utmaning* in: *Kommissionens utmaningen blir att övertyga parlamentet att den kan skapa dessa garantier*. The suggested change was graded as 'Good' since the *-s* in *Kommissionens*, meaning *the commission's*, marks genitive and for Swedish the rule is that any subsequent noun or noun phrase should use the indefinite form.

An example of a correction that is 'Bad' is given in *Istället är det mer logiskt om varje pigfarmer sätter upp sin egen reserv, d.v.s om alla pigfarmer har sin egen spargris.* Google Translate does not recognise the word *pigfarmer* and transfers it as it is into the Swedish translation. In Swedish *farm* is an English loanword with the plural ending *-er*. MS Word identifies that *pigfarmer* is a compound noun in plural with *farm* as its head. But *varje*, meaning *each*, should be followed by a noun in singular so the grammar checker suggests a correction from the plural form *pigfarmer* into the singular *pigfarm*. Instead of correcting agreement MS Word changes the meaning of the sentence.

In the following translation the underlined *vill*, present and indicative form of *want*, is superfluous; *Och jag vill*

SMT	BLEU	Gram. Check.		
		System	BLEU	Change
Moses	22.18	Granska	22.34	0.16 (0.7%)
Google	29.95	MS Word	29.99	0.04 (0.1%)
Moses	19.44	Granska	20.12	0.68 (3.5%)
Google	23.90	MS Word	24.28	0.38 (1.6%)

Table 1: The BLEU scores for the different systems.

SMT	Gram. check.	Good	Neutral	Bad
Moses	Granska	73	8	19
Google	MS Word	76	3	21

Table 2: The outcome of the manual evaluation of the proposed grammar corrections.

än en gång vill uppriktigt tacka mina kollegor i utskottet för deras samarbete. The suggestion made by MS Word is to replace *vill* with the infinitive form *vilja*. Since changing the word form neither improves nor worsens the fluency the correction was labeled as 'Neutral'.

Just as the above examples suggest, the grammar corrections concerned agreement between adjacent words. In Table 2 the evaluation is presented together with the figures reported by Stymne and Ahrenberg. Since the manual inspections are conducted by different authors the figures are not comparable.

3.3 Analysis by Human Translator

We asked a professional translator between English and Swedish to analyse the translations and the grammatical corrections: *When evaluating the performance of the translator or the grammar checker it is easy to miss the bigger picture. Preserving the intentions of the source text is more than agreement between subject and verb.*

In fact, small improvements as agreement do not make up for the increase in human effort needed to ensure that the grammar checker does not get it wrong. The grammar checker adds a new layer of uncertainty on top of the machine translator's approximation of a translation. The result is that we no longer know where problematic sentences arose. They could be the result of a poor translation by the machine translator, the grammar checker getting it wrong or a combination of the both. Look at the 'Bad' example above. The correction made by the grammar checker hides that pigfarmer *was unknown to the machine translator. A human translator working on the post-processed text could easily miss the mis-interpretation.*

Most importantly, you should never guess! If you are in doubt on how to translate a text you should always get in touch with the customer. Getting it wrong means both a loss of customers and reputation.

4. Discussion

It does not seem that the impact of post-processing the candidate translations with a grammar checker is captured by the BLEU-metrics. Three out of four suggested changes improve the fluency of the translations but for these sentences the increase in BLEU is in our case less than 2%. Our interpretation is supported by the results of Stymne and Ahrenberg as well as by a similar study done by Huet et al. (2010). The latter had an increase from 27.5 on the BLEU-scale to 28.0 after applying a sequence of different post-processing techniques, among them grammar correction.

5. Acknowledgements

The authors would like to thank Carl-Magnus Olsson, Malmö University, and Peter Ljunglöf, University of Gothenburg.

6. References

Johan Carlberger, Rickard Domeij, Viggo Kann, and Ola Knutsson. 2004. The development and performance of a grammar checker for Swedish : A language engineering perspective. *Natural Language Engineering*, 1(1).

Stéphane Huet, Julien Bourdaillet, Alexandre Patry, and Philippe Langlais. 2010. The RALI Machine Translation System for WMT2010. In *ACL 2010 Joint 5th Workshop on Statistical Machine Translation and Metrics MATR*, pages 109–115, Sweden, Uppsala.

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 *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand. Asia-Pacific Association for Machine Translation.

Nitin Madnani. 2011. iBLEU: Interactively Debugging and Scoring Statistical Machine Translation Systems. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 213 –214, September.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Dependency-Based Automatic Evaluation for Machine Translation. In *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 80–87, Rochester, New York, April. Association for Computational Linguistics.

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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sara Stymne and Lars Ahrenberg. 2010. Using a Grammar Checker for Evaluation and Postprocessing of Statistical Machine Translation. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, May.