ReWE: Regressing Word Embeddings for Regularization of Neural Machine Translation Systems

Inigo Jauregi Unanue[1,2], Ehsan Zare Borzeshi[*,2], Nazanin Esmaili[2], Massimo Piccardi[1]
[1] University of Technology Sydney, Sydney, Australia
[2] Capital Markets Cooperative Research Centre, Sydney, Australia

{ijauregi,ezborzeshi,nesmaili}@cmcrc.com
massimo.piccardi@uts.edu.au

Abstract

Regularization of neural machine translation is still a significant problem, especially in low-resource settings. To mollify this problem, we propose regressing word embeddings (ReWE) as a new regularization technique in a system that is jointly trained to predict the next word in the translation (categorical value) and its word embedding (continuous value). Such a joint training allows the proposed system to learn the distributional properties represented by the word embeddings, empirically improving the generalization to unseen sentences. Experiments over three translation datasets have showed a consistent improvement over a strong baseline, ranging between 0.91 and 2.54 BLEU points, and also a marked improvement over a state-of-the-art system.

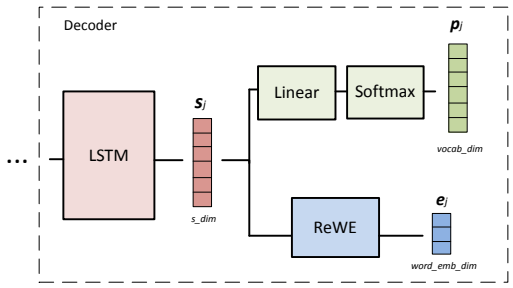

Figure 1: The proposed regularizer: the hidden vector in the decoder, s_j, transits through two paths: 1) a linear and a softmax layers that output vector v_j (vocab_dim) which is used for predicting the target word as usual, and 2) a two-layer network (ReWE) that outputs a vector, e_j, of word embedding size (word_emb_dim). During training, e_j is used in a regressive loss with the ground-truth embedding.

1 Introduction

The last few years have witnessed remarkable improvements in the performance of machine translation (MT) systems. These improvements are strongly linked to the development of neural machine translation (NMT): based on encoder-decoder architectures (also known as seq2seq), NMT can use recurrent neural networks (RNNs) (Sutskever et al., 2014; Cho et al., 2014; Wu et al., 2016), convolutional neural networks (CNNs) (Gehring et al., 2017) or transformers (Vaswani et al., 2017) to learn how to map a sentence from the source language to an adequate translation in the target language. In addition, attention mechanisms (Bahdanau et al., 2015; Luong et al., 2015) help soft-align the encoded source words with the predictions, further improving the translation.

NMT systems are usually trained via maximum likelihood estimation (MLE). However, as pointed out by (Elbayad et al., 2018), MLE suffers from two obvious limitations: the first is that it treats all the predictions other than the ground truth as equally incorrect. As a consequence, synonyms and semantically-similar words — which are often regarded as highly interchangeable with the ground truth — are completely ignored during training. The second limitation is that MLE-trained systems suffer from "exposure bias" (Bengio et al., 2015; Ranzato et al., 2015) and do not generalize well over the large output space of translations. Owing to these limitations, NMT systems still struggle to outperform other traditional MT approaches when the amount of supervised data is limited (Koehn and Knowles, 2017).

In this paper, we propose a novel regularization technique for NMT aimed to influence model learning with contextual properties. The technique — nicknamed ReWE from "regressing word embedding" — consists of modifying a conventional seq2seq decoder to jointly learn to a) predict the next word in the translation (categorical value), as

[*] The author has changed affiliation to Microsoft after the completion of this work. His new email is: ezborzeshi@microsoft.com

usual, and b) regress its word embedding (numerical value). Figure 1 shows the modified decoder. Both predictions are incorporated in the training objective, combining standard MLE with a continuous loss function based on word embeddings. The rationale is to encourage the system to learn to co-predict the next word together with its context (by means of the word embedding representation), in the hope of achieving improved generalization. At inference time, the system operates as a standard NMT system, retaining the categorical prediction and ignoring the predicted embedding. We qualify our proposal as a regularization technique since, like any other regularizers, it only aims to influence the model's training, while leaving the inference unchanged. We have evaluated the proposed system over three translation datasets of different size, namely English-French (en-fr), Czech-English (cs-en), and Basque-English (eu-en). In each case, ReWE has significantly outperformed its baseline, with a marked improvement of up to 2.54 BLEU points for eu-en, and consistently outperformed a state-of-the-art system (Denkowski and Neubig, 2017).

2 Related work

A substantial literature has been devoted to improving the generalization of NMT systems. Fadaee et al. (2017) have proposed a data augmentation approach for low-resource settings that generates synthetic sentence pairs by replacing words in the original training sentences with rare words. Kudo (2018) has trained an NMT model with different subword segmentations to enhance its robustness, achieving consistent improvements over low-resource and out-of-domain settings. Zhang et al. (2018) have presented a novel regularization method that encourages target-bidirectional agreement. Other work has proposed improvements over the use of a single ground truth for training: Ma et al. (2018) have augmented the conventional seq2seq model with a bag-of-words loss under the assumption that the space of correct translations share similar bag-of-words vectors, achieving promising results on a Chinese-English translation dataset; Elbayad et al. (2018) have used sentence-level and token-level reward distributions to "smooth" the single ground truth. Chousa et al. (2018) have similarly leveraged a token-level smoother.

In a recent paper, Denkowski and Neubig (2017) have achieved state-of-the-art translation accuracy by leveraging a variety of techniques which include: dropout (Srivastava et al., 2014), lexicon bias (Arthur et al., 2016), pre-translation (Niehues et al., 2016), data bootstrapping (Chen et al., 2016), byte-pair encoding (Sennrich et al., 2016) and ensembles of independent models (Rokach, 2010).

However, to our knowledge none of the mentioned approaches have explicitly attempted to leverage the embeddings of the ground-truth tokens as targets. For this reason, in this paper we explore regressing toward pre-trained word embeddings as an attempt to capture contextual properties and achieve improved model regularization.

3 Model

3.1 Seq2seq baseline

The model is a standard NMT model with attention in which we use RNNs for the encoder and decoder. Following the notation of (Bahdanau et al., 2015), the RNN in the decoder generates a sequence of hidden vectors, $\{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$, given the context vector, the previous hidden state $\mathbf{s}_{j-1}$ and the previous predicted word $\mathbf{y}_{j-1}$:

$$\mathbf{s}_j = dec_{rnn}(\mathbf{s}_{j-1}, \mathbf{y}_{j-1}, \mathbf{c}_j) \quad j = 1, \ldots, m \quad (1)$$

where y_0 and s_0 are initializations for the state and label chains. Each hidden vector $\mathbf{s}_j$ (of parameter size S) is then linearly transformed into a vector of vocabulary size, V, and a softmax layer converts it into a vector of probabilities (Eq. 2), where W (a matrix of size $V \times S$) and b (a vector of size $V \times 1$) are learnable parameters. The predicted conditional probability distribution over the words in the target vocabulary, $\mathbf{p}_j$, is given as:

$$\mathbf{p}_j = softmax(\mathbf{W}\mathbf{s}_j + \mathbf{b}) \quad (2)$$

As usual, training attempts to minimize the negative log-likelihood (NLL), defined as:

$$NLL_{loss} = -\sum_{j=1}^{m} \log(\mathbf{p}_j(\mathbf{y}_j)) \quad (3)$$

where $\mathbf{p}_j(\mathbf{y}_j)$ notes the probability of ground-truth word $\mathbf{y}_j$. The NLL loss is minimized when the probability of the ground truth is one and that of all other words is zero, treating all predictions different from the ground truth as equally incorrect.

3.2 ReWE

Pre-trained word embeddings (Pennington et al., 2014; Bojanowski et al., 2017; Mikolov et al., 2013) capture the contextual similarities of words, typically by maximizing the probability of word w_{t+k} to occur in the context of center word w_t. This probability can be expressed as:

$$p(w_{t+k}|w_t), \quad -c \leq k \leq c, k \neq 0 \quad t = 1, \ldots, T \tag{4}$$

where c is the size of the context and T is the total number of words in the training set. Traditionally, word embeddings have only been used as input representations. In this paper, we instead propose using them in output as part of the training objective, in the hope of achieving regularization and improving prediction accuracy. Building upon the baseline model presented in Section 3.1, we have designed a new "joint learning" setting: our decoder still predicts the probability distribution over the vocabulary, $\mathbf{p}_j$ (Eq. 2), while simultaneously regressing the same shared $\mathbf{s}_j$ to the ground-truth word embedding, $e(\mathbf{y}_j)$. The ReWE module consists of two linear layers with a Rectified Linear Unit (ReLU) in between, outputting a vector $\mathbf{e}_j$ of word embedding size (Eq. 5). Please note that adding this extra module adds negligible computational costs and training time. Full details of this module are given in Appendix A in the supplementary material.

$$\begin{aligned} \mathbf{e}_j &= ReWE(\mathbf{s}_j) \\ &= \mathbf{W}_2(ReLU(\mathbf{W}_1\mathbf{s}_j + \mathbf{b}_1)) + \mathbf{b}_2 \end{aligned} \tag{5}$$

The training objective is a numerical loss, l (Eq. 6), computed between the output vector, $\mathbf{e}_j$, and the ground-truth embedding, $e(\mathbf{y}_j)$:

$$ReWE_{loss} = l(\mathbf{e}_j, e(\mathbf{y}_j)) \tag{6}$$

In the experiment, we have explored two cases for the $ReWE_{loss}$: the minimum square error (MSE)[1] and the cosine embedding loss (CEL)[2]. Finally, the NLL_{loss} and the $ReWE_{loss}$ are combined to form the training objective using a positive trade-off coefficient, λ:

$$Loss = NLL_{loss} + \lambda ReWE_{loss} \tag{7}$$

[1]https://pytorch.org/docs/stable/nn.html#torch.nn.MSELoss

[2]https://pytorch.org/docs/stable/nn.html#torch.nn.CosineEmbeddingLoss

Dataset	Size	Sources
IWSLT16 en-fr	219,777	TED talks
IWSLT16 cs-en	114,243	TED talks
WMT16 eu-en	89,413	IT-domain data

Dataset	Validation set	Test set
en-fr	TED test 2013+2014	TED test 2015+2016
cs-en	TED test 2012+2013	TED test 2015+2016
eu-en	Sub-sample of PaCo	IT-domain test

Table 1: Top: parallel training data. Bottom: validation and test sets.

As mentioned in the Introduction, at inference time we ignore the ReWE output, $\mathbf{e}_j$, and the model operates as a standard NMT system.

4 Experiments

We have developed our models building upon the OpenNMT toolkit (Klein et al., 2017)[3]. For training, we have used the same settings as (Denkowski and Neubig, 2017). We have also explored the use of sub-word units learned with byte pair encoding (BPE) (Sennrich et al., 2016). All the preprocessing steps, hyperparameter values and training parameters are described in detail in the supplementary material to ease reproducibility of our results.

We have evaluated these systems over three publicly-available datasets from the 2016 ACL Conference on Machine Translation (WMT16)[4] and the 2016 International Workshop on Spoken Language Translation (IWSLT16)[5]. Table 1 lists the datasets and their main features. Despite having nearly 90,000 parallel sentences, the eu-en dataset only contains 2,000 human-translated sentences; the others are translations of Wikipedia page titles and localization files. Therefore, we regard the eu-en dataset as very low-resource.

In addition to the seq2seq baseline, we have compared our results with those recently reported by Denkowski and Neubig for non-ensemble models (2017). For all models, we report the BLEU scores (Papineni et al., 2002), with the addition of selected comparative examples. Two contrastive experiments are also added in Appendix C.

4.1 Results

As a preliminary experiment, we have carried out a sensitivity analysis to determine the optimal value of the trade-off coefficient, λ (Eq. 6), using the

[3]All our code will be released publicly after the anonymity period, and it is also available to the reviewers as supplementary material.

[4]WMT16: http://www.statmt.org/wmt16/

[5]IWSLT16: https://workshop2016.iwslt.org/

Models	en-fr		cs-en		eu-en	
	Word	BPE	Word	BPE	Word	BPE
(Denkowski and Neubig, 2017)	33.60	34.50	21.00	22.60		
(Denkowski and Neubig, 2017) + Dropout	34.5	34.70	21.4	23.60		
(Denkowski and Neubig, 2017) + Lexicon	33.9	34.80	20.6	22.70		
(Denkowski and Neubig, 2017) + Pre-translation	N/A	34.90	N/A	**23.80**		
(Denkowski and Neubig, 2017) + Bootstrapping	34.40	35.20	21.60	23.60		
Our baseline	34.16	34.09	20.57	22.69	12.14	17.17
Our baseline + ReWE (CEL) ($\lambda = 20$)	**35.52**	**35.22**	**21.83**	23.60	**13.73**	**19.71**

Table 2: BLEU scores over the test sets. Average of 10 models independently trained with different seeds.

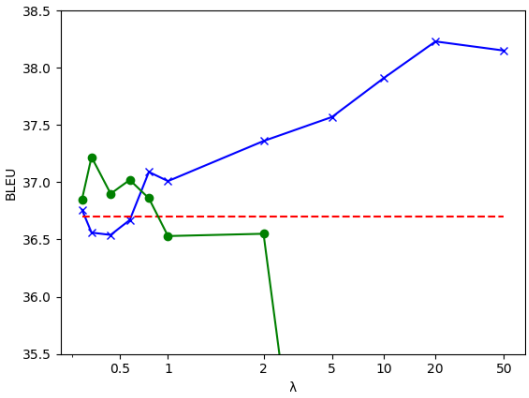

Figure 2: BLEU scores of three models over the en-fr validation set for different λ values: baseline (**red**), baseline + ReWE (MSE) (**green**), baseline + ReWE (CEL) (**blue**). Each point in the graph is an average of 3 independently trained models.

en-fr validation set. The results are shown in Figure 2, where each point is the average of three runs trained with different seeds. The figure shows that the MSE loss has outperformed slightly the baseline for small values of λ (< 1), but the BLEU score has dropped drastically for larger values. Conversely, the CEL loss has increased steadily with λ, reaching 38.23 BLEU points for $\lambda = 20$, with a marked improvement of 1.53 points over the baseline. This result has been encouraging and therefore for the rest of the experiments we have used CEL as the $ReWE_{loss}$ and kept the value of λ to 20. In Section 4.3, we further discuss the behavior of CEL and MSE.

Table 2 reports the results of the main experiment for all datasets. The values of our experiments are for blind runs over the test sets, averaged over 10 independent runs with different seeds. The results show that adding ReWE has significantly improved the baseline in all cases, with an average of 1.46 BLEU points. In the case of the eu-en dataset, the improvement has reached 2.54 BLEU points. We have also run unpaired t-tests between our baseline and ReWE, and the differences have

Src:	Hautatu Kontrol panela → Programa lehenetsiak , eta aldatu bertan .
Ref:	Go to Control Panel → Default programs , and change it there .
Baseline:	Select the Control Panel → program , and change .
Baseline + ReWE:	Select the Control Panel → Default Program , and change it .

Table 3: Translation example from the eu-en test set.

proved statistically significant (p-values < 0.05) in all cases. Using BPE has proved beneficial for the cs-en and eu-en pairs, but not for the en-fr pair. We speculate that English and French may be closer to each other at word level and, therefore, less likely to benefit from the use of sub-word units. Conversely, Czech and Basque are morphologically very rich, justifying the improvements with BPE.

Table 2 also shows that our model has outperformed almost all the state-of-the-art results reported in (Denkowski and Neubig, 2017) (dropout, lexicon bias, pre-translation, and bootstrapping), with the only exception of the pre-translation case for the cs-en pair with BPE. This shows that the proposed model is competitive with contemporary NMT techniques.

4.2 Qualitative comparison

To further explore the improvements obtained with ReWE, we have qualitatively compared several translations provided by the baseline and the baseline + ReWE (CEL), trained with identical seeds. Overall, we have noted a number of instances where ReWE has provided translations with more information from the source (higher adequacy). For reasons of space, we report only one example in Table 3, but more examples are available in Appendix B. In the example, the baseline has chosen a generic word, "program", while ReWE has been capable of correctly predicting "Default Program" and being specific about the object, "it".

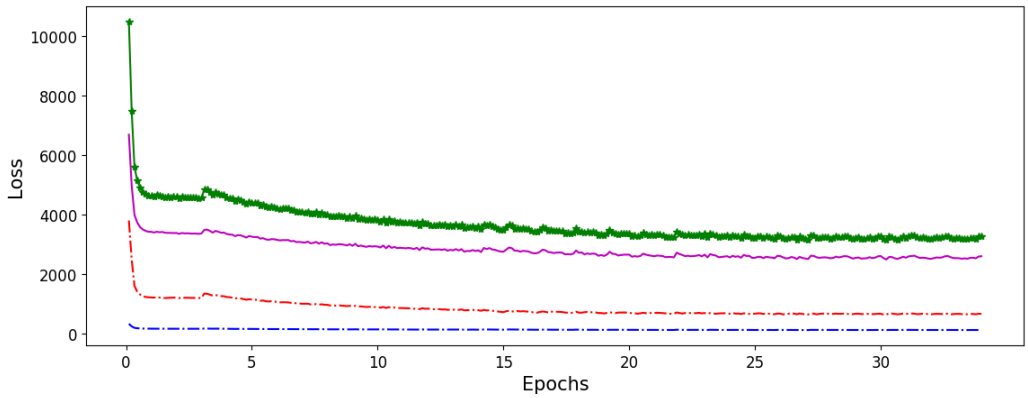

Figure 3: Plot of the values of various loss functions during training of our model over the en-fr training set: **green**: training loss (NLL + ($\lambda = 20$) ReWE (CEL); Eq.7); **red**: NLL loss; **blue**: ReWE (CEL) loss; **magenta**: ReWE (CEL) loss scaled by $\lambda = 20$. Each point in the graph is an average value of the corresponding loss over 25,000 sentences.

4.3 Discussion

To further explore the behaviour of the ReWE loss, Figure 3 plots the values of the NLL and ReWE (CEL) losses during training of our model over the en-fr training set. The natural values of the ReWE (CEL) loss (**blue** curve) are much lower those of the NLL loss (**red** curve), and thus its contribution to the gradient is likely to be limited. However, when scaled up by a factor of $\lambda = 20$ (**magenta** curve), its influence on the gradient becomes more marked. Empirically, both the NLL and ReWE (CEL) losses decrease as the training progresses and the total loss (**green** curve) decreases. As shown in the results, this combined training objective has been able to lead to improved translation results.

Conversely, the MSE loss has not exhibited a similarly smooth behaviour (Appendix D). Even when brought to scale with the NLL loss, it shows much larger fluctuations as the training progresses. In particular, it shows major increases at the re-starts of the optimizer for the simulated annealing that are not compensated for by the rest of the training. It is easy to speculate that the MSE loss is much more sensitive than the cosine distance to the changes in the weights caused by dropout and the re-starts. As such, it seems less suited for use as training objective.

5 Conclusion

In this paper, we have proposed a new regularization technique for NMT (ReWE) based on a joint learning setting in which a seq2seq model simultaneously learns to a) predict the next word in the translation and b) regress toward its word embedding. The results over three parallel corpora have shown that ReWE has consistently improved over both its baseline and recent state-of-the-art results from the literature. As future work, we plan to extend our experiments to better understand the potential of the proposed regularizer, in particular for unsupervised NMT (Artetxe et al., 2018; Lample et al., 2018).

6 Acknowledgments

We would like to acknowledge the financial support received from the Capital Markets Cooperative Research Centre (CMCRC), an industry-led research initiative of the Australian Government. We would also like to thank Ben Hachey, Michael Nolan and Nadia Shnier for their careful reading of our paper and their insightful comments. Finally, we are grateful to the anonymous reviewers for all their comments and suggestions.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Empirical Methods in Natural Language Processing*, pages 1557–1567.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly

learning to align and translate. In *International Conference on Learning Representations*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, pages 135–146.

Wenhu Chen, Evgeny Matusov, Shahram Khadivi, and Jan-Thorsten Peter. 2016. Guided alignment training for topic-aware neural machine translation. *arXiv preprint arXiv:1607.01628*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing*, pages 1724–1734.

Katsuki Chousa, Katsuhito Sudoh, and Satoshi Nakamura. 2018. Training neural machine translation using word embedding-based loss. *arXiv preprint arXiv:1807.11219*.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 18–27. Empirical Methods in Natural Language Processing.

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2018. Token-level and sequence-level loss smoothing for rnn language models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2094–2103.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 567–573.

Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. 2017. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 123–135.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

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. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 66–75.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *Empirical Methods in Natural Language Processing*, pages 5039–5049.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Empirical Methods in Natural Language Processing*, pages 1412–1421.

Shuming Ma, Xu Sun, Yizhong Wang, and Junyang Lin. 2018. Bag-of-words as target for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 332–338.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-translation for neural machine translation. *arXiv preprint arXiv:1610.05243*.

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 on association for computational linguistics*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, pages 1532–1543.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. In *Advances in Neural Information Processing Systems*.

Lior Rokach. 2010. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39.

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*, pages 1715–1725.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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 *Advances in Neural Information Processing Systems*, pages 5998–6008.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhirui Zhang, Shuangzhi Wu, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Regularizing neural machine translation by target-bidirectional agreement. *arXiv preprint arXiv:1808.04064*.