Transformer-Based Models for Automatic Identification of Argument Relations: A Cross-Domain Evaluation

Ramon Ruiz-Dolz, Jose Alemany, Stella M. Heras Barberá, and Ana García-Fornes, *Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Valencia, 46022, Spain*

Argument mining is defined as the task of automatically identifying and extracting argumentative components (e.g., premises, claims, etc.) and detecting the existing relations among them (i.e., support, attack, rephrase, no relation). One of the main issues when approaching this problem is the lack of data, and the size of the publicly available corpora. In this work, we use the recently annotated US2016 debate corpus. US2016 is the largest existing argument annotated corpus, which allows exploring the benefits of the most recent advances in natural language processing in a complex domain like argument (relation) mining. We present an exhaustive analysis of the behavior of transformer-based models (i.e., BERT, XLNET, RoBERTa, DistilBERT, and ALBERT) when predicting argument relations.Finally, we evaluate the models in five different domains, with the objective of finding the less domain-dependent model. We obtain a macro F1-score of 0.70 with the US2016 evaluation corpus, and a macro F1-score of 0.61 with the Moral Maze cross-domain corpus.

Computational argumentation has proved to be a very solid way to approach several problems such as fake news detection,[5] recommendation systems,[14] or debate analysis[4] among others. However, in almost every domain, it is of great importance to be able to automatically extract the arguments and their relations from the input source. Argument mining (AM) is the natural language processing (NLP) task by which this problem is addressed. The Transformer model architecture[17] and its subsequent pretraining approaches have been a turning point in the NLP research area. Thanks to its architecture, it has been possible to capture longer-range dependencies between input structures and, thus, the performance of systems developed for the most general NLP tasks (i.e., translation, text generation, or language understanding) improved significantly. Therefore, the Transformer architecture has laid the foundations on which newer models and pretraining approaches have been proposed, defining the state of the art in NLP. In this work, we analyze the behavior of BERT,[3] XLNET,[20] RoBERTa,[9] DistilBERT,[15] and ALBERT[6] when facing the hardest AM task: identifying relational properties between arguments.

AM was formally defined in the paper by Palau and Moens[13] as the task that aims to automatically detect arguments, relations, and their internal structure. As pointed out in the paper by Lawrence and Reed,[8] due to the complexity of AM, the whole task can be decomposed into three main subtasks depending on their argumentative complexity. First, the identification of argument components consists in distinguishing argumentative propositions from nonargumentative propositions. This allows to segment the input text into arguments, making it possible to carry out the subsequent subtasks. Second, the identification of clausal properties is the part of AM that focuses on finding premises or conclusions among the argumentative propositions. Third, the last subtask is the identification of relational properties. Two different argumentative propositions are considered at a time, and the main objective is to identify which type of relation links both propositions. Different relations can be observed in argument analysis, from the classical attack/support binary analysis,[2] to the identification of complex

1541-1672 © 2021 IEEE
Digital Object Identifier 10.1109/MIS.2021.3073993
Date of publication 19 April 2021; date of current version 17 December 2021.

patterns of human reasoning (i.e., argumentation schemes[19]). Therefore, the identification of argumentative relations is the most complex part of AM,[8] but its complexity may vary depending on how the problem is instantiated.

One of the main problems when addressing any AM task is the lack of high-quality annotated data. In fact, as the argumentative complexity of the task increases, it gets harder to find large enough corpus to do experiments that match the latest NLP advances. An important feature that characterizes the transformer-based models is that large corpora are needed to achieve the performance improvement mentioned above. Recently, in the paper by Visser *et al.*,[18] a new argument annotated corpus of the United States 2016 debate (*US2016*) was published. This corpus contains data from the transcripts of the televised political debates and from Internet debates generated around the same context. This is the first publicly available corpus with enough data to begin exploring the benefits of the most recent contributions in NLP, when applied to the identification of argumentative relations. Additionally, the *US2016* corpus has been annotated using Inference Anchoring Theory (IAT[a]), a standard argument annotation guideline that provides more information than the classic attack/support binary annotation. Learning a model to automatically annotate with the use of IAT makes it possible to evaluate its performance not only with the test samples of the corpus, but also with other different corpus already analyzed and tagged using this standard (e.g., *Moral Maze* corpus). This way, it is our objective to both: evaluate the performance of these new models in the identification of argument relations task; and to find out which one is more robust to variations in the application domain.

In this work, we explore the benefits of the most recent advances in NLP applied to relation prediction in the AM domain. For this purpose, we use the recently published *US2016* corpus, since it is to the best of our knowledge, the largest annotated corpus containing information about argumentative relations, and the *Moral Maze* cross-domain corpus. Then we do: (i) a preprocessing of the corpus in order to clean and structure the data for the requirements of our experiments; (ii) an analysis of the performance of the most relevant transformer-based models (i.e., BERT, XLNET, RoBERTa, DistilBERT, and ALBERT) when learning to predict the relations between argumentative propositions defined by the IAT standard; and (iii) an evaluation of the obtained models in five different domains (*Moral Maze* corpus) with the objective of analyzing the domain dependence of the transformer-based models when facing this AM task.

[a]https://typo.uni-konstanz.de/add-up/wp-content/uploads/2018/04/IAT-CI-Guidelines.pdf

RELATED WORK

AM is one of the main research areas in Computational Argumentation. AM has caught the attention of many researchers since it is considered to be the first step toward autonomous argumentative systems. We identified many different approaches to the Argument (relation) mining problem, which depend on the proposed methods [i.e., parsing algorithms, textual entailment suites, logistic regression, support vector machines (SVMs), and neural networks], and the available corpus at each moment. Initial research on automatic identification of argument relations was done in the paper by Palau and Moens,[13] where parsing algorithms were used to determine the type of relation existing between two argument propositions. Some years later, AM started to gain relevance in the NLP community. We can observe the popularization of machine learning techniques for NLP purposes in the papers by Naderi and Hirst,[11] Stab and Gurevych,[16] and Menini *et al.*[10] SVMs seemed to be the best performing machine learning technique for the purpose of argument relation identification. With the advent of neural networks (NNs), a performance gap between previous works and this new approach could be observed. In the paper by Cocarascu and Toni,[2] the empirical results obtained by recurrent neural network (RNN) models for AM were significantly better. However, there is an interesting observation to make emphasis on, which makes it hard to compare AM works. As it can be pointed out after looking at the results depicted in works like the papers by Niculae *et al.*[12] or Cocarascu and Toni,[2] the corpus used in each work has a lot of influence in the results. This is due to many different factors such as class distributions, variable language complexity (e.g., use of irony, enthymemes, etc.) or the own size of the corpus. Therefore, misleading results may be observed if the generalization of the model is not properly evaluated. On the other hand, deep learning algorithms require much more data to significantly increase its performance compared to classic neural, machine learning, or statistical methods. Therefore, from all these past years of argument relation identification works, the performance has been improved not only with the use of new models or techniques, but also with the creation of better corpora. In the paper by Visser *et al.*,[18] a new argument annotated corpus (*US2016*) was

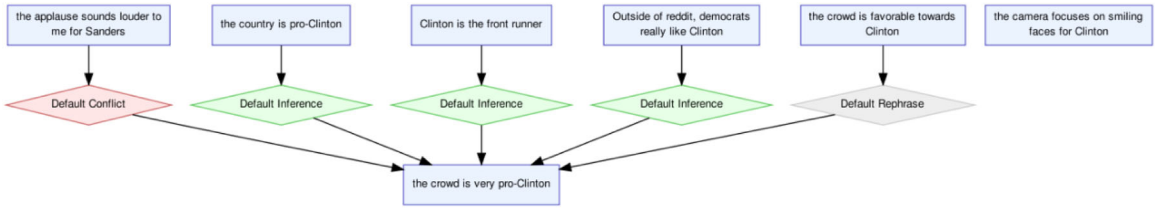

FIGURE 1. *US2016* Argument Map Sample. Argument Discourse Units (ADUs) are bounded by rectangles. Relation types are contained in the rhombuses.

published, with enough data to begin exploiting the benefits of the most recent advances in NLP (i.e., transformer-based models) in the AM domain. To the best of our knowledge, this is the first work addressing the Argument (relation) Mining problem using Transformer-based models in more than a unique domain.

DATA

Two different corpora have been used in this work: the *US2016* debate corpus and the *Moral Maze* multidomain corpus. Both corpora can be downloaded from the Argument Interchange Format Database (AIFdb), an initiative of researchers from the ARG-tech[b] with the objective of creating a standard formatted argument corpus database.[7] This database contains 193 different argumentative corpora structured using the AIF standard. Each corpus is divided into several argument maps (see Figure 1), and each argument map contains a set of Argument Discourse Units (ADUs) with its argumentative relations annotated using the IAT. This annotation method considers the most important three argumentative relations: inference (RA), conflict (CA), and rephrase (MA). An inference relation between two propositions determines that one is used to support or justify the other; a conflict relation indicates that two propositions have contradictory information; and a rephrase between two propositions means that they are equivalent from an argumentative point of view.

In order to adapt the AIFdb corpus to the needs of this task, we did some preprocessing. Each argument map is stored in a JSON file, and represented as a graph following the AIF standard. We generate a unique tab-separated values file per corpus containing three different values: proposition1, proposition2, and label. In addition to the existing IAT relation labels, we decided to generate an additional relation: the no relation (NO) label. Since most of the pairs of propositions found in a debate are not related, we decided to generate a 65% of samples belonging to this new class. For this purpose, we mixed up the propositions that were not annotated with any of the IAT relation classes. This way, the resulting model will also be able to discriminate between related or not related propositions.

US2016 Debate Corpus

The *US2016*[c] corpus is an argument annotated corpus of the electoral debate carried out in 2016 in the United States. It contains both, transcriptions of the different rounds of TV debate, and discussions from the Reddit forums as detailed in the paper by Visser *et al.*[18] The class distribution of the processed *US2016* corpus is depicted in Table 1. Since it is the largest publicly available argument annotated corpus in the literature, we used it to train the models. We decided to split the corpus with the 80% of the proposition pairs for training, and the remaining 20% for the evaluation.

Moral Maze Multidomain Corpus

The *Moral Maze*[d] multidomain corpus is an argument annotated corpus obtained from the transcriptions of the 2012 Moral Maze BBC discussion show. This corpus has been built from samples collected in five different broadcasts. The class distribution of the processed *Moral Maze* corpus is depicted in Table 2. This corpus is used to evaluate the domain robustness of the trained models across five different domain corpus: Bank (B), Empire (E), Money (M), Problem (P), and Welfare (W); each one focused on a specific debate topic and with a different distribution of classes.

AUTOMATIC IDENTIFICATION OF RELATIONAL PROPERTIES

The problem addressed in this article can be seen as an instance of the sentence pair classification

[b]https://www.arg-tech.org/

[c]http://corpora.aifdb.org/US2016
[d]http://corpora.aifdb.org/mm2012

TABLE 1. Class distribution of the *US2016* corpus, train, and test partitions.

	US2016	Train	Test
RA	2,744	2,195	549
CA	888	710	178
MA	705	564	141
NO	8,055	6,444	1,611
Total	12,392	9,913	2,479

TABLE 2. Multidomain evaluation corpus (moral maze) class distribution.

	MM2012	B	E	M	P	W
RA	833	128	121	205	192	187
CA	200	26	36	30	45	63
MA	156	3	25	48	41	39
NO	2,209	292	339	526	517	537
Total	3,398	449	521	810	795	826

problem. The sentence pair classification problem consists of assigning the most likely class to two text inputs at a time. In AM, after segmenting the text and defining the argument components, the argument graph must be built by identifying the relational properties between every two argument components. Therefore, given two argument components (i.e., sentences): $x_1^N = x_1, x_2, \ldots, x_N$ of length N, where x_n is each word of the first component; and $y_1^M = y_1, y_2, \ldots, y_M$ of length M where y_m is each word of the second component, the classification problem can be modeled as defined in the following:

$$\hat{c} = arg\,max_{c \in C} p(c|x_1^N, y_1^M) \quad (1)$$

where $C = [\mathrm{RA}, \mathrm{CA}, \mathrm{MA}, \mathrm{NO}]$, so the four different relation types existing in the IAT labeling are considered: inference (RA), conflict (CA), rephrase (MA), and no relation (NO). To approach this problem, we decided to use transformer-based neural architectures. The most recent works in the literature tackle the AM problem using RNNs (e.g., LSTMs, BiLSTMs, etc.). However, the Transformer architecture presents several interesting improvements with respect to the RNNs. The Transformer architecture uses multiple attention modules, which allow to capture longer range dependencies between words in a sentence. Given the nature of this work's task, we expect to have long input sentences since argumentative text is, generally, more complex than others. Therefore, we think attention mechanisms can be very useful for the identification of relational properties between argument components.

In this work, we apply Inductive Transfer Learning combined with different Transformer pretraining methods that allow us to learn our task not from scratch but using previously calculated weights. We decided to use the pretraining methods that performed the best in other NLP tasks such as Natural Language Understanding, Question Answering or Text Generation: BERT,[3] XLNET,[20] RoBERTa,[9] DistilBERT,[15] and ALBERT.[6] All these models have in common that they are based on the Transformer architecture; however, different approaches have been considered in order to compute the initial weights. BERT, also known as Bidirectional Encoder Representations from Transformers, is pretrained on the masked language model and next sentence prediction tasks. The model is designed to be able to fine-tune its weights on other different tasks by adding an additional output layer. XLNet is proposed after identifying a potential problem in BERT: the language modeling of the existing dependencies between the masked positions. XLNet combines both auto-regressive language modeling and auto-encoding techniques in order to overcome the detected potential issues. RoBERTa is an strong optimization of the BERT pretraining approach. After researchers did a thorough analysis on the impact of the most important hyperparameters, this new model was able to obtain interesting results in most of the evaluated tasks. Finally, both DistilBERT and ALBERT were proposed as smaller and faster versions of the previous approaches. We find it interesting to also analyze and evaluate the behavior of these smaller versions, which have been designed to democratize the use of transformer-based pretraining methods without significant loss of performance.

EVALUATION

Experimental Setup

All the experiments carried out in this work have been run in a double NVIDIA Titan V computer with an Intel Xeon W-2123 CPU and 62 Gb of RAM. This way, we can evaluate not only the performance of the models in the classification task, but also their training computational cost in our specific task. The number of parameters of each model is directly related to the training computational cost. Table 3 summarizes the most relevant features that define each Transformer architecture considered in this research. The Transformer blocks (TBlocks) stand for the number of layers; the

TABLE 3. Transformer-based architectures configuration.

Model	TBlocks	HSize	AH	Params.
BERT-base[3]	12	768	12	110 M
BERT-large[3]	24	1,024	16	340 M
XLNet-base[20]	12	768	12	110 M
XLNet-large[20]	24	1,024	16	340 M
RoBERTa-base[9]	12	768	12	125 M
RoBERTa-large[9]	24	1,024	16	335 M
DistilBERT-base[15]	6	768	12	66 M
ALBERT-base[6]	12	768	12	11 M
ALBERT-xxlarge[6]	12	4,096	64	223 M

hidden size (HSize) represents the number of hidden states in each layer; the attention heads (AH) indicate the number of pointers used by the attention layers; finally, the last feature is the total number of parameters (Params.) of each architecture.

In our experiments, we explore the benefits of transfer learning applied to the argument relation mining task. For that purpose, during our training phase, we use the pretrained encoder of each model with a linear layer on its top. The output size of the linear layer coincides with the number of classes considered in our instance of the problem (i.e., 4). With the *softmax* function, we are able to model the probability of belonging to one class or another for each pair of arguments (1).

We adapted the maximum sequence length and the batch size of our inputs in each experiment. These parameters were configured in order to use the whole available GPU memory. When training BERT-base models, we defined a maximum sequence length of 256 and a batch size of 64. When training BERT-large models, we halved those values to a maximum sequence length of 128 and a batch size of 32. We trained XLNet-base with a maximum sequence length of 256 and a batch size of 32, and XLNet-large with a maximum sequence length of 256 and a batch size of 8. RoBERTa-base was trained with a maximum sequence length of 256 and batch size of 32, and for training RoBERTa-large we used the same maximum sequence length but a batch size of 16. For DistilBERT we used a maximum sequence length of 256 and a batch size of 128. Finally, ALBERT-base was trained defining a maximum sequence length of 256 and batch size of 64, but in order to fit ALBERT-xxlarge in our available memory we had to define a maximum sequence length of 128 and a batch size of 4. We trained all these models for 50 epochs in our corpus. The best results (depicted in the following section) were obtained with a 1e-5 learning rate.

Results

In this section, we present the empirical results obtained after running the experiments on all the previously defined models. In addition to the Transformer-based architectures, we have also trained a RNN as a baseline in our task. We used the best performing RNN architecture in argument relation mining proposed in the paper by Cocarascu and Toni,[2] consisting of two long short-term memory (LSTM) networks working in parallel with each pair of arguments. We trained the baseline model for 50 epochs in our data, as the authors did in the original publication. In order to measure the performance of the different models, we have evaluated them using the macro F1-score metric. Due to the huge class imbalance in our corpora, the use of the macro F1-score makes possible to avoid misleading results during the evaluation. Additionally, we also measured the training time required by each model when learning the task proposed in this work, in order to analyze if it can be worthwhile to sacrifice their performance in pursuit of faster training times or availability in lower resource environments.

The macro F1-scores obtained by each model are depicted in Table 4. In the first column, we can see every trained model. The second column represents the macro F1 obtained by each model when evaluated with the test partition of the same corpus used for training (i.e., *US2016*). The third column contains the scores obtained when the evaluation is performed on a different corpus (i.e., *MM2012*) containing a mixture of five domains. Finally, the last five columns are the macro F1-scores of the models when using each one of the five domain specific corpora (i.e., *Bank*, *Empire*, *Money*, *Problem,* and *Welfare*) for evaluation.

With most of the models, we achieved the state-of-the-art macro F1-scores for relation identification in AM.[1] Here, it is important to make emphasis that the way we considered to represent argumentative relations (i.e., IAT labeling) make this task harder than most of the previous work (i.e., attack/support) in this area. We obtained a 0.70 macro F1-score with *RoBERTa-large*, outperforming the LSTM baseline used as a reference of previous research in argument relation identification. Furthermore, in order to have a more strong reference to compare with previous published results, we carried out an experiment using the same parameters but considering a binary instance of the problem (i.e., only attack and support relations). This

TABLE 4. Performance of the models in the automatic identification of argument relations, given in macro F1-scores.

Experiment	*US2016*-test	MM2012	Bank	Empire	Money	Problem	Welfare
LSTM (baseline)	0.26	0.24	0.25	0.22	0.24	0.25	0.23
BERT-base-cased	0.62	0.53	0.40	0.45	0.54	0.47	0.53
BERT-base-uncased	0.65	0.56	0.42	0.48	0.54	0.50	0.54
BERT-large-cased	0.61	0.55	0.45	0.49	0.53	0.47	0.51
BERT-large-uncased	0.66	0.57	0.47	0.49	0.56	0.49	0.57
XLNet-base	0.65	0.56	0.44	0.49	0.51	0.54	0.55
XLNet-large	0.69	0.57	0.44	0.51	0.53	0.53	0.54
RoBERTa-base	0.68	0.58	0.51	0.52	0.54	0.52	0.58
RoBERTa-large	**0.70**	**0.61**	**0.53**	0.53	**0.59**	**0.56**	**0.59**
DistilBERT	0.55	0.42	0.33	0.39	0.40	0.43	0.39
ALBERT-base-v2	0.60	0.54	0.49	0.45	0.53	0.47	0.51
ALBERT-xxlarge-v2	0.67	0.59	0.50	**0.54**	0.56	0.48	**0.59**

way, *RoBERTa-large* achieved a macro F1-score of 0.81 highlighting the mentioned complexity gap between the two instances of the same problem. In general, we can observe that *RoBERTa* has performed very well in this task. When looking at the cross-domain evaluation, *RoBERTa-large* has also performed the best. We obtained a 0.61 macro F1-score when doing the evaluation with a different domain corpus. Moreover, the model has been able to keep a good performance with each one of the five domain specific corpora, even having different class distributions. With *ALBERT-xxlarge-v2*, it has been possible to obtain a slightly better performance when evaluating with the *Empire* corpus. It is possible to observe how the scores obtained on the *Bank* and *Empire* corpus are slightly lower than the rest. This is mainly due to their smaller size, combined with the strong imbalance between classes. We also did experiments with cased and uncased models, in order to see the relevance of cased text in the relation identification task. As we can observe, the uncased models performed significantly better than the cased models, so we can point out that cased text did not help to improve the performance of the models in our task.

On the other hand, we obtained the worst results with *DistilBERT* and *ALBERT-base-v2*, as one might expect. We decided to use these models in order to see if the observed performance sacrifice was worthwhile in exchange for more feasible training times. Table 5 contains the training times required by each model under our experimental setup. With *DistilBERT*, it was possible to achieve a significant reduction of training time in exchange for a huge drop in performance. However, with *ALBERT-base-v2,* we could not observe a significant reduction of training time. From our experiments, we have not seen any significant advantage in using these *lite* models. We also observed that the computational cost of training *XLNet-large* and *ALBERT-xxlarge-v2* in our task was very expensive. *XLNet-large* was 5.1 times slower to train than *BERT-large*. As for *ALBERT-xxlarge-v2*, the training time was 7.1 times higher than *BERT-large*. This is due to its hidden size of 4,096 with respect to the 1,024 sized large models. Thus, observing the performance of the models in means of their macro F1-score and the required training time, we still think that *RoBERTa* is the best approach to tackle both, domain-

TABLE 5. Training time of 50 epochs running in a double NVIDIA titan v computer.

Experiment	Time
BERT-base	39 m 11 s
BERT-large	2 h 19 m 57 s
XLNet-base	1 h 52 m 38 s
XLNet-large	11 h 51 m 09 s
RoBERTa-base	43 m 17 s
RoBERTa-large	4 h 44 m 33 s
DistilBERT	16 m 15 s
ALBERT-base-v2	38 m 04 s
ALBERT-xxlarge-v2	16 h 20 m 22 s

TABLE 6. Distribution of the misclassified samples per class using the *roberta-Large* model.

Pred. \ Real	RA	CA	MA	NO
RA	-	0.512	**0.603**	**0.730**
CA	0.200	-	0.138	0.226
MA	0.100	0.075	-	0.044
NO	**0.700**	0.412	0.259	-

Each column indicates the real class of the samples, each row indicates the assigned class by our model.

specific and cross-domain identification of relational properties between arguments. Even the *RoBERTa-base* version performed well in this task and it was 6.6 times faster than its large version on training.

Error Analysis

In an effort to conduct a thorough evaluation, we decided to analyze the errors made by *RoBERTA-large*, the best performing model. For this purpose, we measured the volume of misclassifications found on each one of the four classes considered in this work. Table 6 describes the error distribution detected when analyzing the results. Two important remarks can be pointed out when looking at the obtained error distributions. First of all, it is possible to observe how most of the misclassified argument pairs labeled with an inference relation were assigned the no relation class. Similarly, most of the misclassified argument pairs without relation were assigned the inference class by our model. We observed that many of these errors were due to a loss of contextual information. In an argumentative discourse, it is very common to refer to past concepts without explicitly mentioning them (i.e., enthymemes) or simplifying them with the use of pronouns. The lack of dialogical context can make the automatic identification of argument relations a harder task. For a better understanding of this problem, we present the following example with two argument components labeled with inference relation.

P1: *I think* **it***'s not going to help change the culture.*
P2: *In banking* **we***'ve a totally different situation.*

Our system classified the pair as no related samples. In fact, by only reading the sentence pair, one may think there is not any argument relation between them. In these situations, it is evident that the key to avoid any possible error is to give additional information about the uttered propositions. In this case, depending on the background meaning of the "**it**" and "**we**" pronouns, the sentences may be related or not. The only way of considering this proposition pair related as an inference, is assuming that the **it** pronoun refers to the banking system. We also detected that in these situations the softmax outputs of our model gives very close probabilities for both, *RA* and *NO* classes. Another indicators of the existing model misunderstandings presented before are the similar error distributions that conflicting arguments show with both inference and no relation classes. On the other hand, we also pointed out that the rephrased argument pairs were mainly misclassified as inference-related arguments. However, when analyzing them, we observed that most of the relations could also be considered as inference-related arguments depending on the interpretation. For example:

- **We** *do need curriculum reform.*
- **RUBIO too** *believes in curriculum reform*

In this case, the sentence pair can be interpreted as a rephrase, assuming that "**We**" and "**RUBIO too**" are equivalent subjects. But it can also be interpreted as an argument from authority, with P2 supporting (inference relation) P1. In some situations the line that differences rephrase from inference may not be as clear as desired, and both types of relation can be considered correct. Additionally, with these second type of significant detected errors, it is also possible to observe the problem mentioned before. Therefore, the loss of information caused by the use of pronouns or enthymemes in the discourse can be determinant when approaching a task of this complexity.

CONCLUSION

The automatic identification of argument relations is an essential task in the whole computational argumentation process. It allows to automatically generate the argumentative structure from argument discourse units. In this work, we present how the automatic identification of argument relations, based on IAT labeling, can be approached using the latest advances in NLP. To the best of our knowledge, this is the first work using transformer-based pretrained models to learn this task. For this purpose, we have used the largest publicly available argument annotated corpus to the date. Most of the trained models have been able to outperform the state-of-the-art baselines in argument relation mining,[1] even with a more complex instance of the task. We observed a significant better

performance with *RoBERTa* than other models, the best results were achieved with *RoBERTa-large*. We also made a cross-domain evaluation of the models, in order to find out their domain robustness. Even there was a small drop in performance (most probably because of the significant variations of linguistic and class distributions between different domain corpora), the scores on different domains were still close to previous AM reports on a unique domain. This way, it is our objective to contribute on paving the way for finding models that do a better generalization of this task. Finally, we analyzed the errors made by our best performing model. We have seen that two important groups of errors are caused by the loss of contextual information. We also pointed out that another important group of errors made by the model was due to possible multiple interpretations of the relations. We think that significant improvements in model performance can be achieved after analyzing the most common errors detected in this work. As future work, we propose the following modifications to the automatic identification of argument relations task: (i) pronoun replacement, to solve the loss of contextual information in some propositions; (ii) consider the possible classification ambiguity, in some cases, by accepting multiple correct relations if the interpretation leads to this conclusion; and (iii) incorporation of external information. In argumentation theory, an enthymeme is known as the omission of a claim or a support of an argument. In order to make the discourse more fluid, it is very common to use enthymemes in situations where the omitted information is considered to be known by all the participants. Therefore, without external information, the model may not be able to fully understand relations between enthymemes.

ACKNOWLEDGMENTS

This work was supported in part by the Spanish Government project under Grant TIN2017-89,156-R, in part by the FPI under grant BES-2015-074,498, and in part by the Valencian Government project under Grant PROMETEO/2018/002. The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPUs used for this research.

REFERENCES

1. O. Cocarascu, E. Cabrio, S. Villata, and F. Toni, "A dataset independent set of baselines for relation prediction in argument mining," *CoRR*, vol. abs/2003.04970, 2020.
2. O. Cocarascu and F. Toni, "Identifying attack and support argumentative relations using deep learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1374–1379, doi: 10.18653/v1/d17-1144.
3. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2019, pp. 4171–4186, doi: 10.18653/v1/n19-1423.
4. R. Jha, F. Belardinelli, and F. Toni, "Formal verification of debates in argumentation theory," *35th Symp. Appl. Comput.*, in C.-C. Hung, T. Cerný, D. Shin, A. Bechini, Eds., pp. 940–947, 2020, doi: 10.1145/3341105.3373907.
5. N. Kotonya and F. Toni, "Gradual argumentation evaluation for stance aggregation in automated fake news detection," in *Proc. 6th Workshop Argument Mining*, 2019, pp. 156–166, doi: 10.18653/v1/w19-4518.
6. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *Proc. 8th Int. Conf. Learn. Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=H1eA7AEtvS
7. J. Lawrence and C. Reed, "AIFdb Corpora," in *Proc. Comput. Models Argument*, 2014, vol. 266, pp. 465–466, doi: 10.3233/978-1-61499-436-7-465.
8. J. Lawrence and C. Reed, "Argument mining: A survey," *Comput. Linguistics*, vol. 45, no. 4, pp. 765–818, 2019, doi: 10.1162/coli_a_00364.
9. Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.
10. S. Menini, E. Cabrio, S. Tonelli, and S. Villata, "Never retreat, never retract: Argumentation analysis for political speeches," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4889–4896. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16393
11. N. Naderi and G. Hirst, "Argumentation mining in parliamentary discourse," in *Proc. Princ. Pract. Multi-Agent Syst.*, 2015, pp. 16–25, doi: 10.1007/978-3-319-46218-9_2.
12. V. Niculae, J. Park, and C. Cardie, "Argument mining with structured SVMs and RNNs," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 985–995, doi: 10.18653/v1/P17-1091.
13. R. M. Palau and M.-F. Moens, "Argumentation mining: The detection, classification and structure of arguments in text," in *Proc. 12th Int. Conf. Artif. Intell. Law*, 2009, pp. 98–107, doi: 10.1145/1568234.1568246.
14. A. Rago, O. Cocarascu, and F. Toni, "Argumentation-based recommendations: Fantastic explanations and how to find them," in Proc. 27th Int. Joint Conf. Artif. Intell. Main track, 2018, pp. 1949–1955, doi: 10.24963/ijcai.2018/269.

15. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *CoRR*, vol. abs/1910.01108, 2019.
16. C. Stab and I. Gurevych, "Parsing argumentation structures in persuasive essays," *Comput. Linguistics*, vol. 43, no. 3, pp. 619–659, 2017, doi: 10.1162/COLI_a_00295.
17. A. Vaswani*et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
18. J. Visser, B. Konat, R. Duthie, M. Koszowy, K. Budzynska, and C. Reed, "Argumentation in the 2016 us presidential elections: Annotated corpora of television debates and social media reaction," *Lang. Resour. Eval.*, vol. 54, pp. 123–154, 2019, doi: 10.1007/s10579-019-09446-8.
19. D. Walton, C. Reed, and F. Macagno, *Argumentation Schemes*. Cambridge, U.K.: Cambridge Univ. Press, 2008. [Online]. Available: http://www.cambridge.org/us/academic/subjects/philosophy/logic/argumentation-schemes
20. Z. Yang *et al.*, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5754–5764. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

RAMON RUIZ-DOLZ is a Researcher with the Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Polytechnic University of Valencia (UPV), Valencia, Spain and is currently working toward the Ph.D. degree in computer science. His research interests are focused on argument mining, computational argumentation and persuasion technologies. He received the master's degree in artificial intelligence, pattern recognition and digital imaging from the (UPV), Valencia, Spain. Contact him at raruidol@dsic.upv.es.

JOSE ALEMANY is a Postdoctoral Researcher with the Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Polytechnic University of Valencia (UPV), Valencia, Spain. He is also an Assistant Professor with Florida Universitria, Valencia, Spain. His research interests include information dissemination, privacy preserving, content analysis, and complex networks. He received the Ph.D. degree in computer science from the UPV. Contact him at jalemany1@dsic.upv.es.

STELLA M. HERAS BARBERÁ is a Researcher with the Valencian Research Institute for Artificial Intelligence (VRAIN), Valencia, Spain and an Associate Professor with the Department of Languages and Computer Systems, Polytechnic University of Valencia (UPV), Valencia, Spain. Her research area is focused on the development of artificial intelligence systems (computational argumentation, persuasion technologies, educational recommender systems). She received the Ph.D. degree in computer science (extraordinary prize Cum Laude) from the Polytechnic University of Valencia (UPV), Valencia, Spain. Contact her at stehebar@upv.es.

ANA GARCÍ-FORNES is a Full Professor with the Department of Information Systems and Computation, Universitat Polytechnic University of Valencia (UPV), Valencia, Spain and as a Researcher with the Valencian Research Institute for Artificial Intelligence (VRAIN). Her research interests focus on real-time systems, multiagent systems, agreement technologies, and privacy in social media. She received the Ph.D. degree in computer science from the UPV. Contact her at agarcia@dsic.upv.es.