Erratum for
Dual-Path Phrase-Based Statistical Machine Translation

Susan Howlett and Mark Dras
Centre for Language Technology
Macquarie University
Sydney, Australia
susan.howlett@students.mq.edu.au, mark.dras@mq.edu.au

December 6, 2011

We have discovered an error in the experiment run for one of the systems reported in Howlett and Dras (2010). This document first outlines the effects of the error on the argument of the paper, then gives a complete list of the corrections that should be applied. We have published the configuration files used with the reruns below, along with the erratum, at `http://www.showlett.id.au`.

Outline of effects

In Howlett and Dras (2010), we introduced a lattice input to a phrase-based SMT system. The lattice encodes two versions of a sentence, with and without a reordering preprocessing step, which allows the PSMT system to choose whether the reordering is useful for the given sentence. We compared the performance of this system, both plain (LATTICE) and with extra features (+FEATURES), to a plain PSMT baseline (MOSES) and a reordering baseline (REORDER), which were respectively trained on only the original and only the reordered sentences. We also included an oracle experiment that consulted the reference translation to select the better of the two baselines for each sentence, to provide an approximate upper bound.

We have now found that an error occurred during tuning of the reordering baseline system, causing it to converge to a different point and produce a worse overall result. Specifically, the training of each system involved jobs distributed across a cluster, and in the case of REORDER, it appears that in one iteration of the tuning phase, the n-best list of translation candidates was slow in returning from one of the ten splits, and was not made available for the next round of tuning. The error was not flagged by the translation system, and was only uncovered by a manual inspection of the intermediate files after we had failed to replicate the result in a later experiment. We expect that the error would have been caught if we had done repeated runs of each system at the time, and as such we heartily endorse the suggestion by Clark et al. (2011) that replication be carried out as a standard procedure.

We have now rerun all five systems from the paper twice. Results are given in Table 1, along with the results originally reported. Note that the only systems substantially affected are REORDER and the approximate oracle which relied upon its output.

	BLEU score			
System	Original run	Rerun 1	Rerun 2	Rerun average
MOSES	20.77	20.59	20.83	20.71
REORDER	20.04	20.87	20.99	20.93
Approx. oracle	22.45	22.70	23.00	22.85
LATTICE	21.39	21.30	21.36	21.33
+FEATURES	21.10	21.10	21.23	21.16

Table 1: Corrected results for all systems. This should replace Table 3 of the original paper. The "original run" column gives the results reported in the original paper; "rerun 1" and "rerun 2" give the results from running the systems again. The final column gives the average score achieved in the two reruns.

With the updated results, we find that REORDER outperforms MOSES by a narrow margin, the lattice with extra features outperforms REORDER by a similar amount, and the plain lattice system is better again, by a similar amount. A large gap remains between the lattice system and the oracle result, suggesting room for improvement with a better selection of features.

List of corrections

- Abstract, sentence 4:

  > In German-to-English translation, our best system achieves a BLEU score of 21.39, an improvement of 0.62.

  should become

  > In German-to-English translation, our best system achieves an average BLEU score of 21.33, an improvement of 0.40 over the reordering baseline average.

- Section 1, paragraph 6, sentence 1:

  > Our results (§5) do not replicate the finding of Collins et al. (2005) that the preprocessing step produces better translation results overall.

  should become

  > Our results (§5) support the finding of Collins et al. (2005) that the preprocessing step produces better translation results overall.

- Section 1, paragraph 6, sentence 2:

  > However, results for our dual-path PSMT system do show an improvement, with our plain system achieving a BLEU score (Papineni et al., 2002) of 21.39, an increase of 0.62 over the baseline.

  should become

  > However, results for our dual-path PSMT system show a further improvement, with our plain system achieving an average BLEU score (Papineni et al.,

2002) of 21.33, an increase of 0.62 over the baseline average and 0.40 over the reordering baseline average.

- Table 3 should be replaced by Table 1 of this document.
- Section 5, paragraph 2, sentence 1:

  > Interestingly, our reimplementation of the Collins et al. (2005) baseline does not outperform the plain PSMT baseline.

  should become

  > Interestingly, our reimplementation of the Collins et al. (2005) baseline achieves only a small improvement over the plain PSMT baseline.

- Section 5, paragraph 2, sentence 3:

  > It may also be that the inconsistency of improvement noted by Collins et al. (2005) is the cause; sometimes the reordering produces better results and sometimes the baseline, with the effect just by chance favouring the baseline here.

  should become

  > It may also be that the inconsistency of improvement noted by Collins et al. (2005) is the cause; sometimes the reordering produces better results and sometimes the baseline, with the effect just by chance favouring the baseline to a greater extent here.

- Section 5, paragraph 3, sentences 1–2:

  > In our experiment, the oracle preferred the baseline output in 848 cases and the reordered in 1,070 cases. 215 sentences were identical between the two systems, while in 392 cases the sentences differed but had equal numbers of n-gram overlaps.

  should become

  > In the two runs of our experiment (run1;run2), the oracle preferred the baseline output in (863;929) cases and the reordered in (984;968) cases. (248;226) sentences were identical between the two systems, while in (430;402) cases the sentences differed but had equal numbers of n-gram overlaps.

- Section 5, paragraph 3, sentence 3:

  > The BLEU score for the oracle is higher than that of both baselines; from this and the distribution of the oracle's choices, we conclude that the difference between our findings and those of Collins et al. (2005) is at least partly due to the inconsistency that they identified.

  should become

> The BLEU score for the oracle is higher than that of both baselines; from this and the distribution of the oracle's choices, we conclude that the difference in magnitude between our findings and those of Collins et al. (2005) could be at least partly due to the inconsistency that they identified.

- Section 5, paragraph 3, sentence 4:

  > It is especially interesting to note that the reordered system's translations are preferred by the oracle more often even though its overall performance is lower.

  should be deleted.

- Section 6, paragraph 1, sentence 2:

  > We find that providing the system with this choice results in improved translation performance, achieving a BLEU score of 21.39, 0.62 higher than the baseline.

  should become

  > We find that providing the system with this choice results in improved translation performance, achieving an average BLEU score of 21.33, 0.62 higher than the average baseline and 0.40 higher than the average reordering baseline.

- Section 6, paragraph 3, sentence 1:

  > While our reordering step is a reimplementation of the Collins et al. (2005) system, contrary to their findings we do not see an improvement using the reordering step alone.

  should become

  > While our reordering step is a reimplementation of the Collins et al. (2005) system, we see a smaller improvement than they reported using the reordering step alone.

- Section 6, paragraph 3, sentence 2:

  > This provides evidence against the idea that reordering improves translation performance absolutely.

  should be deleted.

- Section 6, paragraph 3, sentence 3:

  > However, our success with the lattice system highlights the fact that it *is* useful for some sentences, and that syntactic confidence features may provide a mechanism for identifying which sentences, thus incorporating syntactic information into phrase-based statistical machine translation in a useful way.

  should become

However, the oracle's selections and our success with the lattice system highlight the fact that the reordering *is* useful for some sentences, and that syntactic confidence features may provide a mechanism for identifying which sentences, thus incorporating syntactic information into phrase-based statistical machine translation in a useful way.

References

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, 2011.

Michael Collins, Philipp Koehn, and Ivona Kucerova. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 531–540, 2005.

Susan Howlett and Mark Dras. Dual-path phrase-based statistical machine translation. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 32–40, 2010.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.