

THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Evaluating MT for massive open online courses

Citation for published version:
Castilho, S, Moorkens, J, Gaspari, F, Sennrich, R, Way, A & Georgakopoulou, P 2018, 'Evaluating MT for massive open online courses', *Machine Translation*. https://doi.org/10.1007/s10590-018-9221-y

Digital Object Identifier (DOI):
10.1007/s10590-018-9221-y

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Machine Translation

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 19. Mar. 2024

Machine Translation manuscript No.
(will be inserted by the editor)

Evaluating MT for Massive Open Online Courses

A multifaceted comparison between PBSMT and NMT systems

Sheila Castilho · Joss Moorkens · Federico Gaspari · Rico Sennrich · Andy Way · Panayota Georgakopoulou

Received: date / Accepted: date

Abstract This article reports a multifaceted comparison between statistical and neural machine translation (MT) systems that were developed for translation of data from Massive Open Online Courses (MOOCs). The study uses four language pairs: English to German, Greek, Portuguese, and Russian. Translation quality is evaluated using automatic metrics and human evaluation, carried out by professional translators. Results show that neural MT is preferred in side-by-side ranking, and is found to contain fewer overall errors. Results are less clear-cut for some error categories, and for temporal and technical post-editing effort. In addition, results are reported based on sentence length, showing advantages and disadvantages depending on the particular language pair and MT paradigm.

Keywords Neural MT · Statistical MT · Human MT evaluation · MOOCs

1 Introduction

1.1 Background

Since the inception of machine translation (MT), new techniques have regularly generated high expectation, often followed by disappointing results. Qualitative improvements have tended to be incremental rather than exponential. Statistical machine translation (SMT) became the dominant MT paradigm for comprehension and production purposes following its adoption by Internet MT providers and the appearance of the first commercial SMT systems in the 2000s (Gaspari and Hutchins, 2007). Since then, MT for production has become mainstream, to the extent that by 2016, 56% of language service

S. Castilho
ADAPT Centre - Dublin City University
Tel.: +353 17006719
E-mail: sheila.castilho@adapcentre.ie

providers offered an integrated, client- or project-specific MT system as part of their workflows, using in-house or third-party engines trained by one of the many dedicated MT providers (Lommel and DePalma, 2016). As MT has become more popular, pressure on costs has increased, making these providers increasingly eager for a leap in MT quality (Moorkens, 2017), such that new or improved techniques reported from MT research are quickly deployed in commercial systems.

This has been the case recently, as Neural MT (NMT) has emerged as a promising new MT paradigm, raising interest in academia and industry by outperforming phrase-based statistical MT (PBSMT) systems despite many years of SMT development, based on impressive results in automatic evaluation (Bahdanau et al, 2014; Bojar et al, 2016; Sennrich et al, 2016a) and well-publicised research and deployment by Google (Wu et al, 2016). Resultant claims that NMT is producing translation "as good as human translation" or has reached "human parity" (Hassan Awadalla et al, 2018) have appeared in the media, often along with related predictions of the end of human translation as a profession [1]. Many subsequent studies have been more circumspect, reporting increases in quality when comparing NMT with PBSMT using either automatic metrics (Bahdanau et al, 2014; Jean et al, 2015), or small-scale human evaluations (Bentivogli et al, 2016; Wu et al, 2016).

1.2 Objectives and Structure of the Paper

While initial experiments concerning NMT have shown impressive results and promising potential, so far there have been a limited number of large-scale human evaluations of NMT output. The key question addressed in this work is whether NMT results also surpass those of SMT when using human evaluation. To this end, the article reports detailed results of a quantitative and qualitative comparative evaluation of SMT and NMT carried out using automatic metrics and a small number of professional translators, considering the translation of educational texts in four language pairs, from English into German (DE), Portuguese (PT), Russian (RU) and Greek (EL). We employ a variety of metrics, including side-by-side ranking, rating for accuracy and fluency, error annotation, and measurements of post-editing effort. Preliminary findings of this work were reported in Castilho et al (2017a) together with a few other use-cases, and with more details in Castilho et al (2017b), but this paper extends those results in several ways, particularly by comparing the performance of PBSMT and NMT depending on sentence length, and by relating error categories in the output to scores of adequacy and fluency for all the language pairs under consideration. One of the key contributions of this paper

[1] SDL has recently claimed to have cracked the Russian to English NMT. See https://www.sdl.com/about/news-media/press/2018/sdl-cracks-russian-to-english-neural-machine-translation.html?utm_content=bufferdc3aa&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer_solution_pillars

is that, to the best of our knowledge, it is the first study to systematically correlate fluency and adequacy ranking with error types for NMT systems.

The rest of this article is structured as follows: in Section 2, we review previous work comparing SMT and NMT output. We explain the motivation behind our experiments in Section 3. We describe our MT systems and the experimental setup in Section 4, and the results of human and automatic evaluations in Section 5. Finally, we draw the main conclusions of the study and outline promising avenues for future work in Section 6. This large-scale evaluation was part of the work of TraMOOC,[2] a European-funded project focused on the translation of Massive Open Online Courses (MOOCs), which established a replicable semi-automated methodology for high-quality MT of educational data. As such, the MT engines tested are built using generic and in-domain data from educational resources, as detailed in Section 4.1.

2 Related Work

2.1 Previous Studies Comparing SMT and NMT

A number of papers have been published recently comparing various aspects of SMT and NMT, and the following is a review of a selection of the most interesting work in this area. Bentivogli et al (2016) asked five professional translators to carry out light post-editing on 600 segments of English TED talks data translated into German. These comprised 120 segments each from one NMT and four SMT systems. Using HTER (Snover et al, 2006) to estimate the fewest possible edits from pre- to post-edit, they found that technical post-editing effort (in terms of the number of edits) when using NMT was reduced on average by 26% when compared with the best-performing SMT system. NMT output showed substantially fewer word-order errors, notably with regard to verb placement (which is particularly difficult when translating into German), and fewer lexical and morphological errors. The authors concluded that NMT has "significantly pushed ahead the state of the art", especially for morphologically rich languages and language pairs that are likely to require substantial word reordering (Bentivogli et al, 2016).

Wu et al (2016) used BLEU (Papineni et al, 2002) scores and human side-by-side assessment of 500 Wikipedia segments that had been machine-translated from English into Spanish, French, Simplified Chinese, and vice-versa. Results from this paper again show that the NMT systems strongly outperformed other approaches and improved translation quality for morphologically rich languages, with human evaluation ratings that were closer to human translation than PBSMT. The authors noted that some additional 'tweaks' would be required before NMT would be ready for real data, and NMT engines subsequently went live on Google Translate for the language pairs tested shortly after this paper was published (Schuster et al, 2016). Bur-

[2] http://tramooc.eu/

chardt et al (2017) noted a "striking improvement" in English to German translation quality after this move to NMT.

Results of the 2016 Workshop on Statistical Machine Translation (WMT16)[3] (Bojar et al, 2016) found that NMT systems were ranked above SMT and online systems for six of 12 translation directions for the news translation task, according to the official results based on human ranking. Direct assessment of six translation directions also showed strong performance of NMT in terms of fluency, with NMT systems top-ranked for 4 translation directions, but less so in terms of adequacy, with NMT systems being top-ranked for only 2 translation directions. In addition, for the automatic post-editing task, neural end-to-end systems were found to represent a "significant step forward" over a basic statistical approach.

Toral and Sánchez-Cartagena (2017) compared NMT and PBSMT for nine language pairs (English to/from Czech, German, Romanian and Russian, plus English to Finnish), with engines trained for the news translation task at WMT16. BLEU scores were higher for NMT output than PBSMT output for all language pairs, except for Russian-English and Romanian-English. NMT and SMT outputs were found to be dissimilar, with a higher inter-system variability between NMT systems. NMT systems appear to perform more reordering than SMT systems, resulting in more fluent translations (taking perplexity of MT outputs on neural language models as a proxy for fluency). The authors found that the tested NMT systems performed better than SMT for inflection and reordering errors in all language pairs. However, using the chrF1 automatic evaluation metric (Popović, 2015), which they argue is more suited to NMT, they found that SMT performed better than NMT for segments longer than 40 words. We revisit the impact of sentence length on NMT and PBSMT output quality in a detailed comparison presented in Section 3.1.

Klubička et al (2017) followed this by carrying out a manual evaluation of pure and hierarchical PBSMT (Chiang, 2005) and NMT in English to Croatian using a taxonomy adapted from the multidimensional quality metrics (MQM) (Lommel et al, 2014). The factored PBSMT system resulted in fewer accuracy errors, with NMT producing the most omission errors. However, NMT output was considered more fluent and more grammatical than that from the other systems, with fewer fluency errors and fewer errors overall found by the two annotators.

Popović (2017) analysed PBSMT and NMT output for English-German and German-English. She found that NMT produced fewer overall errors, and that NMT output contained improved verb order and verb forms, with fewer verbal omissions. NMT also performed strongly regarding morphology and word order, particularly on articles, phrase structure, English noun collocations, and German compound words. Her finding that NMT was less successful in translating prepositions, ambiguous English words, and continuous English verbs led to a suggestion of a possible hybrid SMT and NMT approach. Bur-

[3] http://www.statmt.org/wmt16/

chardt et al (2017) also suggested the investigation of hybrid systems, this time with rule-based MT, as a possible future research topic.

The speed and quality of NMT systems are improving as new techniques are discovered for computational efficiency and for calculation of the best possible output (e.g. Sennrich et al (2017a)). High computational and power consumption costs have so far confined the application of newly discovered techniques (and deployment of NMT more generally) to large and well-funded projects run by leading research groups.[4] Given the progress in NMT development in the last few years, it seems safe to assume that in the foreseeable future NMT systems will continue to be adopted by more users, such as multilingual institutions and organisations as well as language service providers. This prospect increases the need for comparative evaluations of the respective strengths and weaknesses of NMT vs. PBSMT for multiple language pairs and in diverse domains.

3 Methodological Rationale

3.1 Sentence Length and Machine Translation Quality

The average sentence length has been shown to be related to the register, text type or genre of texts (Kucera and Francis, 1967; Westin, 2002), and is considered a key feature for the identification of an author's style (Biber and Conrad, 2009; Lehtonen, 2015).

With specific reference to MT, a number of papers have researched the relationship between sentence length and the quality of MT output. Stymne and Ahrenberg (2012) conducted an error analysis of the output of two English-Swedish SMT systems, dividing the sentences into two sets: short (with a maximum length of 20 words) and random (any length, with average sentence length of 21.9 words). Based on this distinction and on error annotation performed by two individuals, a particularly valuable finding of this study was that the errors found in the short sentences were not representative of the errors in the whole data set, both in nature and quantity.

In NMT, translation quality has also been found to heavily depend on sentence length. Early sequence-to-sequence models without attention (Cho et al, 2014; Sutskever et al, 2014) performed poorly for long sentences. Cho et al (2014) showed that NMT output tends to suffer from a clear drop in quality (in terms of BLEU score) for longer sentences (which they defined as those with over 20 tokens), observing in particular that the problem becomes especially severe when the input sentences are longer than those used for training. The introduction of the attention model (Bahdanau et al, 2014) has mitigated the problem somewhat, but recent studies report that attentional NMT systems underperform PBSMT systems on long sentences (Koehn and Knowles, 2017).

[4] Britz et al (2017), for example, used 250,000 GPU hours, equivalent to roughly 75,000kWh for GPU power consumption alone, when testing various methods of building and extending NMT systems.

In light of this, consistently with the relevant literature, we take 20 tokens as a convenient cut-off point to differentiate long from short input sentences, and in Section 5 we compare the performance of PBSMT and NMT in terms of sentence length on this basis. In addition, since there is ample evidence that text type and genre can be described, among other criteria, according to their average sentence length, we consider this as a particularly relevant parameter to organise our experiments; while our evaluation of MT quality is conducted on user-generated educational data, we anticipate that the length-based results will help to inform the choice of MT users when selecting PBSMT or NMT for their own purposes, depending on which engine type offers the best performance for texts with the sentence length of specific interest to them.

3.2 MT Errors vis-à-vis Fluency and Adequacy

Extending the findings reported in Castilho et al (2017b), this section investigates the relationship between the types of errors identified in the MT output, based on the predefined taxonomy described in Section 4.5, and the evaluation of adequacy and fluency carried out by professional translators. Elliott et al (2004) is an example of early work analysing the correlations between error frequencies based on an *ad hoc* categorisation scheme and human evaluations of fluency and adequacy. Their automated MT evaluation system involved a hierarchical classification scheme of fluency errors in English MT output (with French as the source language), to identify error types and their frequencies in samples representing various text types. Interestingly, on the basis of the evaluation carried out by three judges and concerning four MT systems, Elliott et al (2004, p. 66) remarked that "having set out to annotate fluency errors, adequacy errors were also detectable as contributors to disfluency". Koponen (2010) started from the assumption that MT quality can be considered from different complementary perspectives such as accuracy, fluency, and fitness for purpose, emphasising, however, that preserving semantic accuracy is most important in real-world scenarios. The study therefore attempted to uncover criteria for assessing MT quality specifically in terms of the accuracy of the semantic content in the output, and showed how an analysis of the different error types can help to identify such criteria for the specific evaluation of semantic accuracy.

Stymne (2013) evaluated an English-Swedish SMT system by applying an extended error typology derived from a grammar checker and then related the error types to fluency and adequacy ratings. The author concluded that the majority of the errors made by the SMT system concerned fluency, emphasising that their automatic identification, and perhaps also correction, were possible at least in principle. Federico et al (2014) analysed the impact of different MT error types on output quality, considering various language pairs involving translations into Chinese, Arabic and Russian. They conducted two sets of experiments, one exploring the relationship between errors and human quality judgements, and the other concerning the relationship between errors

and automatic metrics. Although they did not formally differentiate between adequacy and fluency judgements as a function of the errors identified in the output, they concluded in particular that the frequency of errors of a given type did not correlate with human preference, and that errors with the strongest impact could be isolated.

Costa-jussà and Farrús (2015) investigated how linguistically motivated human evaluation (morphology, syntax, semantics, and orthography) correlated with the different types of MT systems (rule-based vs. statistical), as well as with adequacy and fluency ratings. The results showed not only that adequacy and fluency were related to the total quantity of errors made by the system, but also (more specifically) that adequacy was strongly correlated with semantics, moderately with syntax, and not at all with orthography and morphology. In contrast, fluency judgements correlated with all error categories to varying degrees, in the following descending order of importance: semantics, syntax, orthography, and morphology.

Ljubešić et al (2010) evaluated a Croatian SMT system in terms of automatic metrics and adequacy and fluency, using a randomly selected 200-sentence sample examined by a single evaluator. Segments that did not receive the highest score were error-tagged according to these four coarse categories (with their relative percentages found in the data set): all lexical items were correct, but meaning was changed by word order or punctuation errors (39.7%); lexical items were translated incorrectly (34.5%); unknown words in the source (17.2%); and, finally, typing errors in the source (8.6%). Even though the authors did not investigate in detail the connection between such errors and human ratings of adequacy and fluency, they concluded that BLEU, NIST (Doddington, 2002) and METEOR (Banerjee and Lavie, 2005) correlated very strongly with each other.

Finally, Štajner et al (2016) tackled English to Portuguese MT in the IT domain. In a scenario in which only limited domain-specific parallel data was available, they compared a PBSMT and a hybrid system, and tested three strategies to overcome the problem of data sparsity, along with a systematic in-depth error analysis. One of the objectives of this work was to explain how different fine-grained error types affect fluency and adequacy, relating the relevant human assessments to the detailed error analysis. In terms of adequacy, when both outputs were rated as similarly good, PBSMT performed slightly worse than the hybrid MT system for untranslated words and additions. In contrast, when the output of both systems was rated as similarly bad, the hybrid system made twice as many word-sense errors as PBSMT, while the errors in common concerned mistranslations, untranslated words, additions and omissions. With regard to fluency, they observed that in general, the hybrid system failed more often for capitalisation and missing punctuation marks, while PBSMT made more reordering errors, with untranslated and missing words. In particular, when fluency received low scores, errors in word order and verb mood were more harmful for PBSMT, while capitalisation and punctuation errors as well as wrong prepositions were more prominent in the hybrid MT output.

4 Experimental Setup

As part of this study, PBSMT and NMT systems were created and evaluated for four translation directions: English to German, Greek, Portuguese, and Russian. Evaluation was performed using both human and state-of-the-art automatic evaluation metrics. Professional translators carried out side-by-side ranking, adequacy and fluency rating, post-editing and error annotation based on a predefined taxonomy.

4.1 MT Systems

4.1.1 Training Data

The MT engines used in the TraMOOC project were trained on large amounts of data from various sources; we used the training data from the WMT shared translation tasks and OPUS (Tiedemann, 2012) as mixed-domain data, and as in-domain training data we used TED from WIT3 (Cettolo et al, 2012), the QCRI Educational Domain Corpus (QED) (Abdelali et al, 2014), a corpus of Coursera MOOCs, and our own collection of educational data. The amount of parallel training data used is shown in Table 1, broken down by target language.

Mixed-domain:

- Europarl v7 (Koehn, 2005)
- JRC-Acquis 3.0 (Steinberger et al, 2006)
- DGT's Translation Memory (Steinberger et al, 2012) as distributed in OPUS (Tiedemann, 2012)
- OPUS European Central Bank (ECB)
- OPUS European Medicines Agency (EMEA)
- OPUS EU Bookshop
- OPUS OpenSubtitles[5]
- WMT News Commentary
- WMT CommonCrawl
- SETimes (Tyers and Alperen, 2010)
- Yandex English-Russian Parallel Corpus[6]
- Wikipedia names and titles (English-Russian)
- Bootstrapped pseudo-indomain training data

In-domain:

- TED from WIT3 (Cettolo et al, 2012)
- QCRI Educational Domain Corpus (QED) (Abdelali et al, 2014)
- Additional in-domain data collected in the TraMOOC project

[5] `http://www.opensubtitles.org`

[6] `https://translate.yandex.ru/corpus`

Table 1 Detailed statistics of the parallel training corpora for the EN→* translation directions (number of sentences, in millions).

Parallel training corpus	**DE**	**EL**	**PT**	**RU**
Europarl	1.88	1.20	1.92	—
JRC-Acquis (OPUS)	0.58	1.12	1.18	—
DGT (OPUS)	3.12	3.10	3.09	—
ECB (OPUS)	0.11	0.10	0.19	—
EMEA (OPUS)	1.10	1.06	1.07	—
EU Bookshop (OPUS)	9.15	3.88	4.02	—
OpenSubtitles (OPUS)	5.27	20.04	20.50	18.71
SETimes (OPUS)	—	0.23	—	—
News Commentary (WMT16)	0.21	—	—	0.22
CommonCrawl (WMT16)	2.37	—	—	0.87
Yandex (WMT16)	—	—	—	1.00
Wiki Headlines (WMT16)	—	—	—	0.51
TED	0.18	0.07	0.18	0.18
QED	0.05	—	0.15	0.05
TraMOOC	0.04	0.07	0.25	2.07
Total (mixed-domain)	23.80	30.73	31.97	21.30
Total (in-domain)	0.26	0.13	0.58	2.30
Total	24.06	30.86	32.55	23.60

The PBSMT systems used the target side of all parallel corpora for language modelling, plus the News corpora from WMT as additional monolingual data. Statistics are shown in Table 2.

Table 2 Additional monolingual data size for the EN→* translation directions (number of sentences, in millions).

Target language	DE	EL	PT	RU
News (WMT)	159.7	—	—	56.2

4.1.2 Phrase-based SMT

The PBSMT system used is Moses (Koehn et al, 2007), with MGIZA (Gao and Vogel, 2008) used to train word alignments, and KenLM (Heafield, 2011) for training and scoring the language models.

The MT model is a linear combination of various features, including standard Moses features such as phrase translation probabilities, phrase and word penalty, and 5-gram language model with modified Kneser-Ney smoothing (Kneser and Ney, 1995), as well as the following advanced features: a hierarchical lexicalised reordering model (Galley and Manning, 2008), a 5-gram operation sequence model (Durrani et al, 2013), sparse features indicating phrase pair frequency, phrase length, and sparse lexical features, and, for English-Russian, we employ a transliteration model for unknown words (Durrani et al, 2014). Feature weights are optimised to maximise BLEU with batch MIRA

(Cherry and Foster, 2012) on an in-domain tuning set. For Greek and Portuguese, the in-domain tuning set consists of English MOOC data provided by Iversity and VideoLectures.NET, which were translated by professional translators within the TraMOOC project (2687 sentences). For German and Russian, the in-domain tuning sets are the development portions of the QED corpus (646 and 744 sentences, respectively).

Adaptation to the MOOC domain is performed via three mechanisms: sparse domain indicator features in the phrase table, learning of feature weights on the in-domain tuning set, and linear interpolation of language models. As component models for the language model interpolation, a separate language model is trained on each of the listed parallel corpora; the monolingual news data is divided by year, and a separate language model is trained on each portion.

4.1.3 Neural MT

The NMT systems are attentional encoder-decoder networks (Bahdanau et al, 2014), trained with Nematus (Sennrich et al, 2017b). We generally follow the settings used by Sennrich et al (2016a), namely word embeddings of size 500, hidden layers of size 1024, minibatches of size 80, and a maximum sentence length of 50. We train the models with Adadelta (Zeiler, 2012). The model is regularly validated via BLEU on a validation set, and we perform early stopping for single models. Decoding is performed with beam search with a beam size of 12.

To enable open-vocabulary translation, words are segmented via byte-pair encoding (BPE) (Sennrich et al, 2016c). For Portuguese, German, and Russian, the source and target sides of the training set for learning BPE are combined to increase consistency in the segmentation of the source and target text. For each language pair, we learn 89,500 merge operations.

For domain adaptation, we first train a model on all available training data, then fine-tune the model by continued training on in-domain training data (Luong and Manning, 2015; Sennrich et al, 2016b). Training is continued from the model that is trained on mixed-domain data, with dropout and early stopping. The models are an ensemble of 4 neural networks with the same architecture. We obtain the ensemble components by selecting the last 4 checkpoints of the mixed-domain training run, and continuing training each on in-domain data.

4.2 The MOOCs Domain

As this evaluation was intended to identify the best-performing MT system for the TraMOOC project, test sets were extracted from real MOOC data. These data included explanatory texts, subtitles from video lectures, and user-generated content (UGC) from student forums or the comment sections of e-learning resources. One of the test sets was UGC from a business development

course and the other three were transcribed subtitles from medical, physics, and social science courses. The UGC data was often poorly formulated and contained frequent grammatical errors. The other texts presented more standard grammar and syntax, but contained specialised terminology and, in the case of the physics text, non-contextual variables and formulae.

4.3 Materials, Evaluators, and Methods

For the purposes of this study, four English-language datasets consisting of 250 segments each (1K source sentences in total) were translated into German, Greek, Portuguese, and Russian using our PBSMT and NMT engines. The evaluation methods included two conditions: (i) side-by-side ranking and (ii) post-editing, assessment of adequacy and fluency, and error annotation. Both conditions were assessed by professional translators. More specifically, the ranking tasks consisted of only a subset (100 source segments) with their translations from PBSMT and NMT which were randomised. The assessments were carried out by 3 experienced professional translators (4 of them in the case of Greek). The ranking was performed using Google Forms.

For the second condition (ii), all the datasets (1K source sentences) were translated and the MT output (from both NMT and PBSMT) was mixed in each dataset, and the tasks were assigned in random order to the translators. The segments were presented sequentially, so as to maintain as much context as possible. These tasks were carried out by 3 experienced professional translators (2 in the case of English-German) using PET (Aziz et al, 2012) over a two-week period. The participants were sent the PET manual and given PET installation instructions, a short description of the overall TraMOOC project and of the specific tasks, and requested to (in the following order) (i) post-edit the MT output to achieve publishable quality in the final revised text, (ii) rate fluency and adequacy (defined as the extent to which a target segment is correct in the target language and reflects the meaning of the source segment) on a four-point scoring scale for each segment, and (iii) perform error annotation using a simple taxonomy (more details are provided in Section 4.5). This setup had the advantage that measurements of two of Krings' (2001) categories of post-editing effort could be drawn directly from PET logs, namely temporal effort (time spent post-editing) and technical effort (edit count).

4.4 Automatic Evaluation with State-of-the-Art Metrics

The BLEU, chrF3 and METEOR automatic evaluation metrics are used in this study, and two post-edits are used as references for each segment, as Popović et al (2016) suggest that the use of a single post-edited reference from the MT system under evaluation tends to introduce bias. In addition, the HTER metric (Snover et al, 2006) was used to estimate the fewest possible edits between pre- and post-edited segments.

4.5 Human Evaluation Performed by Professional Translators

Ranking: The professional translators were asked to tick a box corresponding to their preferred translation of an English source sentence for the side-by-side ranking task. PBSMT and NMT output was mixed and presented to participants using Google Forms. Two to three segments, where PBSMT and NMT output happened to be identical, were excised for each language pair, as the judges did not have the option to indicate a tie. The remaining tasks were carried out within the PET interface.

Post-editing and error annotation: The professional translators were asked to post-edit the MT segments to publishable quality, and then to highlight issues found in the MT output based on a simple error taxonomy comprising inflectional morphology, word order, omission, addition, and mistranslation. Each sentence could be annotated with more than one error category, but each error category could be assigned only once. Therefore, a segment could contain all the errors, some of the errors, as well as no errors (no issues), but if the translator found that the segment contained two mistranslation errors, for example, the mistranslation category would be assigned only once to that segment. Our expectation was that there would be fewer morphology and word order errors with NMT, especially for short segments. Section 5 investigates the relationship between these categories of errors found in the output and judgements of adequacy and fluency.

Adequacy and fluency rating: The judges were asked to rate adequacy in response to the question 'How much of the meaning expressed in the source fragment appears in the translation fragment?'. To avoid centrality bias, a scoring scale of one to four was used, where one was 'none of it' and four was 'all of it'. Similarly, fluency was rated on a one to four scale, where one was 'no fluency' and four was 'native'. Again, our expectation was that NMT would be rated positively for fluency, with possible degradation for adequacy, especially for longer segments (Cho et al, 2014; Neubig et al, 2015). The impact of sentence length is analysed in detail in Section 5.

5 Results and Discussion

The results presented in Castilho et al (2017a) and Castilho et al (2017b) have shown that NMT presents mixed results with human evaluation, while it receives consistently higher scores with automatic metrics. In this work, we summarise the previous results for automatic metrics, ranking, and post-editing, focusing on more fine-grained results of error annotation and their relationship with adequacy and fluency when making distinctions between short (20 tokens or fewer) and long sentences (more than 20 tokens). To compute statistical significance, a one-way ANOVA pairwise comparison was performed, where $p<.05$ indicates statistically significant results.

Automatic Metrics: Table 3 shows that BLEU and METEOR scores considerably increase for German, Greek, and Russian with NMT when compared

to the PBSMT scores (marginally for Portuguese).[7] HTER scores show that more post-editing was performed when using the output from the PBSMT system for German, Greek and Russian. Finally, the chrF3 scores also show good improvement for NMT over PBSMT for German and Russian, but very similar results for Greek and Portuguese. Statistically significant results are marked with ‡.

Ranking: Participants preferred NMT output across all language pairs, with a particularly marked preference for English-German (Table 4). We also applied short/long distinctions for the NMT system and found that there was a 53% preference for NMT for short segments, and a 61% preference for NMT for long segments. Even though the experimental setup does not allow us to strictly relate the individual domains to the results obtained for MT system types across all four language pairs considered in this study, we surmise that the text genres in which fluency appears to be more important (i.e. business and marketing) have yielded better evaluations for NMT, as opposed to medicine and physics; this applies also to longer segments, on the assumption that for the latter domains translators would tend to adopt a more 'literal' translation.

Post-Editing: The number of segments considered by participants to require interventions during the MT post-editing task was lower for NMT when compared to the PBSMT system. Temporal effort (seconds/segment) was marginally improved for German (PBSMT 74.8 : NMT 72.8), Greek (PBSMT 77.7 : NMT

[7] Results were statistically significant in a one-way ANOVA pairwise comparison ($p<.05$).

Table 3 Automatic Evaluation Scores.

Lang.	System	BLEU	METEOR	chrF3	HTER
DE	PBSMT	41.5	33.6	0.66	49.0
	NMT	61.2‡	42.7‡	0.76	32.2
EL	PBSMT	47.0	35.8	0.65	45.1
	NMT	56.6‡	40.1‡	0.69	38.0
PT	PBSMT	57.0	41.6	0.76	33.4
	NMT	59.9	43.4	0.77	31.6
RU	PBSMT	41.9	33.7	0.67	44.6
	NMT	57.3‡	40.65‡	0.73	33.9

Table 4 Results of human side-by-side ranking of PBSMT and NMT.

Evaluation (No. segments)	preference for PBSMT	NMT
EN-DE (300)	61 20.3%	239 79.7%
EN-EL (400)	174 43.5%	226 56.5%
EN-PT (300)	115 38.3%	185 61.7%
EN-RU (300)	110 36.7%	190 63.3%

Table 5 Words per Second (WPS).

Lang.	PBSMT	NMT
DE	0.21	0.22
EL	0.22	0.24
PT	0.29	0.30
RU	0.14	0.14

Table 6 WPS: long vs short segments.

	Lang.	PBSMT	NMT
Short (up to 20 tokens)	DE	0.21	0.26
	EL	0.24	0.27
	PT	0.33	0.38
	RU	0.15	0.13
Long (more than 20 tokens)	DE	0.21	0.20
	EL	0.20	0.22
	PT	0.26	0.25
	RU	0.13	0.14

70.4) and Portuguese (PBSMT 57.7 : NMT 55.19) when doing post-editing of NMT, whereas for Russian, PBSMT showed a slightly better performance (PBSMT 104.6 : NMT 105.6). Temporal effort is also expressed in words per second in Table 5. Technical post-editing effort (number of edits performed in keystrokes/segment) was reduced for NMT in all language pairs: German (PBSMT 5.8 : NMT 3.9), Greek (PBSMT 13.9 : NMT 12.5), Portuguese (PBSMT 3.8 : NMT 3.6) and Russian (PBSMT 7.5 : NMT 7.2) (cf. the HTER scores in Table 3).

The distinction between long and short segments was also considered for post-editing effort, where in terms of words per second (WPS) (see Table 6), the NMT system performs better with short sentences for German, Greek and Portuguese when compared to the PBSMT system, with the Portuguese language nearly reaching the average professional rate reported for English-German in Moorkens and O'Brien (2015). Interestingly, the Russian output shows a slightly better WPS average for the PBSMT system for short sentences. Regarding long sentences, Greek and Russian show fewer WPS for NMT, but Portuguese and German show fewer WPS for the PBSMT system.

Even though the results for the post-editing assessment were not statistically significant, they suggest that the PBSMT segments that were edited required more cognitive effort than NMT segments for German, Greek, and Portuguese, but the Russian output shows lower effort with PBSMT.

5.1 Error Annotation

As mentioned previously, for the error-annotation tasks professional translators annotated each target sentence according to five error categories: inflectional morphology, word order, addition, mistranslation, and omission. In this section, we first summarise the general results found in Castilho et al (2017a) and Castilho et al (2017b), and then we detail the results when distinguishing long from short sentences. Moreover, we go one step further and divide the error category into fluency and adequacy, according to the MQM typology,[8] where inflectional morphology and word order are errors related to *fluency*, while addition, mistranslation and omission are related to *adequacy*.

[8] http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html

Previous results have shown that for all the language pairs, the total number of issues found at the segment level was higher for the PBSMT system (DE=1687, EL=1277, PT=1082 and RU=1654) compared to the NMT system (DE=1234, EL=1086, PT=1003 and RU=1346). The percentage of sentences not annotated (that is, sentences considered to have no issues) was greater for NMT compared to PBSMT for all language pairs (DE: NMT=18.9%, PBSMT=6%; EL: NMT=16.8%, PBSMT=9%; PT: NMT=23.6%, PBSMT=19.7%; and RU: NMT=19.5%, PBSMT=10%).

Across all languages, the NMT output contained fewer sentences with word-order and inflectional morphology errors compared to the PBSMT output. On average, NMT also had fewer sentences with omission and addition errors than PBSMT, and about the same number with mistranslation errors. For English-Greek, however, the PBSMT output contained fewer sentences with errors of omission, addition, or mistranslation than the NMT output. For English-Russian, PBSMT contained fewer sentences with mistranslations.

Inflectional morphology, word order, and mistranslation were the most frequent problems found in both types of MT systems.

Short sentences vs. long sentences: For the experiments reported here the distinction between short and long sentences used in the error-annotation task (i.e. assuming 20 tokens as the cut-off point) is based on the work reviewed in Section 3.1. In total, from our dataset of 1K sentences, 676 sentences contained 20 tokens or fewer (short sentences), and 324 sentences contained more than 20 tokens (long sentences). Table 7 shows the total error count per sentence type, as well as the percentage of sentences annotated as containing that type of error, calculated as the number of errors per category divided by the total number of sentences.

When looking at the German language, we notice that the PBSMT system presents a slightly higher percentage of short sentences with *omission* errors (13%) than in long sentences (11%), while the NMT system shows a lower percentage of short sentences with errors for all categories when compared to the long sentences. When comparing the results for the two MT systems, the NMT system for German consistently shows a lower percentage of sentences annotated with errors when compared to the PBSMT system. This result is statistically significant for sentences with *word order* error type (‡). In long sentences, however, this difference for *word order* was only moderate (p=.07). We observe that in long sentences, both systems show similar results for *omission* and *addition* errors. Although not statistically significant, these results correlate well with results reported in Castilho et al (2017b) in which we observed that results for the adequacy assessment were less consistent, where the German PBSMT system showed slightly higher means for adequacy than the NMT system (more details are given in Section 5.2).

Regarding the Greek language, both PBSMT and NMT systems show a higher percentage of short sentences with *omission* errors (6% both systems) when compared to long sentences (PBSMT 3%, NMT 5%), which is more evident for the PBSMT system. These results were only fair (p=.10). While NMT

shows lower percentages of short sentences with *inflectional morphology* and *word order* errors compared to the PBSMT system, it shows a slightly higher percentage for sentences with *mistranslation* errors than PBSMT. Interestingly, PBSMT and NMT show the same percentage of sentences with *omission* and *addition* errors. We observe again that NMT outperforms PBSMT for long sentences with *inflectional morphology* and *word order* errors, while for sentences with *omission*, *addition* and *mistranslation* NMT shows higher error percentages than PBSMT. These results, however, were not statistically significant. Overall, these results correlate with previous findings for adequacy (Castilho et al (2017b), where no distinctions of long/short was made) when both Greek systems were found to perform equally.

For the Portuguese language, although none of the results were statistically significant, we note that the PBSMT system shows a lower percentage of long sentences with *omission* and *addition* errors compared to its results in short sentences, while NMT shows the same percentage of short and long sentences with *omission* errors, but lower percentages of errors for the other categories. When compared to PBSMT in short sentences, the Portuguese NMT system shows a lower percentage of sentences with *word order*, *addition*, *mistranslation* and *omission* errors, but, interestingly, a higher rate of sentences with *inflectional morphology* errors. When compared to PBSMT in long sentences, NMT shows higher percentages of *omission* and *addition* and a slightly higher percentage of long sentences with *mistranslation* errors than PBSMT, but fewer cases of sentences with fluency errors (*inflectional morphology* and *word order*).

Finally, regarding the Russian language, even though no statistical significance was found, we observe a lower percentage of errors for short sentences in all categories for both MT systems when compared to long sentences. Russian also shows the highest percentage of *omission* errors compared to all the other languages for both systems, in short and long sentences. When compared to PBSMT, the NMT output shows a lower percentage of errors for all categories apart from *mistranslations*, both in short and long sentences, which correlates with the somewhat mixed results obtained in the adequacy assessment reported in previous work. Interestingly, Russian has the highest number of the *addition* (in short and long sentences) and *omission* (for long sentences) types of error among all the languages. These results are also evident in Figure 1.

The percentage of sentences annotated with 'no issues' (Table 8) was higher for NMT across all target languages when compared to PBSMT. Statistically significant results between PBSMT and NMT are marked with ‡. For all languages, NMT showed a lower percentage of short sentences with errors when compared to long sentences (significance marked with *). For the PBSMT system, this difference was only statistically significant for EL, PT and RU (marked with †). We therefore observe that the specific types of errors displayed by NMT and PBSMT output are to some extent dependent on the particular language pairs involved, and are clearly influenced by the specific morphosyntactic features of the target language.

Table 7 Total error count and percentage of sentences containing errors.

SHORT SENTENCES	DE		EL		PT		RU	
($\leq$ 20 words)	PBSMT	NMT	PBSMT	NMT	PBSMT	NMT	PBSMT	NMT
Inflectional Morphology	456	375	247	168	224	230	413	295
	67%	55%	37%	25%	33%	34%	61%	44%
Word Order	202‡	84‡	170	110	98	79	89	54
	30%	12%	25%	16%	14%	12%	13%	8%
Addition	87	21	10	14	47	26	90	75
	4%	3%	2%	2%	7%	4%	14%	11%
Mistranslation	220	178	289	298	208	200	239	245
	33%	26%	43%	44%	31%	30%	35%	36%
Omission	31	47	38	40	41	38	107	83
	13%	7%	6%	6%	6%	6%	16%	12%
LONG SENTENCES	DE		EL		PT		RU	
($>$ 20 words)	PBSMT	NMT	PBSMT	NMT	PBSMT	NMT	PBSMT	NMT
Inflectional Morphology	276	233	196	139	180	148	282	211
	85%	72%	60%	43%	56%	46%	87%	65%
Word Order	180	96	133	98	118	102	108	68
	56%	30%	41%	30%	36%	31%	33%	21%
Addition	10	10	6	9	6	10	85	68
	5%	5%	4%	5%	4%	5%	28%	23%
Mistranslation	181	145	170	185	140	142	146	159
	56%	45%	52%	58%	43%	44%	45%	49%
Omission	36	37	10	17	12	20	87	80
	11%	11%	3%	5%	4%	6%	27%	25%

Table 8 Total count of sentences containing errors and percentage of sentences with "no issues".

SHORT SENTENCES	DE		EL		PT		RU	
(676 in total)	PBSMT	NMT	PBSMT	NMT	PBSMT	NMT	PBSMT	NMT
sentences with error(s)	618‡	501‡*	590‡†	527‡*	496†	469*	579‡†	498‡*
percentage of "no issues"	9%	26%	13%	22%	27%	31%	14%	26%
LONG SENTENCES	DE		EL		PT		RU	
(324 in total)	PBSMT	NMT	PBSMT	NMT	PBSMT	NMT	PBSMT	NMT
sentences with error(s)	321	310*	320†	305*	307†	295*	320†	307*
percentage of "no issues"	1%	4%	1%	6%	5%	9%	1%	5%

We note that due to the limitations of the annotation process, a metric such as the number of errors of each type per word would have allowed for a more in-depth analysis of error counts and types, but our results are still able to show the type of errors that each sentence contained.

5.2 Fluency and Adequacy

In previous work, we noted that NMT was rated as more fluent than PBSMT for all language pairs when looking at the percentage of scores assigned a 3-4 fluency value (Near Native or Native); in addition, the NMT systems appear to have fewer problems when compared against the PBSMT systems for all the

Table 9 Means for Fluency and Adequacy.

		DE		EL		PT		RU	
		PBSMT	NMT	PBSMT	NMT	PBSMT	NMT	PBSMT	NMT
SHORT	Adequacy	2.80	2.84	3.49	3.53	3.78	3.82	3.14	3.18
	Fluency	2.73	3.01	2.95	3.22	3.18	3.46	2.89	3.17
LONG	Adequacy	2.82	2.85	3.32	3.35	3.65	3.68	2.81	2.83
	Fluency	2.47	2.69	2.61	2.83	2.78	3.00	2.48	2.71

languages when looking at the percentage of scores assigned a 1-2 fluency value (No or Little Fluency). Castilho et al (2017b) discussed a few examples of how NMT improved the fluency of sentences, such as making correct use of the imperative form in German, or choosing a better semantic fit for Portuguese, correctly using the quantifying objects for Russian and successfully using the genitive form and generic masculine in Greek.

The results for adequacy, however, are less consistent. While NMT output received the highest mean ratings for all other language pairs, when considering 3-4 rankings ('Most of It' and 'All of It') as well as 1-2 rankings ('None of It' and 'Little of It'), English-German PBSMT was ranked higher, and English-Greek systems performed equally well. A few examples of errors made by both systems were given in Castilho et al (2017b), such as mistranslations and word-order errors in which polysemous terms seemed to pose the main problem for the NMT system especially for Greek and Russian.

Short sentences vs. long sentences: Table 9 shows the means for fluency and adequacy for short and long sentences. When looking at the results for short sentences, we observe that the NMT system has higher mean results for *fluency* when compared to the PBSMT system in all language pairs (only a moderate statistical significance was found (p=.09)). The difference for *adequacy*, however, is not large and not statistically significant. The same holds true for long sentences in which, again, the results for fluency highlight a marked difference between NMT and SMT (only a moderate significance was found (p=.09)). Similarly to short sentences, the means of adequacy are not as different between the systems.

These findings correlate with results presented previously for the error annotation, in which we showed that addition and omission have similar counts and percentages for NMT and PBSMT in all target languages. Similarly, in line with our expectations, long sentences show lower ratings for fluency and adequacy for both MT systems for all languages (with the exception of adequacy in German where ratings are almost identical for long and short sentences). However, it is interesting to note that German, Greek and Portuguese show similar adequacy ratings for both short and long sentences (for both NMT and PBSMT systems), whereas the difference in adequacy means for Russian is larger between short and long sentences.

Figure 1 visualises these differences, with errors per sentence calculated as the total number of errors per category (see Table 7) divided by the total number of sentences that contain errors (see Table 8). We observe that for

German, Greek and Portuguese, the most common errors for both MT systems appear to be those related to *fluency* (inflectional morphology and word order) as well as mistranslation. For Russian, however, the word-order type of error is not as prominent, and can be closely compared to addition and omission.

The mixed results previously observed for adequacy ratings might be explained when looking at the addition (both in short and long sentences), mistranslation and omission error types for all the languages in Figure 1. While for German the NMT system shows fewer long sentences with mistranslation errors, we observe close numbers to PBSMT for sentences with omissions and additions. The results for short sentences are either very close, or NMT shows fewer sentences with errors than PBSMT. For Greek, NMT shows a slightly higher number of both short and long sentences with adequacy errors than PBSMT, although the difference is more evident in long sentences, especially for the mistranslation category. For Portuguese, again the results seem very close, but we note that in long sentences adequacy errors appear more often than in the NMT system, whereas short sentences with adequacy errors seem to be less frequent for both systems. Interestingly for Russian, sentences with mistranslations seem to be the biggest problem for the NMT system when compared against the PBSMT system. These results for Greek and Russian correlate with findings reported in Castilho et al (2017b), in which it was noted that polysemous words seem to pose a big problem to NMT for these two target languages. Moreover, even when producing sentences with fewer errors on average, the number of times one error appears in a sentence may influence translators when rating fluency and adequacy. It would seem that the number of times adequacy errors appear in the sentences marked as containing issues (as seen in Figure 1) directly influences the mixed ratings for adequacy as assessed by translators.

Finally, the NMT system shows fewer sentences (both short and long) with problems than PBSMT when it comes to fluency-related errors. These results follow those for the fluency rating (see Table 9), in which we observed that the NMT system shows higher means across all languages, in both short and long sentences. Again, we believe that the number of times fluency errors appear in the sentences marked as containing issues directly influences the high ratings for fluency for the NMT system as assessed by translators.

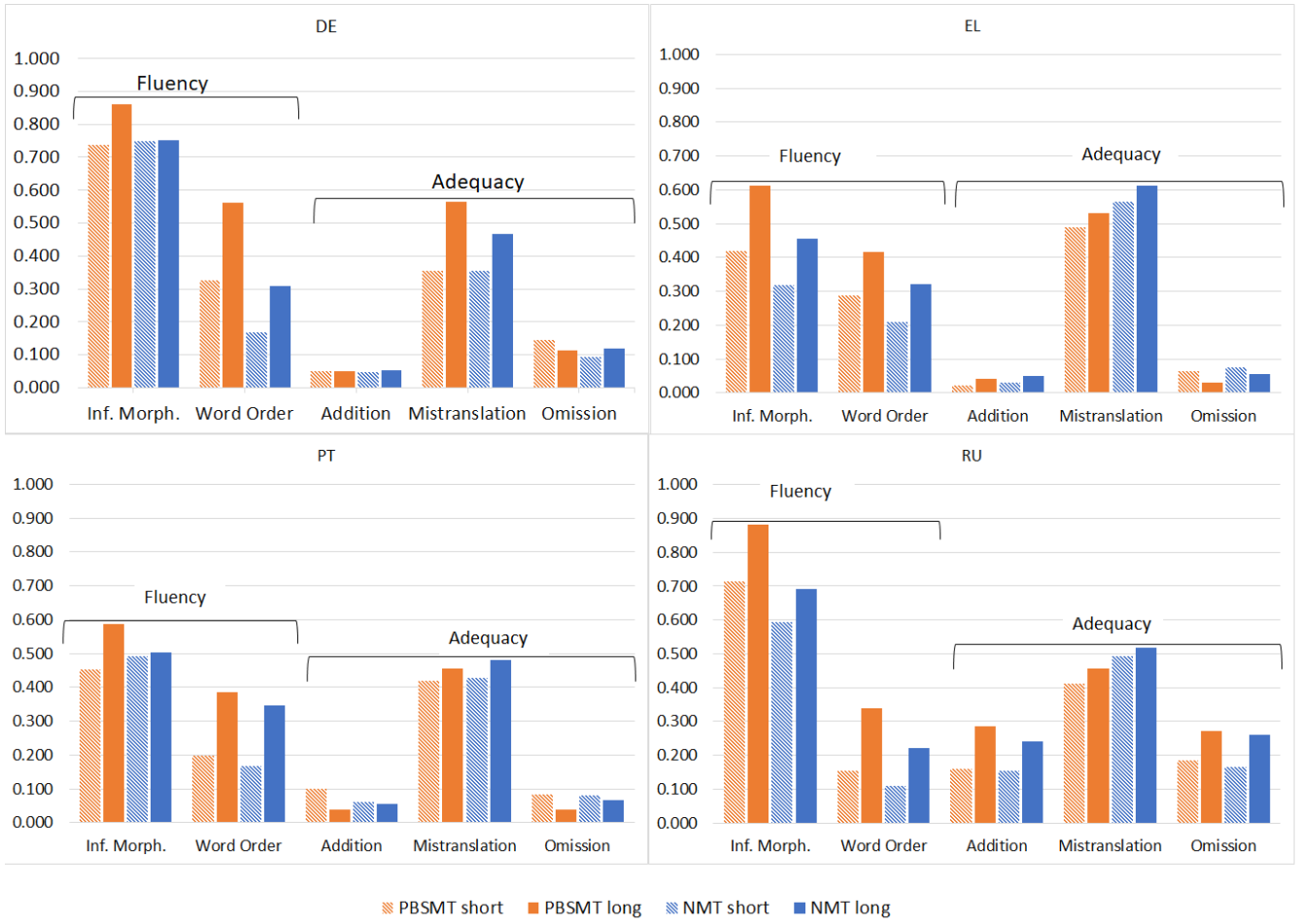


Fig. 1 Errors per sentence.

6 Conclusion and Future Work

One of the limitations of the study is the fact that it was not possible to tag an error type more than once per sentence, as the tool used did not allow for word-level tagging. However, the data analysis still shows the types of errors contained in each sentence. In work being carried out currently, we have fixed that limitation and allowed for errors to be tagged as many times as they appear in a sentence. In addition, due to lack of space, we have not discussed in detail the occasionally contradictory results obtained from automatic metrics, ranking and post-editing effort evaluations, focusing instead on the clearer, and more consistent, findings; further in-depth research into these complex areas is required to avoid offering hasty explanations that at this stage would be mostly based on speculation.

Furthermore, because of time constraints, training with participant translators was not possible. We provided a detailed guide on how to use the tool, along with guidelines on how to post-edit and assess the sentences. It is important to note that the choice of tool, evaluation data set size, and number of translators, as well as the opportunities for training, were impacted by the short amount of time available to perform the comparison, as we needed to move the TraMOOC project forward.

Nonetheless, we contend that the contributions of the study are valuable. We performed a fine-grained comparative evaluation of PBSMT and NMT systems adding to the results reported previously. We have extended the evaluation by distinguishing between long and short sentences, and correlating the types of errors annotated with fluency and adequacy rankings; to the best of our knowledge, this is the first study to correlate fluency and adequacy ranking with error types for NMT systems, in particular.

According to the chosen subset of the MQM error typology, fluency errors consisted of inflectional morphology and word order, while adequacy errors comprised cases of addition, mistranslation, and omission. The results show that rating for fluency correlates well with error annotation, in which NMT outperforms PBSMT across all language pairs in both short and long sentences. These results match our expectation that NMT would produce fewer sentences with inflectional morphology and word order errors, as it has higher fluency ratings. However, the initial hypothesis that NMT would do especially poorly on long sentences (worse than PBSMT) was not confirmed in our results for fluency. Regarding adequacy, the results for rating and error annotation are not so clear-cut, with mixed results between NMT and PBSMT. Although some differences in error type can be seen for different target languages, in general, mistranslation seems to be the biggest problem in terms of adequacy, with high frequency for both systems across all languages.

We note that NMT is currently a far more active research area than PBSMT, and that NMT systems continue to improve at a relatively steep rate according to automatic metrics. In future work, we intend to investigate to what extent this improvement in automatic metric scores is also reflected in

measurable improvements in human evaluation and increased post-editing efficiency.

Acknowledgements The TraMOOC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº644333. The ADAPT Centre for Digital Content Technology at Dublin City University is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. We would also like to thank Maja Popović for invaluable brainstorming.

References

Abdelali A, Guzman F, Sajjad H, Vogel S (2014) The AMARA Corpus: Building Parallel Language Resources for the Educational Domain. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, pp 1856–1862

Aziz W, Castilho S, Specia L (2012) PET: a Tool for Post-editing and Assessing Machine Translation. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, pp 3982–3987

Bahdanau D, Cho K, Bengio Y (2014) Neural Machine Translation by Jointly Learning to Align and Translate. CoRR abs/1409.0473

Banerjee S, Lavie A (2005) METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Ann Arbor, Michigan, vol 29, pp 65–72

Bentivogli L, Bisazza A, Cettolo M, Federico M (2016) Neural versus phrase-based machine translation quality: a case study. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, pp 257–267

Biber D, Conrad S (2009) Register, Genre, and Style. Cambridge, Cambridge University Press

Bojar O, Chatterjee R, Federmann C, Graham Y, Haddow B, Huck M, Jimeno Yepes A, Koehn P, Logacheva V, Monz C, Negri M, Neveol A, Neves M, Popel M, Post M, Rubino R, Scarton C, Specia L, Turchi M, Verspoor K, Zampieri M (2016) Findings of the 2016 Conference on Machine Translation. In: Proceedings of the First Conference on Machine Translation, Berlin, Germany, pp 131–198

Britz D, Goldie A, Luong M, Le QV (2017) Massive exploration of neural machine translation architectures. CoRR abs/1703.03906, URL `http://arxiv.org/abs/1703.03906`

Burchardt A, Macketanz V, Dehdari J, Heigold G, Peter JT, Williams P (2017) A linguistic evaluation of rule-based, phrase-based, and neural MT engines. The Prague Bulletin of Mathematical Linguistics 108(1):159–170

Castilho S, Moorkens J, Gaspari F, Calixto I, Tinsley J, Way A (2017a) Is neural machine translation the new state of the art? The Prague Bulletin of Mathematical Linguistics 108(1):109–120

Castilho S, Moorkens J, Gaspari F, Sennrich R, Sosoni V, Georgakopoulou P, Lohar P, Way A, Miceli Barone AV, Gialama M (2017b) A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In: MT Summit 2017, Nagoya, Japan, pp 116–131

Cettolo M, Girardi C, Federico M (2012) Wit^3: Web inventory of transcribed and translated talks. In: Proceedings of the 16^{th} Conference of the European Association for Machine Translation (EAMT), Trento, Italy, pp 261–268

Cherry C, Foster G (2012) Batch tuning strategies for statistical machine translation. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal, Canada, pp 427–436

Chiang D (2005) A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Ann Arbor, Michigan, pp 263–270

Cho K, van Merrienboer B, Bahdanau D, Bengio Y (2014) On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. CoRR abs/1409.1259, URL `http://arxiv.org/abs/1409.1259`

Costa-jussà MR, Farrús M (2015) Towards human linguistic machine translation evaluation. Digital Scholarship in the Humanities 30(2):157–166

Doddington G (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, HLT '02, pp 138–145

Durrani N, Fraser A, Schmid H (2013) Model With Minimal Translation Units, But Decode With Phrases. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL, Atlanta, GA, USA, pp 1–11

Durrani N, Sajjad H, Hoang H, Koehn P (2014) Integrating an Unsupervised Transliteration Model into Statistical Machine Translation. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, Gothenburg, Sweden, pp 148–153

Elliott D, Hartley A, Atwell E (2004) A fluency error categorization scheme to guide automated machine translation evaluation. In: Machine Translation: From Real Users to Research: Proceedings of the 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, USA, Berlin and Heidelberg, Springer, pp 64–73

Federico M, Negri M, Bentivogli L, Turchi M (2014) Assessing the impact of translation errors on machine translation quality with mixed-effects models. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp 1643–1653

Galley M, Manning CD (2008) A simple and effective hierarchical phrase reordering model. In: Proceedings of the Conference on Empirical Methods in

Natural Language Processing, Honolulu, Hawaii, EMNLP '08, pp 848–856
Gao Q, Vogel S (2008) Parallel Implementations of Word Alignment Tool. In: Software Engineering, Testing, and Quality Assurance for Natural Language Processing (SETQA-NLP '08), Columbus, OH, USA, pp 49–57
Gaspari F, Hutchins WJ (2007) Online and Free! Ten Years of Online Machine Translation: Origins, Developments, Current Use and Future Prospects. In: Proceedings of MT Summit XI, Copenhagen, Denmark, pp 199–206
Hassan Awadalla H, Aue A, Chen C, Chowdhary V, Clark J, Federmann C, Huang X, Junczys-Dowmunt M, Lewis W, Li M, Liu S, Liu TY, Luo R, Menezes A, Qin T, Seide F, Tan X, Tian F, Wu L, Wu S, Xia Y, Zhang D, Zhang Z, Zhou M (2018) Achieving human parity on automatic chinese to english news translation. URL `https://www.microsoft.com/en-us/research/uploads/prod/2018/03/final-achieving-human.pdf`
Heafield K (2011) KenLM: Faster and Smaller Language Model Queries. In: Proceedings of the 6th Workshop on Statistical Machine Translation, Edinburgh, Scotland, UK, pp 187–197
Jean S, Firat O, Cho K, Memisevic R, Bengio Y (2015) Montreal Neural Machine Translation Systems for WMT'15. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, pp 134–140
Klubička F, Toral A, Sánchez-Cartagena VM (2017) Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation. The Prague Bulletin of Mathematical Linguistics 108(1):121–132
Kneser R, Ney H (1995) Improved Backing-Off for M-gram Language Modeling. In: Proceedings of the Int. Conf. on Acoustics, Speech, and Signal Processing, Detroit, MI, USA, vol 1, pp 181–184
Koehn P (2005) Europarl: A Parallel Corpus for Statistical Machine Translation. In: Proceedings of the tenth Machine Translation Summit, Phuket, Thailand, pp 79–86
Koehn P, Knowles R (2017) Six Challenges for Neural Machine Translation. In: Proceedings of the First Workshop on Neural Machine Translation, Vancouver, BC, Canada, pp 28–39
Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the ACL-2007 Demo and Poster Sessions, Association for Computational Linguistics, Prague, Czech Republic, pp 177–180
Koponen M (2010) Assessing machine translation quality with error analysis. vol 4, pp 1–12
Krings HP (2001) Repairing texts: empirical investigations of machine translation post-editing processes. Kent State University Press, Kent, Ohio
Kucera H, Francis WN (1967) Computational analysis of present-day American English. Brown University Press, Providence, RI
Lehtonen M (2015) On sentence length distribution as an authorship attribute. In: Kim KJ (ed) Information Science and Applications, Springer, Berlin, Heidelberg, pp 811–818

Ljubešić N, Bago P, Boras D (2010) Statistical machine translation of Croatian weather forecast: How much data do we need? In: Proceedings of the ITI 2010 32nd International Conference on Information Technology Interfaces, SRCE University Computing Centre, Zagreb, pp 91–96

Lommel A, DePalma DA (2016) Europe's Leading Role in Machine Translation: How Europe Is Driving the Shift to MT. Tech. rep., Common Sense Advisory, Boston, USA

Lommel A, Uszkoreit H, Burchardt A (2014) Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. Tradumàtica: tecnologies de la traducció 12:455–463

Luong MT, Manning CD (2015) Stanford Neural Machine Translation Systems for Spoken Language Domains. In: Proceedings of the International Workshop on Spoken Language Translation 2015, Da Nang, Vietnam, pp 76–79

Moorkens J (2017) Under pressure: translation in times of austerity. Perspectives: Studies in Translation Theory and Practice 25(3), DOI 10.1080/0907676X.2017.1285331

Moorkens J, O'Brien S (2015) Post-editing evaluations: Trade-offs between novice and professional participants. In: Proceedings of European Association for Machine Translation (EAMT), Antalya, Turkey, pp 75–81

Neubig G, Morishita M, Nakamura S (2015) Neural Reranking Improves Subjective Quality of Machine Translation: NAIST at WAT2015. CoRR abs/1510.05203

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, pp 311–318

Popović M (2015) chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the 10th Workshop on Machine Translation (WMT 2015), Lisbon, Portugal, pp 392–395

Popović M (2017) Comparing language related issues for NMT and PBMT between German and English. The Prague Bulletin of Mathematical Linguistics 108(1):209–220

Popović M, Arcan M, Lommel A (2016) Potential and Limits of Using Post-edits as Reference Translations for MT Evaluation. Baltic Journal of Modern Computing 4(2):218–229

Schuster M, Johnson M, Thorat N (2016) Zero-Shot Translation with Google's Multilingual Neural Machine Translation System. URL `https://research.googleblog.com/2016/11/zero-shot-translation-with-googles.html`

Sennrich R, Haddow B, Birch A (2016a) Edinburgh Neural Machine Translation Systems for WMT 16. In: Proceedings of the First Conference on Machine Translation (WMT16), Berlin, Germany, pp 371–376

Sennrich R, Haddow B, Birch A (2016b) Improving Neural Machine Translation Models with Monolingual Data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), Berlin, Germany, pp 86–96

Sennrich R, Haddow B, Birch A (2016c) Neural Machine Translation of Rare Words with Subword Units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), Berlin, Germany, pp 1715–1725

Sennrich R, Birch A, Currey A, Germann U, Haddow B, Heafield K, Miceli Barone AV, Williams P (2017a) The University of Edinburgh's Neural MT Systems for WMT17. In: Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark, pp 389–399

Sennrich R, Firat O, Cho K, Birch A, Haddow B, Hitschler J, Junczys-Dowmunt M, Läubli S, Miceli Barone AV, Mokry J, Nadejde M (2017b) Nematus: a toolkit for neural machine translation. In: Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, pp 65–68

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, Cambridge, Massachusetts, pp 223–231

Steinberger R, Pouliquen B, Widiger A, Ignat C, Erjavec T, Tufis D, Varga D (2006) The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy, pp 2142–2147

Steinberger R, Eisele A, Klocek S, Pilos S, Schlüter P (2012) DGT-TM: A freely available translation memory in 22 languages. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey, pp 454–459

Stymne S (2013) Using a grammar checker and its error typology for annotation of statistical machine translation errors. In: Proceeedings of the 24th Scandinavian Conference of Linguistics, pp 332–344

Stymne S, Ahrenberg L (2012) On the practice of error analysis for machine translation evaluation. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, pp 1785–1790

Sutskever I, Vinyals O, Le QV (2014) Sequence to Sequence Learning with Neural Networks. CoRR abs/1409.3215, URL `http://arxiv.org/abs/1409.3215`

Tiedemann J (2012) Parallel Data, Tools and Interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'2012), Istanbul, Turkey, pp 2214–2218

Toral A, Sánchez-Cartagena VM (2017) A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, pp 1063–1073

Tyers FM, Alperen MS (2010) South-East European Times: A parallel corpus of Balkan languages. In: Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages, Malta, pp 49–53

Štajner S, Querido A, Rendeiro N, Rodrigues JA, Branco A (2016) Use of domain-specific language resources in machine translation. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Paris, France, pp 592–598

Westin I (2002) Language Change in English Newspaper Editorials. Amsterdam and New York, Rodopi

Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser L, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J (2016) Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. CoRR abs/1609.08144, URL `http://arxiv.org/abs/1609.08144`

Zeiler MD (2012) ADADELTA: An Adaptive Learning Rate Method. CoRR abs/1212.5701