Can Statistical Post-Editing with a Small Parallel Corpus Save aWeak MT Engine?

Marianna J. Martindale

Center for Applied Machine Translation
U.S. Dept. of Defense, Ft. Meade, MD
marianna.j.martindale@ugov.gov

Abstract

Statistical post-editing has been shown in several studies to increase BLEU score for rule-based MT systems. However, previous studies have relied solely on BLEU and have not conducted further study to determine whether those gains indicated an increase in quality or in score alone. In this work we conduct a human evaluation of statistical post-edited output from a weak rule-based MT system, comparing the results with the output of the original rule-based system and a phrase-based statistical MT system trained on the same data. We show that for this weak rule-based system, despite significant BLEU score increases, human evaluators prefer the output of the original system. While this is not a generally conclusive condemnation of statistical post-editing, this result does cast doubt on the efficacy of statistical post-editing for weak MT systems and on the reliability of BLEU score for comparison between weak rule-based and hybrid systems built from them.

Keywords: machine translation, machine translation evaluation, statistical post-editing

1. Introduction

Statistical Machine Translation relies on large aligned parallel corpora to produce acceptable results. Resources such as the Europarl corpus (Koehn, 2005) are considered a bare minimum with tens of millions of words (up to 55M) and in the constrained track of the NIST MT-09 evaluation approximately 200M words of parallel Arabic-English data was available. State of the art systems may use considerably more data to maximize results, with BLEU score improvements as the corpus size increases beyond the hundreds of millions and even billions of words (Brants et al., 2007).

For most less-commonly taught languages, parallel text does not exist and is difficult and expensive to create. Where efforts have been undertaken to obtain or create parallel corpora for these languages, the resulting corpora may be only a fraction of the size of those expected for training SMT systems, such as the parallel corpora in the "language packs" created by Simpson et al. (2008). The largest of these are about 2M words but most are in the 200-500k word range—much smaller than the accepted minimum for training a general purpose SMT system from scratch. One possible solution is to create or use existing weak rule-based systems and use the limited available parallel data to improve them through techniques such as statistical post-editing.

Statistical post-editing has recently been shown to improve scores on automated metrics for mature rule-based MT engines with relatively small parallel corpora (e.g., Dugast et al. (2007), Simard et al. (2007a), Simard et al. (2007b)), and Voss et al. (2008) and de Ilarraza et al. (2008) used statistical post-editing for weak MT engines for low-resource languages with similarly successful results. These systems are a type of hybrid, treating the output of a rule-based system as the source language for a statistical system. The statistical system is trained on output from the rule-based system aligned with reference translations of the original source text.

Because the SMT step is an n-gram based process, it is not surprising that it yields vast improvements on the n-gram based BLEU metric. BLEU has been demonstrated to be capable of producing scores that do not correlate with human judgments, with potential variability increasing when translation quality is low (Callison-Burch et al., 2006). Given this variability and the inherent bias of BLEU, can we trust the increase in BLEU score to indicate an overall improvement?

To answer this question, we compared a baseline lexicon transfer-based MT system with statistical and statistical post-editing systems built from various sizes of parallel data as described in section 2. The systems were evaluated using BLEU score and human evaluation as described in section 3. Section 4 gives the results of the evaluation, and section 5 discusses the implications on future research.

2. MT Systems

For this experiment, we chose a language for which sufficient parallel text is available to build a minimal statistical system but that is not among the most frequently demonstrated languages, namely, Czech. We used the existing Czech capability in the baseline system and built new statistical and statistical post-editing systems using off-the-shelf tools as described below.

2.1. Baseline System (GST)

The baseline translation system was Gister (GST). Gister is a U.S. Government-produced lexical transfer-based MT system from the CyberTrans MT package produced by the U.S. Department of Defense's Center for Applied Machine Translation (CAMT). CyberTrans provides automated tools for translation, language and encoding identification, pre- and post-processing, and related language tools for the U.S.

Sentences	Czech	English	Gister
1k	12.8k	14.5k	13.2k
5k	66.7k	75.5k	69.4k
10k	146k	173k	152k
15k	210k	252k	219k
25k	297k	358k	310k
50k	422k	495k	441k
75k	539k	621k	565k
100k	700k	803k	733k
200k	1.51M	1.72M	1.57M
alldata	5.57M	6.41M	5.85M

Table 1: Number of words in the Czech, English, and Gister output for each training set by number of sentences

Government. The standard distribution of CyberTrans contains two U.S. Government-produced systems: Gister and a newer, more advanced, system, MoTrans. Gister provides neither morphological nor syntacic parsing and performs no reordering. Morphological complexity is crudely handled using wildcards or regular expressions and reordering is either ignored or selectively hard-coded using phrasal lexicon entries. The majority of languages available in the standard distribution have not yet been upgraded to MoTrans, especially those languages for which CAMT is particularly lacking in resources or for which there is less customer demand. Czech is one of the languages that is only available through the Gister translation engine.

Gister provides translation from over 65 languages into English, most of which are low-resource languages. The largest lexicons have hundreds of thousands of entries and the smallest have less than 20k. The Czech lexicon used in this experiment contained over 68k words and phrases and used wildcards to greedily match over affixes.

2.2. Statistical System (SMT)

The statistical (SMT) and statistical post-editing (SPE) systems were built using off-the-shelf toolkits SRILM (Stolke, 2002) and Moses (Koehn et al., 2007). Due to version and resource constraints, we used basic phrase-based settings with unfactored models. The SMT system was trained on the one-to-one aligned sentences of the CZeng 0.7 corpus (Bojar and Žabortský, 2006) excluding the Acquis Communautaire portion. The remaining 410k sentences provide 5.6M words of parallel text—twice as much as the largest of the language packs. To see how the systems would perform with various sizes of training data, we randomly selected 10 training sets of increasing size beginning with 1,000 sentences up to all 410k sentences with 5,000 sentences reserved for evaluation. The data sets are summarized in Table 1. Moses was trained on each training set for a total of 10 SMT systems.

2.3. Statistical Post-Editing System (SPE)

To build the statistical post-editing system, Gister was first used to translate the Czech sentences from the same training data used in the SMT system. Moses was then trained on the Gister output for each training set and the corresponding English sentences, yielding a "Gister-to-English" translator for a total of 10 statistical post-editors.

3. Evaluation

The 5,000 evaluation sentences were translated using each of the 21 translation systems. BLEU (Papineni et al., 2002) scores were calculated for the output of each system and a human evaluation was conducted.

3.1. BLEU Score Evaluation

To calculate BLEU scores, the 5,000 test sentences were treated as a document and scored using a Java port of NIST mteval v11[1] integrated into the CyberTrans suite.

3.2. Human Evaluation

For the human evaluation, 33 users of the CyberTrans Machine Translation suite were asked to judge randomly selected sentences based on how they typically use the system. CyberTrans is intended to provide translations sufficient to perform topic identification and filtering to determine if text is worth passing on to human translator for proper translation. The average CyberTrans user is familiar with the domain of the text being translated but at most minimally familiar with the language.

The evaluators were shown the source sentence, the reference translation, and the unlabeled translations from Gister (GST), statistical post-editing (SPE) and statistical MT (SMT) in random order where the SPE and SMT were trained on the same size data. They were given the option to mark a sentence as misaligned and move on to another sentence. Sentences marked as misaligned were disregarded. For each translation the evaluators were asked if the translation was sufficient for filtering and selection—if the translation was "Good Enough". The percent of sentences labeled as "Good Enough" for each system was the GoodEnough score. For each sentence they were also asked to rank the three translations from best to worst. A strict ordering was not enforced, making the ranking essentially a score on a three-point scale, but encouraging scoring relative to each other rather than relative to a subjective ideal. The Rank score for each system was calculated by normalizing the average rank according to the following formula:

$$Rank = 1 - \frac{\sum \text{rank}(n)}{3 * n}$$

4. Results

As in previous work, statistical post-editing led to a dramatic increase in BLEU score. However, the results of the human evaluation indicate that in this case the increased BLEU score may not imply a corresponding improvement in quality.

4.1. BLEU Score Results

As expected, the Gister output had a very low BLEU score (0.111). The first SMT system trained on only 1k sentences had an even lower BLEU score (0.074) and the 1k SPE already showed BLEU score improvement (0.130). The learning curves for the systems are shown in Figure 1. The results show a similar pattern to the results in Simard et

[1] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-11b.pl

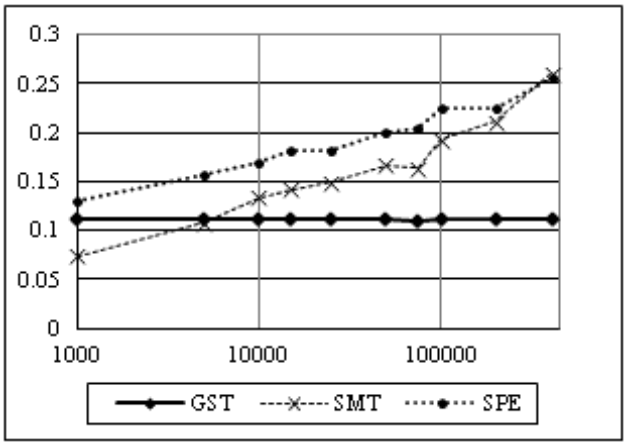

Figure 1: BLEU score vs training data size (in sentences).

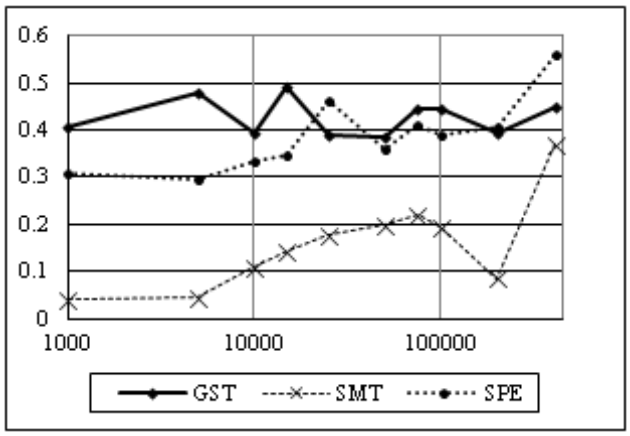

Figure 2: GoodEnough score vs training data size.

al (2007b): the statistical post-editing gives an immediate win over the rule-based system and even over the SMT, although in this case where the rule-based system is not a mature system it takes less data for the SMT to overtake the rule-based system.

4.2. Human Evaluation Results

The learning curves as scored in the human evaluation are shown in Figures 2 and 3. Only SPE with the full data set was able to surpass 50% GoodEnough, which is far below the performance we would hope to see. Surprisingly, the SMT had the lowest GoodEnough score—not even reaching 25% until the full data set was used. SPE performed similarly to Gister as measured by the GoodEnough score. Although the Gister translations remained the same, Gister's GoodEnough scores fluctuated between 39-49%. This is most likely due to sampling error. Although the inter-annotator agreement for Gister across data sizes was reasonable as calculated following Carletta (1996) yielding a Kappa of 0.734, the sentences for evaluation were drawn from too large a pool for too few evaluators so the majority of sentences were not evaluated at all and many were evaluated for only one of the data sizes. This allowed the per sentence quality variability to have a greater effect than it should have. The GoodEnough score was more affected by this problem than the Rank score because it was particularly subjective and perhaps not defined well enough.

In the Rank score, SPE performed slightly worse than Gister until the full data set was used. SMT was by far the least favorite, improving until the full data set was used when it surpassed Gister. One predicts that with more data the SMT would countinue to improve as observed in previous studies such as Brants et al. (2007).

5. Conclusions and Future Work

Examining the output of the systems reveals that each system had some strengths and weaknesses. Table 2 shows some examples of output from some of the systems in which at least one of the outputs was considered "Good Enough".

The Gister output was considered "Good Enough" and ranked highest on segments where nearly all of the words

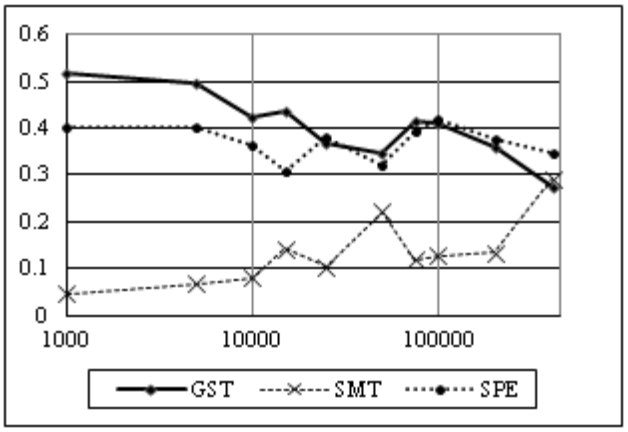

Figure 3: Rank score vs training data size.

translated and where the SPE introduced errors and where the SMT translated fewer words and/or introduced additional error. In the first example, the SPE trained on 15k introduced error by changing *it is needed disengage* to *need to liquid*. Although the original word choice was highly disfluent, the replacement is still disfluent and further from the meaning. The SMT failed to translate three of the content words—not surprising, given that it was trained on such a small data set—effectively removing all connection to the meaning of the source sentence. In the second sentence, both the SMT and SPE omitted a critical negation. Even with the full training data some of these problems still occurred, as in the third example.

The SPE successfully improved on Gister and outranked the SMT when Gister translated most of the words but the translation was inaccurate or disfluent. In the fourth example, even with only 15k sentences of training data the SPE was able to replace confusing, disfluent phrases such as *protection consumption* with more appropriate phrases such as *consumer protection*). The fifth example demonstrates that SPE was sometimes successful in disambiguating when Gister provided a slashed gloss for polysemous words.

The SMT output was considered "Good Enough" and ranked highest on segments where the SMT was able to

Systems	Outputs
(1) Source:	Potlačovanou energii je třeba uvolnit, ale pomalu a obezřetně.
Gister:	Oppressed energy it is needed disengage, but slowly and guardedly.
SMT15k:	Potlačovanou energii need to gain my but slowly and obezřetně.
SPE15k:	Oppressed energy need to liquid, but slowly and guardedly.
Reference:	Pent-up energy had to be released, but cautiously.
(2) Source:	Má hnědé oči a není vysoká.
Gister:	Has brown eyes and isn't high.
SMT25k:	Má hnědé eyes and is high.
SPE25k:	Has brown eyes and is high.
Reference:	She has brown eyes, and she's not tall.
(3) Source:	Každý držitel úřadu musí prokázat, zda je více tajemníkem než generálem.
Gister:	Every owner office must demonstrate, whether more secretary than/before General.
SMTall:	Každý american office must demonstrate whether or not it is more secretary than general.
SPEall:	if europe's Every must demonstrate, whether or not it is more secretary than General.
Reference:	Each holder of the office must demonstrate whether he is more Secretary than General.
(4) Source:	V politikách Unie je zajištěna vysoká úroveň ochrany spotřebitele.
Gister:	in politician/politics union is security high level protection consumption.
SMT15k:	V policies Unie is zajištěna 2. members of spotřebitele.
SPE15k:	in the policies of the union is ensured a high level of consumer protection.
Reference:	Union policies shall ensure a high level of consumer protection.
(5) Source:	Použít jako výchozí kalendář
Gister:	use as/like departure/primal calendar/diary
SMT15k:	Použít as default calendar
SPE15k:	use as a default calendar
Reference:	Use as Default Calendar
(6) Source:	Táta uměl vyčíst z našich tváří, že se něco děje.
Gister:	TÁTA artificial VYČÍST from of our face, that something action.
SMT15k:	Táta might vyčíst from our face that something was going on.
SPE15k:	TÁTA could VYČÍST from our faces, something that happens.
Reference:	Dad could read signs of trouble on our faces.
(7) Source:	Najednou gorila vydá hrozivý bojovný křik.
Gister:	All together gorilla VYDÁ ominous/imminent fighting clamour.
SMT15k:	it shall Najednou gorilla aggressive raises a scream.
SPE15k:	All together VYDÁ terrible fighting the gorillas.
Reference:	Suddenly the gorilla roars a ferocious battle cry.

Table 2: Sample output from some of the systems for seven segments from the test corpus. The Gister translations of segments 1—3 were rated "Good Enough" and Gister ranked highest. Likewise, translations 4—5 for SMT and 6—7 for SPE were rated "Good Enough" with their respective engine ranked highest.

translate most of the words and the Gister output was highly disfluent. In the sixth example, the SMT trained on only 15k sentences was able to capture the meaning of most of the sentence in a fairly fluent manner. Gister, on the other hand, provided a translation that bears little resemblance to the meaning of the source. The SPE was able to improve the translation, but not enough to provide as much meaning as the SPE. When Gister missed a critical word, as in the seventh example, both the Gister translation and the SPE were very far from accurate. In this example, the SMT trained on the full data set still missed a word but was able to provide a usable transla-tion.

The evaluators' collective preference for the rudimentary Gister translator over SPE and strong dislike of the SMT in this evaluation, while partly indicative of errors in the output of those two systems, is also due in part to the context of the evaluation. As stated earlier, CyberTrans is intended for triage, filtering and selection tasks. This means that the usefulness of a translation is based mostly on accuracy rather than fluency, with fluency playing a role only when it inhibits understanding. Users of CyberTrans have most likely built a tolerance for the types of disfluencies often seen in Gister output such as word choice and word order. As a result, evaluators were more forgiving of this type of error than of errors that affected accuracy such as untranslated words, omitted words, and inserted words. That said, it is worth repeating that even these relatively forgiving evaluators did not rate any engine as GoodEnough more than 50% of the time.

While statistical post-editing did show an impressive improvement in BLEU score for weak MT engines with even a small amount of data, the largely cosmetic changes from statistical post-editing were as likely to cause degradation as improvement when the initial translation was already

weak. Because of the flaws in this study, we cannot conclude with certainty that similar results would be obtained in other evaluation scenarios with other translation systems. However, the results of this study provide further evidence that BLEU score is insufficient for judging between weak MT systems. Further research is needed to determine conclusively whether statistical post editing can be of help for weak MT systems in general and how mature an MT system must be to see consistent quality improvements based on human judgments rather than BLEU score alone.

6. References

Ondřej Bojar and Zdeněk Žabokrtský. 2006. Czeng: Czech-english parallel corpus, release version. *Prague Bulletin of Mathematical Linguistics*, 86:59—-62.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning-Prague Bulletin of Mathematical Linguistics*, pages 858–867, Prague, Czech Republic.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *Proceedings of EACL-2006*, pages 249–256.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254.

Arantza Diaz de Ilarraza, Gorka Labaka, and Kepa Sarasola. 2008. Statistical post-editing: A valuable method in domain adaptation of rbmt systems for less-resourced languages. In *MATMT 2008: Mixing Approaches to Machine Translation*, pages 27—-34, Donostia-San Sebastian, Spain.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2007. Statistical post-edition on systran rule-based translation system. In *Proceedings of the Second Workshop On Statistical Machine Translation*, pages 220—223, 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 Annual Meeting of the ACL, Demonstration and session*, pages 177—-180, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the tenth MT Summit*, pages 79—-86, Phuket, Thailand.

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 Association for Computational Linguistics*, pages 311—-318, Philadelphia, PA.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007a. Statistical phrase-based post-editing. In *Proceedings of NAACL-HLT 2007*, pages 508—-515, Rochester, NY.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Ro land Kuhn. 2007b. Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 203—-206, Prague, Czech Republic.

Heather Simpson, Christopher Cieri, Kazuaki Maeda, Kathryn Baker, and Boyan Onyshkevych. 2008. Human language technology resources for less commonly taught languages: Lessons learned. In *Proceedings of the LREC 2008 Workshop on Collaboration: interoperability between people in the creation of language resources for less-resourced-languages*, pages 7–11, Marrakesh, Morocco.

Andreas Stolke. 2002. Srilm - an extensible language modelling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901—-904, Denver, CO.

Claire Voss, Matthew Aguirre, Jeffrey Micher, Richard Chang, Jamal Laoudi, , and Reginald Hobbs. 2008. Boosting performance of weak mt engines automatically: Using mt output to align segments & build statistical post-editors. In *Proceedings 12th EAMT Conference*, pages 192—-201, Hamburg, Germany.