On Post-Editability of Machine Translated Texts

CH RAM ANIRUDH &
KAVI NARAYANA MURTHY

Abstract

Machine Translated texts are often far from perfect and postediting is essential to get publishable quality. Post-editing may not always be a pleasant task. However, modern machine translation (MT) approaches like Statistical MT (SMT) and Neural MT (NMT) seem to hold greater promise. In this work, we present a quantitative method for scoring translations and computing the post-editability of MT system outputs. We show that the scores we get correlate well with MT evaluation metrics as also with the actual time and effort required for post-editing. We compare the outputs of three modern MT systems namely phrase-based SMT (PBMT), NMT, and Google translate for their Post-Editability for English to Hindi translation. Further, we explore the effect of various kinds of errors in MT outputs on postediting time and effort. Including an Indian language in this kind of post-editability study and analyzing the influence of errors on postediting time and effort for NMT are highlights of this work.

Keywords: Machine Translation, Post-Editing, Statistical Machine Translation, Neural Machine Translation.

1. Introduction

Translation involves the conversion of texts from one language to another, preserving certain attributes of the source text. Most importantly, meaning must be preserved, while other properties such as style, ability to produce specific effects on the minds of the readers, etc. may also be required to be preserved. It is also generally expected that translated texts sound natural and fluent in the target language. As such, translation is a hard problem even for expert translators, and translations produced by machines often fall short of these high expectations. There are two important use-cases of Machine Translation (MT) outputs (Koehn 2009): Dissemination where the output of an MT system should be of publishable quality, Assimilation where the output is just good enough to get an idea of what the source language text conveys even if the translation is poor in quality. Outputs of Machine Translation systems are rarely good enough for Dissemination, that is, for deploying for direct use in any kind of end application. For example, a school textbook, translated from one language to another, cannot be directly used as a textbook by school children in the target language. Some degree of manual checking and editing, called post-editing, is inevitable. Overall, the primary goal of building usable MT systems should be to make post-editing an easy and pleasant task. More specifically, the goal should be to minimize the post-editing time and effort. This paper presents a systematic and quantitative study of the Post-Editability of MT systems.

If the translations produced by machines are so poor that post-editors prefer to translate from scratch rather than struggle to fix all the mistakes made by the computer, such systems should be considered not post-editable at all (Specia & Farzindar 2010). It is not just the time and effort required in filling the gaps and correcting the mistakes, the whole process can be psychologically quite taxing and unpleasant. Machines make strange mistakes that can mislead, sidetrack, or confuse the readers. MT systems are usable in practice only if the translations they produce are easily post-editable.

There are two kinds of post-editing, based on whether post-editors refer to source language text or not (Nitzke 2016). In monolingual post-editing, post-editors need not be aware of source-language text, they just post-edit the MT output without looking at the source text. In bilingual post-editing, post-editors should be bilinguals and they have to refer to source language text to post-edit. Throughout this paper, we presume bilingual post-editing unless specified otherwise. That is, post-editors read and try to understand both the source language input sentence and the translation produced by the machine before editing the output. In this process, post-editors may look for correcting the lexical substitution errors, word order errors, spelling errors, proper handling of ambiguity, register, and style. Intuitively, Post-editors may find it easier if words and expressions are properly translated, instead of finding unrelated words or too many unknown words in the MT output. Re-ordering of words, correcting typos, style, and register, etc. may be easier provided lexical substitution is good and intended meaning can be easily understood.

Understanding the source language sentences may be quick and easy or difficult and time-consuming, depending upon the complexity of syntactic structures used, whether rare, strange or unknown words and expressions are used, whether the words, expressions, and structures used are straightforward or highly ambiguous and confusing, etc. Multiple interpretations may be possible. Understanding the machine-produced translations can also be difficult and time-consuming, perhaps more so compared to understanding the source sentences. Editing the machine-produced translations can also take a considerable amount of time and effort, for example, when we cannot easily find suitable equivalent words or expressions. In general, some sentences may be translated quite well while others may be difficult to post-edit. Therefore, instead of defining Post-Editability as a binary yes or no question, we propose a 4-point scale to rate the overall degree of Post-Editability. The Post-Editability scores for a particular MT system, averaged over many sentences, can be taken as a Post-Editability score for the system itself. This way, we can compare MT systems for their Post-Editability. In an absolute sense, an MT system is usable if the time taken for translation and postediting is less than time required for manual translation. In a relative sense, we can check which MT systems produce more easily posteditable translations and are thus better.

Modern MT systems such as Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) systems are data-driven. They model MT as a machine-learning problem, learning from large-scale parallel corpora. A parallel corpus is a collection of source language segments and translations of each of these in the target language (Koehn 2009). A segment may be a sub-sentence unit like a word or a phrase, a sentence, or even a unit longer than a sentence. For example, a parallel corpus may include translations of signboards, sub-headings, etc. which are not complete sentences. Hereafter, we use the term segment to mean a unit of translation. Recent SMT and NMT systems are able to beat traditional rule-based systems both in terms of lexical substitution and word order. In the early days, SMT systems used words (tokens delimited by spaces) as basic lexical units (Peter F, Brown; Vincent J, Della Pietra; Stephen A, Della Pietra; Robert L, Mercer 1993). These word-based MT models (also known as IBM Models) were later surpassed by phrase-based MT systems (Philipp, Koehn; Franz, Josef Och; Daniel, Marcu 2003), where a phrase is any sequence of words, not necessarily a linguistically valid phrase. Today, the term SMT mostly means phrase-based SMT (PBMT). Word-based models are almost obsolete. Neural MT (NMT) ((Ilya Sutskever; Oriol Vinyals; Quoc V Le 2014), (Dzmitry, Bahdanau; Kyunghyun, Cho; Yoshua, Bengio 2015) is a relatively new paradigm in which the MT system is trained using neural networks. In NMT, training proceeds in a sequence-to-sequence (segments in a parallel corpus) fashion, contrary to PBMT in which the segments are split into phrases and processed. NMT systems are hence more complex, computationally intense, and data-hungry. These modern MT systems (SMT and NMT) have started producing much better results compared to traditional rule-based MT systems, and it is time to check them for Post-Editability afresh. In this work, we compare the outputs of three modern MT systems namely PBMT, NMT and Google translate for their Post-Editability for English to Hindi translation. PBMT and NMT systems are trained using Moses (Philipp, Koehn; Hieu, Hoang; Alexandra, Birch; Chris, Callison-Burch; Marcello, Federico; Nicola, Bertoldi; Brooke, Cowan; Wade, Shen; Christine, Moran; Richard, Zens; others 2007)[1] and Open NMT[2] respectively, on IITB English-Hindi parallel corpus (Anoop, Kunchukuttan; Pratik, Mehtal; Pushpak, Bhattacharyya 2018). Google translate is available as a free service online[3]. Google translate is also an NMT system, but the corpus used is unknown as it is not disclosed by Google.

The quality of MT output is best measured using manual methods. Humans can read and understand the texts and check if the meaning of the source language text is properly and completely preserved or not, but the manual evaluation is expensive and time-consuming. Automatic evaluation methods are therefore widely used, although they are crude (for example, simply based on to what extent n-grams in the MT output match n-grams in reference translations). Manual methods typically involve grading MT outputs on a numerical scale, for example, from 1 to 5, 1 for the worst output, and 5 for best output. These scores often turn out to be subjective and difficult to judge. Adequacy, Fluency, comprehensibility are a few manual evaluation measures (White & O’Connell 1993). Automatic methods are objective, fast, and cheap. Most

[1] http://www.statmt.org/moses/

[2] https://opennmt.net/

[3] https://translate.google.com/

of the automatic methods require reference translations to test the quality. They are generally based on string matching techniques. Bilingual Language Evaluation Understudy (BLEU) (Kishore, Papineni; Salim, Roukos; Todd, Ward; Wei-Jing, Zhu 2002), Meteor (Banerjee & Lavie 2005) and Human mediated Translation Edit Rate (HTER) (Matthew, Snover; Bonnie, Dorr; Richard, Schwartz; Linnea, Micciulla; John, Makhoul 2006) are some of the widely used automatic metrics. Automatic and manual evaluation methods only provide a comparison between various MT system outputs and their quality. They do not provide information about what kind of mistakes or errors are made by the MT systems, which is valuable information for MT system developers. Error analysis is usually done to find out which errors are frequent in an MT system output. Errors are broadly classified based on lexical substitution errors and re-ordering errors. These are subjective and sometimes language-dependent. There are various manual annotation methods described in the literature (David, Vilar; Jia, Xu; D’Haro Luis, Fernando; Hermann, Ney 2006 & Popović 2018). Automatic error identification methods require reference segments similar to automatic evaluation methods and they identify errors based on edit-distance and linguistic cues (Popović & Ney 2011).

The main theme of this paper is to check if the outputs produced by the state-of-the-art MT systems are Post-Editable, if yes to what extent, compare the Post-Editability of outputs of various MT systems and probe into the errors in MT outputs influencing the post-editing effort. Post-editing effort is measured in terms of time taken for post-editing (in seconds) and the number of keystrokes required for post-editing the MT output. In practice, the time taken for translation using MT is negligible. Hence, time taken for post-editing alone is considered. Some kinds of errors in MT outputs may be tolerable and easier to correct for post-editors, while some kinds of errors may be annoying or even a put-off. We explore the influence of various types of errors in MT outputs on post-editing effort, in order to find out which kinds of errors affect post-editing the most.

In this work, post-editing and scoring for Post-Editability have been carried out with the help of four freelance professional translators using the Post-Editing Tool (PET) from (Wilker, Aziz; Sheila, Castilho; Lucia, Specia 2012). Inter-Annotator Agreement is computed on a sample data set before proceeding for actual experiments. We compare the time taken for Post-Editing with the time required for manual translation. Post-Editability scores of the three MT systems are compared and correlated with three widely used metrics for automatic evaluation of MT outputs, namely BLEU, Meteor, and HTER. We explore the effect of segment length on post-editing effort. We also look at errors in the machine-translated segments and how they affect post-editing effort. For this purpose, linear mixed-effects models (LMM) (Douglas Bates; Martin Mächler; Ben Bolker; Steve Walker 2015) are used, to model errors in MT outputs as predictors of post-editing effort and find out which errors are significant predictors of post-editing effort. The same model is used to find out if segment length is a significant predictor.

We find that the time taken for manual translation is significantly higher compared to post-editing any MT system output.

NMT and Google have got significantly better Post-Editability scores compared to PBMT. As expected, the time taken to post-edit and the number of keystrokes required for post-editing correlate negatively with the Post-Editability scores. Post-Editability scores correlate positively with BLEU and Meteor and negatively with HTER.

We find that missing words and lexical choice errors significantly influence the post-editing effort. Re-ordering errors affect only the number of keystrokes.

2. Review of Literature

In a paper entitled “The present state of research on mechanical translation” in 1951, Bar-Hillel (1951) claimed that if the machine could resolve grammatical ambiguities and re-arrange the target language words in an appropriate order, post-editing would be an easy task. This was of course only a conjecture.

For many decades that followed, there was general displeasure expressed towards the task of post-editing by human translators. The task was done on pen-and-paper for many years until the arrival of word processors in the late 1970s and early 1980s (Evans 1986).

Rapid post-editing was proposed by SYSTRAN (Wagner 1985), which was useful for getting translations of low quality quickly. Such translations were useful only to get the meaning (assimilation), but not for dissemination. The decision whether to use this service or not lies with the translation user and the user is warned about the quality. Rapid post-editing was further emphasized by Senez (1998) with a discussion on what is expected from post-editor, what kind of texts are suitable, and what the end-users should expect. Use cases reported in the literature are scientific manuals, product manuals, and other technical documents for internal working purposes in labs, companies, etc.

Generally, it was observed that translators got frustrated when they tried to post-edit MT outputs, due to the high expectations they had on the MT systems (Lavorel 1982). Post-editors expressed dissatisfaction at the mistakes that machines made, as they had anticipated that MT output would be like human translators' output (Lavorel 1982; Schäfer 2003). Green (1982) noted that post-editors who are sympathetic towards MT often tended to make a minimum of alterations, accepting lower standard outputs, whereas post-editors who are unsympathetic often get annoyed at the output, and preferred translating from scratch. Translators also do not like the repetitive, mechanical changes that need to be made during post-editing. Schäfer (2003) suggests that post-editing requires special training; otherwise professional translators fail to understand the importance of MT post-editing. O'Brien (2002) made a proposal for a course that trains post-editing to professional translators. The author described skills required for post-editing and how they are different from skills a professional translator has.

Yamada (2015) opines that expert human translators often seem skeptical about accepting post-editing as a worthwhile task. They believe that the task requires less skill than manual translation and continuing to do post-editing may deteriorate their translation skills. (Green, Spence; Jeffrey, Heer; & Christopher D, Manning 2013) also say that often translators express an intense dislike for working with MT output. This leads to a lack of post-editors in the language service market, as professional translators do not come forward to do post-editing. To find out whether non-professionals are capable of post-editing MT outputs, Yamada (2015) got college students with translation as major to do the post-editing task. The author concluded that some students showed an acceptable aptitude for post-editing, although their outputs did not meet the professional quality standards.

Recently, there is an increased demand in the market for professional post-editors. Many Language Service Providers (LSP) hire post-editors as full-time employees. There is a general acceptance of post-editing as an important stage in translation by the translation industry, which is increasingly adapting MT into the pipeline (Garcia 2011). According to a survey conducted by (Federico, Gaspari; Hala, Almaghout; & Stephen, Doherty 2015) on MT users, 38% of them said they always post-edited MT outputs, while 14% of users used post-editing often, 12% used it occasionally and 6% rarely used. 30% of the users said they never resort to post-editing.

Ideas on developing dedicated MT workstations with good support for post-editing have emerged over time (Jäppinen & Kulikov 1991). There are frameworks that adapt and improve the performance of MT via manual post-editing (Michael, Denkowski; Chris, Dyer; Alon, Lavie 2014a), (Michael, Denkowski; Alon, Lavie; Isabel, Lacruz; Chris, Dyer 2014b), (Patrick Simianer; Joern Wuebker; John DeNero 2019).

Post-editing activity was thoroughly investigated first by Krings (2001). He classified post-editing effort indicators into temporal, technical, and cognitive efforts. While temporal and technical efforts are estimated using time and keylogging data, the cognitive effort is usually measured using gaze data or pause analysis.

O’Brien (2004) conducted a study for English-German with post-editing time, processing speed (number of words per second), and a few other measures as indicators of post-editing effort, against translatability indicators (TI) which are linguistic features in source language text known to be problematic for MT. These TIs are assigned numerical weights to give relative importance of indication of translatability. She found that long noun phrases and gerunds in the English language take a longer time compared to TI like abbreviations and proper nouns. There are many other possible TIs that are not considered in this study. The study was very limited in coverage of TIs as well as in terms number of sentences reported (only 40).

Garcia (2011) conducted a study on English-Chinese language pairs and observed a statistically significant reduction in time by post-editing MT outputs compared to translation from scratch. Further, he also found that post-editing produced better quality translations compared to translation from scratch in 54% of the cases.

Spence Green; Jeffrey Heer; Christopher D Manning (2013) conducted the first controlled analysis of post-editing for three language pairs: English to Arabic, French, and German. They too found that post-editing reduces time and improves quality. They modeled various post-editing effort indicators using linear mixed-effects models. They have shown that part of speech (POS) of words in source language texts are significant predictors of post-editing time. They found that percentage of Nouns in source language text is a significant major effect, influencing post-editing time. Post-editors were found to spend more time on nouns in source language according to mouse hover data. Based on a user opinion study on ranking POS in decreasing order of difficulty (Adverb, Verb, Adjective, Other, Noun), authors claim that post-editors often underestimated the difficulty of translating Nouns.

Joke Daems; Sonia Vandepitte; Robert J Hartsuiker; Lieve Macken (2017) have studied the impact of MT errors on post-editing efforts. They considered seven post-editing effort indicators (average duration per word, average fixation duration, average number of fixations, average number of production units, pause ratio, average pause ratio, human-mediated translation edit rate (HTER)), and found that various MT error types affect various effort indicators significantly as shown below. Duration is influenced most by coherence, while fixation duration is influenced by other meaning shifts. Four issues namely, coherence, other meaning shifts, grammar, and structural issues were influencing most of the effort indicators.

They also considered experience (student/professional) as a predictor in their model. Students are influenced by grammatical and lexical issues, while professionals are influenced by coherence and structural issues. The authors mentioned that the study was performed on Google PBMT outputs alone, as Google NMT was not yet available at the time of working. See (Joke Daems; Sonia Vandepitte; Robert J Hartsuiker; Lieve Macken 2017) for more details.

Dimitar Shterionov; Riccardo Superbo; Pat Nagle; Laura Casanellas; Tony O'Dowd; Andy Way (2018) have compared NMT and PBMT using manual and automatic evaluation metrics, for five language pairs (English to German, Chinese, Japanese, Italian and Spanish). They also conducted experiments on post-editing and reported that post-editors are more productive when using NMT outputs compared to PBMT outputs, in terms of the number of words translated per hour. Further, they found that automatic evaluation metrics show higher performance for PBMT whereas manual evaluation metrics show that NMT performs better, which is in line with our own findings in this work.

Yanfang Jia; Michael Carl; Xiangling Wang (2019) have conducted a comparison of NMT and PBMT exclusively for post-editing for English-Chinese language pair. They concluded that the translation output of NMT is better in terms of accuracy and fluency compared to PBMT. Reduced technical, cognitive, and temporal efforts have been observed with post-editing NMT compared to post-editing PBMT. They further reported that complexity measures tailored for human translation (HT) affect HT only, not MT output, and Post-Editing effort.

In the present work, we compare PBMT, NMT and Google translate in terms of Post-Editability for English-Hindi translation. We find MT error types that influence the post-editing effort in terms of post-editing time and the number of keystrokes. Dimitar Shterionov; Riccardo Superbo; Pat Nagle; Laura Casanellas; Tony O'Dowd; Andy Way (2018) stated a possibility of future work including the influence of MT error types on PBMT and NMT in line with work done by Joke Daems; Sonia Vandepitte; Robert J Hartsuiker; Lieve Macken (2017). Joke Daems; Sonia Vandepitte; Robert J Hartsuiker; Lieve Macken (2017), Dimitar Shterionov; Riccardo Superbo; Pat Nagle; Laura Casanellas; Tony O'Dowd; Andy Way (2018), Yanfang Jia; Michael Carl; Xiangling Wang (2019) have mentioned prospective future study on other language families. To the best of our knowledge, perhaps this is the first work reporting a study on post-editing involving an Indian language, the first work on error analysis of MT outputs involving an Indian language, and also on finding the errors influencing post-editing effort across various MT (PBMT and NMT) systems.

3. Post-Editability

On lines similar to Specia & Farzindar (2010), Wilker Aziz; Sheila Castilho; Lucia Specia (2012), we define a subjective four-point scale as shown in table 1 for Post-Editability. A score is assigned manually for each translated segment. Then we report Post-Editability for an MT system as the average of the scores for a given set of translations produced by that MT system.

Score	**Description**
1	Cannot be Post-Edited (better to translate from scratch)
2	Can be Post-Edited with difficulty
3	Easily Post-Edited (minimal editing)
4	No need for editing (perfect translation)

Table 1: Post-Editability Score

Score 3 is given to segments when the post-editing is minimal, such as: handling missing plural markers, spelling errors, resolving case syncretism, adding required punctuation, etc. The meaning of the segment is easily understood. Score 2 is given to segments where some time and effort is spent for Post-Editing but Post-Editing MT output appears to be better than translating from scratch. Consulting a dictionary, choosing the correct sense of a word, obtaining proper syntactic structures by re-ordering the words, handling untranslated words, etc. may be required, in addition to dealing with minor problems in spelling and grammar as in the previous case. An example for each score is given in table 2.

4. Automatic MT Evaluation Metrics

Evaluating the quality of Machine Translation outputs is a challenging task, as there can be many possible translations that are equally good. While manual evaluation is the only way to check if meaning and other required attributes of the source language sentence are fully and properly preserved and/or transferred, manual evaluation requires the time and effort of expert translators who know both the source and target languages. As a more practicable alternative, several methods have been devised to automatically evaluate the quality of translations. Here the aim is not really to check if the meaning is preserved or not but to get a comparative feel between different MT systems or different versions of a given MT system. Automatic evaluation is done by comparing, in some crude sense, the actual translations produced, and reference translations provided by human translators. See Chris Callison-Burch; Miles Osborne; Philipp Koehn (2006) for critical evaluation of BLEU, Kaushal Kumar Maurya; Renjith P. Ravindran; Ch Ram Anirudh; Kavi Narayana Murthy (2020) for a comparison of automatic and manual evaluation metrics.

Score	Example	
4	SL	People remained in their homes to avoid the cold.
	Ref	शीतलहर से बचने के लिए लोग घरों में दुबके रहे।
	MT	ठंड से बचने के लिए लोग अपने घरों में रहे।
3	SL	Which party did what.
	Ref	किस दल ने क्या किया।
	MT	किस पार्टी क्या किया।
2	SL	Everything is subsidized in Germany, from coal to cars and farmers.
	Ref	जर्मनी में सब कुछ रियायती है, कोयले से लेकर, कार, किसानों तक ।
	MT	सब कुछ जर्मनी में कोयले से लेकर और किसानों तक है।
1	SL	Libertarians have joined environmental groups in lobbying to allow the government to use the little boxes to keep track of the miles you drive, and possibly where you drive them - then use the information to draw up a tax bill.
	Ref	आपने द्वारा ड्राइव किए गए मील, तथा संभवतः ड्राइव किए गए स्थान का विवरण रखने - और फिर इस सूचना का उपयोग टैक्स बिल तैयार करने के लिए - सरकार को इन ब्लैक बॉक्स का उपयोग करने की अनुमति देने के पक्ष में समर्थन जुटाने के लिए लिबरेटेरियन पर्यावरणीय समूहों के साथ मिल गए हैं।
	MT	स्वतंत्रता में पर्यावरण समूहों का उपयोग करने की अनुमति देने के लिए सरकार ने आखिर छोटे बक्से का ट्रैक रखने और तुम्हें ड्राइव दूर है , तो आप उन्हें का

		उपयोग करने के बारे में जानकारी प्राप्त कर लेते हैं ।

Table 2: Examples of MT outputs and Post-Editability scores. References are included from the test data.

We have used BLEU, Meteor, and HTER for comparing MT systems as well as for correlating with Post-Editability in our experiments. These metrics are described briefly below.

BLEU

Bilingual Evaluation Understudy (BLEU) (Kishore Papineni; Salim Roukos; Todd Ward; Wei-Jing Zhu 2002) is based on matching of n-gram sequences of words in MT outputs and Reference Translations. BLEU score is the product of the geometric mean of n-gram precision scores with brevity penalty. The precision score used in BLEU is called the modified precision score. It is computed as follows: first, the maximum number of times an n-gram occurs in any single reference translation is counted (n-gram_max_ref); second, the number of times the n-gram occurs in the MT output is counted (n-gram_count); third, the minimum of n-gram_max_ref and n-gram_count is called count; finally, the counts for all n-grams are added and divided by the number of n-grams in the MT output. Brevity penalty is used to penalize the scores if the output segment is shorter than the reference segments. For a detailed explanation, readers may refer to (Kishore, Papineni; Salim, Roukos; Todd, Ward & Wei-Jing Zhu 2002). BLEU has been shown to correlate well with manual evaluation when evaluated at the system level. Drawbacks are: it gives equal weightage to all words; it fails if exact n-grams are not present in the reference translations and it cannot bring out the problems in translation quality to improve the MT systems further (Chris, Callison-Burch; Miles, Osborne & Philipp, Koehn 2006). It has been a useful evaluation resource for tuning statistical MT systems (Och

2003), (Wolfgang, Macherey; Franz, Josef Och; Ignacio, Thayer & Jakob Uszkoreit 2008).

Meteor

Meteor (Banerjee & Lavie 2005) is based on the matching of unigrams between MT output texts and Reference Translations. If it fails to match exact unigrams, it searches for morphological variants based on the stems of the words. If this also fails, it tries to match synonyms. This requires linguistic resources such as morphological analyzers for stemming and WordNet (Miller 1995) for synonyms. Based on the number of matches found, the precision and recall of unigram matches are calculated. A weighted harmonic mean of the precision and recall are computed. Often, Meteor has been found to be correlating well with human judgments better than BLEU. However, its usage is limited by the availability of linguistic resources.

HTER

Translation Edit (Error) Rate (TER) (Olive 2005) calculates the minimum number of editing operations required to transform the MT output segment to a reference translation. TER is the ratio of the number of edits to the average number of words in reference. Matthew Snover; Bonnie Dorr; Richard Schwartz; Linnea Micciulla; John Makhoul (2006) proposed a modification to this, called Human mediated TER (HTER), in which they calculate the minimum number of edits required by a human to transform the MT output (called a hypothesis) into fluent target language segment that is nearest in meaning to the reference translation. This showed a high correlation with both human judgments and automatic evaluation metrics. Often, HTER is also reported as post-editing effort in literature, since this measure depicts human effort involved.

5. Error Analysis

MT evaluation methods described in the previous section help us to know the quality of a given MT system or to compare various MT systems. Often, MT developers and researchers may need additional information like: which aspects of language is the MT system failing in? Which aspects is a system good at? Such insights guide researchers in improving MT systems and in combining various MT systems for boosting performance. Error analysis of MT outputs helps in understanding which errors are significantly affecting the performance of an MT system. David, Vilar; Jia, Xu; D'Haro Luis Fernando & Hermann, Ney (2006) proposed a framework for error analysis and classification of phrase-based MT outputs. The error taxonomy had five major classes: missing words, word order errors, incorrect words, unknown words, and punctuation errors. These errors were manually tagged and hence, the process is difficult, expensive, and takes a lot of time. Popović & Ney (2011) have proposed a framework for counting the errors in the output automatically when reference translations are provided. The idea is to use the standard edit-rate measures namely Word Error Rate (WER) and Position-independent word Error Rate (PER) in combination with linguistic knowledge like base forms and POS tags to identify the errors. They focused on the following types of errors:

- Inflectional Errors
- Re-ordering Errors
- Missing Words
- Extra Words
- Incorrect Lexical Choices

WER is based on the Levenshtein distance algorithm (Levenshtein 1966), which returns the number of editing operations namely, insertions, deletions, and substitutions, of words required for transforming the hypothesis into a reference translation. PER is further classified into two: recall-based ReferencePER (RPER) and precision-based Hypothesis-PER (HPER). Words that appear in the reference but do not appear in the hypothesis are called RPER errors. Words that appear in the hypothesis but do not appear in the reference are HPER errors. Once The WER, HPER, and RPER errors have been identified, errors are classified in the following manner, using base forms of the words:

- Inflectional error: a word that is marked as an HPER/RPER error, but base forms are the same in hypothesis and reference
- Re-ordering error: a word which occurs in both reference and hypothesis, thus not contributing to HPER/RPER, but marked as WER error
- Missing word: a word that is identified as a deletion error in WER, as well as an RPER error, without sharing the base form with any hypothesis error
- Extra word: a word that is identified as an insertion error in WER, as well as an HPER error, without sharing the base form with any reference error
- Incorrect lexical choice: a word that is neither an inflection error nor a missing or extra word is classified as a lexical error

The procedure suggested above has been implemented and shared by the authors via a tool named hjerson (Popović 2011). It is implemented in python and shared under the GNU General Public License[4]. Two examples of errors identified by hjerson along with the source language sentence are given in table 3.

[4] https://github.com/cidermole/hjerson

English: The rain and cold wind on Wednesday night made people feel cold.

ref-err-cats: बुधवार~~x रात~~x की~~lex बारिश~~lex और~~x सर्द~~lex हवा~~x से~~x लोगों~~x को~~x ठंड~~x लगी~~infl |~~lex

hyp-err-cats: बुधवार~~x रात~~x को~~x वर्षा~~lex और~~x ठंडी~~lex हवा~~x से~~x लोगों~~x को~~x ठंड~~x लगती~~infl है~~ext ।~~lex

English: Everything is subsidized in Germany, from coal, to cars and farmers.

ref-err-cats: कोयला~~infl से~~reord लेकर~~reord कार~~miss तक~~reord, ~~miss और~~reord कृषकों~~miss तक~~lex, ~~lex जर्मनी~~x में~~x सब~~reord कुछ~~reord आर्थिक~~lex -~~lex सहायता~~lex प्राप्त~~lex ह~~x | है~~lex

hyp-err-cats: सब~~reord कुछ~~reord जर्मनी~~x में~~x कोयले~~infl से~~reord लेकर~~reord और~~reord किसानों~~lex तक~~reord है~~x ।~~lex

Table 3: Two examples of MT outputs (hypothesis) and references with errors identified by hjerson. x-means no error.

6. Setup of the Experiments

6.1 Corpus

Center for Indian Language Technology (CFILT) at Indian Institute of Technology Bombay (IIT-B), has compiled an English-Hindi parallel corpus[5] and made it publicly available in the year 2018 (Anoop, Kunchukuttan; Pratik, Mehta &

[5] http://www.cfilt.iitb.ac.in/iitb_parallel/

Pushpak, Bhattacharyya 2018). This is a compilation of previously publicly available corpora as well as corpora developed at CFILT. Version 1.0 contains 1.49 million segments. Development and Test sets have 520 and 2507 segments respectively. This corpus is available for non-commercial use under the Creative Commons Attribution-NonCommercial-ShareAlike License, which allows us to use the data for research purposes (non-commercial). Hindi monolingual corpus is also made available by the same group and is used for language model training in our PBMT system. This corpus has 45 million sentences and 844 million tokens.

6.2 Post-Editors and Data

In our experiments, Post-Editing has been done by four freelance professional translators, labeled T1, T2, T3, and T4. The Mother tongue of all the translators is Hindi. In our first experiment, we find inter-annotator agreement among the participants. Each post-editor is given a sample of 30 segments for postediting and scoring for Post-Editability. The second experiment is to find the Post-Editability of each MT system output. For this, each post-editor is given 100 segments translated using any one MT system, for post-editing and scoring. No two post-editors get outputs of the same MT system.

Data for the experiments are taken from the test data (2507 segments) of the IIT Bombay English-Hindi parallel corpus mentioned above. We randomly pick 30 segments, of which 10 segments are further picked randomly for translation using PBMT, 10 for NMT and the remaining using Google NMT. These 30 segments and their translations are used for finding inter-annotator agreement. From the remaining segments (2477), 100 segments are randomly selected for the next experiment. Outputs of PBMT, NMT, and Google translate are given to T1, T2, and T3 respectively. T4 translates 100 segments completely manually.

6.3 PET - Post-Editing Tool

For all the experiments involving post-editors, the Post-Editing Tool (PET), developed by Wilker Aziz; Sheila Castilho; Lucia Specia (2012) is used. This tool mainly collects the implicit and explicit effort indicators while performing the post-editing task. These indicators include time taken for post-editing, number of keystrokes required in post-editing, quality rating by post-editors, and HTER. PET can also be used for completely manual translation. It is developed using Java-6 and works on any platform installed with the Java virtual machine. The tool also allows adding glossaries, dictionaries, etc. for supporting post-editors while post-editing. We do not use any such aids; however, we allow post-editors to use any of the online dictionaries such as www.shabdkosh.com for reference. Using the tool is pretty easy. The post-Editors are supported in installing and using the tool through a tutorial video and a tutorial document. PET is shared by the developers under GNU general public license (GPL).

6.4 MT Systems

6.4.1 Phrase-Based SMT System

For the experiments in this paper, we use the baseline system mentioned in Anoop Kunchukuttan; Abhijit Mishra; Rajen Chatterjee; Ritesh Shah; Pushpak Bhattacharyya's writings (2014). Training is done using the Moses[6] system, with the options set to grow-diag-final-and for extracting phrases and msdbidirectional-fe for lexicalized reordering. Tuning is done using Minimum Error Rate Training (MERT) with default parameters (100 best lists, max 25 iterations). Language model

[6] http://www.statmt.org/moses/

(5-gram) is trained on Hindi monolingual corpus using KenLM (Heafield 2011) (available with Moses) with Kneser-Ney smoothing.

6.4.2 Neural MT System

Baseline NMT with attention method, as specified in the 6th Workshop on Asian Translation (WAT2019) (Toshiaki, Nakazawa; Nobushige, Doi; Shohei, Higashiyama; Chenchen, Ding; Raj, Dabre; Hideya, Mino; Isao, Goto; Win, Pa Pa; Anoop, Kunchukuttan; Shantipriya, Parida; Ondřej, Bojar & Sadao, Kurohashi 2019), for OpenNMT[7] is used. Configuration is given below:

```
encoder_type = brnn
brnn_merge = concat
src_seq_length = 150
tgt_seq_length = 150
src_vocab_size = 100000
tgt_vocab_size = 100000
src_words_min_frequency = 1
tgt_words_min_frequency = 1
```

6.4.3 Google Translate

Google Translate[8] started in the year 2006 as a free translation service, with Statistical Machine Translation in the back-end. In November 2016, Google announced that it shifted to the Neural Machine Translation paradigm. Google possesses data that is two to three decimal orders greater in magnitude compared to the state of the art, for the language pairs like

[7] https://opennmt.net/OpenNMT/

[8] https://translate.google.com/

English-German, English-French, English-Spanish (Yonghui Wu; Mike Schuster; Zhifeng Chen; Quoc V. Le; Mohammad Norouzi; Wolfgang Macherey; Maxim Krikun; Yuan Cao; Qin Gao; Klaus Macherey; Jeff Klingner; Apurva Shah; Melvin Johnson; Xiaobing Liu; Łukasz Kaiser; Stephan Gouws; Yoshikiyo Kato; Taku Kudo; Hideto Kazawa; Keith Stevens; George Kurian; Nishant Patil; Wei Wang; Cliff Young; Jason Smith; Jason Riesa; Alex Rudnick; Oriol Vinyals; Greg Corrado; Macduff Hughes; Jeffrey Dean 2016). It is well known that NMT is data-hungry and Google MT could be giving better results in comparison with other systems simply because of the extremely large data they may have used for training. Therefore, outright comparisons cannot be made with other MT systems we use in our work. How much data is used for English-Hindi translation is not known. Here we include Google Translate in our experiments just to get a general comparative idea.

7. Experiments and Results

The first experiment is to find the inter-annotator agreement among Post-Editors. The pair-wise inter-annotator agreement is reported using Cohen’s kappa coefficient (Cohen 1960). Agreement among all the annotators is reported using the Fleiss kappa coefficient (Fleiss 1971). The second experiment is to measure the Post-Editability of MT system outputs. The Post-Editability scores are presented alongside various automatic evaluation metrics. Correlation between Post-Editability and various automatic evaluation metrics are presented. Further, we probe into the influence of errors in MT output on post-editing effort indicators.

7.1 Inter-Annotator Agreement

The pair-wise inter-annotator agreement is computed using Cohen’s kappa (κ) (Cohen 1960). Interpretation of Cohen’s kappa (Landis & Koch 1977) coefficient value is as follows: κ

< 0.00-Poor agreement, 0.00 ≤ κ ≤ 0.20-Slight agreement, 0.21 ≤ κ ≤ 0.40-Fair agreement, 0.41 ≤ κ ≤ 0.60-moderate agreement, 0.61 ≤ κ ≤ 0.80-Substantial agreement, 0.81 ≤ κ ≤ 1.00-Almost perfect agreement. It is customary to report the average of pair-wise Cohen's kappa scores when the number of annotators is more than two. Table 4 shows pair-wise Cohen's kappa values for all translator pairs. The average of the pairwise Cohen's kappa is found to be 0.258, which shows that there is a fair agreement between annotators. All the pairs of post-editors except T1 and T2 have shown a fair agreement. Fleiss kappa coefficient (Fleiss 1971) value, for assessing agreement among all the post-editors turns out to be 0.229. Fleiss kappa also shows a fair agreement between the annotators (interpretation is the same as Cohen's kappa). In fact, in MT literature, it is very common to find fair agreement among annotators when it comes to evaluation methods involving methods similar to Likert[9] scales (Chris, Callison-Burch; Cameron, Fordyce; Philipp, Koehn; Christof, Monz & Josh Schroeder 2007).

Post-Editor Pair	Cohen's Kappa
T1,T2	0.183
T1,T3	0.300
T1,T4	0.300
T2,T3	0.233
T2,T4	0.250
T3,T4	0.283
Mean	0.258

Table 4: Cohen's kappa for pair-wise Inter Annotator Agreement.

[9] Likert scale is a psychometric scale commonly used in research work employing questionnaires. An example is one in which a user asked about whether one agrees or disagrees with a statement in a graded scale: disagree, weakly agree, cannot say, agree, strongly agree.

7.2 Time taken: Post-editing vs. Manual Translation

The time taken for translation by post-editing MT system outputs is compared here with the time taken for completely manual translation (Table 5). We can observe that the time taken for translation using an MT system and later post-editing the outputs is less than the time taken for completely manual translation. This is an important observation since post-editing may be altogether useless if it does not speed up the task of translation. Post-editing Google translate’s output shows the highest reduction (17.5%) in time whereas PBMT (11.6%) has shown the least. To confirm whether this difference in times happened by chance or it is statistically significant, we perform an independent sample t-test between the time taken by post-editing and completely manual translation, as well as post-editing time between different systems. In comparison with completely manual translation, we observe that there is a significant difference in post-editing NMT ($t(198) = 1.67,p < 0.05$)[10] and Google translate ($t(198) = 2.22,p < 0.05$) outputs, whereas it is not statistically significant in comparison with PBMT. Also, we observe that there is no significant difference in post-editing time given any two MT systems. Figure 1 further substantiates this observation. There is an 11%-17% reduction in time taken using MT systems for translation. Machine Translation saves time but not much more than this.

Method	Post-Editing Time (%Reduction in time)
PBMT	98.36 (11.6%)
NMT	95.40 (14.3%)

[10] $t(df)$ stands for the t−*statistic*; df is degrees of freedom and equal to (n^1-1+n^2-1) where n_1 and n_2 are sizes of sample 1 and sample 2 respectively; this is a standard format of reporting hypothesis tests in American Psychological Association (APA) style.

Google	91.83 (17.5%)

Table 5: Time taken in minutes to post-edit 100 segments. Time taken for completely manual translation is 111.30 minutes.

7.3 Post-Editability

Based on the scores given by the post-editors, we present the Post-Editability of the three MT systems in table 6. BLEU, Meteor, and HTER scores are also presented alongside. It can be seen that HTER (lower the better) is decreasing with an increase in Post-Editability. NMT gets a lesser score compared to PBMT and Google MT in BLEU and Meteor. This may be due to a greater number of unknown (out-of-vocabulary (OOV)) words in NMT, resulting in failure of n-gram matches, leading to lesser scores. This is in line with the observation made by Dimitar Shterionov; Riccardo Superbo; Pat Nagle; Laura Casanellas; Tony O'Dowd; Andy Way (2018) that automatic evaluation scores indicate that PBMT is better compared to NMT, whereas manual evaluation scores show that NMT is better than PBMT in performance.

Correlation of Post-Editability with automatic metrics is shown in table 7 using Pearson correlation coefficient (Pearson 1900) and Polyserial correlation coefficient (Ulf Olsson; Fritz Drasgow; Neil J Dorans 1982). Pearson correlation coefficient should ideally be used to find a correlation between two continuous variables, although it is resorted to in literature for other kinds of variables also.

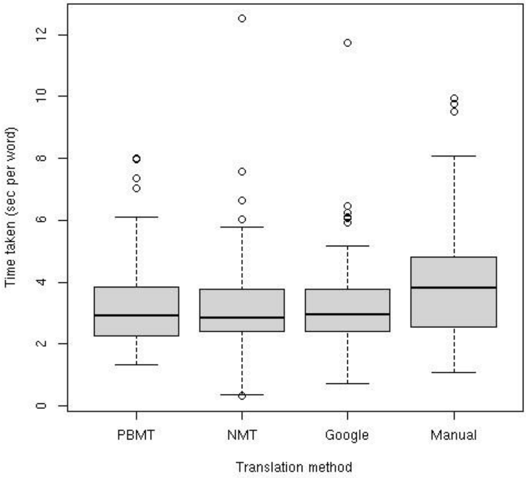

Figure 1: Box-plot indicating time taken for post-editing various MT system outputs and completely manual translation. Clearly, the medians are not significantly different among MT systems, but it is a little better for manual translation.

MT-System	BLEU	Meteor	HTER	Post-Editability Score
PBMT	14.58	0.343	0.817	2.130
NMT	13.02	0.311	0.711	2.400
Google	17.87	0.411	0.5652	2.440

Table 6: Automatic Evaluation Scores and Post-Editability Scores

In settings where one of the variables is ordinal, Polyserial coefficient is more appropriate. This is because Pearson correlation coefficient assumes variables as continuous and fails to model the reduced variance in data due to restricted values of ordinals (Francisco, Pablo Holgado-Tello; Salvador, Chacón-Moscoso; Isabel, Barbero-García & Enrique, Vila-Abad 2010). It is expected that BLEU and Meteor should positively correlate while HTER should negatively correlate with the post-editability scores. The results of our experiments are as expected. We can see those magnitudes of Polyserial correlation coefficient values are usually greater than Pearson correlation coefficient. Meteor correlated with Post-Editability better than BLEU for PBMT and NMT, which is in line with earlier findings in literature (Lavie & Agarwal 2007; Kaushal, Kumar Maurya; Renjith, P. Ravindran; Ch, Ram Anirudh & Kavi, Narayana Murthy 2020).

Metric	Pearson r			Polyserial P		
	PBMT	NMT	Google	PBMT	NMT	Google
BLEU	0.228	0.454	0.293	0.253	0.497	0.334
Meteor	0.262	0.490	0.123	0.291	0.537	0.141
HTER	-.091	-0.523	-0.485	-0.101	-0.573	-0.55

Table 7: Correlation of Post-Editability Scores with different MT evaluation metrics

The frequencies of scores for each MT system are shown in table 8 and figure 2. In PBMT outputs, 56% of segments have got score 2 and 16% of segments have got score 1, summing up to 72% of segments, whereas this sum is significantly lower for NMT and Google translate outputs (54% for both). Among NMT and Google, NMT has more segments with score 1 (10%) compared to Google (4%) which indicates the poorer performance of NMT in comparison with Google. We perform sample t-test for independent means to know whether the difference between the Post-Editability scores for the three systems is statistically significant or not. The difference is significant between PBMT-NMT($t(198) = -2.7233, p < 0.01$) and PBMT-Google ($t(198) = -3.407, p < 0.01$). The difference is not significant between NMT and Google.

It is expected that post-editing effort indicators like post-editing time and the number of keystrokes correlate negatively with the Post-Editability of a system. Higher the Post-Editability, lesser should be the time taken to postedit and lesser the number of keystrokes that are required to post-edit. We see that correlation values come out as expected (table 9). This may probably be an indication that post-editing effort as perceived by post-editors reported in the form of scores is in agreement with the actual effort involved.

MT	1	2	3	4
PBMT	16	56	27	1
NMT	10	44	42	4
Google	4	50	44	2

Table 8: Frequencies of Scores for different MT systems

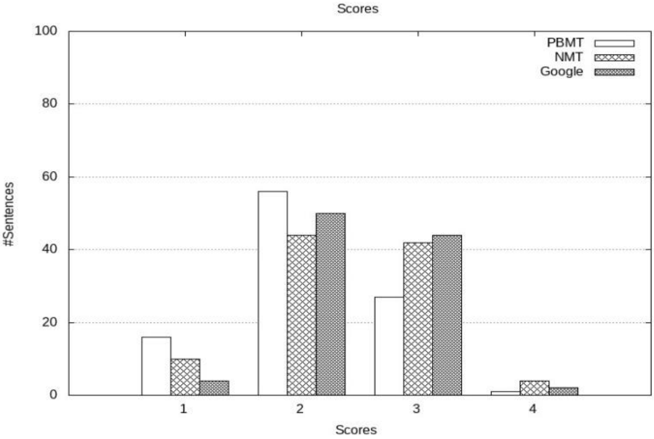

Figure 2: Number of Segments for different scores

Quantity	Pearson r			Polyserial P		
	PBMT	NMT	Google	PBMT	NMT	Google
post-editing time	-0.290	-0.299	-0.408	-0.200	-0.339	-0.154
Number of keystrokes	-0.298	-0.239	-0.396	-0.331	-0.261	-0.451

Table 9: Correlation of Post-Editability with post-editing time and number of keystrokes

7.4 Effect of MT Errors on Post-Editing Effort

Errors in the MT outputs are identified using the tool hjerson. The input to hjerson is a hypothesis (MT output), reference, and base forms (roots/stems) of both hypothesis and reference. Post-edited segments are used as references here. Stems of hypotheses and references are obtained using a Hindi stemmer[11], based on the algorithm from Ramanathan & Rao (2003). These stems are used as base forms. The tool identifies five types of errors as stated in section 5. The number of errors under each category of errors, summed over all segments for each MT system is given in table 10. It can be observed that PBMT has shown the highest number of errors and Google has shown the least number of errors. Re-ordering errors are highest in Google. Missing word errors are highest in NMT. Lexical choice errors, inflectional errors, and extra words are highest in PBMT.

Error class	PBMT	NMT	Google
Inflectional	64	58	72
Re-ordering	396	207	497
Missing words	225	437	163
Extra words	119	92	106
Lexical choice	902	696	499
Total	1706	1490	1337

Table 10: Error Types in the outputs of MT systems

Using this error analysis data and the post-editing effort indicators namely, time and number of keystrokes, we try to find the influence of the different types of errors on the post-editing effort indicators. This is done using linear mixed-

[11] https://research.variancia.com/hindi_stemmer/

effects models provided in R package lme4 (Douglas, Bates; Martin, Mächler; Ben, Bolker & Steve Walker 2015). Linear mixed-effects models allow for including random effects alongside fixed effects in linear models. Analysis of Post-editing effort indicators using linear mixed-effects modeling has been done earlier too ((Joke, Daems; Sonia, Vandepitte; Robert, J Hartsuiker & Lieve Macken 2017; Yanfang, Jia; Michael, Carl; & Xiangling Wang 2019; Spence, Green; Jeffrey Heer & Christopher D Manning 2013). Joke Daems; Sonia Vandepitte; Robert J Hartsuiker; Lieve Macken (2017) modeled the influence of errors in MT output and have shown that post-editing duration is influenced mostly by coherence errors in MT output. Yanfang Jia; Michael Carl; Xiangling Wang (2019) modeled text type (complex and simple texts) in interaction with the task for post-editing time, keystroke, and cognitive effort in terms of pause duration and found that NMT performed better than PBMT in all cases. Post-editing time and keystroke information are considered in this, as the tool used is not equipped with the facility for recording cognitive indicators such as pause and gaze data.

Here two models are built, one with post-editing time (pet) as the dependent variable and another one with the number of keystrokes (keys) as the dependent variable. pet and keys values are calculated for each segment by dividing the time taken and the number of keystrokes by the length of the corresponding source language segment. Thus, the units are post-editing time in seconds per word and the number of keystrokes per word. Error classes defined in Popović (2011) (section 5) are regrouped into three broad classes: lexical_substitution:missing and unknown word errors (ls-unk), lexical_substitution:grammatical_errors (ls-gram) and re_ordering errors (order). ls-unk are the missing words and lexical choice errors, ls-gram is the inflectional errors and order are the re-ordering errors. ls-unk, ls-gram, and order are the three fixed effects in our models. These values are calculated for each segment using hjerson and the count of each error is divided by the source segment length. Segment id (id) is chosen as a random effect. Earlier works chose subjects also as a random effect, but we do not choose since the number of subjects is too small, which may lead to overfitting. The data from different MT systems are combined into a table. MT system is defined as a categorical variable with three categories: PBMT, NMT, and Google. This is done in order to model the MT system in interaction with the errors. To assess the statistical significance of each kind of error (fixed effect) on the dependent variables, a likelihood ratio test is used. In each experiment, the alternate hypothesis is the model including all the fixed effects and the null hypothesis is the one without the fixed effect being tested. The results of the significance tests are given in table 11 for pet and in table 12 for keys.

It may be observed that pet is significantly influenced by ls-unk. keys are significantly influenced by ls-unk and order. Evidence is not enough to see if ls-gram has any significant influence on either pet or keys. The order has a significant influence on keys. Lexical choices influence both post-editing time and the number of keystrokes. Effect plots showing pet and keys versus various errors in interaction with the three MT systems are shown in figures 3 to 8. ls-unk has a positive slope against both pet and keys. Interpreting the ls-gram effect may be difficult since we have already seen that the evidence is insufficient, and the behavior seen in plots may be due to chance. As order error increases, keys (fig. 8) increases, and the effect is more on NMT and PBMT compared to Google.

Type of Error	$\chi2$ statistic and significance
ls-unk	$\chi2(3) = 66.089 (p < 0.001)$

ls-gram	$\chi2(3) = 2.0522$ (Not Significant)
order	$\chi2(3) = 3.4738$ (Not Significant)

Table 11: Significance of errors influencing post-editing time (pet)

Type of Error	$\chi2$ statistic and significance
ls-unk	$\chi2(3) = 96.28(p < 0.001)$
ls-gram	$\chi2(3) = 2.3149$ (Not Significant)
order	$\chi2(3) = 9.4342(p < 0.05)$

Table 12: Significance of errors influencing the number of keystrokes (keys)

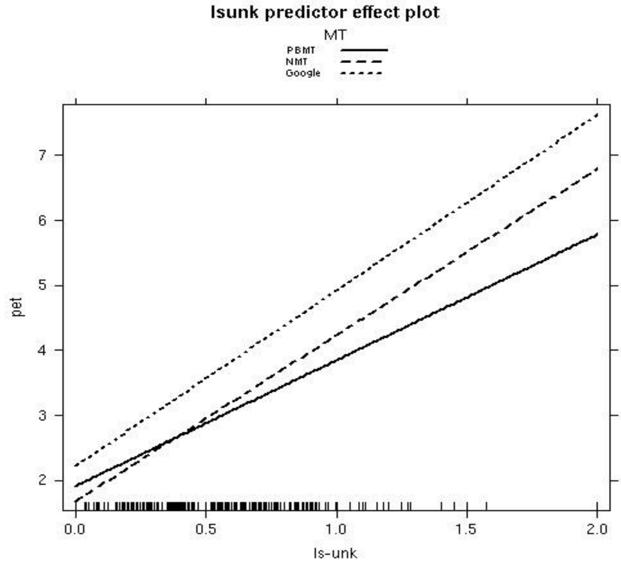

Figure 3: Effect of ls-unk on pet

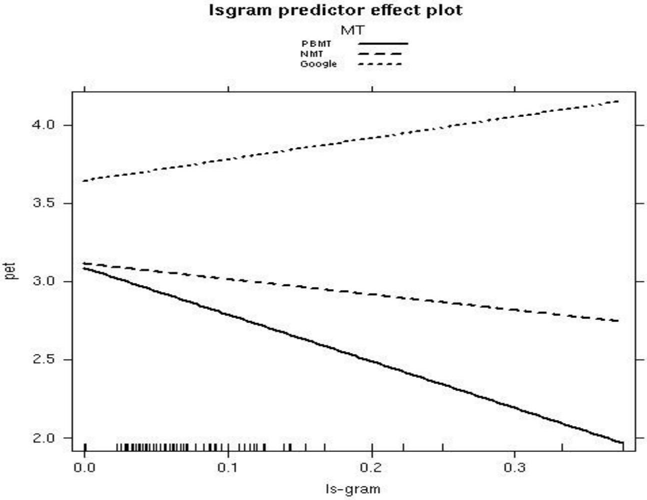

Figure 4: Effect of ls-gram on pet

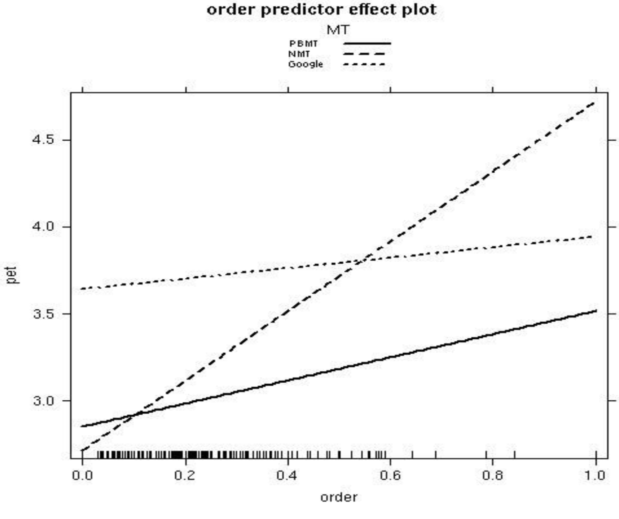

Figure 5: Effect of order on pet

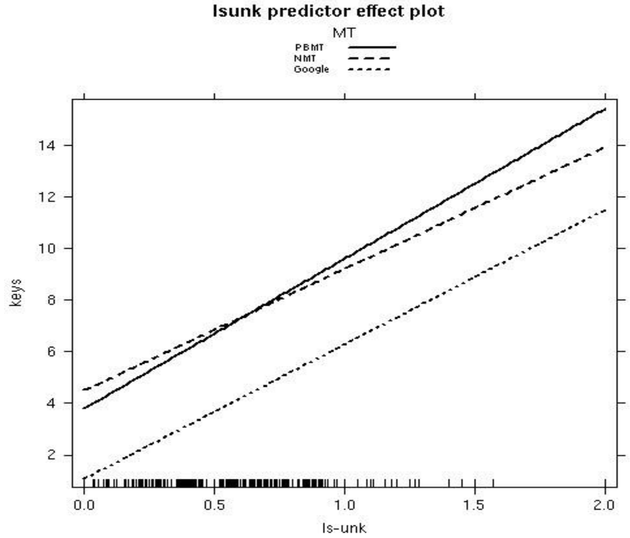

Figure 6: Effect of ls-unk on keys

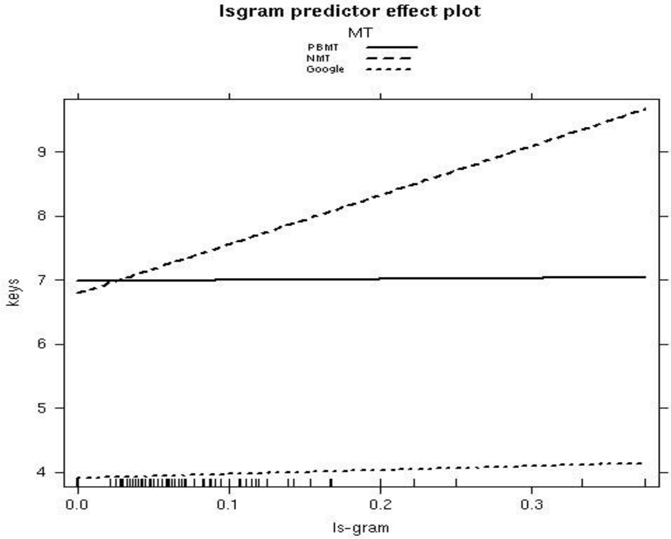

Figure 7: Effect of ls-gram on keys

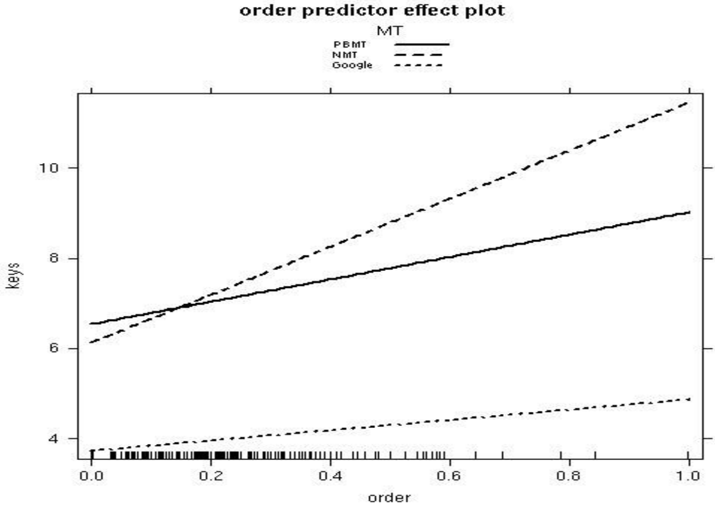

Figure 8: Effect of order on keys

7.5 Effect of Source Segment Length on Post-Editing Effort

The source segment length (slen) is now added as another fixed effect to the models defined in the previous section. The result of the significance test for modeling segment length with post-editing time and the number of keystrokes is shown in table 13. We see that segment length has a significant influence on the number of keystrokes. We present the effect plot (figure 9) for keys and slen, which shows that number of keystrokes required for post-editing increases as the segment length increases.

Post-editing Effort	χ2 statistic and significance
Number of keystrokes	χ2(1) = 6.6096(p < 0.05)
Post-editing time	χ2(1) = 2.7695 (Not significant)

Table 13: Significance of source segment length when modeled as a fixed effect on post-editing effort indicators.

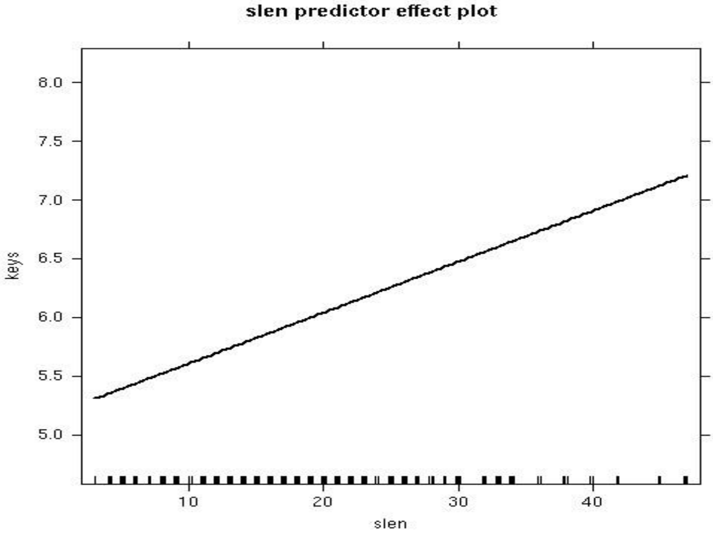

Figure 9: Effect of slen on keys

8. Discussion and Conclusions

In this paper, we have quantitatively assessed the Post-Editability of MT system outputs. We have shown that the scores we get are correlating well with MT evaluation metrics as also with the actual time and effort required for post-editing. We have explored the effect of various kinds of errors in MT outputs on post-editing time and effort. We summarize below some of the salient observations:

1. Time taken for completely manual translation is significantly (statistically) higher compared to the time required to post-edit any MT output. We found an 11% −17% reduction in overall translation time using MT. Evidence from the data is not sufficient to show any statistical significance in the pair-wise difference for the time taken for different MT systems.
2. There is a significant difference in Post-Editability scores between PBMT and NMT, as well as between

PBMT and Google. Post-editing PBMT outputs may not be as pleasant an experience as post-editing NMT or Google translate outputs. In our experiments, PBMT has got the least Post-Editability score (2.13) and Google translate has got the highest (2.44).

3. Post-Editing time is significantly influenced by missing words and lexical choice errors.
4. The number of keystrokes required to post-edit is significantly influenced by missing words and lexical choice errors (ls-unk), and re-ordering (order) errors.
5. Source segment length has a significant influence on the number of keystrokes.

We have seen that post-editing PBMT may be more difficult compared to post-editing the outputs of the other two systems. This is in agreement with the findings of Yanfang Jia; Michael Carl; Xiangling Wang (2019) and Dimitar Shterionov; Riccardo Superbo; Pat Nagle; Laura Casanellas; Tony O'Dowd; Andy Way (2018). We have observed that in NMT outputs, missing word errors are highest and re-ordering errors are the lowest. In PBMT, re-ordering errors are highest and missing word errors are lower than NMT. This shows that these two systems are perhaps complementary. These findings are in line with claims made by Popović (2018), that while PBMT is relatively better when compared to NMT in terms of out-of-vocabulary (or unknown word) errors, word order is handled very well by NMT. Based on our findings, NMT and Google may also be good candidates for combination.

Post-editing effort is influenced by missing word errors and lexical choice errors. This again highlights the importance of OOV problem. For example, Xiaoqing Li; Jiajun Zhang; Chengqing Zong (2016) showed an improvement of 4 BLEU points compared to the baseline attention-based NMT model, by handling the OOV words. Various other author’s works like Habash (2008), Biman Gujral; Huda Khayrallah; Philipp Koehn (2016), Minh-Thang Luong; Ilya Sutskever; Quoc Le; Oriol Vinyals; Wojciech Zaremba (2015) also show an increase in BLEU score by handling OOV. In another experiment on Kannada-Telugu MT (Anirudh & Murthy 2017), where both the languages have a similar syntactic structure and require minimal re-ordering, an improvement in quality of MT outputs (measured by comprehensibility) by 40% has been observed when the system databases are manually updated for OOV.

While post-editing effort is mostly influenced by missing words and lexical errors, NMT, which has the highest number of errors of this type, still managed to get good Post-Editability scores compared to PBMT in our experiments. It appears that properly ordered target language segments are more acceptable to post-editors compared to poorly ordered segments even with a smaller number of lexical errors.

References

ANIRUDH, CH RAM, & KAVI NARAYANA MURTHY. 2017. Strategies for Development of Machine Translation Systems. *Language in India* 17. 215– 222.

AZIZ, WILKER, SHEILA CASTILHO, & LUCIA SPECIA. 2012. Pet: A Tool for Post-Editing and Assessing Machine Translation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12).* 3982–3987. Istanbul, Turkey: European Language Resources Association (ELRA).

BAHDANAU, DZMITRY, KYUNGHYUN CHO, & YOSHUA BENGIO. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *3rd International Conference on Learning Representations, ICLR 2015*. San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

BANERJEE, SATANJEEV, & ALON LAVIE. 2005. Meteor: An Automatic Metric for Mt Evaluation with Improved Correlation with Human Judgments. *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, 65–72, University of Michigan, Ann Arbor. Association for Computational Linguistics.

BAR-HILLEL, YEHOSHUA. 1951. The Present State of Research on Mechanical Translation. *American Documentation* 2.229–237.

BATES, DOUGLAS, MARTIN MÄCHLER, BEN BOLKER, & STEVE WALKER. 2015. Fitting Linear Mixed-Effects Models Using Lme4. *Journal of Statistical Software* 67.1–48.

BROWN, PETER F, VINCENT J DELLA PIETRA, STEPHEN A DELLA PIETRA, & ROBERT L MERCER. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational linguistics* 19.263–311.

CALLISON-BURCH, CHRIS, CAMERON FORDYCE, PHILIPP KOEHN, CHRISTOF MONZ, & JOSH SCHROEDER. 2007. (Meta-) Evaluation of Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation, StatMT '07*, p. 136–158, USA. Association for Computational Linguistics.

CALLISON-BURCH, CHRIS, MILES OSBORNE, & PHILIPP KOEHN. 2006. Re-Evaluation the Role of Bleu in Machine Translation Research. *EACL2006: 11th Conference of the European Chapter of the Association for Computational Linguistics*, 249–256, Trento, Italy. Association for Computational Linguistics.

COHEN, JACOB. 1960. A Coefficient of Agreement For Nominal Scales. *Educational and psychological measurement* 20.37–46.

DAEMS, JOKE, SONIA VANDEPITTE, ROBERT J HARTSUIKER, & LIEVE MACKEN. 2017. Identifying the Machine Translation

Error Types with the Greatest Impact on Post-Editing Effort. *Frontiers in Psychology* 8.1282.

DENKOWSKI, MICHAEL, CHRIS DYER, & ALON LAVIE. 2014a. Learning from Post-Editing: Online Model Adaptation for Statistical Machine Translation. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 395–404, Gothenburg, Sweden. Association for Computational Linguistics.

DENKOWSKI, MICHAEL, ALON LAVIE, ISABEL LACRUZ, & CHRIS DYER. 2014b. Real Time Adaptive Machine Translation for Post-Editing with cdec and TransCenter. *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, 72–77, Gothenburg, Sweden. Association for Computational Linguistics.

EVANS, ANDREW D.C. 1986. Systran - The Translator's Viewpoint. *Terminologie et Traduction*, number 1, 17–23.

FLEISS, JOSEPH L. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological bulletin* 76.378.

GARCIA, IGNACIO. 2011. Translating By Post-Editing: Is it the Way Forward? *Machine Translation* 25.217.

GASPARI, FEDERICO, HALA ALMAGHOUT, & STEPHEN DOHERTY. 2015. A Survey of Machine Translation Competences: Insights for Translation Technology Educators and Practitioners. *Perspectives* 23.333–358.

GREEN, ROY. 1982. The MT Errors which Cause Most Trouble to Posteditors. *Practical experience of machine translation*, ed. by Veronica Lawson, 101–104. Oxford: North-Holland Publishing Company.

GREEN, SPENCE, JEFFREY HEER, & CHRISTOPHER D MANNING. 2013. The Efficacy of Human Post-Editing for Language Translation. *Proceedings of the SIGCHI conference on human factors in computing systems*, 439–448, New York, NY, USA. Association for Computing Machinery.

GUJRAL, BIMAN, HUDA KHAYRALLAH, & PHILIPP KOEHN. 2016. Translation of Unknown Words in Low Resource Languages. *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, 163–177, Austin, USA. Association for Computational Linguistics.

HABASH, NIZAR. 2008. Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation. *Proceedings of ACL-08: HLT*, Short Papers, 57–60, Columbus, Ohio. Association for Computational Linguistics.

HEAFIELD, KENNETH. 2011. Kenlm: Faster and Smaller Language Model Queries. *Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT '11*, p. 187–197, USA. Association for Computational Linguistics.

HOLGADO-TELLO, FRANCISCO PABLO, SALVADOR CHACÓN-MOSCOSO, ISABEL BARBERO-GARCÍA, & ENRIQUE VILA-ABAD. 2010. *Polychoric Versus Pearson Correlations in Exploratory and Confirmatory Factor Analysis of Ordinal Variables.* Quality & Quantity 44.153.

JÄPPINEN, & KULIKOV. 1991. Evaluation of Machine Translation Systems: A System Developer's Viewpoint. *Proceedings of the Evaluators' Forum*, 143–156, Les Rasses, Vaud, Switzerland.

JIA, YANFANG, MICHAEL CARL, & XIANGLING WANG. 2019. Post-Editing Neural Machine Translation Versus Phrase-Based Machine Translation for English-Chinese. *Machine Translation* 33.9–29.

KOEHN, PHILIPP, HIEU HOANG, ALEXANDRA BIRCH, CHRIS CALLISONBURCH, MARCELLO FEDERICO, NICOLA BERTOLDI, BROOKE COWAN, WADE SHEN, CHRISTINE MORAN, RICHARD ZENS, & OTHERS. 2007. Moses: Open-Source Toolkit for Statistical Machine Translation. *Proceedings of the 45th annual meeting of the ACL on* *interactive poster and demonstration sessions*, 177–180, Prague, Czech Republic. Association for Computational Linguistics.

KOEHN, PHILIPP. 2009. Statistical Machine Translation. Cambridge University Press.

KOEHN, PHILIPP, FRANZ JOSEF OCH, & DANIEL MARCU. 2003. Statistical Phrase-based Translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology,* Volume 1, 48–54, Edmonton. Association for Computational Linguistics.

KRINGS, HANS P. 2001. Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes, volume 5. Kent State University Press.

KUNCHUKUTTAN, ANOOP, PRATIK MEHTA, & PUSHPAK BHATTACHARYYA. 2018. The IIT Bombay English-Hindi Parallel Corpus. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018),* 3473–3476, Miyazaki, Japan. European Language Resources Association (ELRA).

KUNCHUKUTTAN, ANOOP, ABHIJIT MISHRA, RAJEN CHATTERJEE, RITESH SHAH, & PUSHPAK BHATTACHARYYA. 2014. Shata-anuvadak: Tackling Multiway Translation of Indian Languages. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* 1781–1787, Reykjavik, Iceland. European Language Resources Association (ELRA).

LANDIS, J. RICHARD, & GARY G. KOCH. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33.159–174.

LAVIE, ALON, & ABHAYA AGARWAL. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. *Proceedings of the second workshop on statistical machine translation*, 228–

231, Prague, Czech Republic. Association for Computational Linguistics.

LAVOREL, BERNARD. 1982. Experience in English-French Post-Editing. *Practical experience of machine translation*, ed. by Veronica Lawson, 105–109. Oxford: North-Holland Publishing Company.

LEVENSHTEIN, VLADIMIR I. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet physics doklady* 10.707–710.

LI, XIAOQING, JIAJUN ZHANG, & CHENGQING ZONG. 2016. Towards Zero Unknown Word in Neural Machine Translation. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2852–2858, New York, New York, USA. AAAI Press.

LUONG, MINH-THANG, ILYA SUTSKEVER, QUOC LE, ORIOL VINYALS, & WOJCIECH ZAREMBA. 2015. Addressing the Rare Word Problem in Neural Machine Translation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 11–19, Beijing, China. Association for Computational Linguistics.

MACHEREY, WOLFGANG, FRANZ JOSEF OCH, IGNACIO THAYER, & JAKOB USZKOREIT. 2008. Lattice-based Minimum Error Rate Training for Statistical Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP ’08*, p. 725–734, USA. Association for Computational Linguistics.

MAURYA, KAUSHAL KUMAR, RENJITH P. RAVINDRAN, CH RAM ANIRUDH, & KAVI NARAYANA MURTHY. 2020. Machine Translation Evaluation: Manual versus Automatic - A Comparative Study. *Data Engineering and Communication Technology*, ed. by K. Srujan Raju, Roman Senkerik, Satya

Prasad Lanka, & V. Rajagopal, 541–553, Singapore. Springer Singapore.

MILLER, GEORGE A. 1995. Wordnet: A Lexical Database for English. *Commun. ACM* 38.39–41.

NAKAZAWA, TOSHIAKI, NOBUSHIGE DOI, SHOHEI HIGASHIYAMA, CHENCHEN DING, RAJ DABRE, HIDEYA MINO, ISAO GOTO, WIN PA PA, ANOOP KUNCHUKUTTAN, SHANTIPRIYA PARIDA, ONDŘEJ BOJAR, & SADAO KUROHASHI. 2019. Overview of the 6th Workshop on Asian Translation. *Proceedings of the 6th Workshop on Asian Translation*, 1–35, Hong Kong, China. Association for Computational Linguistics.

NITZKE, JEAN. 2016. Monolingual Post-editing: An Exploratory Study on Research Behaviour and Target Text Quality. *Eyetracking and applied linguistics*, 2, 83-108. Berlin: Language Science Press.

OCH, FRANZ JOSEF. 2003. Minimum Error Rate Training in Statistical Machine Translation. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL ’03, p. 160–167, USA. Association for Computational Linguistics.

OLIVE, JOSEPH. 2005. Global Autonomous Language Exploitation (GALE). *DARPA/IPTO Proposer Information Pamphlet*.

OLSSON, ULF, FRITZ DRASGOW, & NEIL J DORANS. 1982. The Polyserial Correlation Coefficient. *Psychometrika* 47.337–347.

O’BRIEN, SHARON. 2002. Teaching Post-editing: A Proposal for Course Content. In 6th EAMT Workshop Teaching *Machine Translation*, 99–106.

O’BRIEN, SHARON. 2004. Machine Translatability and Post-editing Effort: How do they Relate. *Translating and the Computer* 26.

PAPINENI, KISHORE, SALIM ROUKOS, TODD WARD, & WEI-JING ZHU. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

PEARSON, KARL. 1900. On The Criterion That A Given System Of Deviations From The Probable In The Case Of A Correlated System Of Variables Is Such That It Can Be Reasonably Supposed To Have Arisen From Random Sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.157–175.

POPOVIĆ, MAJA. 2011. Hjerson: An Open-Source Tool For Automatic Error Classification Of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics* 96.59–67.

POPOVIĆ, MAJA. 2018. Error Classification and Analysis for Machine Translation Quality Assessment, 129–158. *Translation Quality Assessment*. Cham: Springer International Publishing.

POPOVIĆ, MAJA. 2018. Language-Related Issues for NMT and PBMT for English–German And English–Serbian. *Machine Translation* 32.237–253.

POPOVIĆ, MAJA, & HERMANN NEY. 2011. Towards Automatic Error Analysis of Machine Translation Output. *Computational Linguistics* 37.657– 688.

RAMANATHAN, ANANTHAKRISHNAN, & DURGESH D RAO. 2003. A Lightweight Stemmer for Hindi. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL),* Budapest. Association for Computational Linguistics.

SCHÄFER, FALKO. 2003. MT Post-Editing: How to Shed Light on the "Unknown Task", Experiences Made at SAP.

Proceedings of the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, 15–17. European Association for Machine Translation.

SENEZ, DOROTHY. 1998. The Machine Translation Help Desk and the Postediting Service. *Terminologie et Traduction* 1.289–295.

SHTERIONOV, DIMITAR, RICCARDO SUPERBO, PAT NAGLE, LAURA CASANELLAS, TONY O'DOWD, & ANDY WAY. 2018. Human Versus Automatic Quality Evaluation of NMT and PBSMT. *Machine Translation* 32.217–235.

SIMIANER, PATRICK, JOERN WUEBKER, & JOHN DENERO. 2019. Measuring Immediate Adaptation Performance for Neural Machine Translation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 2038–2046, Association for Computational Linguistics. Minneapolis, Minnesota.

SNOVER, MATTHEW, BONNIE DORR, RICHARD SCHWARTZ, LINNEA MICCIULLA, & JOHN MAKHOUL. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,* 223–231, Cambridge. The Association for Machine Translation in the Americas.

SPECIA, LUCIA, & ATEFEH FARZINDAR. 2010. Estimating Machine Translation Post-Editing Effort with HTER. *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10).* 33–41.

SUTSKEVER, ILYA, ORIOL VINYALS, & QUOC V LE. 2014. Sequence to Sequence Learning with Neural Networks. *Proceedings of Advances in Neural Information Processing Systems 27,* ed. by Z. Ghahramani, M. Welling, C. Cortes,

N. D. Lawrence, & K. Q. Weinberger, 3104–3112. Curran Associates, Inc.

VILAR, DAVID, JIA XU, D'HARO LUIS FERNANDO, & HERMANN NEY. 2006. Error Analysis of Statistical Machine Translation Output. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06),* 697–702, Genoa, Italy. European Language Resources Association (ELRA).

WAGNER, ELIZABETH. 1985. Rapid Post-Editing of Systran. *Tools for the Trade, Translating and the Computer*, ed. by Veronica Lawson, volume 5, 199–213, London. Aslib.

WHITE, JOHN S., AND THERESA A. O'CONNELL. 1993. Evaluation of Machine Translation. *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop Held at Plainsboro*, New Jersey, March 21-24, 1993.

WU, YONGHUI, MIKE SCHUSTER, ZHIFENG CHEN, QUOC V. LE, MOHAMMAD NOROUZI, WOLFGANG MACHEREY, MAXIM KRIKUN, YUAN CAO, QIN GAO, KLAUS MACHEREY, JEFF KLINGNER, APURVA SHAH, MELVIN JOHNSON, XIAOBING LIU, ŁUKASZ KAISER, STEPHAN GOUWS, YOSHIKIYO KATO, TAKU KUDO, HIDETO KAZAWA, KEITH STEVENS, GEORGE KURIAN, NISHANT PATIL, WEI WANG, CLIFF YOUNG, JASON SMITH, JASON RIESA, ALEX RUDNICK, ORIOL VINYALS, GREG CORRADO, MACDUFF HUGHES, & JEFFREY DEAN. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR abs/1609.08144*.

YAMADA, MASARU. 2015. Can College Students be Post-Editors? An Investigation into Employing Language Learners in Machine Translation Plus Post-Editing Settings. *Machine Translation* 29.49–67.

***

Cite This Work:

ANIRUDH, CH RAM & NARAYANAMURTHY, KAVI. 2021. On Post-Editability of Machine Translated Texts. *Translation Today*, Vol. 15(1). 73-122. DOI:10.46623/tt/2021.15.1.ar4