ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Towards Robust Auxiliary Tasks for Language Adaptation

Gil Rocha and Henrique Lopes Cardoso *

Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)
Faculdade de Engenharia, Universidade do Porto, Porto, Portugal

Abstract. To overcome the lack of annotated resources in less-resourced languages, unsupervised language adaptation methods have been explored. Based on multilingual word embeddings, Adversarial Training has been successfully employed in a variety of tasks and languages. With recent neural language models, empirical analysis on the task of natural language inference suggests that more challenging auxiliary tasks for Adversarial Training should be formulated to further improve language adaptation. We propose rethinking such auxiliary tasks for language adaptation.

1 Introduction

Current state-of-the-art approaches to address semantic-level tasks in natural language processing (NLP) rely on supervised learning methods. However, collecting annotated data in different languages is a challenging and time-consuming task, especially for less-resourced languages. To tackle this challenge, methods that are capable of leveraging to a target language the knowledge acquired when trained on a source language have been proposed, some of them without requiring labeled data on the target language.

Natural Language Inference (NLI) [1] has emerged as one of the main tasks to evaluate NLP systems for sentence understanding. To address this task in cross-lingual scenarios, Adversarial Training [2] has been successfully employed using multilingual word embeddings [2, 3]. The goal of Adversarial Training is to obtain representations of the input that are useful to address a specific task, while being agnostic to the input language. The model can then be employed to address the task regardless of the input language. One key advantage of Adversarial Training over other proposed approaches [3] is that we obtain a single encoder that can be employed across different languages.

With the advent of multilingual language models [4], new state-of-the-art results were obtained across different downstream tasks and languages [5]. In this study, we observe that when employing recent multilingual language models Adversarial Training is unable to improve the performance of the model in cross-lingual scenarios, when compared to the baseline Direct Transfer procedure. We investigate this phenomenon and found that the auxiliary task proposed for models employing Adversarial Training with multilingual word embeddings is unsuited when the model employs multilingual language models. We propose

*This work was funded by Fundação para a Ciência e a Tecnologia (FCT): Gil Rocha is supported by a PhD studentship (SFRH/BD/140125/2018), this research is supported by LIACC (FCT/UID/CEC/0027/2020) and DARGMINTS (POCI/01/0145/FEDER/031460).

more challenging auxiliary tasks for Adversarial Training, tailored to improve language adaptation for current state-of-the-art systems.

2 Related Work

Recent deep learning models proposed to address NLP tasks in cross-lingual settings rely on the existence of multilingual word embeddings (MWEs) [6] and, more recently, on multilingual language models [4]. In fact, pre-trained language models lead to impressive improvements on several downstream tasks. Devlin *et al.* [4] introduce the Masked Language Modeling (MLM) task and, by employing a neural network based on the Transformer architecture to predict the masked tokens from large-scale text resources (in an unsupervised setting), propose the BERT model (widely used by the community as a state-of-the-art pretrained language model).

The goal of the NLI task is to determine whether the meaning of the text fragment "Hypothesis" (H) is in an *entailment*, *contradiction* or neither (*neutral*) relation to the text fragment "Text" (T) [1]. To address NLI in a cross-lingual setting, unsupervised language adaptation (ULA) techniques have been explored [7, 3]. One the largest available resources, with data annotated in 15 languages, to study language adaptation approaches for the NLI task is the Cross-Lingual Natural Language Inference corpus (XNLI) [7].

ULA methods aim to leverage the knowledge learned while performing supervised learning on a source language to a given target language, without requiring annotated data in the target language. The most common approaches are Adversarial Training [2], Encoder Alignment [7], and Shared-Private architectures [8].

3 Adversarial Training for Cross-lingual NLI

In this work, we employ Adversarial Training, a promising method for ULA across different languages and tasks [2, 3]. Given the advantages of the method (a single encoder for many languages and no requirements on the availability of parallel sentences) compared to other proposed approaches (Encoder Alignment and Shared-Private), Adversarial Training can have a high impact in less-resourced languages. We illustrate its use, benefits and limitations through experiments conducted on the XNLI corpus.

A neural network employing Adversarial Training [2] is composed of three main components: a feature extractor $\mathcal{F}$ that maps an input sequence x to a shared feature space, a task classifier $\mathcal{P}$ that predicts the label for x given the feature representation $\mathcal{F}(x)$, and a language discriminator $\mathcal{Q}$ that given $\mathcal{F}(x)$ predicts whether x is from the source or from the target language. The goal of Adversarial Training is to minimize both the task classifier and adversarial component losses: $\mathcal{L} = \mathcal{L}_{task} + \lambda\, \mathcal{L}_{adv}$, where $\mathcal{L}_{task}$ is the cross-entropy loss between predicted labels and ground-truth, $\mathcal{L}_{adv}$ is the Wasserstein distance [9] between the feature distributions of input sequences in the source and target

Embeddings	Method	EN	AR	DE	ZH
MWE	Direct Transfer	68.01 ± 0.37	40.96 ± 1.58	41.46 ± 1.56	41.11 ± 1.15
	Adversarial Training		**45.61** ± 0.38	**46.24** ± 0.80	**47.90** ± 0.31
mBERT	Direct Transfer	71.14 ± 0.32	57.11 ± 0.68	**62.67** ± 0.10	**63.16** ± 1.25
	Adversarial Training		**57.44** ± 1.14	61.57 ± 0.52	62.99 ± 0.66

Table 1: Accuracy scores in percentage for XNLI experiments

languages, and λ is a hyper-parameter that weights the importance of the adversarial component.

3.1 Experimental setup

The source language used in our experiments is English (EN). To study the impact of Adversarial Training across different language families, the following target languages are analyzed: Arabic (AR), German (DE), and Chinese (ZH).

To encode each input sequence, we employ (a) conventional MWEs and (b) a state-of-the-art multilingual language model. For (a), each token is initialized with pre-trained 300-dimensional fastText word embeddings[1]. In the $\mathcal{F}$ component we use a BiLSTM with 128 hidden units. For optimization, we use Adam with default parameters. For (b), the $\mathcal{F}$ component employs mBERT [4]. We follow the implementation details suggested by Devlin *et al.* [4]. For optimization, we use Adam but, as suggested by Devlin *et al.* [4], with a learning rate of 2e$-$5. In both approaches, $\mathcal{P}$ and $\mathcal{Q}$ are composed of a single-layer feed-forward neural network with a 128 units, using a dropout rate of 0.2. To encode the relation between T and H, we follow the Siamese-Encoder architecture [10]: two $\mathcal{F}$ layers are employed (to encode T and H) and then merged with the widely used aggregation function $\langle T, H, |T - H|, T * H\rangle$.

3.2 Results

The results obtained from our experiments with XNLI corpus are reported in Table 1. Given that the labels are balanced, we use accuracy as evaluation metric. We report average scores of 3 runs with different random seeds. The “Embeddings” column, divides the results for each of the encoding techniques (MWE and mBERT). For the baseline “Direct Transfer” approach, we evaluate the model directly on the target language after supervised training on the source language. The “EN” column presents the scores obtained after supervised training on the source language. Columns “AR”, “DE”, and “ZH” correspond to the scores obtained in each of these target languages.

With MWEs, Adversarial Training improves the accuracy scores for all target languages (+5.41% on average) compared to the baseline (differences are

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

statistically significant with $p < 0.02$). These results are aligned with prior work [3], which concludes that Adversarial Training coupled with MWEs is a robust technique for ULA in different scenarios. However, with mBERT, Adversarial Training cannot improve the scores on the target languages, performing below Direct Transfer in two of them (differences are not statistically significant). Compared to MWEs, mBERT improves by +3.13% the score obtained on the source language, and the drop between the source and target languages (through Direct Transfer) is much smaller, showing that multilingual language models are a promising approach for cross-lingual scenarios.

3.3 Analysis

The main insights taken from our experimental results are: (a) Adversarial Training coupled with MWEs improves the scores across different target languages, but with mBERT it cannot improve over Direct Transfer; and (b) mBERT closes the gap between source and target languages, suggesting that mBERT provides strong cross-lingual baseline scores. Therefore, we conclude that the sentence-level representations provided by mBERT are closer in the feature space across different languages compared to MWEs.

To validate our hypothesis, 500 sentences were sampled from the EN and DE validation sets. Following Chen *et al.* [2], we employ t-SNE with Principal Component Analysis (PCA) to reduce the representation of the input sequences into a two dimensional feature space. To determine the distance between the set of input sequences in both languages, we use the Averaged Hausdorff Distance (AHD) [11]. An AHD distance of zero means that the set of points in both languages coincide, while higher values indicate that the distance between the two sets of points is greater. With MWEs, we obtain an AHD of 39.96 after supervised training on the source language. The AHD drops to 12.35 after Adversarial Training. Consequently, we conclude that Adversarial Training makes the $\mathcal{F}$ layer more agnostic to the input language. With mBERT, the observed drop in AHD is significantly lower in magnitude, from 7.95 after supervised training to 6.59 after Adversarial Training. This shows that the representations provided by mBERT are closer to language-agnostic, from which only marginal improvements can be obtained with Adversarial Training.

Due to the strong cross-lingual properties of mBERT, the auxiliary task objective employed in Adversarial Training is already close to optimal from the outset. For this reason, Adversarial Training cannot provide further improvements in cross-lingual settings. To counter this, we propose alternative auxiliary tasks designed to improve the transfer of knowledge across languages.

4 Robust Auxiliary Tasks for Language Adaptation

Adversarial Training is a promising approach for ULA; however, when employed with recent multilingual language models, alternative formulations for the auxiliary task are required. Sticking with the same general intuition employed by Adversarial Training for ULA, we propose that the $\mathcal{F}$ layer should be encouraged to produce representations from which a decoder-based model could be able to generate the original input sequence (similar to the auto-encoder formulation [12]). Compared to the original language discrimination task, based on which the $\mathcal{F}$ layer might only be capturing salient properties of the input sequence, we believe that this formulation will require the $\mathcal{F}$ layer to capture more information regarding the input sequences (i.e., the generation of valid sentences in a specific language requires more knowledge than language discrimination). Alternatively, following recent proposals in language modeling, we propose to employ the MLM objective [4] as the decoder objective, instead of predicting the complete input sequence. This is in line with prior work on monolingual settings [13], which concluded that performing fine-tuning for downstream tasks including language modeling as an auxiliary objective can accelerate convergence and improve the generalization capability of the learned model. However, additional challenges are expected in multilingual scenarios. Fine-tuning a pre-trained language model based on labeled data in the source language updates the learning weights of the model specifically for the source language. Given that input sequences in the target language are not available during fine-tuning, it is reasonable to expect that the representations for the target language will become outdated and not specifically tuned to address the task at hand. To employ the proposed decoder-based procedure in a cross-lingual setting (which was not considered by Radford *et al.* [13]), we propose that the auxiliary MLM objective is optimized providing input sequences in both source and target languages. We hypothesize that the task-specific fine-tuning will impact the representations in both languages, by encouraging representations for the source and target languages to be jointly updated based on the MLM objective.

Based on the current formulation, we cannot ensure that the $\mathcal{F}$ layer obtains language-specific or language-agnostic representations. For instance, if the neural network is large enough, it could be divided into two partitions: one specialized in the source language (tuned for the target task and for the MLM on the source language), the other in the target language (only tuned for the MLM objective in the target language). If this occurs, then the representations in the target language will not be aligned with the fine-tuned representations on the source language. To counter this, we propose to combine the losses of both adversarial and MLM tasks: L_{adv} and L_{mlm}. The intuition is to encourage the $\mathcal{F}$ layer to obtain sentence-level representations that are agnostic to the input language (as in the conventional Adversarial Training) but that can also provide enough information to retain the language modeling capabilities in both languages, which are critical to encode the representations for the target task. The final loss would be calculated as follows: $\mathcal{L} = \mathcal{L}_{task} + \lambda\, \mathcal{L}_{adv} + \beta\, \mathcal{L}_{mlm}$. In this formulation, $\mathcal{L}_{mlm}$ is the unsupervised MLM objective for input sequences in both source and target languages, and λ and β are weights that control the interaction of the loss terms.

5 Conclusions

Adversarial Training is an ULA method that has been successfully employed with multilingual word embeddings for different NLP tasks, including the challenging NLI task. Our empirical results show that with state-of-the-art multilingual language models, Adversarial Training cannot improve the scores obtained with the baseline Direct Transfer approach. A detailed analysis shows that the conventional language discrimination task proposed in Adversarial Training is trivially solved when we employ recent multilingual language models.

To improve the cross-lingual transfer of state-of-the-art language models, we propose alternative formulations for the adversarial component, tailored to take advantage of recent advancements in language modeling. We believe that our analysis and proposals can pave the way to more robust cross-lingual models.

References

[1] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proc. of Empirical Methods in NLP (EMNLP)*, pages 632–642, Lisbon, Portugal, September 2015. ACL.

[2] Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Q. Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570, 2018.

[3] Gil Rocha and Henrique Lopes Cardoso. A comparative analysis of unsupervised language adaptation methods. In *Proc. of Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo)*, pages 11–21, Hong Kong, China, November 2019. ACL.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota, June 2019. ACL.

[5] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proc. of EMNLP-IJCNLP*, pages 833–844, Hong Kong, China, 2019. ACL.

[6] Sebastian Ruder. A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902, 2017.

[7] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proc. of EMNLP*, pages 2475–2485, Brussels, Belgium, 2018. ACL.

[8] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Proc. of Int. Conf. on Neural Information Processing Systems*, NIPS'16, pages 343–351, USA, 2016. Curran Associates Inc.

[9] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proc. of International Conference on Machine Learning*, pages 214–223, Sydney, Australia, 06–11 Aug 2017. PMLR.

[10] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. of EMNLP-IJCNLP*, pages 3982–3992, Hong Kong, China, November 2019. ACL.

[11] MD. Shapiro and MB. Blaschko. On Hausdorff distance measures. Technical report, Department of Computer Science, University of Massachusetts Amherst, August 2004.

[12] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991.

[13] Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. *Technical report, OpenAI*, 2018.