Learning an Interactive Attention Policy for Neural Machine Translation

Samee Ibraheem* sibraheem@berkeley.edu
Nicholas Altieri* naltieri@berkeley.edu
John DeNero denero@berkeley.edu
**Equal contribution.*

Abstract

Interactive machine translation research has focused primarily on predictive typing, which requires a human to type parts of the translation. This paper explores an interactive setting in which humans guide the attention of a neural machine translation system in a manner that requires no text entry at all. The system generates a translation from left to right, but waits periodically for a human to select the word in the source sentence to be translated next. A central technical challenge is that the system must learn when and how often to request guidance from the human. These decisions allow the system to trade off translation speed and accuracy. We cast these decisions as a reinforcement learning task and develop a policy gradient approach to train the system. Critically, the system can be trained on parallel data alone by simulating human guidance at training time. Our experiments demonstrate the viability of this interactive setting to improve translation quality and show that an effective policy for periodically requesting human guidance can be learned automatically.

1 Introduction

Despite rapid advances in neural machine translation, human input is still needed to meet the translation quality requirements of many applications. Interactive machine translation seeks to combine the quality of human translation with the speed and lexical coverage of machine translation. This paper explores an interactive setting in which the human translator does not type at all, but instead guides the attention of a neural machine translation system by selecting relevant source words as the system translates. While we should not expect that the resulting translations will be as accurate as those produced by predictive typing, this interactive approach could provide fast and accurate draft translations that could later be improved by post-editing. Moreover, source word selection enables new user interface options because it can be performed using a wide variety of input devices, including a mouse, a touch screen, or an eye tracker, which may be used in tandem with traditional text entry methods.

We first address the question of whether guiding the attention of a neural machine translation system can provide enough useful information to improve translation quality. Rather than experimenting directly with human subjects, we compute an experimental upper bound on the accuracy gains from guided attention. For each word that the system is meant to generate, we find an oracle attention that maximizes the probability of generating that word. We find that guiding attention toward this oracle provides a great deal of information to the translation system, yielding substantial gains in translation quality.

Second, we define an interactive translation process in which the system generates a translation left-to-right, but pauses on occasion to request guidance from a human collaborator. Ideally, the system would not pause after every word; if the system can generate some portion of the translation accurately without human intervention, then it would be wasteful for it to solicit human input. Therefore, an ideal system must learn to trade off between translating accurately and requiring as little human input as possible.

However, it is difficult to predict the long-term consequences of choosing whether or not to pause at any given position. The value of receiving human guidance is not only that it may improve the prediction of the next word, but that it may improve predictions of all subsequent words. Therefore, pausing early for human input might allow the system to require less guidance in later parts of a sentence. Our primary technical contribution is to cast the sequence of decisions about when to request human guidance as a reinforcement learning problem that properly accounts for the system's uncertainty about all the downstream effects of requesting human intervention. We apply a policy gradient method to this problem and show that the system is able to learn an effective interaction policy. This policy estimates when, during the process of translation, human guidance is likely to provide enough long-term benefit to justify the cost of pausing.

We evaluate our approach using an English-German neural machine translation system trained for the WMT 2016 news translation task. We show that the whole system, including the learned interaction policy, can be trained fully automatically by approximating human input using simulated guidance.

2 Related Work

Interactive machine translation involves human translators working collaboratively with a machine translation system to produce high quality output efficiently (Foster and Lapalme, 2002). Several interactive interfaces to machine translation systems have been designed and evaluated in the research community, such as TransType (Langlais et al., 2000), Thot (Ortiz-Martínez et al., 2010), and Caitra (Koehn, 2009). Green et al. (2014) investigates the trade-off between human effort and translation quality within the paradigms of post-editing and interactive MT.

A growing line of research has explored the use of neural machine translation with attention (Bahdanau et al., 2014) in an interactive setting. Wuebker et al. (2016) compares the performance of neural and statistical machine translation models for interactive prediction, and shows that neural models are substantially more accurate. Knowles and Koehn (2016) also demonstrates that neural models provide more accurate interactive predictions than statistical models and addresses efficiency challenges. Hokamp and Liu (2017) describes a search algorithm for neural models that specifically targets a typical interactive workflow in which the terms in a bilingual lexicon must be prioritized over alternatives.

Werling et al. (2015) investigates the trade-off between the cost of human intervention and accuracy for three other tasks: named-entity recognition, sentiment classification, and image classification. That work also proposes an approach to decision making that considers the uncertain long-term consequences of actions.

Mi et al. (2016) demonstrates the usefulness of providing additional attention information to a fully automated neural machine translation system. In this work, the authors add an additional loss to the translation model which encourages the attention computed by the NMT system to resemble alignments predicted by an IBM word alignment model.

3 Guided Attention

Neural machine translation with attention (Bahdanau et al., 2014) is a variant of the seq2seq model (Sutskever et al., 2014) that incorporates attention over the source encodings into the decoder. The attention is a distribution over source positions that can be interpreted as a soft indicator of what part of the source sentence will be translated next. We propose to replace the attention predicted by the model with a *guided* attention distribution that is provided directly by a human selecting a source word. In this paper, we simulate the human selection using the source word that is most helpful in the translation decision, described in detail below.

3.1 Neural Machine Translation with Attention

Given a source sentence $x = x_1, \ldots, x_n$ and a target sentence $y = y_1, \ldots, y_m$, the model first encodes x to form input representations $z_1, \ldots, z_n$. To predict the target labels y, the model conditions on a concatenation of two vectors, one being a hidden representation of the output generated so far, and the other being the input representations weighted by the attention: $\sum_i \alpha_i^{(t)} z_i$, where $\alpha_i^{(t)}$ is the attention computed at time t for the ith word in the source sentence. The input representations and hidden decoder states can be defined using an LSTM (Bahdanau et al., 2014) or convolution (Gehring et al., 2017) over word embeddings.

The attention vector is a distribution over source positions: $\sum_i \alpha_i^{(t)} = 1$ and $\alpha_i^{(t)} \geq 0$. To compute $\alpha_i^{(t)}$, a feed-forward neural network is used that takes in as inputs (z_i, h_t) where h_t is the hidden decoder state at time t. Finally, given the attention, hidden decoder state, and input representations, the label y_t is predicted using a learned distribution $p(y_t|h_t, \sum_i \alpha_i^{(t)} z_i)$.

3.2 Simulated Attention

Instead of using human input to train the model, we attempt to simulate the behavior of an accurate human, allowing for faster and cheaper training. We do this by, at each time step, calculating the distribution over the target vocabulary $p(y_t|h_t, z_i)$ for each i, which is equivalent to evaluating a one-hot attention vector for each source sentence word. We then provide the one-hot attention for the source word that had the highest predictive probability for the correct next target word to be translated. That is, if $i^* = \arg\max_i p(y^*|h_j, z_i)$, where y^* is the correct target word, then

$$\alpha^{(t)} = e_{i^*} \implies \sum_i \alpha_i^{(t)} z_i = z_{i^*}.$$

4 Learning When to Ask for Guidance

Given that we have a method for simulating the guidance that a human would provide, we turn to the problem of deciding when to request guidance at all. Each request for guidance affects the input representation used for predicting a single word. Over the course of a sentence, the system can request guidance multiple times.

4.1 Interaction Policy

To implement our interactive method, we use a greedy decoder. For each predicted word, the model decides whether to translate using guided attention or to translate using the attention predicted by the model. At the end of each iteration, there will be a loss penalty corresponding to the amount of guidance requested as well as the likelihood of the sentence under the model. Guidance improves likelihood by providing more information to each decision, but incurs a penalty for requesting guidance.

4.2 Interactive Machine Translation as Reinforcement Learning

We believe that reinforcement learning is an appropriate framework for our set up, since deciding when to ask for assistance can have long term ramifications on final accuracy that are hard to anticipate before training. We therefore model our framework by a Markov decision process (MDP). In this MDP, our agent is the machine translation system, whose actions are whether or not to request guided attention, and our reward function is the cross-entropy between our prediction of the next word and a distribution that predicts the reference with probability 1.

4.3 Reinforcement Learning

An MDP is a tuple (S, A, T, R). S is the set of all possible states that an agent can be in. A is the set of all possible actions the agent can take. T is the transition function $p(s_{t+1}|s_t, a_t) = T(s_{t+1}|s_t, a_t)$ that is the distribution over the next state given the current state and the action to be taken. Finally, R is the reward function $R(s_{t+1}, a_t, s_t)$ that determines the reward for transitioning into s_{t+1} from s_t with action a_t.

An agent acting in a MDP can be described by a policy function $\pi : S \to A$, that takes in states and returns actions. It is the goal of reinforcement learning to learn a policy that maximizes the expected sum of (discounted) rewards: $\mathbf{E}[\sum_t \gamma^t R(s_{t+1}, \pi(s_t), s_t)]$, where $\gamma \in (0, 1]$ is the discount factor.

In the case of interactive attention in machine translation, a state s_t captures the activation of the translation network just before it would generate the next target word w_t. There are only two possible actions: whether to go ahead and generate w_t or to request guidance. If guidance is requested, then a new activation of the translation network is computed by replacing the model's attention weights with the guide's attention weights, and then a new word w'_t is generated using these new activations. If guidance is not requested, then w_t is generated. In either case, the reward function is the cross entropy sequence loss of the correct translation.

4.3.1 Policy Gradient

Policy gradient is a common reinforcement learning method to learn a policy π_θ parameterized by θ. The policy gradient method aims to perform stochastic gradient ascent on the objective

$$J(\theta) = \mathbf{E}\left[\sum_{t=1}^{T-1} \gamma^t R(s_{t+1}, \pi_\theta(s_t), s_t)\right].$$

Let $\pi_\theta(a_t|s_t)$ be the probability of choosing an action a_t in state s_t according to the policy π_θ. The policy gradient theorem states that if a_t are sampled according to $\pi_\theta(s_t)$, and s_{t+1} are sampled according to $T(\cdot|s_t, a_t)$, then an unbiased estimator of $\nabla_\theta J(\theta)$ is

$$\sum_{t=1}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{\tau=t}^{T-1} \gamma^{(\tau-t)} R(s_{\tau+1}, a_\tau, s_\tau).$$

Although using the above expression is an unbiased estimator, it can have high variance, prompting the use of variance reduction methods. For any function $b(s)$, the following is also an unbiased estimator:

$$\sum_{t=1}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{\tau=t}^{T-1} \gamma^{(\tau-t)} (R(s_{\tau+1}, a_\tau, s_\tau) - b(s_\tau)),$$

And the choice that minimizes variance is

$$b(s_\tau) = \mathbf{E}_{\pi_\theta}\left[\sum_{\tau=t}^{T-1} \gamma^{(\tau-t)} R(s_{\tau+1}, a_\tau, s_\tau)\right].$$

This optimal $b(s_\tau)$ can be approximated by a parameterized function V_ϕ, where we learn V_ϕ by approximately minimizing

$$\mathbf{E}_{\pi_\theta}\left(V_\phi(s_t) - \sum_{\tau=t}^{T-1} \gamma^{(\tau-t)} R(s_{\tau+1}, a_\tau, s_\tau)\right)^2.$$

Finally, a policy gradient algorithm alternates between taking a step of stochastic gradient ascent on $J(\theta)$ and taking multiple gradient steps on V_ϕ.

When using the approximate value function V_ϕ to reduce variance, the inner expression of the gradient is typically called the *advantage function* and denoted $A(s_t)$:

$$A(s_t) = \sum_{\tau=t}^{T-1} \left[\gamma^{(\tau-t)} R(s_{\tau+1}, a_\tau, s_\tau)\right] - V_\phi(s_t).$$

For our value function, we use a feed-forward neural network with two hidden layers of 32 units each, and for our policy function we use a neural network with one 32-unit hidden layer. The input to the former is the standard decoder inputs, which consist of the previously output token and the weighted sum of the hidden representations $\sum_i \alpha_i^{(t)} z_i$. The input to the latter additionally includes the original softmax layer input.

4.4 Action Frequency Regularization

Since our goals are to maximize translation accuracy while minimizing the number of times a human would have to intervene, we introduce an action weight parameter w_a, in order to manage the trade-off between accuracy and human effort. To promote accuracy during training, we have part of the reward at time step t be the negative cross entropy of the predictions at time t. To incorporate the number of times that the system requests guidance, we include not only the probability of requesting guidance, but also whether or not guidance was requested. In addition to these, we incorporate a threshold parameter ρ_a, to ensure that the action probabilities do not exceed the designated value. We thus use the following policy gradient objective:

$$\hat{A}(s_t) \cdot \log p_\theta(a_t) + w_a \cdot \max(0, p(a_t) - \rho_a) \cdot a_t,$$

where a_i is a binary scalar that takes value 1 if guidance was requested, and 0 otherwise, and $\hat{A}$ is the standardized advantage function.

That is,

$$\hat{A}(s_t) = \frac{A(s_t) - \mu(A(s_t))}{\sigma(A(s_t))}.$$

5 Experiments

We evaluate our model on the task of translating from English to German. Specifically, we first train a sequence-to-sequence model with attention, and then continue training using our reinforcement learning model. The baseline neural machine translation model was trained for 508,387 iterations.

5.1 Datasets

We use the English-German WMT 2016 news task dataset, which contains 4.2 million training sentence pairs. We apply BPE with 32,000 merge operations.

5.2 Architecture Details

For our base NMT system, we used Google's large seq2seq system implementation (Britz et al., 2017). For the encoder, we had 512 hidden units. For the decoder, both the GRU and the attention have 512 units. [1]

5.3 Results

We evaluate our approach on all 3000 sentences of the WMT 2016 news-test2013 development set. We first evaluate the baseline fully automatic NMT model, which yields a BLEU (Papineni et al., 2002) score of 19.37. In comparison, our model which asks for guidance with a 100% probability has a BLEU score of 32.51. Thus, requesting guidance indeed improves translation quality for this model. However, requesting guidance for every word would require maximal human effort, as the human translator would be required to click at each time step.

We also evaluate a variety of learned policies on the same data and using the same baseline model. During policy learning, the parameters of the translation model are frozen, and only the parameters of the policy and value functions are learned. Varying the action weight and threshold values yields various guidance frequencies and corresponding BLEU scores. To determine whether the learned policy is requesting guidance efficiently, for each trained policy we also evaluate a random policy that asks for guidance with the same frequency as the reinforcement learning policy (Figure 5.3). The learned policy was able to achieve a BLEU score of 27.25 with observed guidance of about 54%, which improved upon the random policy by almost 2 BLEU and upon the baseline model by about 8 BLEU.

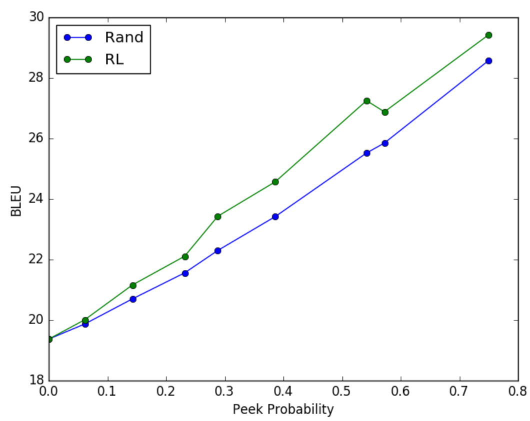

Figure 1: Translation accuracy for a random policy (blue) and a learned policy (green), for different guidance frequencies. More guidance provides higher accuracy. Across a range of guidance frequencies, the learned policy outperforms a policy that makes the same number of guidance requests, but at randomly chosen times.

6 Analysis

We compare the simulated clicks to the attention generated by the neural machine translator. In order to compare them, we compare the optimal word attention location computed by our simulator against the word with the largest weight according to the NMT system. This does provide a problem if the NMT was attending primarily to more than a single object, but nevertheless we believe this method of comparison may still provide useful intuition. In the figure below we

[1] For full specification see: `https://github.com/google/seq2seq/blob/master/example_configs/nmt_large.yml`

only include arrows for which the attended words differ. We note that using the simulated attention seems mostly intuitive with respect to where a human translator would click and corrects some of the NMT system errors. In particular, it makes *von* point to *of* and *zu* point to *counter*. However, there are also a few quirks. For example, it makes *kanishe* point to *Republic* and EOS point to *to*.

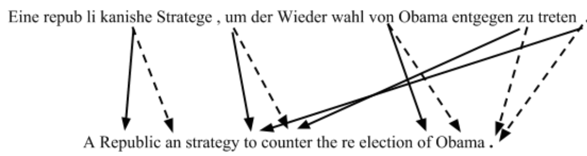

Figure 2: Guided attention (solid) vs NMT attention (dashed)

7 Future work

Our experiments demonstrate that reinforcement learning is an effective framework for requesting human guidance in interactive machine translation. However, we can identify several open questions that merit further investigation. First, we have focused on greedy decoding in this paper, because it is not trivial to apply a more sophisticated search procedure on top of our method. Developing an extension that incorporates beam search could improve performance. Second, during baseline training, the attention mechanism sees soft attention over the entire sentence as opposed to one hot attention over a single word, and the discrepancy between training and testing may limit the performance of the system. In addition, this method assumes that the word that gives the best predictive probability of the next target word is the same word that a human would choose. Another related limitation with our system is that it assumes that the previous system output is the same as the correct translation, and so the best next word to be translated by the system is the same as that of the reference translation.

As our approach is intended to reduce human effort, we look forward to conducting human subject experiments in future work, to see whether the gains we witnessed in simulation carry over to real-world conditions. One interesting direction that our method could provide is investigating whether the behaviors of humans interacting with such a system may be the same as those when interacting with other humans, and if not, to test in which ways human actions might be similar and how they may diverge from expected behavior. Another extension to this work would be incorporating the attention supervision into the main model. Currently, if asked to translate the same sentence twice, the current framework would ask for the same attention help twice, which seems inherently wasteful. Ideally, after getting the supervision, it would be able to incorporate it into the model to reduce redundant queries.

8 Conclusion

We have demonstrated an approach to interactive machine translation that aims to limit the amount of effort required by human translators while maintaining translation quality. We hope that our method inspires further research into this area.

References

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

Britz, D., Goldie, A., Luong, M.-T., and Le, Q. (2017). Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906*.

Foster, G. and Lapalme, G. (2002). *Text prediction for translators*. Université de Montréal.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Green, S., Wang, S. I., Chuang, J., Heer, J., Schuster, S., and Manning, C. D. (2014). Human effort and machine learnability in computer aided translation. In *EMNLP*, pages 1225–1236.

Hokamp, C. and Liu, Q. (2017). Lexically constrained decoding for sequence generation using grid beam search. *arXiv preprint arXiv:1704.07138*.

Knowles, R. and Koehn, P. (2016). Neural interactive translation prediction. *AMTA 2016, Vol.*, page 107.

Koehn, P. (2009). A process study of computer-aided translation. *Machine Translation*, 23(4):241–263.

Langlais, P., Foster, G., and Lapalme, G. (2000). Transtype: a computer-aided translation typing system. In *Proceedings of the 2000 NAACL-ANLP Workshop on Embedded machine translation systems-Volume 5*, pages 46–51. Association for Computational Linguistics.

Mi, H., Wang, Z., and Ittycheriah, A. (2016). Supervised Attentions for Neural Machine Translation.

Ortiz-Martínez, D., García-Varea, I., and Casacuberta, F. (2010). Online learning for interactive statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 546–554. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, Montral, Canada.

Werling, K., Chaganty, A. T., Liang, P. S., and Manning, C. D. (2015). On-the-Job Learning with Bayesian Decision Theory. In *Advances in Neural Information Processing Systems*, pages 3465–3473, Montral, Canada.

Wuebker, J., Green, S., DeNero, J., Hasan, S., and Luong, M.-T. (2016). Models and inference for prefix-constrained machine translation. *54th ACL*, 1:66–75.

A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators

Sheila Castilho[1] sheila.castilho@adaptcentre.ie
Joss Moorkens[1] joss.moorkens@adaptcentre.ie
Federico Gaspari[1] federico.gaspari@adaptcentre.ie
Rico Sennrich[2] rico.sennrich@ed.ac.uk
Vilelmini Sosoni[3] vilelmini@hotmail.com
Panayota Georgakopoulou[4] yota.georgakopoulou@bydeluxe.com
Pintu Lohar[1] pintu.lohar@adaptcentre.ie
Andy Way[1] andy.way@adaptcentre.ie
Antonio Valerio Miceli Barone[2] amiceli@inf.ed.ac.uk
Maria Gialama[4] maria.gialama@bydeluxe.com
[1] ADAPT Centre, Dublin City University, Dublin, Ireland
[2] The University of Edinburgh, Edinburgh, UK
[3] Ionian University, Corfu, Greece
[4] Deluxe Media, Athens, Greece

Abstract

This paper reports on a comparative evaluation of phrase-based statistical machine translation (PBSMT) and neural machine translation (NMT) for four language pairs, using the PET interface to compare educational domain output from both systems using a variety of metrics, including automatic evaluation as well as human rankings of adequacy and fluency, error-type markup, and post-editing (technical and temporal) effort, performed by professional translators. Our results show a preference for NMT in side-by-side ranking for all language pairs, texts, and segment lengths. In addition, perceived fluency is improved and annotated errors are fewer in the NMT output. Results are mixed for perceived adequacy and for errors of omission, addition, and mistranslation. Despite far fewer segments requiring post-editing, document-level post-editing performance was not found to have significantly improved in NMT compared to PBSMT. This evaluation was conducted as part of the TraMOOC project, which aims to create a replicable semi-automated methodology for high-quality machine translation of educational data.

1 Introduction

The industrial use of machine translation (MT) for production has become widespread since statistical machine translation (SMT) established itself as the dominant approach to translating texts automatically. Raw MT is now a viable solution for perishable content (Way, 2013) and post-editing of MT is offered by over 80% of language service providers surveyed by Lommel and DePalma (2016). In the years since the publication of Brown et al. (1993), an ecosystem of tools has grown around PBSMT, including scripts and tools for pre-processing and alignment, enabling incremental improvement in the quality of PBSMT output (Haddow et al., 2015).

More recently, the research community has become increasingly interested in the possibilities of neural machine translation (Bahdanau et al., 2014; Cho et al., 2014) (NMT), which involves building a single neural network that maps aligned bilingual texts and, given input to translate, is trained to "maximize the probability of a correct translation" (Bahdanau et al., 2014) without external linguistic information. This interest is shared by many in the language service industry, where there is a need for improved MT quality and better quality estimation to "help reduce the frustrating aspects of post-editing" (Etchegoyhen et al., 2014). NMT results in the latest shared tasks have quickly matched or surpassed those of PBSMT systems, despite the many years of PBSMT development (Sennrich et al., 2016a; Bojar et al., 2016). Recent studies have reported an increase 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). While these initial experiments with NMT have shown impressive results and promising potential, so far there have been a limited number of human evaluations of NMT output.

This paper reports the results of a quantitative and qualitative comparative evaluation of PBSMT and NMT carried out using automatic metrics and a small number of professional translators, considering the translation of educational texts in four language pairs, i.e. from English into German, Portuguese, Russian and Greek. It employs a variety of metrics, including side-by-side ranking, rating for accuracy and fluency, error annotation, and measurements of post-editing effort. This evaluation is part of the work for TraMOOC,[1] a European-funded project focused on the translation of MOOCs, which aims to create 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 3.1.1. The remainder of this paper is organized as follows: In Section 2 we review previous work comparing MT output using the statistical and neural approaches. We describe our MT systems and the experimental methodology in Section 3, and the results of human and automatic evaluations in Section 4. Finally, we draw the main conclusions of the study and outline promising avenues for future work in Section 5.

2 Previous Work Comparing PBSMT and NMT

A number of papers have been published recently which compare specific aspects of PBSMT and NMT. 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 PBSMT 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 PBSMT 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. Bentivogli et al. (2016) 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.

Wu et al. (2016) used BLEU (Papineni et al., 2002) scores and human ranking 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 system strongly outperforms other approaches and improves 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 Google NMT engines subsequently went live for the language pairs tested shortly after this paper was published (Schuster et al., 2016). Junczys-Dowmunt et al.

[1] `http://tramooc.eu`

(2016) also found BLEU score improvements in NMT when compared with PBSMT for as many as 30 language pairs.

Results of the 2016 Workshop on Statistical Machine Translation (WMT16) (Bojar et al., 2016) found that NMT systems were ranked above PBSMT and online systems for six of 12 language pairs for translation tasks. 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 Sanchez-Cartagena (2017) compared NMT and PBSMT for nine language pairs (English to Czech, German, Romanian, Russian and vice-versa, 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 PBSMT outputs were found to be dissimilar, with a higher inter-system variability between NMT systems. NMT systems appear to perform more reordering than PBSMT systems, resulting in more fluent translations (taking perplexity of MT outputs on neural language models as a proxy for fluency). Toral and Sanchez-Cartagena (2017) found that the tested NMT systems performed better than PBSMT 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 PBSMT performed better than NMT for segments longer than 40 words.

Castilho et al. (2017) also reported on three comparative studies of PBSMT and NMT, discussing some of the preliminary results of the current study, highlighting some strengths and weaknesses of NMT, and the danger of hyperbole in discussions of the potential of NMT. Against this background, this paper attempts to shed more light on the emerging picture of the comparison between PBSMT and NMT.

3 Experiments

We built and evaluated PBSMT and NMT systems for four translation directions: English to German, Greek, Portuguese, and Russian. Evaluation was performed with automatic metrics, as well as with professional translators, who performed side-by-side ranking, adequacy and fluency rating, post-editing and error annotation based on a predefined taxonomy.

3.1 MT Systems

3.1.1 Training Data

The MT engines used in the TraMOOC project are trained on large amounts of data from various sources: the training data from the WMT shared translation tasks[2] and OPUS (Tiedemann, 2012) as mixed domain, and as in-domain training data we use TED from WIT3 (Cettolo et al., 2012); QCRI Educational Domain Corpus (QED) (Abdelali et al., 2014); a corpus of Coursera MOOCs; and our own collection of educational data. The amount of training data used is shown in Table 1.

Lang.	DE	EL	PT	RU
mixed domain	23.78	30.73	31.97	21.30
In-domain	0.27	0.14	0.58	2.31

Table 1: Training data size the EN→* translation directions (number of sentence pairs, in millions).

3.1.2 Phrase-based SMT

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

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

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 LM with modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1998), as well as the following advanced features: a hierarchical lexicalized 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 optimized to maximize BLEU with batch MIRA (Cherry and Foster, 2012) on an in-domain tuning set that has been extracted (and held out) from the in-domain training data.

Adaptation to the MOOC domain is performed via three mechanisms: sparse domain indicator features in the phrase table; linear interpolation of LMs with perplexity optimization on the in-domain tuning set; and learning of feature weights on the in-domain tuning set.

3.1.3 Neural MT

The NMT systems are attentional encoder-decoder networks (Bahdanau et al., 2014), which we trained with Nematus (Sennrich et al., 2017). We generally follow the settings used by Sennrich et al. (2016a). We use word embeddings of size 500, and 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 check-points of the mixed-domain training run, and continuing training each on in-domain data.

3.2 The MOOCs Domain

As this evaluation was intended to identify the best-performing MT system for the TraMOOC project, which focuses on high-quality MT for MOOCs, test sets were extracted from real MOOC data. These data included explanatory texts, subtitles from video lectures, 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 specialized terminology and, in the case of the physics text, non-contextual variables and formulae.

3.3 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 randomized and 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 (Post-Editing Tool) (Aziz et al., 2012) over a two-week period. 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 Likert scale for each segment, and iii) perform error annotation using a simple taxonomy (more details are provided in Section 3.5). This set-up had the advantage that measurements of two of Krings' (2001) categories of post-editing effort could be drawn directly from the PET logs, namely temporal effort (time spent post-editing) and technical effort (edit count).

3.4 Automatic Evaluation

The BLEU, chrF3 and METEOR (Banerjee and Lavie, 2005) automatic evaluation metrics are used in this study, with the caveat that two post-edits are used as references for each segment. It should be noted that Popović et al. (2016) suggest that the use of a single post-edited reference from the MT system under evaluation will tend 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.

3.5 Human Evaluation

Ranking: The professional translators were asked to tick a box containing 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.

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 Likert 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'. 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).

Post-editing and error annotation: Participants 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. Again, our expectation was that there would be fewer morphology and word order errors with NMT, especially for short segments.

4 Results and Discussion

4.1 Automatic Evaluation

The automatic metric results using BLEU, METEOR, chrF3 and HTER are shown in Table 2. In particular, the decrease in word order errors in NMT output (as may be seen in Section 4.3) shows an improvement in BLEU and METEOR scores, especially for some language pairs.

Table 2 shows that BLEU, METEOR and chrF3 scores considerably increase for German, Greek and Russian with NMT when compared to the PBSMT scores. These results were statistically significant in a one-way ANOVA pairwise comparison (p<.05) (marked with †). For Portuguese, moderate improvements can be observed, but no statistically significant differences were found.

Regarding the amount of PE that was required, the HTER scores show that more PE was performed when using the output from the PBSMT system for German, Greek and Russian. However, no statistically significant differences for HTER scores were found. The scores for chrF3 also show good improvement for NMT over PBSMT for German and Russian, but very similar results for Greek and Portuguese.

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 2: Automatic Evaluation Results

4.2 Human Evaluation

Fluency and Adequacy: NMT was rated as more fluent than PBSMT for all language pairs. Table 3 shows the mean ratings for Fluency and Adequacy of the target languages for both PBSMT and NMT systems. Although no statistically significant differences were found, the percentage of scores assigned a 3-4 fluency value (Near Native or Native) for German is 68% for NMT as opposed to 54% for the PBSMT system, for Greek 75% and 65%, for Portuguese 80% and 74% , and for Russian 75% and 60%, respectively.

When looking at the percentage of scores assigned a 1-2 fluency value (No or Little Fluency) for each MT system's output, the NMT systems appear to have fewer problems when compared against the PBSMT systems for all the languages (German: 46% PBSMT vs. 32% NMT; Greek: 35% vs. 25%; Portuguese: 26% vs. 21%; and Russian: 40% vs. 25%).

A typical example of improved output for German NMT was the translation of the segment 'Would you send just 10 materials that are the most suitable.'

Lang.	System	Fluency	Adequacy
DE	PBSMT	2.60	2.85
	NMT	2.95	2.79
EL	PBSMT	2.86	3.44
	NMT	3.08	3.46
PT	PBSMT	3.15	3.73
	NMT	3.22	3.79
RU	PBSMT	2.70	2.98
	NMT	3.08	3.12

Table 3: Mean for Fluency and Adequacy

PBSMT: Würden Sie nur 10 Materialien, die am besten geeignet sind.
NMT: Schicken Sie einfach 10 Materialien, die am besten geeignet sind.

The German PBSMT output left out an infinitive verb at the end of the segment (literally 'Would you [polite form] just 10 materials that are the most suitable.'), while NMT produced a correct German translation, using the imperative verb form and retaining the correct register by using the 'Sie' politeness marker.

For Portuguese, one example of improved fluency is the translation of the segment 'I am just making sure that I understand this correctly.'

PBSMT: Estou só para ter a certeza que entendi corretamente.
NMT: Eu estou apenas me certificando de que eu entendo isso corretamente.

The PBSMT system translates 'just' as 'só' (which in Portuguese can mean 'just' but as it is preceded by the verb 'estar', it implies the meaning 'alone'/'lonely'), conveying the misleading meaning of 'I'm alone to be sure if I understood correctly'. The NMT system translates 'making sure' as 'me certificando', which is accurate with the word 'just' translated as 'apenas' or 'só'.

One example for the Russian language is the translation of 'I liked your presentation a lot.'

PBSMT:Я любил свою презентацию много.
NMT: Мне очень понравилась ваша презентация.

While the NMT output is absolutely correct, the PBSMT system mistranslates the possessive pronoun 'your' as 'свою презентацию', which means 'my presentation'. It also translates 'liked' as 'любил', which means 'loved', and, finally, it also translates 'a lot' as 'много', which translates back as 'many' (quantifying adjective).

For Greek, NMT also shows improved fluency for the translation of 'What is the difference between a financial analyst and technical analyst and business analyst?'

PBSMT: Ποια είναι η διαφορά μεταξύ ένας οικονομικός αναλυτής και τεχνική αναλύτρια και οικονομικός αναλυτής·

NMT: Ποια είναι η διαφορά μεταξύ του οικονομικού αναλυτή και του τεχνικού αναλυτή και του επιχειρηματικού αναλυτή·

The NMT output is both semantically accurate and grammatically correct: the terms 'financial analyst', 'technical analyst' and 'business analyst' were rendered accurately in Greek, and, in addition, the nouns correctly appear in genitive form and the generic masculine is used. The PBSMT mistranslates the term 'business analyst' into 'οικονομικός αναλυτής' (i.e.'financial analyst'), and lacks fluency since the nouns are used in the nominative form and the gender of the noun 'technical analyst' appears in the feminine form rather than in the correct generic masculine form.

Regarding adequacy, however, results were overall less consistent (see Table 3) than those for fluency, with higher mean scores for German PBSMT. 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 (73% against 66% for NMT), and English-Greek systems performed equally well (89% of the sentences assessed as 3-4 in terms of adequacy). For Portuguese and Russian, the NMT systems were ranked slightly higher when including 3-4 rankings, with PBSMT scoring 95% against 97% of NMT output in Portuguese, and for Russian the scores were 73% for PBSMT against 78% for NMT. These results are also replicated when the distinction between short and long sentences is made.

One example of adequacy for German where both MT systems committed errors can be seen in:

EN: We begin our exploration today by looking at a particular ad that appeared in on

American magazines in recent years.

PBSMT: Heute beginnen wir unsere Erforschung von einem bestimmten Ad anschaue, die auf amerikanischen Zeitschriften erschienen in den letzten Jahren.

NMT: Wir beginnen unsere Forschung heute mit einer bestimmten Werbung, die in den letzten Jahren in amerikanischen Zeitschriften veröffentlicht wurde.

The NMT output uses the noun 'Forschung', meaning 'research', rather than the correct 'Erforschung' as chosen by the PBSMT system. As a result, the participants rated this segment poorly for adequacy, and actually substituted the word 'Untersuchung' for 'exploration'. While the PBSMT system chose the correct noun, there were other word order and lexical errors that rendered the translation inadequate.

The following is an example of adequacy not being so consistent in translation into Portuguese, but NMT system still performing better:

EN: What we're going to need to do is, we're going to find the initial stretch, excuse me, the final stretch of the spring, the initial stretch of the spring, and subtract the squares.

PBSMT: O que vamos precisar fazer é, vamos encontrar o troço inicial, desculpe-me, o último troço da Primavera, o troço inicial da Primavera, e subtrair os quadrados.

NMT: O que vamos precisar fazer é, vamos encontrar o limite inicial, desculpe-me, o alongamento final da mola, o alongamento inicial da mola, e subtrair os quadrados.

PBSMT mistranslates the two main words of the sentence: 'stretch' (translates into 'stuff') and 'spring' (as the spring season, 'primavera'), thus making the translation unintelligible. NMT translates the term 'stretch' into two different ways ('limite' and 'alongamento'), but the sentence is still adequate and understandable.

For Russian, both MT systems also return errors for adequacy:

EN: We'll be drawing heavily on the field of art history and how interpretation works in that field.

PBSMT: Мы будем рисовать на области истории искусства и как интерпретации работает в этой области.

NMT: Мы будем активно рисоваться на области художественной истории и то, как интерпретация работает в этом поле.

Both systems translate the word 'drawing' as 'draw a picture'. PBSMT, however, retrieves a better translation for the remainder of the sentence, keeping 'история искусства' as a fixed expression, while 'художественной истории' – chosen by the NMT system – is not natural and the meaning is not clear. The translation of the word 'field' is also better in the PBSMT output: 'область' is 'field' in the sense of area (of research/interest), while NMT translates as 'поле', i.e. a farm field or mathematical concept.

Finally for Greek, the NMT system seems to handle adequacy a bit better:

EN: So, what if a resident or student wants to opt out of doing abortions?

PBSMT: Οπότε, τι γίνεται αν ένας κάτοικος ή μαθητής θέλει να εξαιρεθούν από το να κάνει εκτρώσεις·

NMT: Οπότε, τι γίνεται αν ένας κάτοικος ή φοιτητής θέλει να επιλέξει να κάνει εκτρώσεις·

The PBSMT translation has problems both at the level of fluency and at the level of adequacy, while the NMT translation has problems only at the level of adequacy. In both the PBSMT and the NMT translations the term ‘resident’ - which in this context refers to the North American concept of ‘a medical graduate engaged in specialised practice under supervision in a hospital’- is translated as ‘κάτοικος’, that is, a person who lives somewhere permanently or on a long-term basis. The PBSMT translates the word ‘student’ as ‘μαθητής’, which refers to a pupil, when in fact it should be translated as ‘φοιτητής’ (university student). The PBSMT output also suffers at the level of fluency due to the lack of subject to verb correspondence. In the NMT output, apart from the mistranslation of the term ‘resident’, there is one major mistranslation involving the phrasal verb ‘opt out’, as the NMT system translates it as ‘opt’, thus distorting completely the meaning of the source sentence.

Polysemous terms appear to pose the main problem to the NMT system for Greek and Russian languages, as it appears unable to discern semantic differences and choose the equivalent which bears the same meaning as the ST one in the translation. This can pose significant problems during the PE process, as translators may be misled by the inaccurate NMT rendering, and end up transferring the erroneous term in the final translation. For instance, for the translation of ‘This is a magazine and a campaign called Got Milk where several famous figures appeared and they always asked the question, got milk?’, the term ‘figure’ is translated into Greek by the PBSMT as ‘προσωπικότητα’, while it is translated erroneously by the NMT system as ‘φιγούρα’, which is semantically wrong. Another example of polysemous term appears in the Russian translation of ‘Is it free?’, where NMT translated as ‘Свободно ли?’, meaning ‘unoccupied’ (‘is this seat/place free?’), while the PBSMT output includes a more frequent lexical item, ‘Это бесплатно?’, which relates to price (‘free of charge’). For German and Portuguese, however, the polysemous terms are either not handled well by neither systems, or the NMT system provides a better translation.

This small selection of examples demonstrates the types of errors prevalent in the respective MT systems for each language pair studied, with the NMT output generally found to be more fluent and comprehensible, although not without errors. The type and prevalence of these errors throughout the test sets are detailed in Section 4.3.

4.3 Error Annotation

Category	DE		EL		PT		RU	
	PBSMT	NMT	PBSMT	NMT	PBSMT	NMT	PBSMT	NMT
Inflectional Morphology	732	608	443	307	404	378	695	506
	43%	49%	35%	28%	37%	37%	42%	38%
Word Order	382	180	303	208	216	181	197	122
	23%	15%	24%	19%	20%	18%	12%	9%
Omission	126	84	48	57*	53	58*	194	163
	7%	7%	4%	5%	5%	6%	12%	12%
Addition	46	39	24	31*	61	44	183	151
	3%	3%	2%	3%	6%	4%	11%	11%
Mistranslation	401	323	459	483*	348	342	385	404*
	24%	26%	36%	44%	32%	34*%	23%	30%
Total number of issues	1687	1234	1277	1086	1082	1003	1654	1346
Total number of “No Issues”	61	**189**	90	**168**	197	**236**	101	**195**
	6%	**18.9**%	9%	**16.8**%	19.7%	**23.6**%	10%	**19.5**%

Table 4: Error annotation

Table 4 shows the results of the error annotation task for all target languages, the total count of the errors and the percentage of errors of each category.[3] The total number of issues

[3]The percentage of errors is the number of error per category divided by the total number of errors found.

is greater for PBSMT than NMT for all language pairs. Moreover, the number of segments left without error annotations (No issues) is greater for NMT across all language pairs (in bold). NMT output was also found to contain fewer word order errors and fewer inflectional morphology errors in all the target languages. For English-Greek, the PBSMT output contained fewer errors of omission, addition, or mistranslation than NMT output (marked with an asterisk). For English-Portuguese, PBSMT showed fewer omissions and mistranslations, while English-Russian PBSMT contained fewer mistranslations (also marked with an asterisk).

Interestingly, the percentage of errors found in PBSMT and NMT seems to follow a pattern, with inflectional morphology, word order, and mistranslation being the most frequent problems found in both types of MT systems; with exception of the Russian language which presents a bit more mixed results for omission and addition. For German, inflectional morphology errors make up 49% of all the errors found in NMT output, a higher proportion than that found for PBSMT (where it accounts for 43% of the errors).

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. This, in turn, has implications for the post-editing effort involved in bringing the output to publishable quality, which will inevitably vary from one target language to another, also keeping the text type and the domain constant.

4.4 Ranking

For the ranking task, 400 English segments translated into Greek, and 300 segments translated into the other three target languages with NMT and PBSMT were compared side-by-side by professional translators who participated in the evaluation, using Google Forms. Participants preferred NMT output across all language pairs, with a particularly marked preference for English-German, as seen in Table 5. Inter-annotator agreement shows moderate agreement among the annotators (κ=0.60 for DE, κ=0.48 for EL, κ=0.40 for PT and κ=0.61 for RU).

This preference was consistent across all text types, with a 65% preference for NMT in the business analysis forum content, 54% preference for translations of a medical training transcript, 52% for translations of a physics transcript, and 55% for translations of an advertising transcript. Using distinctions from Pouget-Abadie et al. (2014), there was a 53% preference for NMT for short segments (20 tokens or fewer), and a 61% preference for NMT for long segments (over 20 tokens).

Evaluation	**preference for**	
	PBSMT	**NMT**
EN-DE	61	239
(300)	20.3%	79.7%
EN-EL	174	226
(400)	43.5%	56.5%
EN-PT	115	185
(300)	38.3%	61.7%
EN-RU	110	190
(300)	36.7%	63.3%

Table 5: Ranking

We believe that the text genres in which fluency is considered to be more important (i.e. business and marketing) have scored much better for NMT, as opposed to medicine and physics where a translator would tend to follow a more 'literal' translation, as it would typically be more important to translate all the words in the source, so as to ensure that the exact same meaning is preserved, sacrificing fluency if needed. We speculate that, for this reason, NMT may be a good fit for the subtitling domain in general, especially for material that is not particularly specialised.

4.5 Post-editing

Similarly to those segments left without error annotation, fewer NMT segments were considered by participants to require editing during the MT post-editing task. Table 6 shows the number of segments changed and unchanged for all MT systems.

For German, the difference between the number of segments unchanged for NMT when compared with PBSMT output was very statistically significant in a one-way ANOVA pairwise comparison ($p<.05$, where M=.06, SE=.04) (marked with †). Table 7 shows the mean and standard deviation for temporal post-editing effort and Table 8 shows technical post-editing effort in the form of the average number of keystrokes per segment.

Lang.	System	Post-Edited	Unchanged
DE	PBSMT	940	60
	NMT	813	187†
EL	PBSMT	928	72
	NMT	863	137
PT	PBSMT	874	126
	NMT	844	156
RU	PBSMT	930	70
	NMT	848	152

Table 6: Unchanged Segments (out of 1000)

Average throughput or temporal effort was only marginally improved for German, Greek and Portuguese post-editing with NMT, as may be seen at the segment level in Table 7 and expressed in words per second in Table 9, while temporal effort for Russian was lower for PBSMT at the segment level.

Technical post-editing effort was reduced for NMT in all language pairs using measures of actual keystrokes (Table 8) or the minimum number of edits required to go from pre- to post-edited text (cf. the HTER scores in Table 2). Even though these results were not statistically significant, they suggest that those NMT segments that were edited required more cognitive effort than PBSMT segments. Feedback from the participants indicated that they found NMT errors more difficult to identify, whereas word order errors and disfluencies requiring revision were detected faster in PBSMT output.

Lang.	System	Mean	Std. Deviation
DE	PBSMT	74.8	21.12
	NMT	72.8	17.16
EL	PBSMT	77.7	1.85
	NMT	70.4	8.86
PT	PBSMT	57.7	14.23
	NMT	55.19	15.58
RU	PBSMT	104.6	3.62
	NMT	105.6	21.29

Table 7: Temporal Post-Editing Effort (secs/segment)

None of the participants reached the average rate of professional throughput, i.e. 0.39 words per second, found in Moorkens and O'Brien (2015) (Table 9): possibly with the exception of Portuguese, the translators remained quite far from this level of productivity, although it has to be stressed that this is heavily influenced by the type of text being translated as well as by the degree of expertise of the translators, not only with the subject matter at hand, but also, and crucially in the specific case reported here, with PE. This particular result may have also been affected by the unfamiliarity with the interface, the specialised nature of the texts and related research requirements, or perhaps the fact that the rating and annotation tasks carried out after post-editing disturbed the translators' momentum. Productivity is normally achieved with continuous work

Lang.	System	Mean	Std. Deviation
DE	PBSMT	5.8	1.84
	NMT	3.9	1.63
EL	PBSMT	13.9	0.16
	NMT	12.5	1.31
PT	PBSMT	3.8	1.68
	NMT	3.6	1.91
RU	PBSMT	7.5	4.99
	NMT	7.2	5.80

Table 8: Technical Post-Editing Effort (keystrokes/segment)

and translators/editors often report that their productivity peaks half-way into their day.

As for the distinction between long and short segments regarding the decision as to whether post-editing is required, the number of unchanged segments follows the same trend shown in Table 6, where fewer NMT segments were considered to require editing. In terms of words per second (see Table 10), 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 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.

Similarly to the temporal effort results, the technical effort (keystrokes) results show that when distinguishing long and short sentences, German, Greek, and Portuguese present lower PE effort for NMT in short sentences, but the Russian output shows lower effort with PBSMT. For the long sentences, Greek and Russian show lower technical effort for NMT, whereas Portuguese and German show lower effort for the PBSMT system.

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

Table 9: Words per Second (WPS)

	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 (greater than 20 tokens)	DE	0.21	0.20*
	EL	0.20	0.22
	PT	0.26	0.25*
	RU	0.13	0.14

Table 10: WPS: long vs short segments

5 Conclusions

This paper has presented the results of a large-scale comparative evaluation between NMT and PBSMT for four language pairs across several metrics, using complementary methods of human evaluation in addition to state-of-the-art automatic evaluation metrics, thus expanding the understanding of NMT's strengths and weaknesses compared to those of PBSMT. The study, that was conducted as part of the TraMOOC project, used translations of English educational domain data from real-life MOOCs into German, Greek, Portuguese, and Russian. For these language pairs and in this domain, we can conclude that fluency is improved and word order errors are fewer when using NMT, confirming the findings of other recent studies (see Section 2). Fewer segments require post-editing when using NMT, especially due to the lower number of morphological errors. There was, however, no clear improvement with regard to omission and mistranslation errors when moving from PBSMT to NMT. There was also no great decrease in PE effort, suggesting that NMT for production may not as yet offer more than an incremental improvement in temporal PE effort.

While overall NMT produced better results for our domain, expectations are high for NMT and financial pressures mean that the translation industry is eager for a leap forward in MT quality (Moorkens, 2017). At this juncture, however, the neural paradigm is not a panacea. Following on from this study, we intend to compare cognitive post-editing effort using average pause ratio (Lacruz et al., 2012) and to evaluate the effects of added in-domain data on NMT quality and domain specificity.

Acknowledgement The TraMOOC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement №644333. 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.

References

Abdelali, A., Guzman, F., Sajjad, H., and 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.

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

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

Banerjee, S. and 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*, volume 29, pages 65–72.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. *CoRR*, abs/1608.04631.

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., and Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

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

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT3: Web Inventory of Transcribed and Translated Talks. In *Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy.

Chen, S. F. and Goodman, J. (1998). An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge, MA, USA.

Cherry, C. and 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*, NAACL HLT '12, pages 427–436, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Durrani, N., Fraser, A., and 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*, pages 1–11, Atlanta, GA, USA.

Durrani, N., Sajjad, H., Hoang, H., and 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*, pages 148–153, Gothenburg, Sweden.

Etchegoyhen, T., Bywood, L., Fishel, M., Georgakopoulou, P., Jiang, J., Loenhout, G. V., Pozo, A. D., Maucec, M. S., Turner, A., and Volk, M. (2014). Machine Translation for Subtitling: A Large-Scale Evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Galley, M. and Manning, C. D. (2008). A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 848–856, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gao, Q. and Vogel, S. (2008). Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, Columbus, OH, USA.

Haddow, B., Huck, M., Birch, A., Bogoychev, N., and Koehn, P. (2015). The Edinburgh/JHU Phrase-based Machine Translation Systems for WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 126–133, Lisbon, Portugal.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Jean, S., Firat, O., Cho, K., Memisevic, R., and Bengio, Y. (2015). Montreal Neural Machine Translation Systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal.

Junczys-Dowmunt, M., Dwojak, T., and Hoang, H. (2016). Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. In *Arxiv*.

Kneser, R. and Ney, H. (1995). Improved Backing-Off for M-gram Language Modeling. In *Proceedings of the Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, Detroit, MI, USA.

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., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL-2007 Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Krings, H. P. (2001). *Repairing texts: empirical investigations of machine translation post-editing processes*. Kent State University Press, Kent, Ohio.

Lacruz, I., Shreve, G. M., and Angelone, E. (2012). Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, pages 21–30, San Diego, USA.

Lommel, A. R. and DePalma, D. A. (2016). Europe's Leading Role in Machine Translation: How Europe Is Driving the Shift to MT. Technical report, Common Sense Advisory, Boston, USA.

Luong, M.-T. and Manning, C. D. (2015). Stanford Neural Machine Translation Systems for Spoken Language Domains. In *Proceedings of the International Workshop on Spoken Language Translation 2015*, Da Nang, Vietnam.

Moorkens, J. (2017). Under pressure: translation in times of austerity. *Perspectives: Studies in Translation Theory and Practice*, 25(3).

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

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Stroudsburg, PA, USA.

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

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

Pouget-Abadie, J., Bahdanau, D., van Merrienboer, B., Cho, K., and Bengio, Y. (2014). Overcoming the Curse of Sentence Length for Neural Machine Translation using Automatic Segmentation. *CoRR*, abs/1409.1257.

Schuster, M., Johnson, M., and Thorat, N. (2016). Zero-Shot Translation with Google's Multilingual Neural Machine Translation System.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A. V., Mokry, J., and Nadejde, M. (2017). 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*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

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

Sennrich, R., Haddow, B., and 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)*.

Sennrich, R., Haddow, B., and 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.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200(6).

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

Toral, A. and Sanchez-Cartagena, V. M. (2017). A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. In *Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2017. To Appear, Valencia, Spain. ACL.

Way, A. (2013). Traditional and Emerging Use-cases for Machine Translation. In *Translating and the Computer 35*, TC35, London, UK. ASLIB.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., 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., and Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, abs/1609.08144.

Zeiler, M. D. (2012). ADADELTA: An Adaptive Learning Rate Method. *CoRR*, abs/1212.5701.