Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 892–899
Marseille, 20-25 June 2022

Domain Mismatch Doesn't Always Prevent Cross-Lingual Transfer Learning

Daniel Edmiston,[1] **Phillip Keung,**[1] **Noah A. Smith**[2,3]
[1]Amazon, [2]University of Washington, [3]Allen Institute for Artificial Intelligence
danedm@amazon.com, keung@amazon.com, nasmith@cs.washington.edu

Abstract

Cross-lingual transfer learning without labeled target language data or parallel text has been surprisingly effective in zero-shot cross-lingual classification, question answering, unsupervised machine translation, etc. However, some recent publications have claimed that domain mismatch prevents cross-lingual transfer, and their results show that unsupervised bilingual lexicon induction (UBLI) and unsupervised neural machine translation (UNMT) do not work well when the underlying monolingual corpora come from different domains (e.g., French text from Wikipedia but English text from UN proceedings). In this work, we show that a simple initialization regimen can overcome much of the effect of domain mismatch in cross-lingual transfer. We pre-train word and contextual embeddings on the concatenated domain-mismatched corpora, and use these as initializations for three tasks: MUSE UBLI, UN Parallel UNMT, and the SemEval 2017 cross-lingual word similarity task. In all cases, our results challenge the conclusions of prior work by showing that proper initialization can recover a large portion of the losses incurred by domain mismatch.

Keywords: Domain mismatch, cross-lingual transfer, transfer learning, machine translation

1. Introduction

Zero-shot cross-lingual transfer via representation learning has been studied in many recent works spanning a variety of tasks: cross-lingual text classification and named entity recognition (Devlin et al., 2018), unsupervised neural machine translation (Lample et al., 2018a; Artetxe et al., 2018b) and unsupervised bilingual lexicon induction (Conneau et al., 2018; Zhang et al., 2017), among others. Cross-lingual transfer techniques typically assume that the source and target text come from the same domain (e.g., English and French Wikipedia for UNMT), but many recent papers have reported issues in the domain-mismatched setting (e.g., English Europarl and French Wikipedia).

Particularly, domain mismatch has been shown to have a pernicious effect on UBLI, and has been labeled a "core limitation" (Søgaard et al., 2018), with word embeddings pre-trained on domain-mismatched corpora showing markedly degraded scores. In the case of UNMT, mismatched domains between source and target training data have also been shown to cause large reductions in BLEU scores (Marchisio et al., 2020). The results in Table 1 illustrate the severity of the problem.

In this work, we show that cross-lingual transfer can occur *even when there is no overlap between the domains in the same language and no overlap between the languages in the same domain* (Figure 1). Earlier work such as mBERT (Devlin et al., 2018) and XLM (Lample and Conneau, 2019) demonstrated that pre-training contextual embeddings on concatenated multilingual Wikipedia text induces cross-lingual transfer effects. We extend these findings to the domain-mismatched case, where we pre-train our embeddings on concatenated multilingual domain-mismatched text. We compare the effect of initializing with and without joint pre-training for three cross-lingual tasks: MUSE BLI (Conneau et al., 2018), UN Parallel MT (Ziemski et al., 2016), and SemEval 17 cross-lingual word similarity (Camacho-Collados et al., 2017). Contrary to the findings on UBLI and UNMT from recent publications, we find that the availability of domain-matched corpora is not a prerequisite for effective cross-lingual transfer, since the domain mismatch issue can be mitigated by using an appropriate initialization.

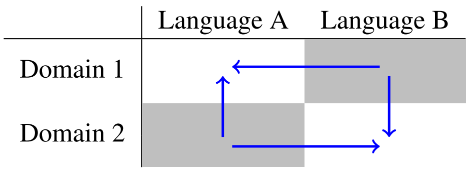

Figure 1: In our zero-shot experiments, training corpora for each language belong to different domains (grey cells). We show joint pre-training induces cross-domain, cross-lingual transfer (arrows) even when no data exists in the double-crossed scenario (white cells).

2. Unsupervised BLI Experiments

2.1. Background

Bilingual lexicon induction refers to a word translation task, with modern methods relying on retrieval in a continuous space shared by both source and target embeddings. BLI has successfully used small seed dictionaries as a form of cross-lingual signal (Mikolov et al., 2013; Duong et al., 2016), but recent unsupervised alternatives have proven competitive with supervised approaches (Artetxe et al., 2018a; Heyman et al., 2019),

	Language Pair	Source/Target Domain	Score	△ Mismatch vs. Match
UBLI			*Accuracy@1*	
Søgaard et al. (2018)	English-Spanish	Europarl/Europarl	61.0	
		Europarl/Wikipedia	0.1	–60.9
	English-Hungarian	Wikipedia/Wikipedia	6.7	
		Wikipedia/Europarl	0.1	–6.6
UNMT			*BLEU*	
Marchisio et al. (2020)	French-English	UN Parallel/UN Parallel	27.6	
		UN Parallel/Common Crawl	3.3	–24.3
	Russian-English	UN Parallel/UN Parallel	23.7	
		UN Parallel/Common Crawl	0.7	–23.0

Table 1: Unsupervised bilingual lexicon induction (UBLI) and neural machine translation (UNMT) results from some previous papers. When the domains of the monolingual text are mismatched, UBLI and UNMT yield retrieval accuracy and BLEU scores very close to 0.

with methods based on adversarial learning (Zhang et al., 2017; Conneau et al., 2018) and point-cloud matching (Hoshen and Wolf, 2018).

We use the MUSE model of Conneau et al. (2018) for all UBLI experiments, with all experiments conducted on AWS p3.16xlarge hosts. MUSE uses an adversarial objective (Goodfellow et al., 2014) to learn a transformation from the word embedding space of the source language to that of the target language, along with a discriminator to distinguish transformed source embeddings from target embeddings. Word translation is achieved using margin-based nearest-neighbors to retrieve target embeddings from transformed source embeddings, and evaluated using the test sets provided with MUSE.

In UBLI with MUSE, word embeddings for the source and target languages are pre-trained independently (e.g., French and English word embeddings are trained separately on Wikipedia text). When the domains are mismatched (e.g., Wikipedia and UN), UBLI retrieval accuracy has been shown to suffer greatly (Søgaard et al., 2018), as mentioned above.

We compare the standard way of pre-training for MUSE UBLI—pre-training word embeddings separately for each language—with joint pre-training, where we train multilingual word embeddings on concatenated domain-mismatched corpora (Lample et al., 2018a). Note that this is a simple form of multilingual joint pre-training, and does not include any *post-hoc* processing steps (cf. Wang et al. (2020), who perform a vocabulary reallocation step to eliminate spurious anchors in the shared embedding space[1]).

We study the following language pairs in both directions: English-French, English-Spanish, and English-Russian.[2] For each experiment, we initialize embeddings via *fastText* (Bojanowski et al., 2016) using Wiki and UN corpora[3] and perform grid search over MUSE hyperparameters, reporting scores for the configuration with the highest CSLS (cross-domain similarity local scaling) score, which is an unsupervised metric discussed in Conneau et al. (2018).[4] We optimize CSLS score for four random seeds (123, 456, 789, and 321), three choices of iterations for Procrustes refinement (1, 3, and 5), and three choices of epochs (1, 3, and 5).

2.2. Results

We present our retrieval accuracies at 1 for UBLI experiments in Table 2. As expected, MUSE works well with domain-matched corpora, while our domain-mismatched experiments show large degradations relative to matched domain baselines in all cases. In particular, scores for Es-En, En-Es, and En-Ru all fall to near 0.0, showing that cross-lingual transfer has failed in these cases. However, in all cases, joint pre-training recovers a large portion of the losses incurred by mismatched corpora, showing cross-lingual transfer is still possible, contrary to the conclusions drawn in Søgaard et al. (2018) and Vulić et al. (2019).

2.3. The Role of Identical Words in UBLI Performance

It is important to note that a large proportion of word pairs in the MUSE test dictionaries are identical (e.g., *Paris-Paris* in Fr-En), and joint-training is able to take advantage of identical spellings, since words with the

[1] Spurious anchors are embeddings in the shared space which result from forms which appear in both languages, but which have different meanings in each language, e.g. *coin* in English and French. In French *coin* means *corner*, and therefore *coin* should not map to a single vector in the shared embedding space.

[2] We lemmatize all Russian data for the UBLI experiments with the *pymorphy2* (Korobov, 2015) morphological analyzer, abstracting out challenges posed by morphologically rich languages like Russian.

[3] We used Wiki dumps from June 2020 and UN corpus v1.0 (Ziemski et al., 2016). To address the disparity in corpus sizes, we sampled 5M lines from each for training. Tokenization was done with Moses (Koehn et al., 2007).

[4] The unsupervised CSLS scores are computed using only the training corpora.

Source Domain-Target Domain	Es-En	En-Es	Fr-En	En-Fr	Ru*-En	En-Ru*
Matched Domain						
Wiki - Wiki	81.8	82.5	81.3	82.2	59.5	64.0
UN - UN	68.7	70.8	74.2	75.2	55.2	56.7
Mismatched Domain						
Wiki - UN	0.1	0.2	35.2	33.8	16.6	0.1
Wiki - UN w/ Joint Pre-training	65.2	56.9	68.3	54.4	28.0	20.1
Δ	+65.1	+56.7	+33.1	+20.6	+11.4	+20.0

Table 2: Retrieval accuracy@1 for UBLI across language pairs on MUSE test dictionaries. *Joint pre-training* uses word embeddings that are jointly pre-trained on concatenated source Wiki and target UN corpora. Ru* denotes lemmatized Russian.

same spelling will always have the same word embedding (Lample et al., 2018a). In Table 3, we show the performance of the 'copying baseline', which simply treats each word as its own translation. This baseline is surprisingly strong; its accuracy at 1 exceeds 40% for English-French in both directions.

Søgaard et al. (2018) explicitly use identical words to create a seed dictionary to improve performance in the cross-domain scenario. While Søgaard et al. (2018)'s best reported score for En-Es with the seed dictionary approach actually falls below the simple copying baseline at 25.5%, we show that the joint pre-training approach yielded 56.9%. A comparison between results from the two approaches and the copying baseline are shown in Table 3.

2.4. The Source-Target Domain Mismatch (STDM) Score

The considerable improvement of joint-training over the unsupervised seed dictionary method as detailed in Table 3 could be the result of the relative distances between source and target domains rather than of the differing techniques. For instance, if the Wikipedia and UN corpora were more similar than the Wikipedia and Europarl corpora, domain mismatch would be more pronounced in the Wikipedia-Europarl case and lower scores would be expected. The recently proposed Source-Target Domain Mismatch (STDM) score of Shen et al. (2021) provides a means of measuring domain similarity between corpora, and we show that the Europarl and UN corpora are not dramatically different from the Wikipedia corpora in all cases. This suggests that the disparity between the results is attributable to the different methods rather than the relative similarity between domains.

The STDM score is computed in the following way. Let $\mathcal{A} = concat(A_1, A_2)$, or the concatenation of two corpora A_1 and A_2, where A_1 and A_2 are the corpora we wish to compare and which consist of n and m documents respectively. Then let $\mathcal{A}_{tfidf} \in \mathbb{R}^{(n+m)\times|V|}$ be the result of applying TF-IDF (Jones, 1972) to the concatenated corpora, thus giving a matrix whose top n rows are representations of the documents from A_1 and the bottom m rows are representations of the documents from A_2. Then the (truncated) SVD decomposition of $\mathcal{A}_{tfidf}$ is $\bar{\mathcal{A}}_{tfidf} = USV = (U\sqrt{S})(\sqrt{S}V) = \bar{U}\bar{V}$, where $\bar{U}$ contains topic representations of the corpora's documents, again with the first n rows ($\bar{U}_1$) being the representations of corpus A_1 and the bottom m rows ($\bar{U}_2$) the representations of corpus A_2. Then define $s_{A,B}$ as in Equation 1:

$$s_{A,B} = \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{j=1}^{m} (\bar{U}_A \bar{U}_B^\top)_{i,j} \tag{1}$$

$s_{A,B}$ then measures the average similarity of documents between corpus A and corpus B. Given $s_{A,B}$ for each combination of corpora, then the STDM score is defined as in Equation 2.

$$\mathit{STDM} = \frac{s_{1,2} + s_{2,1}}{s_{1,1} + s_{2,2}} \tag{2}$$

The STDM score, which in practice ranges from 0 (completely dissimilar) to 1 (identical), thus uses latent semantic analysis (LSA) (Dumais et al., 1988) on a combined corpus to derive topic representations of the individual corpora, using the intuition that similar corpora will have similar documents. We use this score to quantify the similarity between mismatched corpora and rule out relative domain divergence as a causal factor in the disparity between our scores and those of Søgaard et al. (2018).[5]

The STDM scores for the mismatched corpora for each experiment are found in the STDM column of Table 3. The STDM comparisons between corpora show that, while on average the Wikipedia and UN corpora are more similar than the Wikipedia and Europarl corpora, this difference is small and unlikely to account for the

[5]Note that a corpus similarity score based on TF-IDF cannot compare corpora from different languages directly. Shen et al. (2021) work around this issue by comparing the corpora in the target language. For example, to quantify the domain mismatch in an English-Wikipedia to Spanish-Europarl experiment, the corpora of comparison would be the Spanish Wikipedia and the Spanish Europarl, since Spanish and English Wikipedia would cover similar topics.

	Language Pair	Src/Tgt Domain	Acc@1	Copying Baseline	Δ	STDM
Søgaard et al. (2018)	En-Es	Wiki/Europarl	25.5	32.8	-7.3	0.27
	En-Fi		10.1	28.8	-18.7	0.20
	En-Hu		9.2	29.6	-20.4	0.20
Joint pre-training (this paper)	En-Es	Wiki/UN	56.9	32.8	+24.1	0.25
	Es-En		65.2	29.7	+35.5	0.33
	En-Fr		54.4	41.3	+13.1	0.28
	Fr-En		68.3	42.9	+25.4	0.33
	En-Ru		20.1	3.9	+16.2	0.25
	Ru-En		28.0	0.0	+28.0	0.33

Table 3: Accuracies based on the method of Søgaard et al. (2018) and joint pre-training versus a simple copying baseline. Δ refers to difference between copying baseline and Acc@1 score for each method. STDM refers to the Source-Target Domain Mismatch score for each corpus and language pair, as described in Shen et al. (2021).

large disparity in results. On the one point of direct comparison, namely UBLI from English into Spanish, the Wikipedia and Europarl corpora are shown to be more similar (STDM=0.27) than the Wikipedia and UN corpora (STDM=0.25), yet joint-training on the more dissimilar corpora still produces better results (56.9% vs. 25.5%).

2.5. On Domain Mismatch in UBLI

We emphasize three points from the results of these experiments. Firstly, in many cases, initialization by a simple joint-training regimen can largely overcome the deleterious effects of domain mismatch for the UBLI task, challenging the conclusions of Søgaard et al. (2018) and Vulić et al. (2019). This is seen primarily in experiments involving closely related languages (English-Spanish, English-French), where mismatched domain experiments run with joint-training initialization approach the scores of the matched domain experiments.

Secondly, while still beneficial, joint-training initialization is less effective on distantly related languages, such as English-Russian. Improvements from joint-training are considerable, but scores still fall below the matched domain experiments, suggesting that this method does not fully solve domain mismatch in all cases.

Lastly, task-agnostic joint-training initialization performs favorably when compared against the identical-word seed dictionary method of Søgaard et al. (2018) in terms of ameliorating the effects of domain mismatch, as shown by comparison of each method against a copying baseline.

3. Unsupervised NMT Experiments

3.1. Task Description

Unsupervised NMT systems address the paucity of available parallel data for most language pairs, relying only on monolingual data from the source and target languages. The models of Lample et al. (2018a) and Artetxe et al. (2018b) are representative, each employing encoder-decoder architectures with weight-sharing between languages. Models are trained via the dual tasks of sentence reconstruction and back-translation (Sennrich et al., 2016).[6] Follow-up work has incorporated statistical machine translation (SMT) systems (Koehn et al., 2003); Artetxe et al. (2019) and Marie and Fujita (2018) use unsupervised SMT systems to initialize UNMT systems, while Ren et al. (2019) incorporate SMT as a form of posterior regularization.

For all UNMT experiments, we adopt the encoder-decoder model of Lample et al. (2018b), a sequence-to-sequence model with 6 transformer layers for both the encoder and decoder, and use the implementation provided by the authors.[7] We study the English-French and English-Russian language pairs in both directions, training all models for ten epochs. We used 5 million sentences per language in the monolingual data used for UNMT training. We trained two English-French and two English-Russian UNMT models on the following sets of 10 million sentences: (En Wiki, Fr UN), (En UN, Fr Wiki), (En Wiki, Ru UN), and (En UN, Ru Wiki).

In a manner similar to our UBLI experiments, we compare UNMT performance with and without jointly pre-trained contextual embeddings. In the baseline system, we follow the UNMT approach outlined in Lample and Conneau (2019), where the encoder and decoder are initialized with a contextual embedding pre-trained on Wikipedia text only. In the jointly pre-trained case, the encoder and decoder are initialized with a contextual embedding that was pre-trained on a mix of Wikipedia and UN text.

3.2. Results

In Tables 4 and 5, we show the difference in the UN development BLEU scores between a UNMT system with and without joint pre-training. Table 4 shows results for experiments in which the monolingual data contains UN text in the target language, while Table 5 shows results for experiments in which it does not. We

[6] See Wu et al. (2019) and Keung et al. (2020) for alternatives to back-translation.

[7] https://github.com/facebookresearch/XLM

Monolingual Data	Task		Epoch 1	2	3	4	5	6	7	8	9	10
En Wiki, Fr UN	En UN → Fr UN	UNMT Baseline	3.37	3.65	4.01	4.40	4.30	4.22	4.17	3.73	4.16	3.99
		w/ Joint Pre-train.	22.76	24.61	24.83	25.19	24.98	25.09	24.95	24.95	24.70	**25.45**
		Δ	+19.39	+20.96	+20.82	+20.79	+20.68	+20.87	+20.78	+21.22	+20.54	**+21.46**
Fr Wiki, En UN	Fr UN → En UN	UNMT Baseline	4.11	4.01	3.99	4.17	4.35	4.41	4.65	4.87	5.21	5.39
		w/ Joint Pre-train.	19.78	20.20	**20.91**	19.94	19.62	19.86	19.68	19.62	20.54	19.80
		Δ	+15.67	+16.19	**+16.92**	+15.77	+15.27	+15.45	+15.03	+14.75	+15.33	+14.41
En Wiki, Ru UN	En UN → Ru UN	UNMT Baseline	1.50	1.04	1.45	1.46	1.59	1.35	1.29	1.32	1.37	1.37
		w/ Joint Pre-train.	**10.83**	9.14	9.62	8.96	8.82	9.79	10.09	10.37	9.82	9.71
		Δ	**+9.33**	+8.10	+8.17	+7.50	+7.23	+8.44	+8.80	+9.05	+8.45	+8.34
Ru Wiki, En UN	Ru UN → En UN	UNMT Baseline	1.51	1.27	1.52	1.53	1.31	1.22	1.40	1.28	1.23	1.22
		w/ Joint Pre-train.	7.71	**7.95**	7.82	7.43	7.46	7.49	6.45	6.85	6.60	6.77
		Δ	+6.20	**+6.68**	+6.30	+5.90	+6.15	+6.27	+5.05	+5.57	+5.37	+5.55

Table 4: UNMT BLEU scores with and without joint pre-training on the UN Corpus development sets. (Highest BLEU scores in each row are bolded.) Initializing with jointly pre-trained contextual embeddings yields up to a 21 point gain in BLEU. BLEU score improvements are always larger when the monolingual pre-training data contains target language text in the UN domain.

report results for the first 10 epochs (where one epoch consists of 200k sentences), and the tables show BLEU scores as training progresses.

As we saw in previous work with domain mismatch (Table 1), the baseline system without joint pre-training fails to learn to translate; BLEU scores never exceed 6 for any language pair.[8] However, with joint pre-training, we observe as much as 21 points of BLEU improvement. An additional 10 epochs of training (not shown) do not yield a BLEU improvement for the baseline.

The difference between results in Tables 4 and 5 shows that UNMT performance depends strongly on whether the monolingual data contains UN text in the target language. For example, Fr UN → En UN BLEU scores are higher when the monolingual data is (Fr WIKI, En UN) and lower when it is (Fr UN, En Wiki).

Furthermore, while we did not perform matched domain UNMT experiments for comparison due to computational constraints, we can take the French-English and Russian-English experiments from Marchisio et al. (2020) (see again Table 1) to establish the extent to which joint-training ameliorates domain mismatch. Marchisio et al. (2020) report a BLEU score of 27.6 for the Fr UN - En UN experiment, and a BLEU score of 23.7 for the Ru UN - En UN experiment. In our best domain-mismatched experiment for the Fr UN - En UN task, joint-training resulted in a BLEU score of 20.91, making up most of the losses incurred by domain mismatch. However, in our best domain-mismatched experiment for the Ru UN - En UN task, the BLEU score was only 7.95, which is well below the domain matched score of 23.7 reported in Marchisio et al. (2020). Thus, similar to the UBLI experiments above, joint-training results in a pronounced recovery of cross-lingual transfer when the language pair is similar, but results in only modest gains for more distant language pairs.

[8]Lample et al. (2018b) show that in a domain-matched Fr-En setting, the baseline UNMT system can achieve > 24 BLEU on WMT'14.

4. Cross-lingual Semantic Word Similarity Experiments

4.1. Task Description

In addition to investigating the UBLI and UNMT tasks, we also examine cross-lingual transfer via word similarity tasks, and in doing so show that joint-training via concatenation is useful generally, even for non-translation related tasks. The semantic word similarity task consists of evaluating pairs of words (e.g. WS353) (Finkelstein et al., 2001) via a similarity metric (e.g., cosine similarity) on their embeddings, and comparing these scores with human judgments. The SemEval 17 cross-lingual semantic word similarity task (Camacho-Collados et al., 2017) evaluates pairs of words from different languages for similarities[9] in their underlying meaning on a scale from 0–4, with a step size of 0.5.

Given semantic similarity predictions for a list of word pairs constructed per Camacho-Collados et al. (2015), performance is measured as the harmonic mean of Pearson and Spearman correlations with human judgments. Datasets were constructed from English, Farsi, German, Italian, and Spanish. As in the UBLI experiments, we train *fastText* embeddings for each language pair using domain-matched (Wiki-Wiki) and domain-mismatched (UN-Wiki) corpora, and compare separate and joint pre-training. Finally, semantic similarity is computed by cosine similarity.

4.2. Results

Results on the cross-lingual semantic word similarity experiments are shown in Table 6. The Wiki-UN correlation scores are very low without joint pre-training, but the correlation scores improve greatly with joint pre-training and approach scores obtained using matched Wiki-Wiki data, even for a distant language pair like English-Farsi.

[9]The task was designed to distinguish similarity from relatedness (Hill et al., 2015). An example of the distinction in English-German would be, e.g. *dog-Hund*, which are very similar, whereas *leaf-Baum* would be more dissimilar yet still related.

Monolingual Data	Task		Epoch 1	2	3	4	5	6	7	8	9	10
En UN, Fr Wiki	En UN → Fr UN	UNMT Baseline	3.94	3.84	3.75	3.77	3.89	3.87	3.87	4.12	4.21	4.40
		w/ Joint Pre-train.	11.78	12.50	**12.76**	11.73	11.90	11.65	11.20	11.24	11.62	11.59
		Δ	+7.84	+8.66	**+9.01**	+7.96	+8.01	+7.78	+7.33	+7.12	+7.41	+7.19
Fr UN, En Wiki	Fr UN → En UN	UNMT Baseline	2.66	3.79	3.70	4.32	4.11	4.33	4.11	4.14	4.18	4.04
		w/ Joint Pre-train.	13.48	13.82	**14.19**	13.40	13.70	13.39	13.34	13.38	12.74	13.65
		Δ	+10.82	+10.03	**+10.49**	+9.08	+9.59	+9.06	+9.23	+9.24	+8.56	+9.61
En UN, Ru Wiki	En UN → Ru UN	UNMT Baseline	1.03	1.09	1.19	0.89	0.79	0.98	0.79	0.79	0.80	0.81
		w/ Joint Pre-train.	4.47	4.41	4.44	4.40	**4.52**	4.38	4.10	4.36	4.11	3.99
		Δ	+3.44	+3.32	+3.25	+3.51	**+3.73**	+3.40	+3.31	+3.57	+3.31	+3.18
Ru UN, En Wiki	Ru UN → En UN	UNMT Baseline	1.24	0.97	0.99	1.29	1.49	1.47	1.41	1.64	1.57	1.42
		w/ Joint Pre-train.	**7.63**	7.57	6.93	6.98	7.31	6.89	6.70	7.11	7.02	6.56
		Δ	**+6.39**	+6.60	+5.94	+5.69	+5.82	+5.42	+5.29	+5.47	+5.45	+5.14

Table 5: UNMT BLEU scores when the monolingual data *doesn't* contain target language UN text, in contrast to Table 4.

	En-De	En-Es	En-Fa	En-It
Wiki-Wiki	0.45	0.54	0.25	0.49
UN-Wiki w/o Joint PT	0.02	-0.01	0.01	-0.05
UN-Wiki w/ Joint PT	0.43	0.47	0.23	0.45
Δ	+0.41	+0.48	+0.22	+0.50

Table 6: Correlation scores for cross-lingual word similarity (SemEval 2017 Task 2, Subtask 2). Δ refers to the difference between the domain-mismatched scores with and without joint pre-training.

Our experiments on cross-lingual semantic word similarity show that while domain mismatch is a significant obstacle to cross-lingual transfer, joint-training initialization is an effective means of overcoming this issue. Furthermore, unlike the results of the UBLI and UNMT experiments in which joint-training was less effective for the relatively distant language pair of English-Russian in recovering losses compared to matched domain baselines, here joint-training results in virtually identical scores between matched domain and mismatched domain for the distant language pair of English-Farsi.

5. Conclusion

Recent publications on UBLI and UNMT have noted that domain mismatch hinders zero-shot cross-lingual transfer (Braune et al., 2018; Tae et al., 2020; Kim et al., 2020). Our work shows that initialization via joint pre-training can reduce the impact of this mismatch, even when that pre-training doesn't involve post-processing steps such as vocabulary reallocation (Wang et al., 2020). Furthermore, the improvements brought about by this initialization scheme generalize to other cross-lingual tasks such as cross-lingual word similarity. Our results show as much as a 65% absolute increase in UBLI retrieval accuracy, up to a 21 point gain in UNMT BLEU scores, and as much as a 0.5 improvement on word similarity correlation under domain mismatch. It is well-known that pre-training contextual embeddings on unaligned multilingual corpora induces zero-shot cross-lingual transfer learning (e.g., mBERT and XLM), but our work shows that joint pre-training can also induce simultaneous zero-shot cross-domain, cross-lingual transfer, which we expect will be useful guidance for NLP practitioners.

While the results reported here are encouraging, future work should include experimentation with a wider assortment of language pairs and corpus domains, as well as an investigation of how the distance between languages can affect joint-training's ability to mitigate domain mismatch for different tasks. While in all three tasks it proved very effective for closely related language pairs, the UBLI and UNMT improvements for the more distant language pair of English-Russian were less pronounced. Conversely, on the cross-lingual semantic word similarity experiment, language distance seemed less relevant as joint-training on English-Farsi resulted in scores comparable to the domain matched scenario.

6. Bibliographical References

Artetxe, M., Labaka, G., and Agirre, E. (2018a). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, July. Association for Computational Linguistics.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018b). Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Artetxe, M., Labaka, G., and Agirre, E. (2019). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy, July. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Braune, F., Hangya, V., Eder, T., and Fraser, A. (2018). Evaluating bilingual word embeddings on the long tail. In *Proceedings of the 2018 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 188–193, New Orleans, Louisiana, June. Association for Computational Linguistics.

Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2015). A framework for the construction of monolingual and cross-lingual word similarity datasets. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)*, pages 1–7.

Camacho-Collados, J., Pilehvar, M. T., Collier, N., and Navigli, R. (2017). Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2018). Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285.

Duong, L., Kanayama, H., Ma, T., Bird, S., and Cohn, T. (2016). Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295, Austin, Texas, November. Association for Computational Linguistics.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, et al., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Heyman, G., Verreet, B., Vulić, I., and Moens, M.-F. (2019). Learning unsupervised multilingual word embeddings with incremental multilingual hubs. 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 1890–1902, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hill, F., Reichart, R., and Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December.

Hoshen, Y. and Wolf, L. (2018). Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478, Brussels, Belgium, October-November. Association for Computational Linguistics.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Keung, P., Salazar, J., Lu, Y., and Smith, N. A. (2020). Unsupervised bitext mining and translation via self-trained contextual embeddings. *arXiv preprint arXiv:2010.07761*.

Kim, Y., Graça, M., and Ney, H. (2020). When and why is unsupervised neural machine translation useless? *arXiv preprint arXiv:2004.10581*.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Korobov, M. (2015). Morphological analyzer and generator for russian and ukrainian languages. *CoRR*, abs/1503.07283.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018a). Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (2018b). Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Marchisio, K., Duh, K., and Koehn, P. (2020). When does unsupervised machine translation work? *arXiv preprint arXiv:2004.05516*.

Marie, B. and Fujita, A. (2018). Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *CoRR*, abs/1810.12703.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Ren, S., Zhang, Z., Liu, S., Zhou, M., and Ma, S. (2019). Unsupervised neural machine translation with SMT as posterior regularization. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 241–248. AAAI Press.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Shen, J., Chen, P.-J., Le, M., He, J., Gu, J., Ott, M., Auli, M., and Ranzato, M. (2021). The source-target domain mismatch problem in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1519–1533, Online, April. Association for Computational Linguistics.

Søgaard, A., Ruder, S., and Vulić, I. (2018). On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia, July. Association for Computational Linguistics.

Tae, Y., Park, C., Kim, T., Yang, S., Khan, M. A., Park, E., Qin, T., and Choo, J. (2020). Meta-learning for low-resource unsupervised neural machinetranslation. *CoRR*, abs/2010.09046.

Vulić, I., Glavaš, G., Reichart, R., and Korhonen, A. (2019). Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China, November. Association for Computational Linguistics.

Wang, Z., Xie, J., Xu, R., Yang, Y., Neubig, G., and Carbonell, J. G. (2020). Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *International Conference on Learning Representations*.

Wu, J., Wang, X., and Wang, W. Y. (2019). Extract and edit: An alternative to back-translation for unsupervised neural machine translation. 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 1173–1183, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zhang, M., Liu, Y., Luan, H., and Sun, M. (2017). Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada, July. Association for Computational Linguistics.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.