Machine Translation for Machines:
the Sentiment Classification Use Case

Amirhossein Tebbifakhr[1,2], Luisa Bentivogli[1], Matteo Negri[1], Marco Turchi[1]
[1] Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento - Italy
[2] University of Trento, Italy
`{atebbifakhr,bentivo,negri,turchi}@fbk.eu`

Abstract

We propose a neural machine translation (NMT) approach that, instead of pursuing adequacy and fluency ("human-oriented" quality criteria), aims to generate translations that are best suited as input to a natural language processing component designed for a specific downstream task (a "machine-oriented" criterion). Towards this objective, we present a reinforcement learning technique based on a new candidate sampling strategy, which exploits the results obtained on the downstream task as weak feedback. Experiments in sentiment classification of Twitter data in German and Italian show that feeding an English classifier with machine-oriented translations significantly improves its performance. Classification results outperform those obtained with translations produced by general-purpose NMT models as well as by an approach based on reinforcement learning. Moreover, our results on both languages approximate the classification accuracy computed on gold standard English tweets.

1 Introduction

Traditionally, machine translation (MT) pursues a "human-oriented" objective: generating fluent and adequate output to be consumed by speakers of the target language. But what if the intended use of MT is to feed a natural language processing (NLP) component instead of a human? This, for instance, happens when MT is used as a pre-processing step to perform a downstream NLP task in a language for which dedicated tools are not available due to the scarcity of task-specific training data. The rapid growth of cloud-based software-as-a-service offerings provides a typical example of this situation: a variety of affordable high-performance NLP tools can be easily accessed via APIs but often they are available only for a few languages.

Translating into one of these high-resource languages gives the possibility to address the downstream task by: *i)* using existing tools for that language to process the translated text, and *ii)* projecting their output back to the original language.

However, using MT "as is" might not be optimal for different reasons. First, despite the qualitative leap brought by neural networks, MT is still not perfect (Koehn and Knowles, 2017). Second, previous literature shows that MT can alter some of the properties of the source text (Mirkin et al., 2015; Rabinovich et al., 2017; Vanmassenhove et al., 2018). Finally, even in the case of a perfect MT able to preserve all the traits of the source sentence, models are still trained on parallel data, which are created by humans and thus reflect quality criteria relevant for humans.

In this work, we posit that these criteria might not be the optimal ones for a machine (i.e. a downstream NLP tool fed with MT output). In this scenario, MT should pursue the objective of preserving and emphasizing those properties of the source text that are crucial for the downstream task at hand, even at the expense of human-quality standards. To this end, inspired by previous – human-oriented – MT approaches based on *Reinforcement Learning* (Ranzato et al., 2016; Shen et al., 2016) and *Bandit Learning* (Kreutzer et al., 2017; Nguyen et al., 2017), we explore a NMT optimization strategy that exploits the weak feedback from the downstream task to influence system's behaviour towards the generation of optimal "machine-oriented" output.

As a proof of concept, we test our approach on a sentiment classification task, in which Twitter data in German and Italian are to be classified according to their polarity by means of an English classifier. In this setting, a shortcoming of previous translation-based approaches (Denecke, 2008; Balahur et al., 2014) is that, similar to other traits,

sentiment is often not preserved by MT (Salameh et al., 2015; Mohammad et al., 2016; Lohar et al., 2017). Although it represents a viable solution to leverage sentiment analysis to a wide number of languages (Araujo et al., 2016), the translation-based approach should hence be supported by advanced technology able to preserve the sentiment traits of the input. Along this direction, our experiments show that machine-oriented MT optimization makes the classifier's task easier and eventually results in significant classification improvements. Our results outperform those obtained with translations produced by general-purpose NMT models as well as by an NMT approach based on reinforcement learning (Ranzato et al., 2016). Most noticeably, on both languages we are able to approximate the classification accuracy computed on gold standard English tweets.

2 Background and Methodology

During training, NMT systems based on the encoder-decoder framework (Sutskever et al., 2014; Bahdanau et al., 2015) are optimized with maximum likelihood estimation (MLE), which aims to maximize the log-likelihood of the training data. In doing so, they indirectly model the human-oriented quality criteria (adequacy and fluency) expressed in the training corpus. A different strand of research (Ranzato et al., 2016; Shen et al., 2016; Kreutzer et al., 2018) focuses on optimizing the model parameters by maximizing an objective function that leverages either an evaluation metric like BLEU (Papineni et al., 2002) or an external human feedback. These methods are based on *Reinforcement Learning* (RL), in which the MT system parameters θ define a *policy* that chooses an *action*, i.e. generating the next word in a translation candidate $\hat{\mathbf{y}}$, and gets a *reward* $\Delta(\hat{\mathbf{y}})$ according to that action. Given S training sentences $\{\mathbf{x}^{(s)}\}_{s=1}^{S}$, the RL training goal is to maximize the expected reward:

$$\mathcal{L}_{RL} = \sum_{s=1}^{S} E_{\hat{\mathbf{y}} \sim p_\theta(.|\mathbf{x}^{(s)})} \Delta(\hat{\mathbf{y}})$$
$$= \sum_{s=1}^{S} \sum_{\hat{\mathbf{y}} \in \mathbf{Y}} p_\theta(\hat{\mathbf{y}}|\mathbf{x}^{(s)}) \Delta(\hat{\mathbf{y}}) \quad (1)$$

where $\mathbf{Y}$ is the set of all translation candidates. Since the size of this set is exponentially large, it is impossible to exhaustively compute the expected reward, which is thus estimated by sam-

pling one or few candidates from this set. In the MT adaptation (Ranzato et al., 2016) of REINFORCE (Williams, 1992), the closest approach to the one we present, only one candidate is sampled.

$$\hat{\mathcal{L}}_{RL} = \sum_{s=1}^{S} p_\theta(\hat{\mathbf{y}}|\mathbf{x}^{(s)}) \Delta(\hat{\mathbf{y}}), \hat{\mathbf{y}} \sim p_\theta(.|\mathbf{x}^{(s)}) \quad (2)$$

We now focus on how the key elements of RL methods have been adapted to properly work in the proposed "MT for machines" setting. Our novel *Machine-Oriented* approach is described in Algorithm 1.

Algorithm 1 *Machine-Oriented* RL

1: **Input:** $\mathbf{x}^{(s)}$ s-th source sentence in training data, $\mathbf{l}^{(s)}$ the ground-truth label, K number of sampled candidates
2: **Output:** sampled candidate $\hat{\mathbf{y}}^{(s)}$
3: $\mathbf{C} = \emptyset$ ▷ Candidates set
4: **for** $k = 1,...,$K **do**
5:      $\mathbf{c} \sim p_\theta(.|\mathbf{x}^{(s)})$
6:      $\mathbf{f} = P_{class}(\mathbf{l}^{(s)}|\mathbf{c})$ ▷ Feedback from the classifier
7:      $\mathbf{C} = \mathbf{C} \cup (\mathbf{c}, \mathbf{f})$
8: $\hat{\mathbf{y}}^{(s)} = \max_{\mathbf{f}}(\mathbf{C})$ ▷ Best candidate w.r.t. feedback
---

Reward Computation. In current RL approaches to MT, rewards are computed either on a development set containing reference translations (Ranzato et al., 2016; Bahdanau et al., 2017) or by means of a weak feedback (e.g. a 1-to-5 score) when ground-truth translations are not accessible (Kreutzer et al., 2017; Nguyen et al., 2017). In both cases, the reward reflects a human-oriented notion of MT quality. Instead, in our machine-oriented scenario, the reward reflects the performance on the downstream task, independently from translation quality.

In the sentiment classification use case, given a source sentence ($\mathbf{x}^{(s)}$) and its sentiment ground-truth label ($\mathbf{l}^{(s)}$), the reward is defined as the probability assigned by the classifier to the ground-truth class for each translated sentence $\mathbf{c}$ ($P_{class}(\mathbf{l}^{(s)}|\mathbf{c})$ in Algorithm 1, line 6). By maximizing this reward, the MT system learns to produce a translation that has higher chances to be correctly labeled by the classifier. This comes at the risk of obtaining less fluent and adequate output. However, as will be shown in Section 4, this

type of reward induces highly polarized translations that are best suited for our downstream task.

Sampling Approach. A possible sampling strategy is to exploit beam search (Sutskever et al., 2014) to find, at each decoding step, the candidate with the highest probability. Another solution is to use multinomial sampling (Ranzato et al., 2016) which, at each decoding step, samples tokens over the model's output distribution. In (Wu et al., 2018), the higher results achieved by multinomial sampling are ascribed to its capability to better explore the probability space by generating more diverse candidates. This finding is particularly relevant in the proposed "MT for machines" scenario, in which the emphasis on final performance in the downstream task admits radical (application-oriented) changes in the behaviour of the MT model, even at the expense of human quality standards.

To increase the possibility of such changes, we propose a new sampling strategy. Instead of generating only one candidate *token* via multinomial sampling, K candidate *sentences* are first randomly sampled (lines 4-5 in Algorithm 1). Then, the reward is collected for each of them (line 7) and the candidate with the highest reward is chosen (line 8). On one side, randomly exploring more candidates increases the probability to sample a "useful" one, possibly by diverting from the initial model behaviour. On the other side, selecting the candidate with the highest reward will push the system towards translations emphasizing input traits that are relevant for the downstream task at hand. In the sentiment classification use case, these are expected to be sentiment-bearing terms that help the classifier to predict the correct class.

3 Experiments

Our evaluation is done by feeding an English sentiment classifier with the translations of German and Italian tweets generated by:

- A general-purpose NMT system (*Generic*);

- The same system conditioned with REIN-FORCE (*Reinforce*);

- The same system conditioned with our *M*achine-*O*riented method (*MO-Reinforce*).

As other terms of comparison, we calculate the results of:

- The English classifier on the gold standard English tweets (*English*);

- German and Italian classifiers on the original, untranslated tweets (*Original*).

Task-specific data. We experiment with a dataset based on Semeval 2013 data (Nakov et al., 2013), which contains polarity-labeled parallel German/Italian–English corpora (Balahur et al., 2014). For each language pair, the development and test sets respectively comprise 583 (197 negative and 386 positive) and 2,173 tweets (601 negative and 1,572 positive). To cope with the skewed data distribution, the negative tweets in the development sets are over-sampled, leading to new balanced sets of 772 tweets.

NMT Systems. Our *Generic* models are based on Transformer (Vaswani et al., 2017), with parameters similar to those used in the original paper. Training data amount to 6.1M (De-En) and 4.56M (It-En) parallel sentences from freely-available corpora. The statistics of the parallel corpora are reported in Table 1. For each language pair, all data are merged and tokenized. Then, byte pair encoding (Sennrich et al., 2016) is applied to obtain 32K sub-word units.

	De-En	It-En
Europarl	2M	2M
JRC	0.7M	0.8M
Wikipedia	2.5M	1M
ECB	0.1M	0.2M
TED	0.1M	0.2M
KDE	0.3M	0.3M
News11	0.2M	0.04M
News	0.2M	0.02M
Total	6.1M	4.56M

Table 1: Statistics of the parallel corpora used for training the generic NMT systems

To emulate both scarce and sufficient training data conditions, MT systems are trained using 5% and 100% of the available parallel data. In the most favorable condition (i.e. with 100% of the data), the BLEU score of the two models is 30.48 for De-En and 28.68 for It-En.

To condition the generic models and obtain the *Reinforce* and *MO-Reinforce* systems, we use the polarity-labeled German/Italian tweets in our de-

velopment sets, with reward as defined in Section 2. The SGD optimizer is used, with learning rate set to 0.01. In *MO-Reinforce*, the number of sampled candidates is set to $K = 5$.

Classifiers. To simulate an English cloud-based classifier-as-a-service, the pre-trained BERT (Base-uncased) model (Devlin et al., 2019) is fine-tuned with a balanced set of 1.6M positive and negative English tweets (Go et al., 2009). The German and Italian classifiers are also created by fine-tuning BERT on the polarity-labeled tweets composing the source side of our development set (772 tweets).[1] Before being passed to the classifiers, URLs and user mentions are removed from the tweets, which are then tokenized and lower-cased.

4 Results and Discussion

Table 2 shows our classification results, presenting the F1 scores obtained by the different MT-based approaches in the two training conditions. When NMT is trained on 100% of the parallel data, for both languages *Reinforce* produces translations that lead to classification improvements over those produced by the *Generic* model (+0.5 De-En, +0.8 It-En). Although the scores are considerably better than those obtained by the *Original* classifiers (+9.3 De-En, +7.2 It-En), the gap with respect to the *English* classifier is still quite large (-1.4 De-En and -2.3 It-En).

The observed F1 gains over the *Generic* model reflect an improvement in translation quality. Indeed, the BLEU score (not reported in the table) increases for both languages (+0.83 De-En, +1.37 It-En). As suggested by Kreutzer et al. (2018), this can be motivated by the fact that, rather than actually leveraging the feedback, RL mainly benefits from the optimization on in-domain source sentences in the development set. This is confirmed by the fact that *MO-Reinforce*, which uses a different sampling strategy, is able to outperform *Reinforce* on both languages (+0.7 F1 on De-En, +1.7 on It-En), approaching the *English* upper bound.

Unsurprisingly, the BLEU score obtained by *MO-Reinforce* is close to zero. Indeed, the generated sentences are highly polarized and fluent, but

---

	De - En		It - En	
	5%	*100%*	*5%*	*100%*
Generic	79.7	83.2	78.2	81.6
Reinforce	80.4	83.7	77.8	82.8
MO-Reinforce	**80.9**	**84.4**	**80.3**	**84.5**
English	85.1			
Original	74.4		75.6	

Table 2: Classification results (F1) obtained with: *i)* automatic English translations by three models (*Generic*, *Reinforce*, *MO-Reinforce*), and *ii)* gold-standard English (*English*) and untranslated German/Italian (*Original*) tweets.

not adequate with respect to the source sentence. For example, the positive instances in the test set are translated into "*It's good!*", "*I'm happy*" or "*I'm grateful*", and the negative ones into "*it is not good!*" or "*I'm sorry*". This polarization shows that *MO-Reinforce* maximizes the exploitation of the received sentiment feedback thus producing an output that, at the expense of adequacy, can be easily classified by the downstream task.

When reducing the MT training data to only 5%, emulating a condition of parallel data scarcity, all the classification results decrease (on average by 3-4 points). However, also in this case the output of *MO-Reinforce* yields the closest scores to the *English* upper bound. This indicates that our method does not require large data quantities to outperform the other approaches.

To validate the hypothesis that *MO-Reinforce* can better leverage the feedback from the downstream task, Figures 1 and 2 show the average rewards for the De-En and the It-En candidates generated by *Reinforce* and *MO-Reinforce* at each epoch of the adaptation process. In line with the findings of Kreutzer et al. (2018), for both languages, *Reinforce* is not able to leverage the feedback and to generate candidates that increase their reward during training. Indeed, its curves in the two figures show either a stable (It-En) or even a slightly downward trend (De-En) that confirms the known limitations of NMT to preserve sentiment traits of the source sentences (see Section 1). In contrast, the *MO-Reinforce* curves show a clear upward trend indicating a higher capability to exploit the feedback and produce, epoch by epoch, increasingly polarized translations that are easier

---

[1]This strategy is similar to the one proposed in (Eriguchi et al., 2018), where a pre-trained multilingual encoder is used to build a cross-lingual classifier that is then trained on task-specific data. In our case, however, labeled data are available in a small amount.
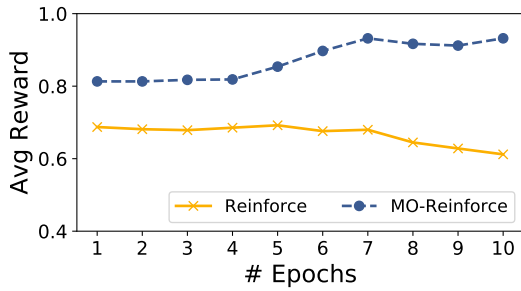
Figure 1: Average rewards for *Reinforce* and *MO-Reinforce* candidates at each training epoch (De-En).
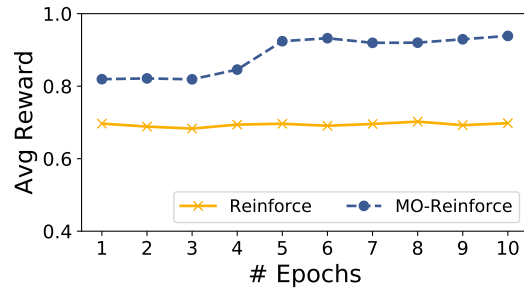

Figure 2: Average rewards for *Reinforce* and *MO-Reinforce* candidates at each training epoch (It-En).
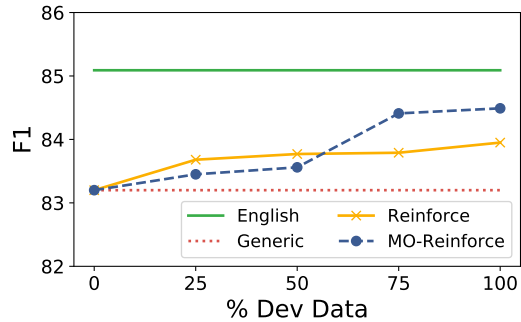

Figure 3: Classification results with increasing amounts of labeled development data (De-En).
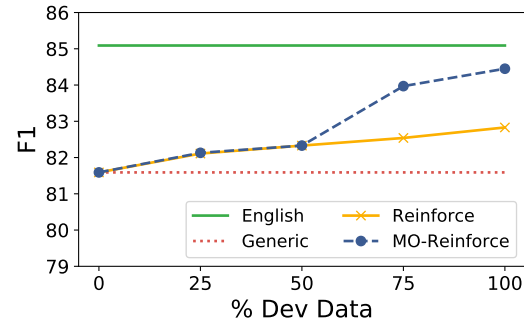

Figure 4: Classification results with increasing amounts of labeled development data (It-En).

to classify.

Another important aspect to investigate is the relation between the size of the development set (i.e. the amount of human-annotated source language tweets needed by the RL methods) and final classification performance. Figures 3 and 4 show the De-En and It-En performance variations of *Reinforce* and *MO-Reinforce* at different sizes of the development set. For comparison purposes, also the *Generic* and *English* results (which are independent from the development set size) are included. In both plots, each point is obtained by averaging the results of three different data shuffles. With limited amounts of data (25% and 50%) *Reinforce* and *MO-Reinforce* have a similar trend. When adding more data, *MO-Reinforce* shows a better use of the labeled development data, with a boost in performance that allows it to approach the *English* upper bound in both language settings. These results suggest that the effort to create labeled data can be minimal, and with 75% of the set (579 points) it is already possible to achieve a better performance than *Reinforce*.

5   Conclusions

We proposed a novel interpretation of the machine translation task, which pursues *machine-*

oriented quality criteria (generating translations that are best suited as input to a downstream NLP component) rather than the traditional *human-oriented* ones (maximizing adequacy and fluency). We addressed the problem by adapting reinforcement learning techniques with a new, exploration-oriented sampling strategy that exploits the results obtained on the downstream task as weak feedback. Instead of generating only one candidate totally randomly via multinomial sampling (i.e. random stepwise selection of each word during generation, as in *Reinforce*), our approach selects K full translation candidates, it computes the reward for each of them and finally chooses the one with the highest reward from the downstream task. As shown by our experiments in sentiment classification, this more focused (and application-oriented) selection allows our "MT for machines" approach to: *i)* better explore the hypotheses space, *ii)* make better use of the collected rewards and eventually *iii)* obtain better downstream classification results compared to translation-based solutions exploiting either general-purpose models or previous reinforcement learning strategies. In future work, we will target new application scenarios, covering multi-class classification and regression tasks.

References

Matheus Araujo, Julio Reis, Adriano Pereira, and Fabricio Benevenuto. 2016. An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, SAC '16, pages 1140–1145, Pisa, Italy. ACM.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. An actor-critic algorithm for sequence prediction. In *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015*.

Alexandra Balahur, Marco Turchi, Ralf Steinberger, José Manuel Perea Ortega, Guillaume Jacquet, Dilek Küçük, Vanni Zavarella, and Adil El Ghali. 2014. Resource creation and evaluation for multilingual sentiment analysis in social media texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 4265–4269.

Kerstin Denecke. 2008. Using sentiwordnet for multilingual sentiment analysis. In *Proceedings of the 24th International Conference on Data Engineering Workshops, ICDE 2008, April 7-12, 2008, Cancún, Mexico*, pages 507–512.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *CoRR*, abs/1809.04686.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12):2009.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada. Association for Computational Linguistics.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can neural machine translation be improved with user feedback? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 92–105, New Orleans - Louisiana. Association for Computational Linguistics.

Julia Kreutzer, Artem Sokolov, and Stefan Riezler. 2017. Bandit structured prediction for neural sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1503–1513, Vancouver, Canada. Association for Computational Linguistics.

Pintu Lohar, Haithem Afli, and Andy Way. 2017. Maintaining Sentiment Polarity in Translation of User-Generated Content. *The Prague Bulletin of Mathematical Linguistics*, 108:73–84.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108, Lisbon, Portugal. Association for Computational Linguistics.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55(1):95–130.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.

Khanh Nguyen, Hal Daumé III, and Jordan Boyd-Graber. 2017. Reinforcement learning for bandit neural machine translation with simulated human feedback. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1464–1474, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages

1074–1084, Valencia, Spain. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016.*

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, Denver, Colorado. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, Brussels, Belgium. Association for Computational Linguistics.