Challenges in Predicting Machine Translation Utility for Human Post-Editors

Michael Denkowski and **Alon Lavie**
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15232, USA
{mdenkows,alavie}@cs.cmu.edu

Abstract

As machine translation quality continues to improve, the idea of using MT to assist human translators becomes increasingly attractive. In this work, we discuss and provide empirical evidence of the challenges faced when adapting traditional MT systems to provide automatic translations for human post-editors to correct. We discuss the differences between this task and traditional adequacy-based tasks and the challenges that arise when using automatic metrics to predict the amount of effort required to post-edit translations. A series of experiments simulating a real-world localization scenario shows that current metrics under-perform on this task, even when tuned to maximize correlation with expert translator judgments, illustrating the need to rethink traditional MT pipelines when addressing the challenges of this translation task.

1 Introduction

Traditionally, machine translation research has focused on improving the adequacy of automatic translations, that is semantic similarity to reference translations as perceived by human judges. Often, automatic metrics such as BLEU are used as stand-ins for human judgments when optimizing or evaluating system performance. This work is motivated by the idea that machine translation should output a reliable finished product that can be immediately read by end users or used as input for other natural language processing tasks. Spurred by shared evaluation tasks such as the ACL Workshops on Statistical Machine Translation (Callison-Burch et al., 2011) and NIST Open Machine Translation Evaluations (Przybocki, 2009), significant effort has gone into the development of stronger models and more effective optimization techniques for improving MT performance as measured by automatic metrics. In turn, more sophisticated metrics have been developed to better predict translation adequacy when compared to reference translations, both in system optimization and evaluation, driven in turn by shared tasks such as the WMT11 Tunable Metrics Task (Callison-Burch et al., 2011) and the 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR (Callison-Burch et al., 2010). As a result, state-of-the-art performance on adequacy-based tasks has improved greatly for many language pairs in recent years.

As MT quality continues to improve, the historically under-explored idea of using automatic translation to assist human translators becomes more attractive. Recent work has explored the possibilities of integrating MT into human translation workflows by providing automatic translation as a starting point for translators to correct, saving time compared to translating source sentences from scratch. While this idea has already taken hold in both research and industry, with some translation service providers already incorporating MT into their translation workflows, the task of improving MT utility for post-editing is still not as well explored as traditional adequacy-driven tasks. This work examines the differences between widely used adequacy-based evaluations and post-editing scenarios with an emphasis on the ability to identify real improvements in quality and reliably predict a translation system's

performance based on automatic metrics, a key requirement in both optimization and evaluation of MT systems. We discuss the challenges of predicting post-editing effort required by human translators and conduct a series of experiments that demonstrate these challenges empirically. We find that current automatic metrics under-perform on this task, illustrating the need for further work to develop utility prediction methods better suited for post-editing applications. We encourage the machine translation community to consider these challenges when developing new techniques targeted at improving performance on this increasingly popular task.

2 Related Work

While not explicitly focused on post-editing applications, the DARPA GALE program (Olive et al., 2011) included the first major machine translation evaluation campaign to use human-targeted translation edit rate (HTER) (Snover et al., 2006) as the primary evaluation metric. This was motivated by the interpretation of HTER as the distance between MT output and the closest conceivable correct translation. Participants in the program noted that techniques developed for adequacy-evaluated tasks do not necessarily carry over to post-editing tasks. This led to the adaptation of traditional MT pipelines to better suit the evaluation objective. For instance, the Z-MERT implementation of minimum error rate training (Zaidan, 2009; Och, 2003) added support for optimizing systems toward TER-BLEU, which is observed to correlate with HTER better than standard BLEU. Versions of the TER-plus and Meteor automatic metrics were also tuned to maximize correlation with HTER (Snover et al., 2009; Denkowski and Lavie, 2010). In the following sections, we discuss the major differences between adequacy and post-editing tasks and illustrate the need for such adaptations. The 2010 ACL Workshop on Statistical Machine Translation (Callison-Burch et al., 2010) featured a post-editing task where human judges were shown MT outputs without source sentences or reference translations and asked to edit them based on perceived meaning, if possible. This is different from post-editing tasks in translation workflows that feature bilingual editors working with source sentences and to our knowledge there has not been any work dealing with predicting usability of MT output for monolingual post-editing.

More recent work has focused on the practical challenges associated with using MT to improve the efficiency of human translators. He et al. (2010) conducted a user study that integrated MT into a translation memory (TM) system used by human translators. For each sentence in a data set, translators were presented with both a human translation from the TM and a MT output and asked to select the translation that was most suitable for post-editing. The authors find that both TM and MT outputs were selected regularly and in some cases, translators were unable to tell the difference. An automatic classifier that uses features from the MT and TM systems in addition to widely used confidence estimation features shows good performance predicting which output will be preferred. While this work focuses on the case were both human translations (aproximately matched) and automatic translations are available, we focus on the case where only automatic translations are available and directly measure utility rather than preference. Specia (2011) evaluates the effectiveness of using MT confidence estimation methods to predict post-editor effort as measured by HTER, editing time, and editor post-assessments. Post-assessments (editors' self-ratings of how much effort was required to correct MT output) appear to be most predictable across language pairs. Particularly useful confidence estimation features include source and target language model scores for both surface word forms and part-of-speech tags. This work bears some similarities to our experiments, though we focus on the challenges of predicting overall MT system utility for future inputs based on automatically evaluated performance on a development set. Hardt and Elming (2010) show the potential benefit of tighter integration of MT into translation workflows by conducting an experiment that simulates incrementally re-training a MT system as translators provide post-edited translations. Retraining is shown to greatly improve performance when input sentences are taken from the same domain. This type of work, focusing heavily on MT system training, especially requires reliable automatic metrics for predicting post-editing effort.

In addition to measuring quantitative performance of MT systems for post-editing, work such as that by

Blain et al. (2011) focuses on $qualitative$ analysis of post-editing effort. The authors introduce a measure based on *post-editing actions* such as NP structure change, verb agreement correction, and multiword expression correction, with the goal of gaining a more linguistic understanding of what translation errors must be corrected. While this work is helpful for analyzing the types and severity of errors made by different MT systems, we focus on quantitative post-editing analysis (determining the total amount of work required by human translators), for which we use HTER as described in Section 3.2.

3 Adequacy Versus Post-editing Utility

Large machine translation evaluation campaigns such as the ACL Workshops on Statistical Machine Translation (Callison-Burch et al., 2011) and NIST Open Machine Translation Evaluations (Przybocki, 2009) focus on improving translation adequacy, the perceived quality of fully automatic translations compared to reference translations. As such, current techniques for MT system building, optimization, and evaluation are largely geared toward improving performance on this task. Automatic metrics such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006), and Meteor (Denkowski and Lavie, 2011), designed to correlate well with adequacy judgments, are often used as stand-ins for actual judgments during optimization and evaluation.

3.1 Adequacy and Ranking

Originally introduced by the Linguistics Data Consortium, adequacy ratings elicit straightforward quality judgments of machine translation output according to numeric scales (LDC, 2005). Traditionally, these judgments were split between $adequacy$, the degree to which MT output captures the meaning of a reference translation, and $fluency$, the degree to which MT output is grammatically correct in the target language. Due to observed high correlation between adequacy and fluency, more recent evaluations such as NIST OpenMT (Przybocki, 2008; Przybocki, 2009) combine the two into a single scale. Recent WMT evaluations (Callison-Burch et al., 2007; Callison-Burch et al., 2011) use $ranking$-based evaluation to further abstract away from concepts such as adequacy and fluency as well as difficult-to-decide numeric ratings. Human judges are simply asked to rank several MT outputs for the same sentence from best to worst according to a reference translation. This can be interpreted as rating $relative\ adequacy$ and consistently achieves higher inter-annotator agreement than absolute adequacy ratings. It is left up to judges to determine the relative severity of different types of translation errors when comparing translations.

3.2 Post-editing Utility

Adequacy and ranking judgments can be interpreted as assessing the $acceptability$ of machine translation output as a final product, either for consumption by end users or as input to other natural language processing tasks such as information extraction or speech synthesis. In contrast, post-editing judgments can be viewed as assessing the $utility$ of MT output as an intermediate step in the translation process. Whereas MT systems targeting adequacy should maximize the semantic similarity of automatic translations with reference translations, systems targeting post-editing utility should minimize the effort required by human translators to correct automatic translations. We measure post-editing effort according to cased human-targeted translation edit rate (HTER) (Snover et al., 2006), which is defined as the minimum edit distance between a translation output and targeted reference translation created by post-editing the output. Edit distance is calculated automatically using the TER metric, which is defined as:

$$\text{TER} = \frac{\#\text{ of edits}}{\#\text{ of reference words}}$$

where possible edits are word insertions, deletions, substitutions, and shifts of multiple-word spans. We choose HTER because it closely reflects actual actions taken by human translators to correct MT output and treats all edits equally, an important quality illustrated in following examples. As TER is an error measure, lower HTER scores are better, indicating less work to post-edit.

The adequacy and post-editing tasks bear some similarities, as automatic translations that have high similarity to reference translations often require minimal post-editing. However, when MT outputs contain errors, as is the usual case, the most ade-

	HTER	Translation
Reference	–	Advocators had hoped this would reduce the USA's dependency on foreign oil supplies.
System 1	0.33	Supporters of the US seeks to reduce dependence on oil supplies from abroad.
Post-edit 1	–	Supporters sought to reduce US dependence on oil supplies from abroad.
System 2	0.14	The advocates by trying to reduce the US dependence on oil supplies from abroad.
Post-edit 2	–	The advocates hoped to reduce the US dependence on oil supplies from abroad.
Reference	–	He was supposed to pay half a million to Luboš G.
System 1	0.27	He had for Luboši G. to pay half a million crowns.
Post-edit 1	–	He had to pay Luboš G. half a million crowns.
System 2	0.09	He had to pay luboši G. half a million kronor.
Post-edit 1	–	He had to pay Luboš G. half a million kronor.
Reference	–	Only the crème de la crème of the many applicants will fly to the USA.
System 1	0.40	Only the crème de la crème from many candidates, it's going to go to the US.
Post-edit 1	–	Only the crème de la crème from many candidates will fly to the US.
System 2	0.20	Only crème de la crème of many customers will travel to the US.
Post-edit 2	–	Only the crème de la crème of many applicants will fly to the US.

Table 1: Examples where lower-ranked MT outputs require less work to post-edit. Translations taken from WMT11 Czech-to-English submissions. For each case, system output 1 is ranked better than system output 2 by human judges. Minimally post-edited acceptable translations follow system outputs. Post-editing and scoring conducted by authors.

quate translations are often not the easiest to post-edit. Table 1 shows ranked system outputs and reference translations from the difficult Czech-to-English translation track of the 2011 EMNLP Workshop on Statistical Machine Translation (Callison-Burch et al., 2011). In addition, we provide minimally post-edited translations and HTER scores. In each case, the translation deemed more adequate by expert judges actually requires more effort to post-edit. In the first example, both translations deliver the general message, though the first is slightly more fluent. However, correcting the translation to be fully fluent requires additional edits, such as correcting verb tense and replacing multiple function words to accommodate moved content words. In the second example, the lower-ranked sentence is more fluent, but fails to capitalize a proper noun, leaving the reader unable to distinguish it from a passed-through foreign word. However, this error is easily fixed in post-editing. In the third example, the preferred sentence correctly delivers the meaning, though it is disfluent and cumbersome to post-edit. These examples illustrate types of errors that have a large impact on sentence meaning but require relatively little work to correct, as well as accumulated minor errors that do not impact meaning, but are cumbersome to correct.

3.3 Automatic Metrics

While shared evaluation tasks often include human evaluation, the majority of research in machine translation relies on automatic metric scores to measure improvement, and even in shared evaluations, most translation systems are optimized toward automatic metrics. Generally, improvement in BLEU score, which measures simple surface word n-gram precision balanced with a brevity penalty (Papineni et al., 2002), is presented as evidence of improvement in translation quality. However, BLEU has been shown to be an insufficient stand-in for adequacy judgments (Callison-Burch et al., 2006). We similarly evaluate its ability to predict post-editing effort.

Table 2 shows BLEU-scored system outputs from the WMT11 Czech-to-English translation track (Callison-Burch et al., 2011), along with reference translations. We additionally provide minimally post-edited translations and HTER scores. In each example, the translation with the higher BLEU score actually requires more effort to post-edit. In the first case, sentence 2 is penalized for using a different

	BLEU	HTER	Translation
Reference	–	–	The problem is that life of the lines is two to four years.
System 1	0.49	0.29	The problem is that life is two lines, up to four years.
Post-edit 1	1.00	–	The problem is that life of the lines is two to four years.
System 2	0.34	0.14	The problem is that the durability of lines is two or four years.
Post-edit 2	0.67	–	The problem is that the life of lines is two to four years.
Reference	–	–	The rate of unemployment in France has remained stable in the 3rd quarter.
System 1	0.26	0.08	The rate of unemployment remains in the third quarter in France stable.
Post-edit 1	0.30	–	The rate of unemployment remains stable in the third quarter in France.
System 2	0.16	0.00	The unemployment rate remains stable in the third quarter in France.
Post-edit 2	0.16	–	The unemployment rate remains stable in the third quarter in France.
Reference	–	–	He was supposed to pay half a million to Luboš G.
System 1	0.34	0.27	He had for Luboši G. to pay half a million crowns.
Post-edit 1	0.21	–	He had to pay Luboš G. half a million crowns.
System 2	0.19	0.09	He had to pay luboši G. half a million kronor.
Post-edit 2	0.21	–	He had to pay Luboš G. half a million kronor.

Table 2: Examples where translations with lower BLEU scores require less work to post-edit. Translations taken from WMT11 Czech-to-English submissions. For each case, minimally post-edited acceptable translations follow system outputs. Post-editing and scoring conducted by authors.

word order from the reference even though it is both more adequate and less work to correct. In the second case, a fully acceptable translation is penalized over an erroneous translation simply because it is phrased differently. In the third case, the brevity penalty unduly penalizes a shorter but more correct translation. Further, several of the post-edited translations receive relatively low BLEU scores, despite the fact that they are considered fully acceptable by human judges. As is the case with adequacy judgments, improvement in BLEU does not necessarily carry over to improvement in post-editing utility.

The inconsistencies between adequacy judgments, BLEU scores, and post-editing effort, as shown in Tables 1 and 2, illustrate the need for caution when applying existing MT techniques to post-editing tasks. Simply choosing the "best" MT system, ranked by human judges or BLEU score, to provide translations for human post-editors will not necessarily yield the most efficient translation workflow.

4 Experiments

To empirically evaluate the behavior of human post-editors and the effectiveness of current MT techniques for predicting translation utility, we conduct a series of experiments to simulate a real-world localization scenario. We selected 5 English–Spanish parallel documents in the software documentation domain, totaling 90 sentences. Each English sentence was translated into Spanish using two translation engines: Microsoft Translator's online service[1] and a phrase-based Moses system representative of the 2011 WMT baseline (Hoang et al., 2007; Callison-Burch et al., 2011). The outputs of the two systems, which have significant lexical differences but are statistically indistinguishable by automatic metrics, were combined into a single data set of 180 translations. The Spanish side of the parallel data provides reference translations.

We employed the assistance of an expert translator and several students of translation studies to obtain two types of annotation for each automatic translation. First, the expert translator assigned a rating from 1 to 4 predicting the degree of post editing that should be required to correct the sentence according to the following scale:

1. No editing required
2. Minor editing, meaning preserved

[1] http://www.microsofttranslator.com/

Reference	BLEU	TER	Met
Gold	0.32	0.49	0.58
Post-edit 1	0.64	0.26	0.79
Post-edit 2	0.76	0.15	0.88
Closest	0.79	0.12	0.90
Closest vs Gold	0.34	0.48	0.59

Table 3: Corpus-level BLEU, TER, and Meteor scores of MT output against gold standard and post-edited references.

r	4-pt	BLEU	TER	Met	Met_o
4-point	–	0.32	0.28	0.33	0.35
HTER 1	0.35	0.20	0.45	0.33	–
HTER 2	0.40	0.21	0.22	0.21	–
HTER	0.49	0.26	0.24	0.27	0.34

Table 4: Sentence-level Pearson's correlation (absolute value) between 4-point ratings, individual and minimum HTER scores, and automatic metric scores. Met_o indicates oracle Meteor scores that maximize correlation with the corresponding type of judgment.

3. Major editing, meaning lost

4. Re-translate

Second, each translation was post-edited by 2 students from a pool of 7 total. It is important to note that just as in a real world localization task, participants saw only English source sentences and Spanish automatic translations. No reference translations were shown.

We also tracked the time each participant spent post-editing each document, though we do not consider this to be a reliable measure of effort for two reasons. First, translation times varied widely between participants post-editing the same translations, making times difficult to compare across documents without normalization. Second, even when times are scaled based on time spent by different translators on the same document, time does not appear to correlate with expert ratings or HTER scores. This indicates that some human translators simply work faster than others.

4.1 Automatic Metric Scores

We use the BLEU (Papineni et al., 2002), TER (Snover et al., 2006), and Meteor (Denkowski and Lavie, 2011) metrics to evaluate MT outputs against different types of reference translations. In Table 3, "gold" refers to the gold standard references from the parallel data (unseen by human editors). "Post-edit" 1 and 2 refer to always selecting either the first or second post-edited reference for each sentence. "Closest" refers to using the post-edited reference with the minimum edit distance to each MT output, as in HTER scoring. Post-edited lines correspond to HTER, "H-BLEU", and "H-Meteor". Finally, the last line of Table 3 scores the closest reference set against the gold standard references. As TER is an error measure, lower scores are better.

These results demonstrate several challenges in predicting post-editing effort. The significantly lower metric scores against gold standard references compared to the post-edited references illustrate the known problem that metrics are good at detecting similar translations, but poor at evaluating sentences with different structure and lexical choice. The large score differences between the the two post-edited references demonstrate the degree to which human translators accept automatic translations. Editors from group 1 apply nearly twice as many edits to the same MT outputs as editors from group 2. Finally, metrics assign nearly the same scores to the closest post-edited references as they do to the raw MT outputs. In other words, currently used metrics are unable to distinguish erroneous MT output from fully fluent and adequate translations as edited by human translators.

To empirically evaluate the impact of the post-editing task difficulties discussed in Sections 3.2 and 3.3, we examine the correlation of sentence-level automatic metric scores against gold standard reference translations against expert predictions and actual HTER scores. As shown in Table 4, automatic metrics have low correlation with both expert 4-point effort predictions and HTER scores. It can be noted that the stabilizing effect of using multiple post-editors to calculate HTER (taking the minimum edit distance at the sentence level) improves correlation with both expert predictions and metric scores.

We also conduct an oracle experiment to determine the best possible performance on this data set using currently available evaluation techniques. We select Meteor, the best performing publicly avail-

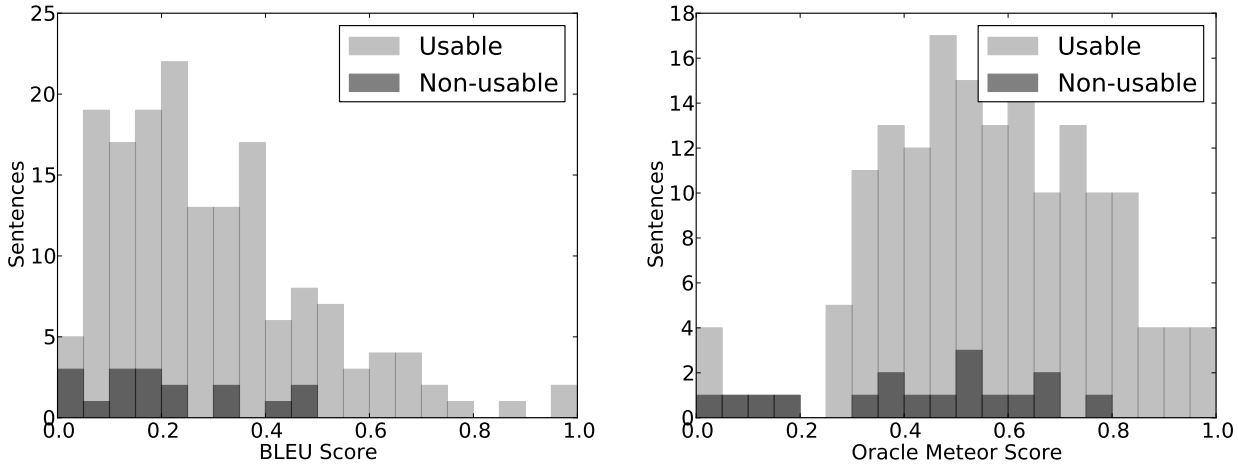


Figure 1: Automatic metric score distributions of usable and non-usable translations when scored against gold standard references. Meteor scores maximize correlation with 4-point ratings on this data set.

able metric for evaluating translations into Spanish as of WMT11[2] and tune the various model parameters described by Denkowski and Lavie (2011) to maximize correlation with the 4-point ratings and HTER scores in this data set. These scores, labeled "Met$_o$" in Table 4, correlate slightly better than other metric scores, but still clearly demonstrate the inability of current metrics to accurately predict post-editing effort even in a best-case scenario. These are the same automatic metrics that are widely used as objective functions in MT system optimization and as criteria for selecting the best system configurations to use in production environments, including post-editing workflows. The low correlation coefficients for BLEU, in addition to the examples in Section 3.3, provide a counterexample to the common notion that improvements in BLEU translate to improvements in quality since BLEU is most often unduly harsh. In reality, increases and decreases in BLEU score are only weakly correlated with translation utility as predicted by experts and scored by HTER.

[2]The AMBER (Chen and Kuhn, 2011) and MPF (Popović, 2011) families of metrics correlate better with human judgments of translations into Spanish but have no publicly available scoring tools.

4.2 Predicting Translation Usability

Reliable prediction of translation usability is one of the most important aspects of incorporating MT into translation workflows. To avoid wasting translators' time, systems should be able to predict when MT output is sufficiently good to serve as a starting point for post-editing or sufficiently bad to require total re-translation and recommend accordingly. In an additional experiment, we simplify the 4-point predictions into two groups: usable (1 and 2) and non-usable (3 and 4), corresponding to whether the expert translator would personally post-edit or re-translate each sentence. MT outputs were largely deemed to be usable (90.6%) with a minority classified as non-usable (9.4%). We examine the distributions of sentence-level BLEU and oracle Meteor scores for each group to determine the feasibility of learning quality thresholds based on existing automatic metrics. It should be noted that even the BLEU result is an oracle experiment as MT systems in real-world translation workflows must predict usability without the aid of reference translations, relying instead on confidence estimation techniques such as those described by Specia (2011). As shown in Figure 1, the BLEU score distributions overlap completely and translations are clustered in the same region. Any quality threshold (visualized as a vertical line on the graph) that removes a substantial number of non-usable transla-

tions also removes a disproportionally large number of usable translations. The Meteor score distributions are slightly more separated, with a quality threshold of 0.2 removing usable and non-usable translations equally. However, the major parts of the distributions are nearly identical subject to scale. As the group of usable translations is substantially larger, it is unclear from these distributions if any quality threshold, even set according to oracle Meteor score, would be of any benefit. This further reinforces the difficuly encountered when applying current techniques to post-editing tasks.

4.3 Human Ability to Predict Editing Effort

We finally examine the expert translator's accuracy when predicting translation usability. Figure 2 shows the distributions of HTER scores for translations rated usable and non-usable. The expert is largely able to detect easily correctable translations, judging nearly all translations with HTER under 0.2 to be usable. Above 0.2, translations requiring comparable numbers of edits are judged to be both usable and non-usable. When determining whether post-editing would save time over retranslation, even expert judgments can be inaccurate for partially correct MT outputs with the types of difficult-to-analyze errors shown in Section 3.2. To maintain top efficiency, a human translator must make a snap judgment as to whether or not a translation is usable, and each misjudgment costs time. These results hint at the possibility of adapting automatic usability predictors that even outperform expert human translators, which would boost translator productivity tremendously.

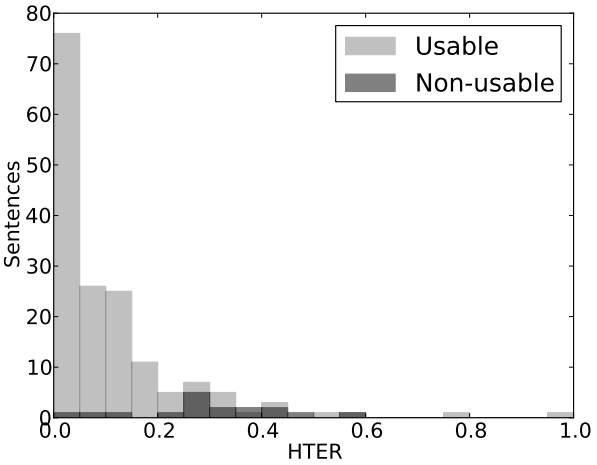


Figure 2: HTER distributions of usable and non-usable translations (lower scores are better).

5 Conclusion

We have presented an analysis of the unique challenges associated with predicting the utility of automatic translations for human post-editors and a series of experiments demonstrating the difficulties encountered when applying current MT techniques to this new task. The traditional approach of using the BLEU metric to optimize system parameters and select the best system configuration quickly breaks down, as the types of translations that require less post-editing can have quite different characteristics from those that receive high BLEU scores, and even from translations that are preferred by humans. This arises from the fundamental differences between MT as a final product and MT as an intermediate step for human translators. Our experimental results show the shortcomings of current automatic metrics when predicting translation utility, leaving a large space for improvement in key components of system optimization and evaluation for post-editing. We encourage the larger machine translation community to consider the challenges we have discussed and the need to address them when working on this increasingly popular translation task.

Acknowledgements

We would like to thank our collaborators at the Institute for Applied Linguistics at Kent State University for providing valuable post-editing and expert translator judgment data.

References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proc. of EACL 2006*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proc. of ACL WMT 2007*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proc. of ACL WMT/MetricsMATR 2010*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan. 2011. Findings of the 2011 Joint Workshop on Statistical Machine Translation. In *Proc. of ACL WMT 2011*.

Boxing Chen and Roland Kuhn. 2011. AMBER: A Modified BLEU, Enhanced Ranking Metric. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Michael Denkowski and Alon Lavie. 2010. Extending the METEOR Machine Translation Metric to the Phrase Level for Improved Correlation with Human Post-Editing Judgments. In *Proc. of NAACL/HLT 2010*.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proc. of the EMNLP WMT 2011*.

Daniel Hardt and Jakob Elming. 2010. Incremental Re-training for Post-editing SMT. In *Proc. of AMTA 2010*.

Yifan He, Yanjun Ma, Johann Roturier, Andy Way, and Josef van Genabith. 2010. Improving the Post-Editing Experience using Translation Recommendation: A User Study. In *Proc. of AMTA 2010*.

Hieu Hoang, Alexandra Birch, Chris Callison-burch, Richard Zens, Rwth Aachen, Alexandra Constantin, Marcello Federico, Nicola Bertoldi, Chris Dyer, Brooke Cowan, Wade Shen, Christine Moran, and Ondřej Bojar. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of ACL 2007*.

LDC. 2005. Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations. Revision 1.5.

Franz Josef Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proc. of ACL 2003*.

Qualitative Analysis of Post-Editing for High Quality Machine Translation. 2011. Frédéric Blain and Jean Senellart and Holger Schwenk and Mirko Plitt and Johann Roturier. In *Proc. of MT Summit XIII*.

Joseph Olive, Caitlin Christianson, and John McCary. 2011. *Handbook of Natural Language Processing and Machine Translation: Darpa Global Autonomous Language Exploitation*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL 2002*.

Maja Popović. 2011. Morphemes and POS tags for n-gram based evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Mark Przybocki. 2008. NIST Open Machine Translation 2008 Evaluation. http://www.itl.nist.gov/iad/mig/tests/mt/2008/.

Mark Przybocki. 2009. NIST Open Machine Translation 2009 Evaluation. http://www.itl.nist.gov/iad/mig/tests/mt/2009/.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proc. of AMTA 2006*.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In *Proc. of ACL WMT 2009*.

Lucia Specia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In *Proc. of EAMT 2011*.

Omar F. Zaidan. 2009. Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems. *The Prague Bulletin of Mathematical Linguistics*.