Cross-Lingual Argumentative Relation Identification: from English to Portuguese

Gil Rocha[*] and **Christian Stab**[†] and **Henrique Lopes Cardoso**[*] and **Iryna Gurevych**[†]
[*] LIACC/DEI, Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[†] Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
https://www.ukp.tu-darmstadt.de/

Abstract

Argument mining aims to detect and identify argument structures from textual resources. In this paper, we aim to address the task of argumentative relation identification, a subtask of argument mining, for which several approaches have been recently proposed in a monolingual setting. To overcome the lack of annotated resources in less-resourced languages, we present the first attempt to address this subtask in a cross-lingual setting. We compare two standard strategies for cross-language learning, namely: projection and direct-transfer. Experimental results show that by using unsupervised language adaptation the proposed approaches perform at a competitive level when compared with fully-supervised in-language learning settings.

1 Introduction

The aim of argument mining (AM) is the automatic detection and identification of argumentative structures contained within natural language text. In general, arguments are justifiable positions where pieces of evidence (premises) are offered in support of a conclusion. Most existing approaches to AM build upon supervised machine learning (ML) methods that learn to identify argumentative content from manually annotated examples. Building a corpus with reliably annotated arguments is a challenging and time-consuming task, due to its complexity (Habernal et al., 2014). Consequently, training data for AM is scarce, in particular for less-resourced languages. To overcome the lack of annotated resources for AM in less-resourced languages, we explore cross-language learning approaches (Xiao and Guo, 2013). The aim of cross-language learning is to develop ML techniques that exploit annotated resources in a source language to solve tasks in a target language. Eger et al. (2018) propose the first attempt to address the identification of argumentative components in a cross-language learning setting. In this paper, we aim to employ existing state-of-the-art cross-language learning techniques to address the task of *argumentative relation identification*, leveraging knowledge extracted from annotated corpora in English to address the task in a less-resourced language, such as Portuguese. As it may be costly to produce small amounts of training data in many different languages, we employ unsupervised language adaptation techniques, which do not require labeled data in the target language.

The aim of argumentative relation identification, the last subtask of the AM process (Peldszus and Stede, 2015), is to classify each argumentative discourse unit (ADU) pair as argumentatively related or not. We assume that the subtask of text segmentation in ADUs is already solved (although no ADU classification is assumed). The task is formulated as a binary classification problem: given a tuple $\langle ADU_s, ADU_t \rangle$, we aim to classify the relation from ADU_s to ADU_t as "support" (where ADU_s plays the role of premise and ADU_t plays the role of conclusion), or "none" (unrelated ADUs). This is a consistent way of formulating the problem (*i.e.* the premise on the left and conclusion on the right side of the tuple), which is an important requirement for the learning process as the relation we aim to capture is a directional relation (*i.e.* ADU_s supports/refutes ADU_t and not on the way around).

We hypothesize that good semantic representations of text, capturing argumentative relations between ADUs, can be independent of the text language. By capturing the semantics of such relations in a higher-level representation (through sentence encoding and aggregation techniques) that is agnostic of the input language, we believe that transfer learning (Pratt and Jennings, 1996) is feasible and, consequently, encouraging results can

Proceedings of the 5th Workshop on Argument Mining, pages 144–154
Brussels, Belgium, November 1, 2018.

be obtained for less-resourced languages. For that, we propose employing cross-language learning techniques, such as *projection* (Yarowsky et al., 2001) and *direct transfer* (McDonald et al., 2011). We show promising results following the approach presented in this paper, by obtaining performance scores in an unsupervised cross-language setting that are competitive (and in some settings better) than fully-supervised in-language ML approaches.

To the best of our knowledge, this is the first approach to consider the task of argumentative relation identification in a cross-lingual setting.

2 Related Work

The full process of AM can be decomposed into several subtasks (Peldszus and Stede, 2015), namely: text segmentation, identification of ADUs, ADU type classification, relation identification, and relation type classification.

Addressing argumentative relation identification in isolation, Nguyen and Litman (2016) adopt a feature-based approach including lexical (unigrams), syntactic (part-of-speech, production rules), discourse indicators (PDTB relations) and topic-context features. Recent works address the task through deep learning architectures. Bosc et al. (2016) employ an encoder-decoder architecture and two distinct LSTMs to identify support and attack relations on tweets. Cocarascu and Toni (2017) follow architectures used for the recognizing textual entailment task, reporting results that substantially improve accuracy as compared to a feature-based ML approach on the same corpus.

Other approaches model the problem jointly with previous subtasks of AM. Stab and Gurevych (2017) follow a feature-based approach employing features at different levels of abstraction and integer linear programming for joint optimization of the subtasks. Eger et al. (2017) propose an end-to-end AM system by framing the task as a token-level dependency parser and sequence tagging problem. Potash et al. (2017) use an encoder-decoder problem formulation by employing a pointer network based deep neural network architecture. The results reported by Potash *et al.* (0.767 macro F1-score) constitute the current state-of-the-art on the Persuasive Essays corpus (Stab and Gurevych, 2017) for the subtask of argumentative relation identification.

Related work aiming to capture relations between elementary units of texts is closely related to our task. For instance, recognizing textual entailment (RTE) also focuses on pair classification (Sammons et al., 2012). 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, followed by a set of hidden layers (including aggregation functions (Chen et al., 2017; Peters et al., 2018) and attention mechanisms (Rocktäschel et al., 2015)). In another line of work, discourse parsing approaches aim to identify the structure of the text in terms of discourse or rhetorical relations between elementary units of text (*e.g.* propositions). Recent work focuses on building good representations of text relying on neural network architectures (Braud et al., 2017). Some attempts exist to address these related tasks in cross-lingual settings. For RTE 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). Typically, these systems explore projection approaches and abstract representations that do not require prior translation, namely bilingual dictionaries, syntactic information, statistical knowledge and external knowledge from lexical resources (*e.g.* ConceptNet, WordNet, BabelNet). More recently, Agic and Schluter (2018) provide multilingual test data for four major languages (Arabic, French, Spanish and Russian) and baseline cross-language RTE models. Preliminary work shows that projection approaches work better in cross-lingual settings than direct transfer.

Despite the similarity between the tasks of argumentative relation identification and RTE, since both tasks are grounded in different conceptual frameworks, the inherent semantic relations that the tasks aim to capture is conceptually different (as detailed by Cabrio and Villata (2013)). In this respect, it is important to notice that the SNLI corpus (Bowman et al. (2015a), the reference corpus for RTE) is composed of literal descriptions of scenes depicted in images, where pairs were manually created. Compared to the Argumentative Essays corpus and, more specifically, to ADU pairs extracted from it, we observe that the latter tend to require higher-level semantic reasoning (this is apparent when comparing the example provided in Table 2 with the following example extracted from the SNLI corpus: "A soccer game with multiple

Lang	Corpus	#Docs	#Rel	#None	#Support	#Attack	Arg. Schema	Type
EN	Argumentative Essays	402	22,172	17,923	3,918	331	Premise, Claim, Major Claim	Essays
PT	ArgMine	75	778	621	153	4	Premise, Claim	Opinion Articles

Table 1: Corpora Statistics

Lang.	Source ADU	Target ADU	Label
EN	Teachers are not just teachers, they are also friends and conseilieurs	In conclusion, there can be no school without a teacher	support
	computers need to be operated by people	no one can argue that technological tools are must-haves for the classroom	none
PT	Durante a última década, a saúde, o meio ambiente, a biodiversidade, assim como a evolução humana tem sido temas recorrentes em todos os meios de comunicação. *(During the last decade, health, environment, biodiversity, as well as human evolution have been recurring topics in all sorts of media)*	O século XXI é sem sombra de dúvida a era da Biologia *(The 21st century is undoubtedly the era of biology)*	support
	Seria da mais elementar prudência não voltar a precisar de lhe pedir dinheiro *(It would be most prudent not to need asking it money again)*	O fluxo de migrantes agravou o peso do euroceptismo nos governos *(The flow of migrants has increased the weight of euroscepticism in governments)*	none

Table 2: Annotated examples extracted from the Argumentative Essays (EN) (Stab and Gurevych, 2017) and ArgMine corpus (PT) (Rocha and Lopes Cardoso, 2017)

males playing." entails "Some men are playing a sport.").

To the best of our knowledge, Eger et al. (2018) present the first work exploiting cross-lingual techniques for argument mining. The authors address component extraction and classification and show that machine translation and (naïve) projection work considerably better than direct transfer. More details regarding cross-language learning techniques are presented in Section 4.3.

3 Corpora

To address the task of argumentative relation identification in a cross-language setting, argument-annotated corpora are required in different languages. Such corpora should, ideally, (a) contain annotations of arguments in different languages, (b) follow the same argumentation theory and (c) belong to the same genre of text and similar domains. Currently, there are resources for English (Stab and Gurevych, 2017) and Portuguese (Rocha and Lopes Cardoso, 2017) that follow the premise-conclusion argumentation model and contain annotations of argumentative relations between ADUs, and thus fulfill the first and the second criteria listed above. However, the corpora collected for this work (Table 1) do not meet the third criterion because they contain annotations from different types of texts: persuasive essays and opinionated articles. We focus our attention on the language adaptation of the models proposed in this paper, even though we are aware that this domain shift might play an important role in the performance of our proposed methods.

3.1 Data Preparation

Since we focus on a specific subtask of AM, argumentative relation identification, we need to generate appropriate datasets from the corpora listed in Table 1. As input, we receive texts annotated with argumentative content at the token level following a specific argumentation theory (*i.e.* premise-conclusion model). For the task at hand, we construct a dataset containing ADU pairs annotated with "none", "support" or "attack". We start by splitting each document into paragraphs, for the following reasons: (a) in all corpora used in this work, arguments are constrained to paragraph boundaries; (b) paragraph splitting reduces the number of "none" relations in the final dataset and, therefore, leads to a less skewed class distribution of the labels.

For each paragraph with ADUs $c_1, ..., c_n$, we generate tuples $\langle c_i, c_j \rangle$, with $i \neq j$ and $i, j \in [1, n]$ as argument component pairs, and label them with "support"/"attack" if the original annotation contains a direct argumentative relation from c_i to c_j, or with "none" otherwise. As shown in Table 1, label distribution is skewed towards "none" relations. Given the low number of "attack" relations,

we disregard them for this paper. Hence, we formulate the task as a binary classification problem: each tuple is classified as "none" or "support".

Table 2 shows an example of the content available in the corpora for each of the labels.

4 Methods

Similarly to approaches that aim to learn universal sentence representations able to capture the semantics of the sentence (Bowman et al., 2015b; Conneau et al., 2017), we explore different deep learning architectures to encode the meaning of ADUs for the task of argumentative relation identification. To help replicate our results, we publish the code used in this work[1]. We propose five neural network architectures that differ in the sentence encoding techniques employed (as described in Section 4.1), to which we add a fully-connected hidden layer with the same dimension as the output of the sentence encoding component, followed by a softmax layer to obtain the final predictions. To prevent the model from overfitting, we apply dropout (Srivastava et al., 2014) in each model after the sentence encoding component.

4.1 Sentence Encoding

We explore different ways of encoding the meaning of ADU pairs.

LSTM. LSTMs (Hochreiter and Schmidhuber, 1997) are recurrent neural networks (RNN) that process each word at a time and decide which information to keep in order to produce a concise representation of the word sequence. We concatenate word embedding representations of the words in ADU_s and ADU_t with a special delimiter token $delim$ (with its embeddings randomly initialized). The role of this delimiter is to indicate the RNN that a transition from ADU_s to ADU_t is being made. Then, the LSTM cell processes the entire sequence. The final hidden state representation is used as the sentence representation.

BiLSTM. Traditional LSTMs process the text in a single direction and do not consider contextual information of future words in the current step. Bidirectional LSTMs use both previous and future context by processing the input sequence in two directions. We follow the same procedure described for LSTM by concatenating ADUs using a special token. The final representation is the concatenation of the forward and backward step.[2]

Conv1D. Both ADUs are encoded separately using a convolutional neural network (CNN) (LeCun et al., 1998), with a fixed kernel size of 2, stride 1 and a max pooling layer to obtain the final fixed-length representation. The motivation for using CNNs is the fact that they can model the sequence of words by processing subsequences in parallel to obtain a final higher-level representation of the sentence. This is a promising approach when dealing with text in different languages, where the order of words are different.

Inner-Att. Inspired by previous successful work using attention (Bahdanau et al., 2014; Stab et al., 2018) in several NLP applications, we propose an attention-based sentence encoding that learns the importance of weighting ADU_t depending on the content of ADU_s. We adopt an inner-attention mechanism as proposed by Wang et al. (2016). First, we encode ADU_s using a LSTM. Then, we determine the importance weighting on the input sequence ADU_t instead of on the hidden states of the $LSTM(ADU_t)$: this has been shown to prevent biased importance weights towards the end of a sequence (Wang et al., 2016). This attention mechanism uses the information encoded in $LSTM(ADU_s)$ to inform which of the words in ADU_t the model should pay more attention to, given ADU_s. By employing this attention mechanism, we obtain a weighted input embeddings representation of ADU_t, represented as $\widetilde{x}_t$ The final hidden state used as the encoding of the tuple is obtained by applying a LSTM over the weighted representation of ADU_t: $LSTM(\widetilde{x}_t)$.

4.2 In-Language Baseline Models

As in-language baselines, we present experiments using the following models: (a) logistic regression employing a bag-of-words encoding (1 to 3 n-grams) for feature extraction based on word counts, without employing weighting techniques[3] (*BoW+LR*); (b) Chen et al. (2017): propose the enhancement of sequential inference models based on chain networks to address the task of RTE. The authors propose two models: a sequential model

[1] https://github.com/GilRocha/emnlp2018-argmin-workshop-xLingArgRelId

[2] For both LSTM and BiLSTM sentence encoding, we also tried to encode ADU_s and ADU_t separately using two distinct RNNs followed by a concatenation of both representations, obtaining a consistently lower performance.

[3] we also tried using TF-IDF encoding, obtaining lower performance metrics consistently.

ESIM and a model that incorporates syntactic parsing information in tree LSTMs, Tree-LSTM. Since the Tree-LSTM requires preprocessing tools to obtain the syntactic parsing information, which we argue are not suited for cross-lingual settings targeting less-resourced languages, we only explore the ESIM model in this work. The neural inference model is composed by three major components: input encoding (based on BiLSTMs), local inference modeling, and inference composition; and (c) Peters et al. (2018): a re-implementation of the widely used decomposable attention model developed by Parikh et al. (2016). At the time of development of this work, models (b) and (c) constitute current state-of-the-art models for RTE. We used the code publicly available for both approaches with small modifications in order to make predictions in our binary classification task[4]. These baseline models were employed to obtain a lower-bound for our task and to determine how well existing approaches perform. Since all baselines were originally developed in a monolingual setting, there is no trivial way to employ them as baselines in a cross-lingual setting.

4.3 Cross-Language Learning Techniques

Several approaches have been presented for cross-language learning, including *projection*, *direct transfer*, and *feature space analysis*. As a convention, L_S denotes the source language (in which most of the annotated data is available) and L_T the target language (in which the capability of the system to perform cross-language adaptation will be evaluated, typically containing few or no labeled data).

In projection approaches (Yarowsky et al., 2001; Hwa et al., 2005), annotated data in L_S is projected (by translation) to L_T. More concretely, the learning instances originally in L_S are translated (*e.g.* using machine translation tools or using parallel data) to L_T and the corresponding labels are projected to the new learning instances in L_T. Then, a ML system is trained and evaluated on the projected data in L_T. Typically, fine-grained word alignment techniques are employed to obtain high quality translations and to better preserve the annotation's token-level boundaries. The majority of cross-language learning approaches follow the projection approach. Recent studies, namely (Eger et al., 2018), point out that the quality of current machine translation systems and word alignment tools provide a good basis for projection approaches.

In a direct transfer approach (McDonald et al., 2011), the system is fully trained on the source language L_S, and then the learned model is used to initialize a new model that will work on the target language L_T. If few or no annotated data is available in L_T, the model is used after updating the embedding layer for the target language (using multilingual word embeddings), to make predictions on L_T (unsupervised direct transfer learning). If enough (according to the task) annotated data is available in L_T, the model can be retrained on L_T (after supervised training in L_S) for better adaptation to the target language (supervised direct transfer learning).

Feature space approaches (Bell et al., 2014) perform subspace analysis to find a feature space that can be employed across different languages and at the same time is suitable for the target language.

In this work, we explore the projection and direct transfer approaches. We leave for future work exploring feature space approaches. Regarding the projection approach, we machine translate the ADUs obtained from the Argumentative Essays corpus (Stab and Gurevych, 2017), originally in English, to the target language (*i.e.* Portuguese) using the Google Translator API[5]. Since we formulated the problem as a classification task given two ADUs, the projection of the labels is trivial (no token level alignment is required). Mandatory for the direct transfer approach is the existence of cross-lingual word embeddings, which are trained to obtain a shared embedding space representation of words in different languages. With them, we are able to employ techniques based on word embeddings across different languages. Similarly to monolingual word embeddings, various approaches for learning cross-lingual word embeddings have been proposed in recent years (Ruder, 2017). In this paper, we use pre-trained multilingual embeddings publicly available (Ferreira et al., 2016). The embeddings were obtained by combining parallel data from the TED Corpus with pre-trained English GloVe embeddings[6]. Each embedding contains 300 dimensions.

[4] RTE considers three labels: "neutral", "entailment", and "contradiction".

[5] https://cloud.google.com/translate/

[6] https://nlp.stanford.edu/projects/glove/

4.4 Dealing with unbalanced datasets

As shown in Table 1, the distribution of labels is skewed towards the "none" class. In the presence of unbalanced datasets, ML algorithms tend to favor predictions of the majority class. Aiming to improve the results for the "support" label (minority class), we explore two widely used techniques to deal with unbalanced datasets in ML problems (He and Garcia, 2009): *random undersampling* and *cost-sensitive learning*.

Random undersampling consists of randomly removing examples from the majority class until a predefined number of examples, determined to obtain a balanced dataset in the end of process. In cost-sensitive learning, each class is assigned a weight that works as a penalty cost. Higher/lower costs are used for examples of the minority/majority class, respectively. The ML model is then trained to minimize the total cost, which will become more sensitive to misclassification of examples in the minority class. To determine the weight matrix for each class we follow the heuristic proposed by King and Zeng (2001).

For all the experiments presented in the following sections, these techniques are only applied to the training set (random undersampling) or during the training phase (cost-sensitive learning).

5 Evaluation

In order to validate the main hypothesis proposed in this paper – that the proposed models can capture argumentative relations between ADUs at a semantic-level that is transferable across languages – we have run a set of in-language and cross-language experiments.

Our cross-language experiments use 80% of ADU pairs originally available in L_S as training data and the remaining 20% as test data. In order to tune the parameters of the model, we sample 10% of the training set as the validation data. All splits of the datasets are made at the document-level (*i.e.*, ADU pairs belonging to document $\mathcal{D}$ are not spread in different partitions) and keeping the original distribution of labels in each partition (stratified splitting). Then, the models are evaluated on the full dataset in L_T without retraining (unsupervised language adaptation).

In-language experiments aim to establish baseline scores for a supervised ML system that can make use of annotated resources in L_T. We perform 5-fold cross-validation for in-language experiments. Final scores correspond to the sum of the confusion matrix from the test set predictions in each fold (Forman and Scholz, 2010). Following this procedure, we obtain final evaluation metrics for the full dataset in L_T that are directly comparable with the scores reported on the full dataset for L_T in cross-language experiments, as the evaluation scores are obtained from exactly the same data in both settings. Cross-validation splits are also at the document-level and keep the original label distribution.

Since reporting single performance scores is insufficient to compare non-deterministic learning approaches (Reimers and Gurevych, 2017), we report average scores of 10 runs with different random seeds. Due to the unbalanced nature of the datasets, evaluation metrics reported in the experiments are average macro F1-scores over all 10 runs. All models are trained using the Adam optimizer, using the default parameters suggested in the original paper (Kingma and Ba, 2014), and cross-entropy loss function. The activation function used in all the layers was ReLU (Glorot et al., 2011). To find the best model in each run, we stop training once the accuracy on the validation set does not improve for 5 epochs (early-stop criterion) or 50 epochs are completed. The batch size used in the experiments was set to 32 learning instances. The dimension of the LSTM cell, used by some of the models, was set to 96 after hyperparameter tunning (we tried with 32, 64, 96 and 128). Finally, to accelerate training, we set the maximum length for all ADUs to 50 tokens[7].

5.1 In-Language Results

Table 3 summarizes in-language results obtained for the Argumentative Essays corpus, which contains essays written in English.

Without using any technique to deal with the unbalanced nature of the dataset (upper part of Table 3), results show that all neural network models outperform the baselines. Surprisingly, state-of-art models adopted from the RTE community, namely Peters et al. (2018) and Chen et al. (2017), perform poorly in our task. These results were unexpected because: (a) the tasks are similar (both approaches aim to classify pairs of propositions in similar classes) and (b) the results reported for RTE are quite impressive, namely 0.893 and 0.886

[7] Only 0.2% of ADUs in ArgEssays (Stab and Gurevych, 2017) and 4.5% of ADUs in ArgMine Corpus (Rocha and Lopes Cardoso, 2017) exceed this length.

of accuracy on the SNLI test set, respectively. We hypothesize that despite the similarity of the tasks, the fact that texts have inherently different genres and the datasets different characteristics (label distribution and number of examples) prevents the models proposed for RTE from generalizing well to our task. Results show that the baseline *BoW+LR* is very competitive compared to the neural network architectures. In this setting, the best performing system is *Conv1D*.

Model	***Macro-F1***	***F1-None***	***F1-Supp***
Random	.447	.625	.269
Peters et al. (2018)	.512	.903	.121
Chen et al. (2017)	.577	.879	.275
BoW+LR	.604	.898	.311
LSTM	.606	.877	.336
BiLSTM	.624	.867	.381
Conv1D	.634	.879	.390
Inner-Att	.621	.882	.360
Cost Sensitive Learning			
BoW+LR	**.641**	.875	.407
LSTM	.616	.822	.410
BiLSTM	.634	.835	**.434**
Conv1D	.631	.832	.430
Inner-Att	.606	.822	.410
Random Undersampling			
BoW+LR	.574	.748	.401
LSTM	.566	.734	.399
BiLSTM	.609	.796	.422
Conv1D	.598	.786	.410
Inner-Att	.586	.775	.397

Table 3: In-Language Scores - Arg. Essays (EN). Bold numbers indicate the highest score in the column.

As expected, the skewed nature of the dataset plays an important role in the reported results: scores for the "support" relation are very low compared with scores for "none". We also report experiments conceived to address the unbalanced nature of the dataset, as explained in Section 4.4. We can observe that using cost-sensitive learning we obtained better results for *BoW+LR*, *LSTM* and *BiLSTM*. It is notable that the simple *BoW+LR* approach obtains better results than more complex neural network techniques. We believe this is due to the fact that the number of examples in the dataset is not sufficient to explore the full capabilities of the neural network techniques proposed here (and that have been successful in many other scenarios). Finally, in the cost-sensitive learning setting we obtain the best performance scores for the "support" label, in all models. Regarding random undersampling, results are consistently below those reported using the cost-sensitive learning approach.

The first column in Table 4 summarizes in-language results on the Portuguese ArgMine corpus. We observe similar results compared to the English results reported above. The only exceptions are: (a) *Inner-att* model obtains better results without using balancing techniques, and (b) random undersampling performs better than cost-sensitive learning.

Existing state-of-the-art work on the Argumentative Essays corpus for the subtask of argumentative relation identification reports, as macro F1-scores, 0.751 (Stab and Gurevych, 2017), 0.756 (Nguyen and Litman (2016), in an initial release of the Argumentative Essays corpus containing 90 essays) and 0.767 (Potash et al., 2017). Finally, Eger et al. (2017) reported a F1-score of 0.455 (100% token level match) and 0.501 (50% token level match), but these scores are dependent on the classification of the components in the previous steps (the problem was modeled differently). Therefore, the results reported in Table 3 are worse than state-of-the-art work. The aim of this work is to address the task for a less-resourced language using cross-language learning approaches. Consequently, the main goal is not to propose a novel approach for argumentative relation identification in a monolingual setting. It is important to notice that some of the previous approaches proposed in a monolingual setting do not comply with the proposed approach in this paper: Stab and Gurevych (2017) and Nguyen and Litman (2016) employ different types of features which we argue not to be suitable for cross-language learning targeting less-resourced languages, as extracting these features requires complex linguistic preprocessing tools which cannot be reliably employed in less-resourced languages; and Eger et al. (2017) and Potash et al. (2017) modeled the problem differently by jointly modeling different subtasks of the argumentation mining process.

5.2 Cross-Language Results

Table 4 includes results obtained for cross-language experiments, exploring unsupervised language adaptation techniques (English to Portuguese). Comparing direct transfer and projection approaches, we can observe that projection performs slightly better in most cases. Comparing the scores obtained in the in-language and cross-language settings, we observe that, in general, performance in the cross-language setting improves

	In-Language			Direct Transfer			Projection		
Model	*Macro*	*None*	*Supp*	*Macro*	*None*	*Supp*	*Macro*	*None*	*Supp*
Random	.448	.613	.283	-	-	-	-	-	-
BoW+LR	.457	.888	.025	-	-	-	-	-	-
Peters et al. (2018)	.485	.887	.082	-	-	-	-	-	-
Chen et al. (2017)	.522	.856	.188	-	-	-	-	-	-
LSTM	.489	.868	.110	.461	.887	.036	.462	.884	.041
BiLSTM	.510	.840	.180	.463	.870	.057	.466	.877	.055
Conv1D	.459	.882	.035	.459*	.880	.038	.462*	.884	.039
Inner-Att	**.534**	.764	.305	.454	.883	.025	.456	.882	.030
Cost Sensitive Learning									
BoW+LR	.520	.846	.193	-	-	-	-	-	-
LSTM	.496	.680	.312	.489	.870	.109	.493	.849	.137
BiLSTM	.523	.786	.259	.485	.861	.109	.503	.845	**.162**
Conv1D	.503	.827	.178	.497	.854	.141	.494	.841	.147
Inner-Att	.479	.637	.321	.477	.867	.088	.484*	.844	.123
Random Undersampling									
BoW+LR	.264	.191	.337	-	-	-	-	-	-
LSTM	.494	.668	.321	.494*	.870	.118	.495*	.859	.131
BiLSTM	.464	.581	.348	**.500***	.856	**.145**	**.512***	.865	.158
Conv1D	.423	.554	.292	.499*	.855	.144	.492*	.849	.134
Inner-Att	.487	.621	**.352**	.482	.878	.087	.495*	.861	.128

Table 4: In and Cross-Language scores on the Portuguese (PT) corpus. Bold numbers indicate the highest score in the column. $\star$ = equal or above in-language scores. All metrics correspond to F1-scores.

for the "none" relation and, conversely, drops for the "support" relation. In general, we can observe that the macro-f1 scores of in-language and cross-language approaches are very similar and, in some settings, cross-language macro F1-scores are equal or above in-language scores (marked with the $\star$ symbol in Table 4). Compared to fully-supervised approaches on the target language, such cross-language approaches are able to perform similarly without any annotated data in the target language. These results suggest that transfer learning across languages is possible using the proposed models and that the hypothesis (*i.e.* the argumentative relations between ADUs can be captured in higher-level representations that are transferable) explored in this work is valid.

Regarding the balancing techniques in a cross-language settings, results show that random undersampling works generally better than cost-sensitive learning. Finally, balancing techniques improved the overall scores for all the models.

Similarly to the findings of Eger et al. (2018), we observed better results following the projection approach. As discussed by the authors, it seems that current neural machine translation models have reached a level that makes approaches relying on automated translations feasible and very promising. In this work, the drop in performance using direct transfer was less severe than that of Eger et al. (2018) and very close to the results obtained using the projection approach.

5.3 Error Analysis

To better understand the errors, in particular in cross-lingual scenarios, we selected 5 documents from the ArgMine Corpus (randomly sampled from the set of documents but manually selected to contain false-positive and false-negative examples), comprising a total of 56 ADU pairs for each setting (in-language and cross-language experiments were manually compared).

We noticed that the ArgMine Corpus lacks linguistic indicators of argumentative content (*e.g.* "therefore", "thus", "firstly") that prevail in the Argumentative Essays corpus. This constitutes a consequence of the domain shift between the corpora with potential impact on the performance loss reported in this work. Furthermore, the ArgMine Corpus contains opinionated news articles, which typically require common-sense knowledge and temporal reasoning to identify relations of support (*e.g.* ADU_s: "Greece, last year, tested the tolerance limits of other European taxpayers" and ADU_t: "The European Union of 2016 is no longer the one of 2011.". This example was manually translated from Portuguese to English).

Finally, we also noticed that our deliberative choice of not distinguishing between *linked* and *convergent* arguments (Peldszus and Stede, 2013) led to the problem of including in our dataset

linked arguments with p premises as p ADU pairs. Linked arguments seem to be more prevalent in the ArgMine corpus, and treating them simply as convergent brings us to a problem of partial argumentative relation detection, for which further premises are needed.

6 Conclusions and Future Work

We have presented the first attempt to address the task of argumentative relation identification in a cross-lingual setting. By performing cross-language learning experiments for Portuguese using two popular transfer learning approaches – projection and direct transfer – we have shown that competitive results can be obtained using unsupervised language adaptation, when compared to a fully-supervised machine learning approach on the target language. Experimental results have shown that the cross-lingual transfer loss is relatively small (always below 10%) and, in some settings, transfer learning approaches achieve better scores than fully supervised in-language approaches. These findings demonstrate that suitable higher-level representations of argumentative relations can be obtained that, combined with cross-lingual word embeddings, can be transferred across languages.

In future work, we aim to evaluate the proposed approaches in other languages and explore feature-space analysis techniques recently proposed to address related NLP tasks. Furthermore, we intend to explore multi-task learning techniques, to leverage the knowledge gathered from related tasks (*e.g.* training the models both in argument relation identification and RTE datasets).

Acknowledgments

The first author was partially supported by an Erasmus Placements grant. Part of this work has been supported by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 03VP02540 (ArgumenText).

References

Zeljko Agic and Natalie Schluter. 2018. Baselines and test data for cross-lingual inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Peter Bell, Joris Driesen, and Steve Renals. 2014. Cross-lingual adaptation with multi-task adaptive networks. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 21–25.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. Tweeties squabbling: Positive and negative results in applying argument mining on social media. In *Computational Models of Argument - Proceedings of COMMA 2016, Potsdam, Germany, 12-16 September, 2016*, pages 21–32.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *EMNLP*, pages 632–642. The 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. Association for Computational Linguistics.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual rst discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2013. A natural language bipolar argumentation approach to support users in online debate interactions. *Argument & Computation*, 4(3):209–230.

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. Association for Computational Linguistics.

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

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. Association for Computational Linguistics.

Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1374–1379.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, volume Volume 1: Long Papers, pages 11–22. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844. Association for Computational Linguistics.

Daniel C. Ferreira, André F. T. Martins, and Mariana S. C. Almeida. 2016. Jointly learning to embed and predict with multiple languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2019–2028. Association for Computational Linguistics.

George Forman and Martin Scholz. 2010. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *SIGKDD Explor. Newsl.*, 12(1):49–57.

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.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining on the web from information seeking perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 26–39. CEUR-WS.

Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.*, 21(9):1263–1284.

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

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.*, 11(3):311–325.

Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political Analysis*, 9:137–163.

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

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 62–72, Stroudsburg, PA, USA. 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 - Volume 1*, HLT '11, pages 1336–1345, Stroudsburg, PA, USA. Association for Computational Linguistics.

Huy Nguyen and Diane J. Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *Int. Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proc. Conf. on Empirical Methods in NLP*, pages 938–948, Lisbon, Portugal. ACL.

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 NAACL: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Here's my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1364–1373. Association for Computational Linguistics.

Lorien Pratt and Barbara Jennings. 1996. A survey of transfer between connectionist networks. *Connection Science*, 8(2):163–184.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348. Association for Computational Linguistics.

Gil Rocha and Henrique Lopes Cardoso. 2017. Towards a relation-based argument extraction model for argumentation mining. In *Statistical Language and Speech Processing: 5th International Conference, SLSP 2017, Le Mans, France, October 23–25*, pages 94–105, Cham. Springer International Publishing.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *CoRR*, abs/1509.06664.

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

Mark Sammons, V.G.Vinod Vydiswaran, and Dan Roth. 2012. Recognizing textual entailment. In Daniel M. Bikel and Imed Zitouni, editors, *Multilingual Natural Language Applications: From Theory to Practice*, pages 209–258. Prentice Hall.

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.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Comput. Linguist.*, 43(3):619–659.

Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. *CoRR*, abs/1802.05758.

Bingning Wang, Kang Liu, and Jun Zhao. 2016. Inner attention based recurrent neural networks for answer selection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1288–1297. Association for Computational Linguistics.

Min Xiao and Yuhong Guo. 2013. A novel two-step method for cross language representation learning. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1259–1267. Curran Associates, Inc.

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*, HLT '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.