Supplementary Material of Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion

Armand Joulin **Piotr Bojanowski** **Tomas Mikolov** **Hervé Jégou** **Edouard Grave**
Facebook AI Research
{ajoulin,bojanowski,tmikolov,rvj,egrave}@fb.com

In this supplementary material, we present additional experimental results. First, we perform an ablation study, to validate the design choices that we made. We then perform additional experiments to check the impact of applying a non-orthogonal mapping on our word vectors. Finally, we report additional results on 28 language pairs from the MUSE benchmark.

1 Ablation study

In this section, we make several experiments to understand the importance of the design choices of our approach as well as the impact of the quality of the embeddings on the alignment.

Impact of the retrieval criterion. Table 1 shows performance on the MUSE benchmark when the CSLS criterion is replaced by the nearest neighbors (NN) criterion. Our approach is still significantly better than Procrustes.

Size of training lexicon. Figure 1 compares the accuracy of our method and Procrustes as a function of the training set size. For small training sets, the difference between our approach and Procrustes is marginal but increases with the training set size.

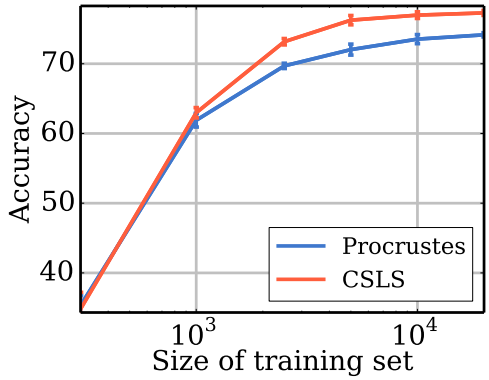

Figure 1: Accuracy as a function of the training set size (log scale) on the en-de pair.

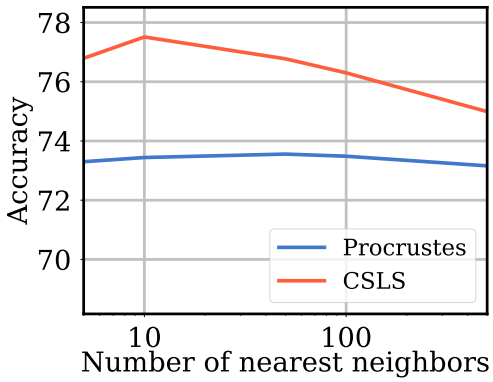

Figure 2: Accuracy as a function of the number of nearest neighbors, averaged over 8 different pairs.

Impact of the number of nearest neighbors. The CSLS criterion and the RCSLS loss are sensible to the number of nearest neighbors. Figure 2 shows the impact of this parameter on both Procrustes and our approach. Procrustes is impacted through the retrieval criterion while our approach is impacted by the loss and the criterion. Taking 10 nearest neighbors is optimal and the performance decreases significantly with a large number of neighbors.

Comparison of alternative criterions. As discussed in the main paper, the dot product in the CSLS terms can be replaced by any convex function of $\mathbf{W}$ and still yield a convex objective. Using a logSumExp function, i.e., $f(x) = \log(\sum_i(\exp(x_i)))$ is equivalent to a "local" logistic regression classifier, or equivalently, to a logistic regression with hard mining. In this experiment, we train our model using the alternative loss and report the accuracy of the resulting lexicon in Table 2. We observe that this choice does not significantly modify the performance. This suggests that the local property of the criterion is more important than the form of the loss.

Method	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en	avg.
Adversarial + refine	79.1	78.1	78.1	78.2	71.3	69.6	37.3	54.3	30.9	21.9	59.9
Procrustes	77.4	77.3	74.9	76.1	68.4	67.7	47.0	58.2	40.6	30.2	61.8
RCSLS	**81.1**	**84.9**	**80.5**	**80.5**	**75.0**	**72.3**	**55.3**	**67.1**	**43.6**	**40.1**	**68.0**

Table 1: Comparison with a nearest neighbor (NN) criterion between RCSLS, Procrustes and the unsupervised approach of Conneau et al. (2017).

	en-es	es-en	en-ru	ru-en
Linear	84.1	86.3	58.0	67.2
logSumExp	84.1	86.3	58.3	67.0

Table 2: Comparison between different functions in CSLS on four language pairs. Linear is the standard criterion, while logSumExp is equivalent to a logistic regression with hard mining.

2 Impact of the input word vectors

	without subword	with subword
en-es	82.8	84.1
es-en	84.1	86.3
en-fr	82.3	83.3
fr-en	82.5	84.1
en-de	78.5	79.1
de-en	74.1	76.3

Table 3: Impact of the quality of the word vectors on the alignment. All the word vectors are trained on the same corpora.

Quality of the embedding model. In this experiment, we study the impact of the quality of the word vectors on the performance of word translation. For this purpose, we trained word vectors on the same Wikipedia data, using skipgram with and without subword information. In Table 3, we report the accuracy for different language pairs when using these two sets of word vectors. Overall, we observe that using subword information improves the accuracy by a few points on all pairs.

Impact of the source of training. Table 4 compares the quality of the alignments as we change the source of training data for the word vectors. Clearly, when the word vectors are trained on a similar data source, like Wikipedia or Common Crawl, we observe similar performance in alignment. The slight drop of performance between Wiki-Wiki and Crawl-Crawl is mostly due to the lower casing of the bilingual lexicon by Conneau et al. (2017). Indeed, the Wikipedia fastText word vectors are trained on a lower cased corpora while the Crawl version is not (Grave et al., 2018), leading to a reduced coverage (shown in brackets). A more interesting observation is that, when we align vectors learned from two different sources, there is a significant drop in performance. This suggests that the alignment is strongly relying on the statistics of the original corpora.

	Wiki-Wiki	CC-Wiki	CC-CC
fr→it	83.5 [100.0]	75.9 [99.6]	82.6 [99.6]
fr→de	76.0 [100.0]	67.9 [98.9]	73.5 [98.9]

Table 4: wiki→CC. On top, CC are the Common Crawl vectors of Grave et al. Wiki are the original fastText vectors. In parenthesis, the coverage of the test set.

3 Impact on word vectors

In the main paper, we study the impact of applying a non-orthogonal mapping on the word vectors, by evaluating them on the word analogy task. For this purpose, we compare the accuracy on the word analogy task of English vectors mapped to various languages with the original vectors. We also evaluate our approach on Cross-lingual word similarities. For all these experiments, we use fastText vectors trained on Wikipedia aligned with the Original MUSE training set.

Impact on English word vectors. We evaluate the impact of a non-orthogonal mapping on the English word analogy task (Mikolov et al., 2013). In Table 5, we report the accuracy on analogies for the raw English word vectors and for vectors mapped to four languages. Regardless of the target language, the mapping does not negatively impact the word vectors. We confirm this finding on the state-of-the-art English word vectors of Mikolov et al. (2018), where aligning to Spanish leads to

	Sem.	Synt.	Tot.
Orig.	79.4	73.4	76.1
en→es	**80.5**	75.8	**78.0**
en→fr	79.8	**75.9**	77.6
en→de	80.0	**75.9**	77.6
en→ru	79.5	74.6	76.8

Table 5: Semantic and syntactic accuracies of original English vectors and mapped English vectors to different languages. On both sides we use the fastText vector of Bojanowski et al. (2017).

	en-es.	en-de	en-it
NASARI baseline	0.64	0.60	0.65
BabylonPartners	0.72	0.69	0.71
MUSE	0.71	0.71	0.71
Ours	0.71	0.71	0.71

Table 6: Cross-lingual word similarity on the NASARI evaluation datasets of Camacho-Collados et al. (2016). We report the Pearson correlation. BabylonPartners, MUSE and Ours uses the same 200k word embeddings from Bojanowski et al. (2017).

an improvement of 1% both for vectors trained on Common Crawl (85% to 86%) and Wikipedia + News (87% to 88%).

Cross-lingual similarity. Finally, we evaluate our method on the task of cross-lingual word similarity in Table 6. We observe that our method obtains similar results to an alignment based on an orthogonal matrix. These experiments concur with the previous observation that a linear non-orthogonal mapping does not hurt the geometry of the word vector space, and even improves it in some cases.

4 Additional results on the MUSE benchmark

Recently, several supervisedly aligned word vectors based on the Wikipedia fastText vectors have been released, noticeably the BabylonPartners (BP)[1] vectors of Smith et al. (2017) and the supervised MUSE vectors of Conneau et al. (2017). In both BP and MUSE, vectors for different languages are aligned in a common space using variations of the Procrustes algorithm. We use our approach with the Frobenius relaxation to align all languages in a common space, using English as an anchor. Table 7 compares the resulting alignments for the 28 languages with a CSLS and NN criterion. We evaluate on the full MUSE test set and a restricted version, where we remove the exact string matches. We note that the gap between our vectors and others is more important with the NN criterion. We also observe that, the performance of all the methods drop when the exact string matches are removed, but the order is roughly the same (but for BP and MUSE). The impaxt of additional training pairs is also reduced, because most of the additional pairs are exact string matches.

[1] https://github.com/Babylonpartners/fastText_multilingual

		Original		Full	
	BP	MUSE	RCSLS	Proc.	RCSLS
with exact string matches					
NN	54.6	57.4	62.4	57.5	**68.5**
CSLS	60.8	63.8	67.4	65.2	**70.2**
without exact string matches					
NN	56.6	55.5	61.4	53.7	**64.3**
CSLS	61.5	60.4	65.4	60.2	**65.7**

Table 7: Comparison with publicly available aligned vectors, averaged over 28 language pairs. All use supervision. Alignements are learned either on the "Original" or "Full" MUSE training. We report performance with the NN and CSLS criterion on either the full MUSE test set or without the exact string matches. BP uses a different training set with 5k words.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. https://fasttext.cc/docs/en/pretrained-vectors.html.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*. http://github.com/facebookresearch/MUSE.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learn-

ing word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*. https://github.com/Babylonpartners/fastText_multilingual.