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

A Training and hyperparameters

In this appendix we provide all the information required to reproduce our results. The models have been implemented by modifying OpenNMT (Klein et al., 2017) and we will release our code publicly immediately after the anonymity period. All the code is already available to the reviewers as supplementary material.

To build a strong and current baseline, we have closely followed the indications of (Denkowski and Neubig, 2017). The baseline uses a single-layer bidirectional LSTM and a unidirectional LSTM as encoder and decoder, respectively. The attention mechanism is that of (Bahdanau et al., 2015). We have set the size of the LSTMs' hidden layer to 1024, the size of the attention layer to the same size, and the size of the word embeddings to 300. We have initialized the word embeddings with the publicly-available pre-trained vectors from fastText[1] for each language. The maximum length of the training sentences has been set to 100 tokens. The model vocabulary has been limited to $50,000$ words for both the source and target languages. Words that are not present in the vocabulary are mapped to an unk token, but are later replaced with the corresponding source word with highest attention, following (Luong et al., 2015). For inference, we have used beam search with a beam size of 5.

We have added ReWE to this baseline, keeping all the aforementioned values unchanged. As mentioned in the paper, ReWE is a stack of two linear layers with a ReLU in between. The first linear layer reduces vector s_j from size 1024 to 200. After the ReLU, the second linear layer expands the vector from size 200 to 300, which is the size of the word embeddings. The value for λ has been selected by evaluating the model over the en-fr validation set (see Section 4.2 in the paper).

All the models have been trained until convergence of the perplexity, using the Adam optimizer (Kingma and Ba, 2015), with a maximum step size of 0.0002, multiple restarts, and learning rate annealing (Denkowski and Neubig, 2017). After three consecutive validation evaluations without perplexity improvement, we halve the learning rate, and we repeate this process 5 times. After the 5-th halving, we stop the training if there is no perplexity improvement over 20 consecutive runs. The batch size is 40 and the model is evalauted every $25,000$ sentences.

We have also trained the models at sub-word level using byte pair encoding (BPE) (Sennrich et al., 2016). We have learned the sub-word models using the concatenated training sets of all datasets, setting the number of merge operations to $32,000$ for en-fr and cs-en, and to $8,000$ for eu-en, given its much smaller size. We have also pre-trained word embeddings of size 300 for the new sub-word vocabularies, and used them for initialization of the word embeddings.

For each model, we have reported the average BLEU score (Papineni et al., 2002) of 10 independent runs, except for the selection of λ where we have averaged only 3 independent runs.

B Translation examples

In this section we showcase more examples of translations made by the model with and without ReWE for all the language pairs evaluated in the paper (en-fr, cs-en and eu-en). In general the translations made by ReWE seem to preserve a higher amount of information from the original source sentence, which is often referred to as higher "adequacy".

[1] https://fasttext.cc/docs/en/crawl-vectors.html

Src:	Even in just the past few years , we've greatly expanded our knowledge of how Earth fits within the context of our universe .
Ref:	Rien qu' au cours des dernières années , nous avons beaucoup appris sur la façon dont la Terre s' intègre dans le contexte de notre univers .
Baseline:	Même ces dernières années , nous avons énormément élargi notre connaissance de la manière dont la Terre s' adapte au sein de notre univers .
Baseline+ReWE:	Même ces dernières années , nous avons grandement élargi nos connaissances sur la manière dont la Terre s' adapte dans le contexte de notre univers .
Src:	So , the first example is " a long time ago . "
Ref:	Donc , le premier exemple est " il y a longtemps " .
Baseline:	Le premier exemple est " il y a longtemps . "
Baseline+ReWE:	Donc , le premier exemple est " il y a longtemps . "
Src:	And let me tell you , kids with power tools are awesome and safe .
Ref:	Laissez-moi vous dire que les enfants sont géniaux et prudents avec des outils électriques .
Baseline:	Et laissez moi vous dire , les enfants avec les outils du pouvoir sont stupéfiantes et sûrs .
Baseline+ReWE:	Laissez-moi vous dire que les enfants avec des outils électriques sont stupéfiantes et sûrs .

Table 1: Translation examples from en-fr test set.

Src:	Nikdy totiž na architekturu neexistovala dobrá zpětná vazba .
Ref:	That's because there's never been a good feedback loop in architecture .
Baseline:	You've never had a good feedback in architecture .
Baseline+ReWE:	It's never been a good feedback in architecture .
Src:	Před tisíci lety jste se museli projít do vedlejší vesnice , abyste se na nějakou budovu podívali .
Ref:	A thousand years ago , you would have had to have walked to the village next door to see a building .
Baseline:	A thousand years ago , you had to go to the side of the village to look at some building .
Baseline+ReWE:	A thousand years ago , you had to go to the next village to look at some building .
Src:	V tomto okamžiku se vám uvnitř hlavy promítá film.
Ref:	Right now you have a movie playing inside your head .
Baseline:	And at that point , I 'm going to give you a film inside a film .
Baseline+ReWE:	In this point , you have a film inside the head .

Table 2: Translation examples from cs-en test set.

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 .
Src:	Hautatu Diapositiba aukerak → Pantaila → Erakutsi ataza barra . Aukeratu ireki nahi duzun programa . Sakatu PowerPoint ikonoa aurkezpenera itzultzeko .
Ref:	Select the Slide Options → Screen → Show Taskbar . Choose a program you 'd like to open . Click the PowerPoint icon to return to the presentation .
Baseline:	Select the Slide Options → Display the Show tasbar . Choose the program you want to open . Click the program to return the presentation to the presentation .
Baseline+ReWE:	Select the Slide Options → Display → Show Screen Bar . Choose the program that you want to open . Press PowerPoint icon to return to the presentation .
Src:	Konektatu gailua energia iturri batera . Sakatu Ezarpenak → Orokorra → Software eguneratzea . Sakatu Deskargatu eta instalatu . Sakatu Instalatu deskarga osatzean .
Ref:	Plug in your device to a power source . Tap Settings → General → Software Update . Tap Download and Install . Tap Install when the download completes .
Baseline:	Connect the device to the power . Tap Settings → General → Software update . Tap Download and install . Click Install to download .
Baseline+ReWE:	Connect the device to a power source . Tap Settings → General → Software update . Tap Download and install it . Click Install when completed Download .

Table 3: Translation examples from eu-en test set.

C Constrastive experiments

To gain further insight on the performance of the proposed technique, we have added two contrastive experiments. The first one (Contrastive A) removes ReWE from the architecture, but still retains the combined loss function (Eq. 7 in the paper). Instead of computing the $ReWE_{loss}$ between the ground-truth embedding and the regressed embedding, we compute it between the

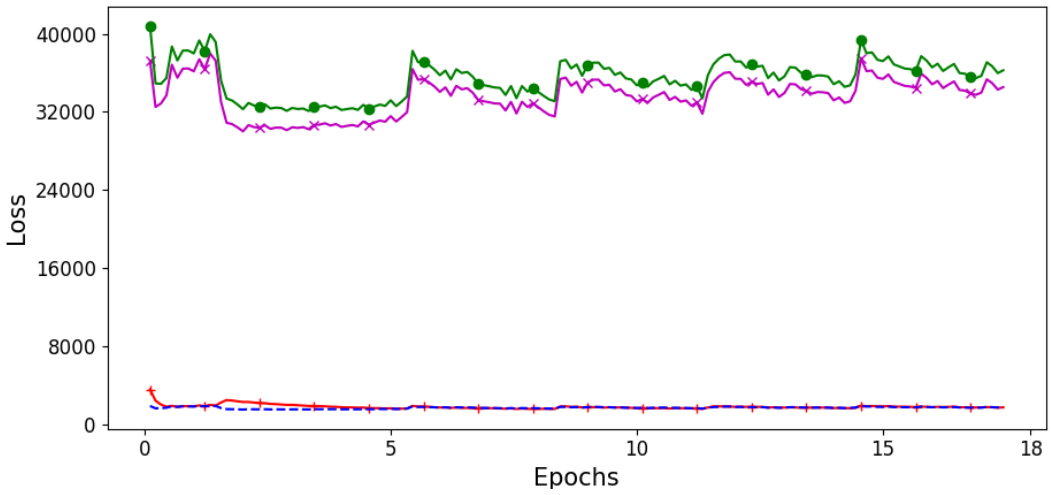

Figure 1: Plot of the values of various loss functions during training of our model over the en-fr training set: green, •: training loss (NLL + ($\lambda = 20$) ReWE (MSE); Eq.7); red, +: NLL loss; blue, dashed: ReWE (MSE) loss; magenta, ×: ReWE (MSE) loss scaled by $\lambda = 20$. Each point in the graph is an average value of the corresponding loss over 25,000 sentences.

Dataset	**BLEU**	
	Word	BPE
en-fr	33.82	33.37
cs-en	20.70	22.53
eu-en	12.15	17.53

Table 4: Results of the Contrastive A experiment ($\lambda = 0.2$; average of 10 models trained independently from different random seeds).

ground-truth embedding and the word embedding of the predicted word, $e(\text{argmax}\, p_j)$. This experiment probes whether the system can leverage the distributional properties of the word embeddings without explicitly predicting them.

The second contrastive experiment (Contrastive B) relies solely on ReWE for both training and inference. Instead of the combined loss function, we only use the $ReWE_{loss}$ for training. At inference time, a search is performed over the embedding space to find the nearest neighbor of the predicted embedding and use it as the predicted word. This experiment aims to explore whether the word embeddings can completely replace the usual categorical prediction.

Table 4 shows the results for the Contrastive A experiment. For this experiment, the value of λ has been specifically tuned over the er-fr validation set (highest score for $\lambda = 0.2$). However, this configuration has rarely improved over our baseline (e.g., on the eu-en dataset), and it has performed considerably worse with the en-fr pair. This shows that, in comparison, the proposed joint learning is a much more effective setting.

In turn, the Contrastive B experiment has achieved much lower BLEU scores. The first experiment over the cs-en dataset reported only 12.71 BLEU points (average of 10 independent runs), approximately half of the other models. Due to this poor result, we have not carried out this experiment further. Our interpretation of this result is that targeting the word embedding is an effective regularizer in the continuous domain, but the conversion of the predicted word embedding to a categorical value is prone to errors from closer neighbors.

D Behaviour of the ReWE (MSE) loss

Figure 1 plots the values of the NLL and ReWE (MSE) losses during training of our model over the en-fr training set. The ReWE (MSE) loss shows large fluctuations as the training progresses, with major increases at the re-starts of the optimizer for the simulated annealing that are not compensated for by the rest of the training.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

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.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

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*.

Minh-Thang Luong, Ilya Sutskever, Quoc V Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics*, pages 11–19.

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*.

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.