Erratum for
Clause Restructuring for SMT Not Absolutely Helpful

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 configuration files for some of the experiments that we reported in Howlett and Dras (2011). 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. The corrected configuration files have been published with this erratum at `http://www.showlett.id.au`.

Outline of effects

In Howlett and Dras (2011), we systematically varied several settings of a reordering-as-preprocessing phrase-based SMT system to explore under what conditions it might underperform relative to the baseline PSMT system. This was motivated by our earlier work (Howlett and Dras, 2010) where we had seen precisely this outcome.

We have identified an error in the configuration files for the Howlett and Dras (2011) experiments evaluating on the news test set (`newstest2009`). This error caused the reordered systems to decode and evaluate on the original test set rather than the reordered version, and thus underperform on this test set. Table 1 gives the corrected results table, which should replace Table 5 of the original paper. The final three columns are the only ones affected; the first four columns remain unchanged.

This error is responsible for the dramatic data effect that we reported. Contrary to our original findings, in the corrected results, the reordered system always outperforms the baseline system. In fact, we find that in many cases the difference between baseline and reordered systems is greater on the news test set than on the Europarl test set, which brings the results closer to agreement with Xia and McCord's (2004) findings.

With the increase in BLEU score of the reordered systems, the oracle scores are correspondingly increased. Thus, as in the original paper, the oracle always outperforms both baseline and reordered systems by a substantial margin, demonstrating that each system is providing better translations for some sentences. To further substantiate this, Table 2 lists the number of times that the oracle selects the baseline and reordered systems for each experiment. These figures clearly show that the baseline system output is being selected in a significant number of cases.

LM	DM	T	Base.	Reord.	Diff.	Oracle
3	dist	–	16.28	*17.27*	*+0.99*	*18.52*
		E	16.43	*17.52*	*+1.09*	*18.92*
		N	17.25	*17.87*	*+0.62*	*19.64*
	lex	–	16.81	*17.46*	*+0.65*	*18.90*
		E	16.75	*17.07*	*+0.32*	*18.88*
		N	17.75	*18.31*	*+0.56*	*19.79*
5	dist	–	16.44	*17.43*	*+0.99*	*18.75*
		E	16.21	*16.83*	*+0.62*	*18.48*
		N	17.27	*18.31*	*+1.04*	*19.64*
	lex	–	17.10	*17.66*	*+0.56*	*19.24*
		E	17.03	*17.58*	*+0.55*	*19.37*
		N	17.73	*18.26*	*+0.53*	*20.01*

Table 1: Corrected results on news test set. This should replace Table 5 of the original paper. Changed entries are italicised. Columns give: language model order, distortion model (distance, lexicalised), tuning data (none (–), Europarl, News), baseline BLEU score, reordered system BLEU score, performance increase, oracle BLEU score.

Our other findings remain unchanged. Our observation about the difference in scores reported by the NIST BLEU scorer and the Moses multi-reference BLEU script (multi-bleu) still holds. That is, the NIST scores were always lower than multi-bleu's on test2008 and higher on newstest2009, by a margin of at most 0.23.

Also, it is still the case that all of the factors tested can affect the reordered system's performance. That is, many settings within the translation system, independent of the reordering process, may erode overall performance gains.

Finally, we note that this correction means that we have no longer managed to reproduce the result in our earlier paper (Howlett and Dras, 2010) that motivated this investigation. We have found that the earlier result is also in error (with the BLEU score for the reordered system out by approximately 0.9), but for an entirely unrelated reason. We are publishing a separate erratum for that work.

With these changes, we stand behind our headline claim that the clause restructuring or reordering-as-preprocessing approach to SMT is not absolutely helpful, but with a slight change of emphasis. Reordering can provide overall improvements, but factors unrelated to the reordering process may erode performance gains, and oracle experiments demonstrate that the baseline output is still to be preferred in a substantial number of cases.

List of corrections

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

> We see that the choice of data can have a profound effect, nullifying or even reversing the overall result, even when the reordering system remains the same. Genre differences are an obvious possibility, but we have demonstrated only a dependence on data set.

			Europarl test set (2000 sents)			News test set (2525 sents)		
LM	DM	T	Neither	Baseline	Reordered	Neither	Baseline	Reordered
3	dist	–	465	636	899	662	758	1105
		E	398	673	929	600	825	1100
		N	355	830	815	540	902	1083
	lex	–	472	662	866	656	803	1066
		E	430	746	824	547	941	1037
		N	455	702	843	640	841	1044
5	dist	–	468	654	878	663	791	1071
		E	407	759	834	581	901	1043
		N	434	656	910	662	816	1047
	lex	–	493	681	826	641	836	1048
		E	411	749	840	520	892	1113
		N	385	750	865	579	918	1028

Table 2: Selections made by the oracle (neither preferred, baseline output preferred, reordered system output preferred) on the Europarl (test2008) and news (newstest2009) test sets.

should be deleted.

- Section 5, paragraph 4, sentence 1:

  > The other factors tested—language model order, lexicalisation of the distortion model, and use of a tuning phase—can all affect the overall performance gain of the reordered system, but less distinctly.

  should become

  > The various factors tested—language model order, lexicalisation of the distortion model, and use of a tuning phase—can all affect the overall performance gain of the reordered system.

- Section 5, paragraph 5, sentence 2:

  > Its [The oracle's] selections show that, in changing test sets, the balance shifts from one system to the other, but both still contribute strongly.

  should become

  > Its selections show that both systems contribute strongly.

- Section 6, paragraph 2, sentence 1:

  > We have systematically varied several aspects of the Howlett and Dras (2010) system and reproduced results close to both papers, plus a full range in between.

  should become

  > We have systematically varied several aspects of the Howlett and Dras (2010) system and reproduced results close to Collins et al. (2005). However, our results do not correspond to those originally reported in Howlett and Dras

(2010), where the baseline system outperformed the reordered system, casting doubt on that result.

- Section 6, paragraph 2, sentences 2–3:

  > Our results show that choices in the PSMT system can completely erode potential gains of the reordering preprocessing step, with the largest effect due to simple choice of data. We have shown that a lack of overall improvement using reordering-as-preprocessing need not be due to the usual suspects, language pair and reordering process.

  should become

  > Our results show that choices in the PSMT system can completely erode potential gains of the reordering preprocessing step and that a lack of overall improvement using reordering-as-preprocessing need not be due to the usual suspects, language pair and reordering process.

References

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.

Susan Howlett and Mark Dras. Clause restructuring for SMT not absolutely helpful. In *Proceedings of the 49th Annual Meeting of the Assocation for Computational Linguistics: Human Language Technologies*, pages 384–388, 2011.

Fei Xia and Michael McCord. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 508–514, 2004.