A Comparative Analysis of Unsupervised Language Adaptation Methods

Gil Rocha and **Henrique Lopes Cardoso**
Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)
Departamento de Engenharia Informática,
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
{gil.rocha, hlc}@fe.up.pt

Abstract

To overcome the lack of annotated resources in less-resourced languages, recent approaches have been proposed to perform unsupervised language adaptation. In this paper, we explore three recent proposals: Adversarial Training, Sentence Encoder Alignment and Shared-Private Architecture. We highlight the differences of these approaches in terms of unlabeled data requirements and capability to overcome additional domain shift in the data. A comparative analysis in two different tasks is conducted, namely on Sentiment Classification and Natural Language Inference. We show that adversarial training methods are more suitable when the source and target language datasets contain other variations in content besides the language shift. Otherwise, sentence encoder alignment methods are very effective and can yield scores on the target language that are close to the source language scores.

1 Introduction

Recently proposed approaches for unsupervised adaptation have been explored in a variety of machine learning domains, including image recognition (Ganin and Lempitsky, 2015; Bousmalis et al., 2016) and natural language processing (Chen et al., 2018; Conneau et al., 2018).

In unsupervised language adaptation, annotated resources on a source language (S) are available, in the form $\langle X_S, Y_S \rangle$. For the target language (T), however, no annotations are assumed to exist for training machine learning models with. The goal is to learn representations that are useful to perform a given task on S while using representations useful to perform the same task in the target language T (or even across multiple languages).

Approaches to unsupervised language adaptation can be divided into those that (a) do not assume any particular kind of inter-language data (Chen et al., 2018), and those that (b) require sentences aligned for the source and target languages, obtained either manually or through machine translation systems (Banea et al., 2008; Zhou et al., 2016).

In this paper, we explore recent proposals from different domains for unsupervised adaptation and employ them to two natural language tasks. To do so without making use of aligned sentences, we explore Adversarial Training (Section 4.1) (Chen et al., 2018). Assuming the availability of parallel data, we also explore approaches that learn the similarities and differences between source and target language. We explore two different approaches that leverage parallel data: a Sentence Encoder Alignment (Section 4.2) (Conneau et al., 2018) and a Shared-Private Architecture (Section 4.3) (Bousmalis et al., 2016). We select these approaches from many recent proposals because they differ on the main axis of our analysis (assumptions made on the availability of unlabeled data resources), they approach the problem using conceptually different methods, and they correspond to state-of-the-art approaches.

To evaluate the proposed approaches, we explore two different cross-lingual tasks: Natural Language Inference (NLI) (also know as Recognizing Textual Entailment) (Dagan et al., 2013) and Sentiment Classification (Socher et al., 2013). Our source language is English, in both cases. For the target language, we constrain our work to Chinese and Arabic, the languages that the both tasks have in common. We believe that the linguistic differences between the source and target languages explored in this work are rich enough to demonstrate the quality of the proposed approaches, in particular in such a challenging setting as unsupervised language adaptation.

The main contributions of this work can be sum-

Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo), pages 11–21
Hong Kong, China, November 3, 2019. ©2019 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

marised as follows: (a) we divide and analyse proposed approaches for unsupervised language adaptation by taking into account their assumptions on available resources; (b) for the natural language inference (NLI) task, we explore adversarial training approaches and provide a new baseline for sentence encoders without requiring parallel data. Moreover, we explore a shared-private architecture that leverages parallel sentences; (c) for the sentiment classification task, we explore recent approaches that use parallel data (sentence encoder alignment and shared-private architecture).

2 Related Work

The Natural Language Inference (NLI) task has emerged as one of the main tasks to evaluate NLP systems for sentence understanding. Given two text fragments, "Text" (T) and "Hypothesis" (H), NLI is the task of determining whether the meaning of H is in an *entailment*, *contradiction* or neither (*neutral*) relation to the text fragment T. Consequently, this task is framed in a 3-way classification setting (Dagan et al., 2013).

State-of-the-art systems explore complex sentence encoding techniques using a variety of approaches, such as recurrent (Bowman et al., 2015a) and recursive (Bowman et al., 2015b) neural networks. To capture the relations between the text and hypothesis, sentence aggregation functions (Chen et al., 2017; Peters et al., 2018) and attention mechanisms (Rocktäschel et al., 2016) have been successfully applied to address the task. On the cross-lingual setting, there has been work using parallel corpora (Mehdad et al., 2011) and lexical resources (Castillo, 2011), as well as shared tasks (Camacho-Collados et al., 2017). Most of these systems rely heavily on the availability of multilingual resources (e.g. bilingual dictionaries) and on machine translation systems to explore projection (Yarowsky et al., 2001) or direct transfer (McDonald et al., 2011) approaches. Recently, a large-scale corpus for NLI for 15 languages was released (details in Section 3) together with multilingual sentence encoders baselines (Conneau et al., 2018). More recently, new methods to train language models provided the ground basis for contextualized word embeddings (Peters et al., 2018), which constitute the new state-of-art in several tasks, including the NLI and XNLI tasks (Devlin et al., 2019; Lample and Conneau, 2019). In this paper, we constraint our work to the conventional (cross-lingual) word embeddings (Ruder, 2017) that have been widely used and focus on a comparative analysis between different approaches for unsupervised language adaptation. We leave the study of the effects of this recent line of work on our analysis as future work.

For Sentiment Classification, several efforts have been made to address the task in a cross-lingual setting. Similarly to the NLI research focus, most of the approaches rely on projection or direct transfer approaches (Wan, 2008; Mihalcea et al., 2007; Banea et al., 2008; He et al., 2010). Some works explore parallel datasets to learn bilingual document representations (Zhou et al., 2016) or to perform cross-lingual distillation (Xu and Yang, 2017). Without requirements for parallel data resources and machine translation systems, Adversarial Deep Averaging Networks (ADAN) (Chen et al., 2018) employing adversarial training have been proposed to address the task in an unsupervised language adaption setting, which we follow in our work.

Crucial for our work is the existence of cross-lingual word embeddings (Ruder, 2017). Similarly to monolingual word embeddings, various approaches to learn cross-lingual word embeddings have been proposed in recent years, leading to existence of several pre-trained cross-lingual embeddings, including fastText embeddings (Joulin et al., 2018; Bojanowski et al., 2017), Multilingual Unsupervised and Supervised Embeddings (MUSE) (Lample et al., 2018), and bilingual word embeddings (BWE) (Zhou et al., 2016).

3 Corpora

In this section we detail on the corpora used to evaluate the unsupervised language adaptation approaches explored in this work.

3.1 Natural Language Inference

The Cross-Lingual Natural Language Inference corpus (XNLI) (Conneau et al., 2018) is a large-scale corpus for the task of NLI that contains annotations for 15 languages. Each pair of sentences is annotated with one of three labels: Entailment, Contradiction or Neutral.

The XNLI corpus is an extension for cross-lingual settings of the Multi-Genre Natural Language Inference (MultiNLI) corpus (Williams et al., 2018). This is a crowd-sourced collection

of 433k sentence pairs annotated with textual entailment information. The corpus is modeled on the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015a), but differs in that it covers a range of genres of spoken and written text, and supports a distinctive cross-genre generalization evaluation. Given that the test portion of the MultiNLI data was kept private, they collect and validate 750 new test set examples from each of the ten text sources. To create the test set for the remaining languages, professional translators were asked to translate it into the ten target languages. The training set for the English portion is the same training data from the MultiNLI corpus. Additionally, in the official repository of the XNLI corpus, machine translations of the English data (including training, validation, and test set) to each of the 15 languages of XNLI are provided.

3.2 Sentiment Classification

For the Sentiment Classification task we follow the work of Chen et al. (2018), and replicate the dataset collection used by the authors.

For the English partition, we use a balanced dataset of $700k$ Yelp reviews from Zhang et al. (2015) with their ratings as labels (scale 1-5). We adopt the same training set of $650k$ reviews, but we randomly split the original $50k$ reviews validation set into $25k$ for the test set and the remaining for the validation set (keeping label distributions unchanged). For the Chinese dataset, $10k$ balanced Chinese hotel reviews from Lin et al. (2015) are used as validation set for model selection and parameter tuning. The results are reported on a separate test set of another $10k$ hotel reviews. Similarly to the English dataset, data is annotated with 5 labels (1-5). For the unlabeled target language data used during the training, we use another $150k$ unlabeled Chinese hotel reviews.

Regarding the Arabic dataset, we use the BBN Arabic Sentiment Analysis dataset (Mohammad et al., 2016) for Arabic sentiment classification. The dataset contains 1200 sentences (600 validation + 600 test) from social media posts annotated with 3 labels ($-$, 0, $+$). Since the label set does not match with the English dataset, we map the 4 and 5 English ratings to $+$ and the 1 and 2 ratings to $-$, while the 3 rating is converted to 0. For the unlabeled target language data used during the training, we use the text from the validation set (without labels) during training (similar to procedure followed by Chen et al. (2018)).

3.3 Parallel Sentence Resources

We use publicly available parallel sentence resources to learn the alignment between English and target language sentence encoders, an approach that is used by Sentence Encoder Alignment (Section 4.2) and Shared-Private Architecture (Section 4.3). To retrieve and preprocess these parallel sentence datasets, we follow the description presented by Conneau et al. (2018). For the target languages addressed in this work, Arabic and Chinese, we use the United Nations (UN) corpus (Ziemski et al., 2016). This parallel corpus consists of manually translated UN documents from the last 25 years (1990 to 2014). In all the experiments reported in this paper, we set the maximum number of parallel sentences to 2 million.

4 Methods

To address the task of unsupervised language adaptation, we explore three approaches: Adversarial Training (Section 4.1), Sentence Encoder Alignment (Section 4.2), and Shared-Private Architecture (Section 4.3). By unsupervised language adaptation we consider that during the training phase the model is fed with labeled data (for the task at hand) on the source language and that no labeled data on target language is available. However, to train the model on a cross-lingual setting, unlabeled data on the source and target language are provided. We study on the assumptions that are made on the availability of unlabeled data for the source and target language.

The first, Adversarial Training, only requires the availability of unlabeled data in both languages, without requiring parallel sentences to perform the language adaptation. The remaining two approaches require parallel sentences for the source and target languages.

4.1 Adversarial Training

In a cross-lingual setting, the aim of adversarial training is to make the neural network agnostic to the input language while learning to address a specific task, following the intuition that if the network learns representations that are useful for the task and at the same time agnostic to language specificities, then such representations can be directly employed to address the task on a target language (unsupervised language adaptation).

A neural network with adversarial training is typically composed of three main components: a *Feature Extractor* $\mathcal{F}$ that maps the input sequence x to a feature space $\mathcal{F}(x)$, a *Task Classifier* $\mathcal{P}$ that given the feature representation $\mathcal{F}(x)$ predicts the labels for the task at hand, and a *Language Discriminator* $\mathcal{Q}$ that also receives $\mathcal{F}(x)$ as input and aims to discriminate the language of the input sequence. $\mathcal{F}$ and $\mathcal{P}$ correspond to the typical components employed to address a text classification task. $\mathcal{Q}$ corresponds to the second objective we want to optimise the neural network for, where the adversarial objective is defined.

The first formulation for an adversarial component following this setting was the Gradient Reversal Layer (GRL) (Ganin and Lempitsky, 2015), where $\mathcal{Q}$ is a binary classifier distinguishing whether the input sequence x comes from the source or target language.

However, training a neural network using the GRL is very unstable, and efforts need to be made to coordinate the adversarial training. To address this issue, Chen et al. (2018) propose to minimise the Wasserstein distance $\mathcal{W}$ between the distribution of the joint hidden features $\mathcal{F}$ for the source $P^{src}_{\mathcal{F}} \triangleq P(\mathcal{F}(x^{src}))$ and, similarly, for the target instances according to the Kantorovich-Rubinstein duality, and demonstrate that this improves the stability for hyperparameter selection. Following Chen et al. (2018), the adversarial component aims to maximize the following loss:

$$\mathcal{L}_{adv} \equiv \max_{\theta_q}(\mathbb{E}_{\mathcal{F}(x^{scr})\sim P^{src}_{\mathcal{F}}}[\mathcal{Q}(\mathcal{F}(x^{src}))] - \mathbb{E}_{\mathcal{F}(x^{tgt})\sim P^{tgt}_{\mathcal{F}}}[\mathcal{Q}(\mathcal{F}(x^{tgt}))]) \quad (1)$$

For the task classifier component $\mathcal{P}$, we want to minimize the negative log-likelihood of the target class for each source language example:

$$\mathcal{L}_{task} = -\sum_{i=0}^{N_{src}} y_i^{src} \cdot \log \hat{y}_i^{src}, \quad (2)$$

where y_i^{src} is the one-hot encoding of the class label for source input i and $\hat{y}_i^{src}$ are the softmax predictions of the model: $\hat{y}_i^{src} = \mathcal{P}(\mathcal{F}(x_i^{src}))$.

Finally, the goal of training the complete neural network is to minimize both the task classifier and adversarial component losses:

$$\mathcal{L}_{ADAN} = \mathcal{L}_{task} + \lambda\, \mathcal{L}_{adv} \quad (3)$$

where λ is a hyper-parameter that balances the importance of the adversarial component in the overall loss computation. Differently from Chen et al. (2018), who use a constant value $\lambda = 0.01$, we employ a λ schedule that increases with the number of epochs. The intuition is to make the adversarial component more important along time, while keeping a good performance on the task. Following Ganin and Lempitsky (2015), λ starts at 0 and is gradually increased up to 1:

$$\lambda_p = \frac{2}{1 + \exp(-\gamma \cdot p)} - 1 \quad (4)$$

where γ was set to 10 and p corresponds to the percentage of training completed given a predefined maximum number of epochs.

4.2 Sentence Encoder Alignment

The Sentence Encoder Alignment method aims to align the encoder for the target language based on a pre-trained encoder on the source language (Conneau et al., 2018). The key idea is that the target encoder learns to copy the source encoder representation based on parallel sentences in both languages. This method relies on the assumption that the representations captured by the source encoder (based solely on source language training for the task at hand) are useful for the target language as well. We hypothesise that in situations where the only variation between task and parallel data is the language shift, this approach can obtain promising results. However, in cases where the language shift is accompanied by other linguistic phenomena discrepancies (*e.g.* differences in domain), sentence encoder alignment might not yield competitive results.

This method includes three steps: (a) source language training using labeled data on the task at hand, (b) aligning sentence encoders with parallel data, and (c) inference on the target language. The architecture has three main components: a *Feature Extractor for the Source Language* $\mathcal{F}_{\mathcal{S}}$ that maps input sequence x^{src} to a feature space $\mathcal{F}_{\mathcal{S}}(x^{src})$, a *Feature Extractor for the Target Language* $\mathcal{F}_{\mathcal{T}}$ that maps the input sequence x^{tgt} to a feature space $\mathcal{F}_{\mathcal{T}}(x^{tgt})$, and a *Task Classifier* $\mathcal{P}$ that given the feature representation $\mathcal{F}(x)$ predicts the labels for the task at hand.

The first step, source language training, follows the typical training on monolingual settings. $\mathcal{F}_{\mathcal{S}}$ and $\mathcal{P}$ are trained using labeled data in the source language. In the next step, the goal is to align a target encoder $\mathcal{F}_{\mathcal{T}}$ based on the source encoder $\mathcal{F}_{\mathcal{S}}$ learned in the previous step.

Given parallel sentences (from resources external to the task at hand) in the source and target language, z^{src} and z^{tgt}, we train $\mathcal{F}_\mathcal{T}$ to represent input sequence z^{tgt} as close as possible in the feature space to the representation produced by $\mathcal{F}_\mathcal{S}$ for the parallel sentence z^{src}. To this end, we follow the alignment loss $\mathcal{L}_{align}$ (Conneau et al., 2018):

$$\begin{aligned}\mathcal{L}_{align} =& dist(\mathcal{F}_\mathcal{S}(z^{src}), \mathcal{F}_\mathcal{T}(z^{tgt})) - \\ & \eta(dist(\mathcal{F}_\mathcal{S}(z^{src}_{neg}), \mathcal{F}_\mathcal{T}(z^{tgt})) + \\ & dist(\mathcal{F}_\mathcal{S}(z^{src}), \mathcal{F}_\mathcal{T}(z^{tgt}_{neg})))\end{aligned} \quad (5)$$

where $(z^{src}_{neg}, z^{tgt}_{neg})$ are contrastive terms obtained using negative sampling (*i.e.* z^{src}_{neg} was randomly sampled from the parallel sentences dataset and does not correspond to a parallel sentence of z^{tgt}; similarly between z^{tgt}_{neg} and z^{src}), and η controls the weight of the negative examples in the loss (we fix $\eta = 0.25$ has suggested by Conneau et al. (2018)). For the distance measure, we use the L2 norm $dist(x, y) = \|x - y\|_2$. During training, we only back-propagate through $\mathcal{F}_\mathcal{T}$ when optimizing $\mathcal{L}_{align}$ such that the target feature extractor is mapped to the source language feature space.

In the last step, the neural network is composed of $\mathcal{F}_\mathcal{T}$ obtained in the second step of this method and $\mathcal{P}$ obtained in the first step. Following this procedure we can directly make inferences on the target language, without requiring any kind of supervision on the target language.

4.3 Shared-Private Architecture

The key idea of a shared-private architecture is to obtain two different representations of the input. The shared representation aims to capture language agnostic features that can be shared across different languages. On the other hand, the private representation aims to capture language specific features. To prevent the shared and private spaces from interfering with each other, two strategies are typically used: adversarial training (Ganin and Lempitsky, 2015; Liu et al., 2017) and orthogonality constraints (Bousmalis et al., 2016).

A neural network following a shared-private architecture designed for a cross-lingual setting is composed of: a *Shared Feature Extractor* $\mathcal{F}_\mathcal{C}$ that maps the input sequence x to a common/shared feature space $\mathcal{F}_\mathcal{C}(x)$, a *Private Feature Extractor* $\mathcal{F}_\mathcal{P}$ that maps the input sequence to a private feature space $\mathcal{F}_\mathcal{P}(x)$, *Task Classifier* $\mathcal{P}$ that given $\mathcal{F}_\mathcal{C}(x)$ predicts the labels for the task at hand, and a *Language Discriminator* $\mathcal{Q}$ that receives $\mathcal{F}_\mathcal{P}(x)$ as input and aims to discriminate the language of the input sequence.

For the task classifier component $\mathcal{P}$, the goal is to minimize the negative log-likelihood of the ground truth class for each source language input sequence x^{src} given the representation obtained from $\mathcal{F}_\mathcal{C}(x^{src})$. The loss used for this component is defined in Equation 2.

For the language discriminator component $\mathcal{Q}$ the main goal is to train the private feature extractor $\mathcal{F}_\mathcal{P}$ to capture language specific phenomena. In the language discriminator component, we aim to minimize the negative log-likelihood of the ground truth language discrimination for each input sequence in x^{mix}, where x^{mix} corresponds to a balanced sample of sentences randomly taken from both source and target language datasets. $\mathcal{Q}$ receives the representation of the input sequence x^{mix} from the private feature extractor $\mathcal{F}_\mathcal{P}(x^{mix})$. Again, we use the loss defined in Equation 2.

The difference loss, $\mathcal{L}_{diff}$, is applied to input sentences of both languages x^{mix} and encourages the shared and private feature extractors to encode different aspects of the input sequences. Following Bousmalis et al. (2016), we define the loss via a soft subspace orthogonality constraint between the private and shared representations, as follows:

$$L_{diff} = \left\| \mathcal{F}_\mathcal{C}(x^{mix})^\top \mathcal{F}_\mathcal{P}(x^{mix}) \right\|^2_F \quad (6)$$

where $\|\cdot\|^2_F$ is the squared Frobenius norm.

The similarity loss, $\mathcal{L}_{sim}$, encourages the representations $\mathcal{F}_\mathcal{C}(x^{src})$ and $\mathcal{F}_\mathcal{P}(x^{tgt})$ to be as similar as possible irrespective of the language. We employ the same loss defined in Equation 5 as similarity loss, *i.e.*, $\mathcal{L}_{sim} = \mathcal{L}_{align}$. However, we emphasise that the training procedure is different. Here the alignment loss is one component of the total loss applied to the neural network, working concurrently with the other components.

Finally, the goal of training the complete neural network is to minimize the following loss:

$$\begin{aligned}\mathcal{L}_{SP} =& \mathcal{L}_{task} + \lambda\, \mathcal{L}_{lang} + \\ & \beta\, \mathcal{L}_{diff} + \gamma\, \mathcal{L}_{sim}\end{aligned} \quad (7)$$

where λ, β and γ are hyper-parameters that balance the importance of each component in the overall loss computation. All these values are parameterized with the same schedule (Eq. 4). We leave for future work finding optimal values for these hyper-parameters.

5 Experiments

To evaluate the methods described in Section 4 for unsupervised cross-lingual settings, we report on experiments performed on two different tasks: Natural Language Inference and Sentiment Classification. On both tasks we consider English (en) as source language and Chinese (zh) and Arabic (ar) as target languages.

5.1 Implementation Details

For the NLI task, we kept most of the architecture details as similar as possible to the initial work (Conneau et al., 2018). More specifically, the Sentence Encoder Alignment architecture is similar to this work. However, some of the parameters were changed to speedup computations on all architectures, so we expect the results to be worst than those reported by Conneau et al. (2018). The main goal of this work is not to provide a new state-of-the-art system for the task, but instead we focus on alternative architectures that explore different assumptions about the data and that are backed up by promising theoretical motivations.

The only pre-processing step required is the tokenization of the input sequence. We use MOSES tokenizer (Koehn et al., 2007) for sentences in English and Arabic, and Stanford segmenter (Chang et al., 2008) for Chinese. Each token is associated to the corresponding word embedding. We use the fastText[1] pre-trained 300 dimensional word vectors computed on Wikipedia, aligned on several languages using the relaxed cross-domain similarity local scaling (RCSLS) method (Joulin et al., 2018; Bojanowski et al., 2017). For the Feature Extractor component $\mathcal{F}$, we use a BiLSTM (Hochreiter and Schmidhuber, 1997) with 128 hidden units, concatenating the initial and final hidden states (Sutskever et al., 2014). For the Task Classifier $\mathcal{P}$ and Language Discriminator $\mathcal{Q}$ we employ a feed-forward neural network with a 128 hidden units hidden layer, regularized with dropout (Srivastava et al., 2014) at a rate of 0.2. As suggested by Chen et al. (2018), the weights of the adversarial component are clipped to $[-0.01, 0.01]$. For optimization, we use Adam (Kingma and Ba, 2014) with default parameters.

To compare the results of the different architectures described in Section 4 on the Sentiment Classification task with existing work, we follow the experimental setup used by Chen et al. (2018). The tokenization is performed using Stanford CoreNLP (Manning et al., 2014) for all languages. Regarding word embeddings, for Chinese we used the pre-trained 50 dimensional Bilingual Word Embeddings (BWE) by Zhou et al. (2016). For Arabic, the 300 dimensional Bil-BOWA BWE (Gouws et al., 2015) trained by Chen et al. (2018) were not available. Instead, we used the pre-trained 300 dimensional word vectors fastText. For the Feature Extractor component $\mathcal{F}$, we use the Deep Averaging Network (DAN) (Iyyer et al., 2015). For each input sequence, DAN calculates the average of the word vectors in the input sequence, then passes this tensor of average values through a feed-forward network with ReLU (Glorot et al., 2011) non-linearities. The feature extractor $\mathcal{F}$ has three fully-connected layers, while both $\mathcal{P}$ and $\mathcal{Q}$ have two. All hidden layers contain 900 hidden units. We also use Adam optimizer for this task, but using a learning rate of 0.0005 as employed by Chen et al. (2018).

For both tasks, to find the best model in each experiment, we stop training once the accuracy on the validation set does not improve for 3 epochs (early-stop criterion) or when 30 epochs are completed. The batch size used in the experiments was set to 96 learning instances.

5.2 Analysis

Experimental results for the NLI task are shown in Table 1. The "Conneau et al. (2018) BiLSTM-last" architecture corresponds to the *BiLSTM-last* multilingual sentence encoders (in-domain) proposed by Conneau et al. (2018); the remaining architectures correspond to those described in sections 4.1, 4.2 and 4.3, respectively. The evaluation metric used is accuracy given that all labels are equally represented (balanced dataset).

Comparing our results with existing state-of-the-art (*e.g.* Conneau et al. (2018)), we can observe that our scores are lower. We attribute this to some parameter choices that were driven by computational efficiency concerns (described in Section 5.1). We focus our work on a comparison between different architectures and, therefore, we aim at a comparative analysis between those architecture in similar settings.

Comparing the architectures presented in Section 4, we can conclude that the Sentence Encoder Alignment architecture yields better results in both

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

Architecture	*en*	*zh*	*ar*
Conneau et al. (2018) BiLSTM-last	71.00	63.70	62.7
Adversarial	68.62	47.29	45.59
Sent Enc Align	68.62	**58.24**	**57.33**
Shared-Private	68.62	49.14	48.80

Table 1: XNLI accuracy scores

	5 labels		**3 labels**	
Architecture	*en*	*zh*	*en*	*ar*
ADAN	-	42.49	-	54.54
Adversarial	60.40	**43.22**	77.68	**52.17**
Sent Enc Align	60.40	35.10	77.68	48.17
Shared-Private	60.40	29.13	77.68	43.50

Table 2: Sentiment Classification accuracy scores

languages. Against our intuition, the Shared-Private Architecture presents a considerable drop of performance when compared with the Sentence Encoder Alignment method even if the sentence encoder alignment procedure is also performed in the former (*i.e.* $\mathcal{L}_{sim} = \mathcal{L}_{align}$). We attribute this to the reduced number of updates that is performed for the alignment procedure in the Shared-Private Architecture (given that we compute a joint loss, the number of iterations is determined by the size of the labeled data for the task at hand). On the other hand, the Sentence Encoder Alignment method can make complete use of the 2 million parallel sentences. We also studied the capability of the shared and private feature extractors to predict the language of a given set of input sequences. After some epochs of training, we observe that the shared feature extractor is unable to distinguish the input sequence language (obtaining 50% of accuracy to distinguish the languages). On the other hand, the private feature extractor masters the task reaching an accuracy of approximately 100%.

Adversarial Training performed considerably worst in both target languages. We emphasise that this architecture relieves the assumption of the availability of parallel sentences in both languages, and therefore removes the expense of acquiring such data. This can be relevant for less-resourced languages, where the availability of such parallel datasets is scarce and where neural machine translation systems perform worst. To the best of our knowledge, this constitutes the first effort to obtain a NLI system in a cross-lingual setting employing adversarial training, and to address the task without making any requirement on the availability of parallel sentences. Therefore, we present here a baseline system in this setting.

The results of the experiments conducted for the Sentiment Classification task are shown in Table 2. The "ADAN" architecture corresponds to the ADAN model (Chen et al., 2018). In the 5 labels setting, the labels are distributed equally (balanced dataset). In the 3 labels setting, the classes are unbalanced in both target languages. We keep using the accuracy metric in order to compare with the current state-of-the-art in this task.

Since in this setting we follow the same component architectures and parameters used in Chen et al. (2018), the results of our implementation using Adversarial Training are close to the scores reported by Chen et al. (2018). From this we can conclude that the differences introduced in this work, namely the dynamic schedule for the λ value, did not influence the overall scores. Even if no substantial differences exist between the scores, we obtain a new state-of-the-art score for the Chinese language. We attribute the small drop of performance in Arabic to the different word embeddings used.

It is interesting to notice that in this task Adversarial Training works substantially better than the remaining architectures. We attribute this to the differences of domain between the source and the target language datasets (for both Chinese and Arabic). Using the Sentence Encoder Alignment in such a setting is not as promising, comparing with the NLI setting, where both source and target languages share the domain (even if the XNLI dataset is composed of different domains, they overlap between the languages). In fact, in the Sentiment Classification task we perform the alignment of the target language feature extractor to the source language feature extractor (i.e. for Yelp related reviews) and then ask the system to perform predictions on a different language and domain (*e.g.* Chinese and hotel reviews, respectively). On the other hand, Adversarial Training aims to obtain representations that are agnostic in respect to an auxiliary task, in our case related with language and domain shift. Consequently, despite the considerable drop of performance of Adversarial Training when compared to the source language, it might be a strong baseline for unsupervised adaption for datasets that differ not only

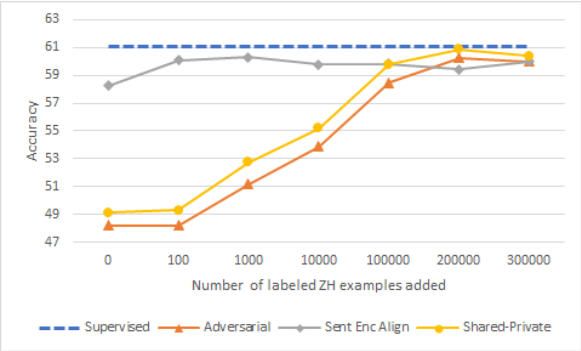

Figure 1: XNLI accuracy scores for Chinese in the semi-supervised setting.

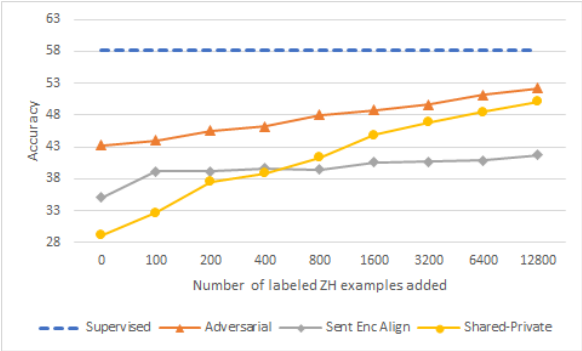

Figure 2: Sentiment Classification accuracy scores for Chinese in the semi-supervised setting.

in language but also in other phenomena (such as domain, genre, style, etc).

5.3 Semi-Supervised Learning

In several scenarios, some annotated data in the target language is available. In this section we study how performance of the methods detailed in Section 4 evolve as some examples in the target language are added to the training set.

For the NLI task, results are shown in Figure 1. The blue dotted line, dubbed "Supervised", corresponds to training the model in a supervised setting on the target language, using the machine translated training set provided by the XNLI corpus. Sentence Encoder Alignment already obtained scores close to the supervised model in the unsupervised language adaptation setting. By adding 100 instances from the target language, scores increase slightly. However, adding more instances does not affect overall performance. For the remaining models, only when we add $1k$ instances the accuracy starts to increase substantially. As we add more target language instances, accuracy keeps increasing at a consistent rate, reaching the Sentence Encoder Alignment and Supervised baseline when we add $200k$ instances.

For the Sentiment Classification task, results are shown in Figure 2. Adversarial Training remains the best model for this task as we increase learning instances from the target language in the semi-supervised setting. Accuracy scores increase as we add more instances. The Sentence Encoder Alignment is the method that less effectively takes advantage of the added data on the target language. On the other side, the Shared-Private Architecture is the method that makes better use of the added target language instances, surpassing the Sentence Encoder Alignment when we add 800 instances and becoming competitive with Adversarial Training when 1600 instances are added.

In both tasks, the Sentence Encoder Alignment is the method that takes less profit from the added supervision in the target language, while Adversarial Training and Shared-Private Architecture can improve the overall accuracy as more supervision is provided.

6 Conclusions and Future Work

We have studied unsupervised language adaptation approaches on two natural language processing tasks, taking into consideration the assumptions made regarding the availability of unlabeled data in the source and target languages.

Our results indicate that the characteristics of the datasets used in the source language (to train the models) and on the target language (to evaluate the cross-lingual approaches) are an important factor to consider when choosing the architecture to employ. When the source and target datasets present other variations in content besides the language shift, adversarial training approaches outperform those that rely on sentence alignment methods. On the other hand, when the source and target language datasets have the same characteristics, sentence alignment approaches are very effective and obtain scores in the target language that are closer to source language scores.

In future work, we aim to explore recent advances made on multilingual contextualized word embeddings and determine whether they impact the results reported in this work. Hyper-parameter tuning of different loss components is a challenging task that we aim to study in more detail.

Acknowledgments

Gil Rocha is supported by a PhD scholarship (SFRH/BD/140125/2018) from Fundação para a Ciência e a Tecnologia (FCT). This research is supported by project DARGMINTS (POCI/01/0145/FEDER/031460), funded by FCT.

References

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 127–135, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 343–351, USA. Curran Associates Inc.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. 2015b. Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 12–21, Beijing, China. Association for Computational Linguistics.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 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, Vancouver, Canada. Association for Computational Linguistics.

Julio Javier Castillo. 2011. A wordnet-based semantic approach to textual entailment and cross-lingual textual entailment. *International Journal of Machine Learning and Cybernetics*, 2(3):177–189.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 224–232, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Q. Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *TACL*, 6:557–570.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA. PMLR.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 748–756, Lille, France. PMLR.

Yulan He, Harith Alani, and Deyu Zhou. 2010. Exploring English lexicon knowledge for Chinese sentiment analysis. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China. Association for Computational Linguistics.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 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, Prague, Czech Republic. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Yiou Lin, Hang Lei, Jia Wu, and Xiaoyu Li. 2015. An empirical study on sentiment classification of Chinese review using word embedding. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*, pages 258–266, Shanghai, China.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2011. Using bilingual parallel corpora for cross-lingual textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1336–1345, Portland, Oregon, USA. Association for Computational Linguistics.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, Prague, Czech Republic. Association for Computational Linguistics.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *J. Artif. Int. Res.*, 55(1):95–130.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

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

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Xiaojun Wan. 2008. Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 553–561, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Vancouver, Canada. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 649–657, Cambridge, MA, USA. MIT Press.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1412, Berlin, Germany. Association for Computational Linguistics.

Micha Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).