Appendix for Unsupervised Cross-lingual Transfer of Word Embedding Spaces

Ruochen Xu, Yiming Yang, Naoki Otani, Yuexin Wu
Carnegie Mellon University
{ruochenx, yiming, notani, yuexinw }@cs.cmu.edu

A Implementation Detail

The input embeddings X and Y are prepossessed to have zero mean and unit variance on each dimension.

In our preliminary experiments, we found that tying the weight matrix of F and G to be the transpose of each other improve the performance. So we keep this configuration in all our experiments. This constraint is well-motivated since $F(G(X))$ could be viewed as an autoencoder where $G(\cdot)$ is the encoder and $F(\cdot)$ is the decoder. And it is a common practice to tie the weights of encoder and decoder in an autoencoder.

We minimized the full objective (6) on mini-batches of size 2048 using RMSprop optimizer (Tieleman and Hinton, 2012) at a learning rate of 0.0005. We run WGAN training for the first 2000 epochs and switched to the Sinkhorn objective. We used a learning rate decay of 0.95 if objective fails to decrease in each epoch and early stopped training if the objective stopped to decrease for 2000 epochs.

For input embeddings with large vocabulary size, we found it is not necessary and sometimes harmful to include all the words into our training procedure. Therefore we input the $10,000$ most frequent words in WE-C for each language in our experiments. For smaller WE-Z, we simply use all available given embeddings. As for word frequencies, we simply assume a uniform distribution of words, i.e. $r = \mathbb{1}_n/n$, $c = \mathbb{1}_m/m$. We also tried using true word frequencies from the corpus in LEX-Z and task 1[1], but no significant performance improvement was observed.

B Ablation Study

We verify the necessity of our system choice by an ablation study. Our first ablated model changes the symmetric objective function to be one-sided. More specifically we let the full objective to be:

$$d_{sh}(G) + \sum_i 1 - cos(x_i, F(G(x_i)))$$

The similar change is also applied to initial point searching process with adversarial training. We refer this model as OneSided.

Our second ablated model uses only the WGAN training procedure described in subsection 3.4. We found it difficult to find stopping criterion for WGAN based on its own objective, therefore we used our full objective (6) to select model during training. We refer this model as WGAN.

The results with LEX-C are shown in figure 2 and figure 1 for the task 1 and task 2 respectively. We can see that WGAN always produces the worst cross-lingual transformation. Although the adversarial training is good at search the parameter space, it is not good at converging to the well-performing local optimal. OneSided is better than WGAN for most of times, but failed to find reasonable initial point sometimes("fa-en" in figure 1 and "sv-en", "en-sv" in figure 2). Overall, our model with full objective achieves the best performance on all tasks compared with ablated models. The only exception is "ca-en" for LEX-C. These results justified our system choice described in our main paper.

C Error Analysis

To gain the insights to further improve our method, we conducted error analysis for en-es and en-bg translation on LEX-C, where our model achieved the higher or on par performance with supervised methods. Although we only used 1,500 query words in our experiments, we consider the 10k most frequent English words[2] as query words to

[1] LEX-C does not contain word frequency information

[2] Released by Conneau et al. (2017)

	Category	Gold	Prediction
1	Grammatical conjugation (es)	permitir [permit/allow]	permitió (preterite form in the third person singular)
	Articles (bg)	интервю [interview]	интервюто [the interview]
2	Numbers	шестдесет [sixty]	десет [ten]
	Singular/plural	общност [community]	общностите [communities]
	Antonym	inicia [to start]	termina [to end]
3	Person	Reyes [Reyes]	Hernández [Hernandez]
	City	дарбишър [Derbyshire]	пазарджик [Pazardzhik]

Table 1: Typical errors on the en-es and en-bg translation tasks.

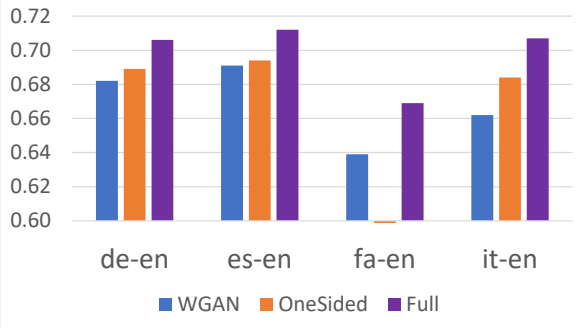

Figure 1: Pearson correlation for cross-lingual semantic word similarity task for ablation study

better understand the errors the model tend to make. We randomly sampled 100 failure cases from each language pair and analyzed each case. Our key observations are three-fold (Table 1):

1. The evaluation results for both the language pairs contain not a few false negatives (14 for en-es and 23 for en-bg), namely correct translations but not included in LEX-C. Such cases for en-es mostly came from the different forms of verbs because Spanish verbs are rich in grammatical conjugation. The false negatives for en-bg are typically nouns because articles are postfixed to nouns but LEX-C does not cover such forms sufficiently.

2. The model often confuses similar- or opposite-meaning entities because their word embeddings are often very similar. This is a well-known problem of word embeddings in general.

3. The evaluation set contains some proper nouns such as person names and city names, and they are typically very difficult to generate exact translations. However, translating proper nouns is not necessarily important in downstream cross-lingual tasks, and we could ignore them when we evaluate the cross-lingual transfer of word embeddings.

References

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

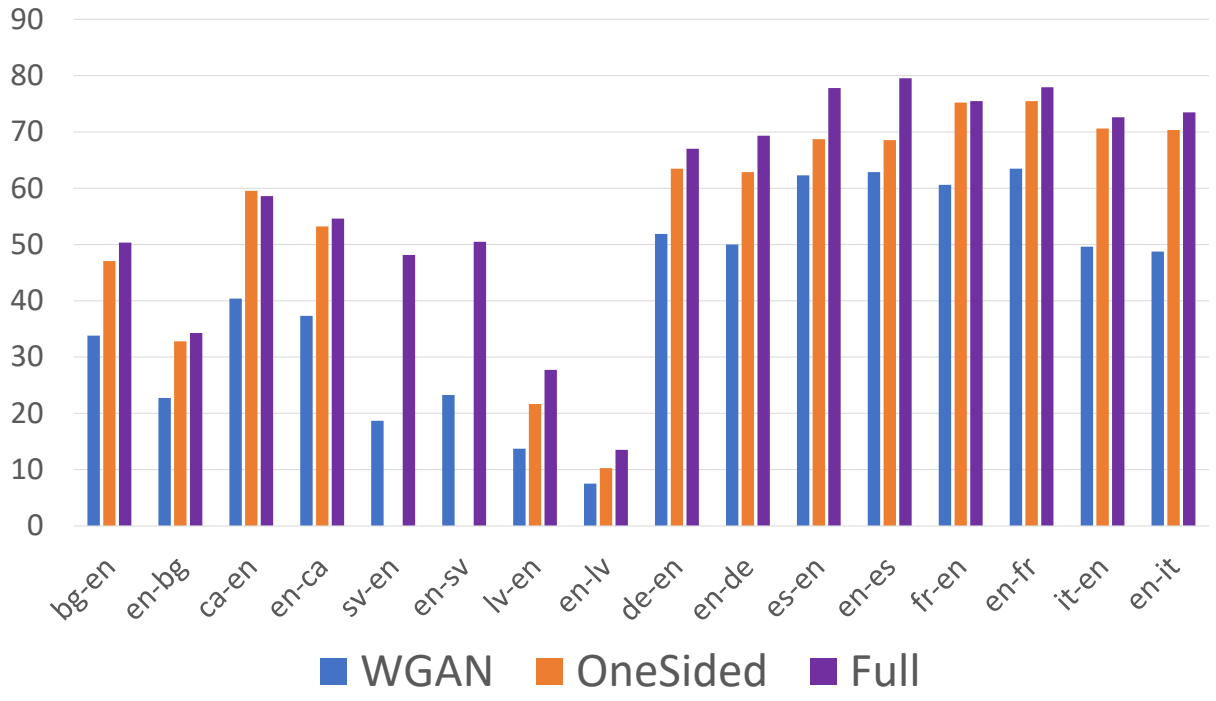

Figure 2: Bilingual lexicon induction accuracy(in %) on LEX-C for ablation study