Learning Strategies for Robust Argument Mining: An Analysis of Variations in Language and Domain

Ramon Ruiz-Dolz[⋆], Chr-Jr Chiu[†], Chung-Chi Chen[‡], Noriko Kando[§], Hsin-Hsi Chen[†]
[⋆]Centre for Argument Technology, University of Dundee, United Kingdom
[†]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
[‡]AIST, Japan
[§]National Institute of Informatics, Japan
rruizdolz001@dundee.ac.uk, ccchiu@nlg.csie.ntu.edu.tw, c.c.chen@acm.org,
kando@nii.ac.jp, hhchen@ntu.edu.tw

Abstract

Argument mining has typically been researched for specific corpora belonging to concrete languages and domains independently in each research work. Human argumentation, however, has domain- and language-dependent linguistic features that determine the content and structure of arguments. Also, when deploying argument mining systems *in the wild*, we might not be able to control some of these features. Therefore, an important aspect that has not been thoroughly investigated in the argument mining literature is the robustness of such systems to variations in language and domain. In this paper, we present a complete analysis across three different languages and three different domains that allow us to have a better understanding on how to leverage the scarce available corpora to design argument mining systems that are more robust to natural language variations.

Keywords: argument relation, argument mining, cross-lingual

1. Introduction

The public availability of annotated data and its quality is one of the major limitations in argument mining. Annotating argumentative labels in natural language is a complex process that requires a large amount of resources and time. Furthermore, different frameworks for argument annotation have been proposed in the literature (Lawrence and Reed, 2014; Stab and Gurevych, 2014; Dusmanu et al., 2017; Naderi and Hirst, 2018; Chen et al., 2021), resulting in a very heterogeneous and limited collection of corpora for argument mining research. The language, the domain, and also the nature of the task (e.g., segmentation, component classification, or relation identification) for which each corpus is created conditions its versatility. This problem is closely related to the high complexity of understanding and analysing argumentation for humans themselves, which is directly reflected in argument mining annotation (Cabrio and Villata, 2018). Therefore, to address this existing heterogeneity in data, we consider that the next step in the argument mining community is to investigate the robustness of the proposed systems from (at least) the language and domain perspectives.

We have observed a recent trend in argument mining research that partially moves into this direction. This is the case of cross-lingual argument mining research. Firstly investigated in Eger et al. (2018), the authors approach the problems of segmentation and classification of arguments considering corpora in three different languages (i.e., English, German, and Chinese) through the use of machine translation techniques. This research was extended in Rocha et al. (2018); Sousa et al. (2021), where the authors approach the segmentation, classification, and identification of argumentative units and relations in a cross-lingual setup where English and machine-translated Portuguese languages are taken into consideration. In Toledo-Ronen et al. (2020), the authors continue with the machine translation approach, and present a multilingual approach to argument mining tasks considering five European languages (i.e., Spanish, French, Italian, German, and Dutch). Recently, in Chung et al. (2021), the authors explore the task of classifying counter narratives in three different European languages. We can observe that most of the cross-lingual research focuses on machine translated versions of the same corpus and consider close languages. However, as stated in Zhao et al. (2021), cross-lingual experiments on languages more distant from English, e.g., Asian languages, may suffer a huge drop in performance. In-depth cross-lingual argument mining research considering a more varied set of languages is still unexplored.

Another line of research that moves into this direction is the cross-domain argument mining. In Ajjour et al. (2017), multiple corpora are combined for both training and testing, and the models are evaluated considering exclusively in-domain data, and considering the cross-domain combination. In Bouslama et al. (2019), the authors propose a similar approach to the cross-domain argument component classification problem, but using a convolutional neural network. A different approach based

LREC-COLING 2024, pages 10286–10292

	Train	Dev./Test	Total	License	Ethical Risks
EN-F	5,521 (3,859/62/1,600)	690 (482/8/200)	6,901 (4,823/78/2,000)	GPL 3.0	Low
CN-F	6,549 (3,676/2,158/684)	819 (460/270/85)	8,187 (4,623/2,710/854)	CC BY-NC-SA 4.0	Low
CAT-D	13,216 (7,566/1,553/4,097)	1,652 (946/194/512)	16,520 (9,458/1,941/5,121)	CC BY-NC-SA 4.0	Low
EN-D	11,710 (6,520/1,676/3,514)	1,464 (815/210/439)	14,638 (8,150/2,096/4,392)	Public	Low
EN-E	4,666 (2,891/175/1,600)	583 (361/22/200)	5,832 (3,613/219/2,000)	Research only	Low
XLD-ARI	41,662 (24,512/5,624/11,495)	5,208 (3,064/704/1,436)	52,078 (30,667/7,044/14,367)	CC BY-NC-SA 4.0	Low

Table 1: Distribution of the argumentative relation samples and classes (support/attack/none) in the XLD-ARI data collection.

on transfer learning is presented in Ruiz-Dolz et al. (2021a), where Transformer-based models are fine-tuned in a single-domain corpus of significantly larger size, and evaluated in five smaller corpora belonging to different domains. The segmentation of arguments from a cross-domain viewpoint has also been approached in Alhamzeh et al. (2021) and Alhamzeh et al. (2022a) where the authors make use of transfer learning techniques and evaluate in corpus outside of the training domain.

This paper provides the first multidimensional analysis on robust argument mining, and presents the following three contributions to the definition of more robust argument mining systems: (1) we propose XLD-ARI, a collection of cross-lingual and cross-domain adapted Argument Relation Identification (ARI) tasks that can be used for measuring the robustness of argument mining systems; (2) we present a complete analysis of different learning strategies aimed at improving both the performance and the robustness of our models; and (3), we release all pre-trained models which can be used as a baseline for future research.

2. Data

We propose a collection of data for exploring model robustness in ARI. Our collection considers the 3-class instance of ARI, consisting of *support*, *attack*, and *none* relations. In addition to the previously existing corpora, we have also included in the collection a new annotated Chinese language corpus belonging to the financial domain for completing our analysis.

2.1. Chinese Financial (CN-F)

Different from the previous financial argument mining corpus using English formal documents (Alhamzeh et al., 2022b), we collected our data from the Chinese financial social media platform, Mobile01[1]. Our corpus includes opinions from a wide spectrum of user backgrounds, professional investors and netizens with little experience in investment. In CN-F, most of the data is short and colloquial, with informal language. For the annotation of the data, we cooperated with an expert working in a securities company. The expert manually annotated all the argumentative pairs in the corpus. It is worth noting that this social media platform had administrators that removed offensive posts manually. Therefore, no offensive posts have been included in this corpus. In the end, we obtained 8,187 pairs with annotations divided into 4,623 supports, 2,710 attacks, and 854 non-related argument pairs. To check their quality, the other annotator with financial background annotated 1,000 randomly-selected pairs. We report a Cohen-Kappa (McHugh, 2012) of 62.21%, which can be interpreted as a substantial agreement (Landis and Koch, 1977) between annotators.

2.2. Cross-Language & Domain for Argument Relation Identification (XLD-ARI)

We release XLD-ARI, a complete collection of annotated corpora structured and pre-processed for providing a consistent environment for evaluating the robustness of argument mining systems. The languages included in XLD-ARI are Chinese (CN), English (EN), and Catalan (CAT). Each of them belongs to a different language family (i.e., Sino-Tibetan, Germanic, and Romance respectively). The selected families present different degrees of similarity, the Romance and the Germanic being closer, while the Sino-Tibetan remains more distant. On the other hand, we included three different domains: Financial (F), Debate (D), and Essay (E). These domains allow us to perform an analysis based on the specificity of each corpus. Therefore, our data collection is composed of the following corpora (from more domain-specific to less): EN-F, CN-F, CAT-D, EN-D, and EN-E. This sorting has been proposed based on the annotation process and data sources used in the creation of each corpus.

The EN-F corpus has been adapted from the FinArg corpus (Alhamzeh et al., 2022b), consisting of 804 different argument-annotated earnings conference calls. The CAT-D corpus has been refined from the VivesDebate corpus (Ruiz-Dolz et al.,

[1] https://www.mobile01.com/topiclist.php?f=793

2021b), that consists of 29 complete spoken debates in Catalan language. The debates belong to a university debate tournament on the ban of surrogacy. The EN-D corpus has been produced from a combination of the US2016 (Visser et al., 2020) and the QT30 (Hautli Janisz et al., 2022) corpora. Both corpora contain argument-annotated political debates in different contexts such as reddit, TV, and radio programs. Finally, the EN-E corpus has been refined from the Argumentative Essays corpus (Stab and Gurevych, 2017), which contains 402 completely annotated argumentative essays. The essays included in this corpus belong to a very heterogeneous set of topics, providing a very varied vocabulary.

A complete summary of the data collection has been depicted in Table 1. The data is released for academic usage and under the CC BY-NC-SA 4.0 license[2], including both publicly available corpora and the annotated CN-F.[3]

3. Methods

In this paper, we explore two different learning techniques for argument mining: Sequential Transfer (ST) learning, and Multi-aspect Learning (MaL). By considering these techniques, it is our objective to improve our understanding on two relevant aspects of low-resource and domain-dependant NLP tasks. First, to define effective cross-corpora learning strategies and leverage limited resources to maximise the improvement. Second, to improve model robustness by considering data belonging to different languages and domains with variable class-distributions.

3.1. Baseline Models

To provide a solid reference that allows us to understand the advantages of the learning strategies analysed in this work, we provide five single task baselines for EN-F, CN-F, CAT-D, EN-D, and EN-E. These baselines have been obtained by fine-tuning the base language model individually in each of the train splits of the five tasks included in XLD-ARI.

3.2. Sequential Transfer Learning (ST)

Under the ST paradigm, we consider two different corpora independently in each experiment, the *source* corpus and the *target* corpus. Therefore, we use the baseline models as the starting point (i.e., *source* model), and fine-tune them with all the possible combinations of the *target* corpora in the XLD-ARI collection. This way, we are able to analyse transfer learning in ARI from a cross-domain (e.g., EN-F $\rightarrow$ EN-D), cross-lingual (e.g., CN-F $\rightarrow$ EN-F), and combined cross-domain and cross-lingual (e.g., CAT-D $\rightarrow$ CN-F) approaches.

3.3. Multi-aspect Learning (MaL)

In the MaL strategy, the complete set of training data is learnt by the model at once rather than learning it in different steps (i.e., ST). With this strategy, the amount of training data is significantly enlarged, but each individual language/domain data distribution also suffers an important variation compared to the data distribution of the corpus combination. Therefore, using the XLD-ARI collection, it is possible to explore multi-domain (MD), multilingual (ML), and a combination of both multi-domain and multilingual (MD & ML) approaches under the MaL paradigm.

4. Experimental Analysis

In order to provide solid results about the robustness, and the learning strategies investigated in this paper, we have run all of our experiments under the same experimental setup with XLM-RoBERTa (Conneau et al., 2020), and use macro F1-score for evaluation.

4.1. Experimental Setup

Given the multilingual nature of the XLD-ARI collection, we have used the XLM-RoBERTa language model (Conneau et al., 2020) as the starting point in all of our experiments. XLM-RoBERTa is a language model pre-trained on a large collection of natural language data containing more than 100 languages, including Chinese, English and Catalan. Therefore, with this language model we can better observe the impact of language and domain variations in performance than if we focus on using specific language models. Even though it is possible that using language models pre-trained on a specific language achieve better results, the availability of such models in underrepresented languages (e.g., Catalan) is very limited. Thus, the purpose of this work is not to beat some specific literature baseline, but to understand how different languages and/or domains can be useful together and detect synergies between these aspects that lead us to achieve significant improvements in situations where obtaining and annotating data is a challenge.

[2]The EN-E license does not allow to publish adapted versions of this data. Instead, we provide a script to generate the EN-E part of the XLD-ARI collection.

[3]XLD-ARI: https://github.com/raruidol/RobustArgumentMining-LREC-COLING-2024

Training Order	EN-F	CN-F	CAT-D	EN-D
Single Task	51.4	**65.0**	65.2	70.2
CN-F → EN-F	49.1	44.4	-	-
EN-F → CN-F	**69.2**	58.5	-	-
CAT-D → EN-D	-	-	58.7	70.2
EN-D → CAT-D	-	-	**68.1**	63.8

Table 2: Cross-Lingual Results.

Training Order	EN-F	EN-D	EN-E
Single Task	51.4	70.2	47.4
EN-F → EN-D	**74.9**	61.7	-
EN-D → EN-F	**57.0**	59.7	-
EN-E → EN-D	-	58.1	39.0
EN-D → EN-E	-	64.6	**59.5**
EN-E → EN-F	50.1	-	40.6
EN-F → EN-E	**88.2**	-	**56.6**

Table 3: Cross-Domain Results.

4.2. Cross-Corpora Learning Strategies

The observed results have been depicted in Tables 2, 3, and 4 for the cross-linugal, cross-domain, and the MaL experiments respectively. A complete summary of the improvement achieved by these learning strategies is presented in Table 5. These results lead us to three interesting observations for low-resource NLP problems. First of all, we can observe that having more training data has not always positive implications. We observed that in these cases where language is very specific to a given domain (e.g., EN-F) it is better to perform a sequential transfer step on more generic language data instead of including all the data for the training process. Conversely, in these cases where our domain contains more generic language (e.g., EN-E), a greater improvement is observed by extending the training data to new domains. Second, we observed that learning across languages within a unique domain in low-resource problems is harder than learning across domains within a unique language. Finally, in most of the cases we observed that model performance on smaller corpora could easily improve from extending the training data with larger corpora, but not vice versa. This behaviour can be caused by the impact on the data distribution that larger corpora have over the smaller ones.

4.3. Model Robustness

For analysing the robustness of models, we have focused on the test partitions of tasks *unseen* during training. Table 6 shows the performance of the best model for each task when evaluated with the non-learnt tasks. We have compared the best performing model in each task against the single task baselines' performance on non-learnt corpora. Note that MD & ML (CAT-D) was trained with all the tasks, and thus we could not compare it with the results of the CAT-D single task baseline model.

	EN-F	CN-F	CAT-D	EN-D	EN-E
Single Task	51.4	65.0	65.2	70.2	47.4
MD	**54.7**	-	-	**70.4**	**61.9**
ML	**52.6**	**72.3**	-	-	-
MD & ML	50.7	**66.2**	**68.4**	67.7	**60.0**

Table 4: Multi-domain and Multilingual Results.

	Best-Performing Model	Improvement Macro-F1	Improvement Ratio
EN-F	ST (Cross-Domain)	+36.8	71.60%
CN-F	MaL (ML)	+7.3	11.23%
CAT-D	MaL (MD & ML)	+3.2	4.91%
EN-D	MaL (MD)	+0.2	0.28%
EN-E	MaL (MD)	+14.5	30.59%

Table 5: Summary of Cross-Corpora Learning.

We can observe a generalised improvement on model robustness for argument mining. What is particularly interesting is the result of the multilingual model that performed the best in CN-F, being able to significantly outperform the CN-F single task baseline on unseen tasks such as the CAT-D, EN-D and EN-E. The CN-F baseline had a particular bad performance in these tasks, mainly due to the language family differences. However, by only including one of the smaller corpora available in the XLD-ARI collection in the training process (i.e., EN-F), the robustness of the model improved substantially.

5. Discussion

5.1. Potential of Learning Strategies

In this section, we demonstrate the efficacy of various learning strategies by comparing their performance with that of prior models. The models EN-F and CN-F participated in the open competition, FinArg, at NTCIR-17 (Chen et al., 2023). Official evaluations were conducted on the outputs of 19 and 18 systems for EN-F and CN-F, respectively. The highest scores achieved in this competition for both datasets were 61.50% for EN-F and 73.94% for CN-F. By employing the learning strategies discussed in this paper, we attained an enhanced performance of 88.20% for EN-F and 72.30% for CN-F. These results surpass previous methods in EN-F and align closely with the performances of methods specifically designed for the CN-F. This outcome underscores the advantage of systematically exploring existing datasets through varied strategies.

	Model	EN-F	CN-F	CAT-D	EN-D	EN-E
EN-F	Single Task	-	10.3	26.5	33.9	36.5
	Cross-Domain	-	**16.6**	**37.1**	**40.7**	-
CN-F	Single Task	17.5	-	25.4	25.4	26.3
	ML	-	-	**40.2**	**46.7**	**40.7**
EN-D	Single Task	41.0	28.0	44.9	-	43.1
	MD	-	**27.7**	**45.5**	-	-
EN-E	Single Task	26.9	20.6	33.6	35.4	-
	MD	-	**27.7**	**45.5**	-	-

Table 6: Robustness of best performing models.

1st Task	2nd Task	SS	PRS	Difference
EN-D	EN-F	59.7	67.5	**13.02%**
	CAT-D	63.8	63.9	**0.06%**
	EN-E	64.6	65.1	**0.77%**
EN-E	EN-F	40.6	40.3	-0.67%
	CAT-D	43.7	43.3	-0.98%
	EN-D	39.0	42.2	**8.12%**
EN-F	CAT-D	81.5	86.7	**6.39%**
	EN-E	88.8	89.4	**0.66%**
	EN-D	74.9	70.3	-6.11%
CAT-D	EN-F	47.5	59.3	**24.76%**
	EN-D	67.8	57.5	-15.19%
	EN-E	53.2	57.5	**8.10%**

Table 7: Comparison of review strategies.

5.2. Review Strategy

Humans learn sequentially and continually, but also require periodic brushing up of learned tasks. Various strategies are employed by individuals to review learned knowledge. However, previous studies have scarcely addressed the review strategies of models. Consequently, this paper concentrates on discerning the effective review strategy during continual learning. We experiment with two strategies: the standard strategy (SS) and the partial review strategy (PRS).

The standard strategy involves updating both the language model (LM) and the multilayer perceptron (MLP) layers during the review of the learned task, as described in the previous sections. This approach necessitates the most extended training time as it requires updating all parameters in the model.

Drawing on human learning experiences, where students often review exercises related to specific subjects before exams—refreshing their memory on the target subject —we designed an alternative review strategy called PRS. During the review step, PRS updates only the last MLP layer, thus refreshing the learned knowledge about the target task. The PRS also in line with the idea in Serra et al. (2018), which froze some neurons when learning a new task. Specifically, during the review step, we freeze the LM encoder, which has already been sequentially trained with different tasks, and initialize a new MLP classifier. This classifier integrates the encoder knowledge from all previously learned tasks and optimizes it specifically for the early learned task. In this manner, the model does not need to update all the parameters, but only those belonging to the MLP classifier.

In accordance with the findings of McCloskey and Cohen (1989), models often encounter a catastrophic forgetting problem, resulting in a decrease in performance on previously learned tasks after training on a new task. Therefore, we placed particular emphasis on evaluating the performance of the first learned tasks across different review strategies, as presented in Table 7.

The initial analysis reveals that the PRS outperforms the SS in eight out of the twelve experiments, indicating that PRS is a more effective review strategy for the ARI task. Notably, PRS consistently achieves superior performance compared to SS when the EN-D dataset is used as the first task, regardless of the learning order. This suggests that PRS should be the preferred choice for minimizing the forgetting rate after training on EN-D.

Furthermore, the performance difference between the review strategies appears to vary depending on the similarity of the tasks. When the second task is more similar to the first task, the performance difference is relatively small. For instance, in the EN-D -> CAT-D task, the accuracy difference between SS and PRS is only 0.06%. However, when the tasks are more dissimilar, the performance difference becomes more significant. For example, in the CAT-D -> EN-F task, PRS achieves a remarkable improvement of 24.76% compared to SS. Interestingly, when the CAT-D dataset is the first task, the difference between SS and PRS becomes noticeably larger compared to other experiments involving datasets of the same language. Additionally, we observed significant performance fluctuations when modifying the learning order, such as in the EN-D -> EN-E and EN-E -> EN-D tasks, regardless of the employed review strategy.

6. Conclusion

The study provides an inaugural examination of robust argument mining models across varied dimensions, namely language and domain. Our findings reveal notable patterns that can enhance learning strategies by considering linguistic families and the unique linguistic attributes of each corpus. This research lays the foundation for subsequent studies aimed at developing more informed learning strategies, optimizing the use of limited data, and crafting resilient argument mining models. All the resulting models have been publicly shared at https://huggingface.co/raruidol, and can be used as the starting point for addressing argument mining tasks in new languages and/or domains which remain as a pending analysis to be conducted in future work.

Acknowledgements

The work of Ramon Ruiz-Dolz was supported by the NII International Internship Program 2019 and by the 'AI for Citizen Intelligence Coaching against Disinformation (TITAN)' project, funded by the EU Horizon 2020 research and innovation programme under grant agreement 101070658, and by UK Research and innovation under the UK governments Horizon funding guarantee grant numbers 10040483 and 10055990. The work of Noriko Kando was supported by the JSPS Grant-in-Aid 19H04420 and 23H03686. The work of Chr-Jr Chiu and Hsin-Hsi Chen was supported by National Science and Technology Council, Taiwan, under grants MOST 110-2221-E-002-128-MY3, NSTC 112-2634-F-002-005 -, and Ministry of Education (MOE) in Taiwan, under grants NTU-112L900901. The work of Chung-Chi Chen was supported in part by JSPS KAKENHI Grant Number 23K16956 and a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Bibliographical References

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128.

Alaa Alhamzeh, Mohamed Bouhaouel, Előd Egyed-Zsigmond, Jelena Mitrović, Lionel Brunie, and Harald Kosch. 2021. A stacking approach for cross-domain argument identification. In *International Conference on Database and Expert Systems Applications*, pages 361–373. Springer.

Alaa Alhamzeh, Előd Egyed-Zsigmond, Dorra El Mekki, Abderrazzak El Khayari, Jelena Mitrović, Lionel Brunie, and Harald Kosch. 2022a. Empirical study of the model generalization for argument mining in cross-domain and cross-topic settings. In *Transactions on Large-Scale Data- and Knowledge-Centered Systems LII*, pages 103–126. Springer.

Alaa Alhamzeh, Romain Fonck, Erwan Versmee, Elod Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022b. It's time to reason: Annotating argumentation structures in financial earnings calls: The finarg dataset. In *Proceedings of the 4th Workshop on Financial Technology and Natural Language Processing*, pages 15–21, Vienna, Austria.

Rihab Bouslama, Raouia Ayachi, and Nahla Ben Amor. 2019. Using convolutional neural network in cross-domain argumentation mining framework. In *International Conference on Scalable Uncertainty Management*, pages 355–367. Springer.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. *From opinion mining to financial argument mining*. Springer Nature.

Chung-Chi Chen, Chin-Yi Lin, Chr-Jr Chiu, Hen-Hsen Huang, Alaa Alhamzeh, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2023. Overview of the ntcir-17 finarg-1 task: Fine-grained argument understanding in financial analysis. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan*.

Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2021. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. 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.

Annette Hautli Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. Qt30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC2022). ACL*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

John Lawrence and Chris Reed. 2014. Aifdb corpora. In *Computational Models of Argument - Proceedings of COMMA 2014*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, pages 465–466. IOS Press.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Nona Naderi and Graeme Hirst. 2018. Automated fact-checking of claims in argumentative parliamentary debates. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 60–65, Brussels, Belgium. Association for Computational Linguistics.

Gil Rocha, Christian Stab, Henrique Lopes Cardoso, and Iryna Gurevych. 2018. Cross-lingual argumentative relation identification: from english to portuguese. In *Proceedings of the 5th Workshop on Argument Mining*, pages 144–154.

Ramon Ruiz-Dolz, Jose Alemany, Stella M Heras Barberá, and Ana García-Fornes. 2021a. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(6):62–70.

Ramon Ruiz-Dolz, Montserrat Nofre, Mariona Taulé, Stella Heras, and Ana García-Fornes. 2021b. Vivesdebate: A new annotated multilingual corpus of argumentation in a debate tournament. *Applied Sciences*, 11(15):7160.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. PMLR.

Afonso Sousa, Bernardo Leite, Gil Rocha, and Henrique Lopes Cardoso. 2021. Cross-lingual annotation projection for argument mining in portuguese. In *EPIA Conference on Artificial Intelligence*, pages 752–765. Springer.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 303–317.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767.

7. Appendix

7.1. Implementation Details

All the experiments reported in this paper have been carried out under the same experimental setup. We fine-tuned the XLM-RoBERTa language model on the XLD-ARI collection following the two learning strategies (i.e., ST and MaL) for 100 epochs with a learning rate of 1e-7 and a weight decay of 0.01 until model convergence. In these cases where the training data was significantly larger (e.g., MaL), we extended the total number of epochs by 100 to allow the resulting model to converge. We used the Hugging Face[4] library for the implementation of our experiments. The performance was evaluated with the macro averaged F1 score, considering the high class imbalance present in our data.

[4] https://huggingface.co/docs