Transfer Learning with Sentence Embeddings for Argumentative Evidence Classification

Davide Liga[1,2][0000−0003−1124−0299] and Monica Palmirani[1][0000−0002−8557−8084]

[1] Alma Mater Studiorum - University of Bologna, Bologna, Italy
{monica.palmirani,davide.liga2}@unibo.it
[2] University of Luxembourg, Luxembourg

Abstract. This work describes a simple Transfer Learning methodology aiming at discriminating evidences related to Argumentation Schemes using three different pre-trained neural architectures. Although Transfer Learning techniques are increasingly gaining momentum, the number of Transfer Learning works in the field of Argumentation Mining is relatively little and, to the best of our knowledge, no attempt has been performed towards the specific direction of discriminating evidences related to Argumentation Schemes. The research question of this paper is whether Transfer Learning can discriminate Argumentation Schemes' components, a crucial yet rarely explored task in Argumentation Mining. Results show that, even with small amount of data, classifiers trained on sentence embeddings extracted from pre-trained transformers can achieve encouraging scores, outperforming previous results on evidence classification.

Keywords: Argumentation Mining · Transfer Learning · Argumentation Schemes

1 Introduction

In the last few years, the use of Transfer Learning methodologies generated in remarkable hype in the State of the Art of many Natural Language Processing tasks. Particularly, the Transformer known as "Bidirectional Encoder Representations from Transformer" (BERT) has shown extremely good results, establishing several new records in terms of metrics results [3]. In 2018, BERT obtained new state-of-the-art results on eleven NLP-related tasks. In a couple of years dozens of variants have been developed, establishing other new records not just in English but also in other languages (e.g., the Italian versions, GilBERTo[3] and umBERTo[4], the French camemBERT [11]).

Despite the high celebrity recently achieved by Transfer Learning techniques, these methodologies have been applied relatively few times in Argumentation Mining [12, 14]. To the best of our knowledge, this is the first work that explicitly assesses Transfer Learning performances with the aim of discriminating

[3] https://github.com/idb-ita/GilBERTo

[4] https://github.com/musixmatchresearch/umberto

CMNA'20 - F. Grasso, N. Green, J. Schneider, S. Wells (Eds) - 8th September 2020

argumentative components related to Argument Schemes [17]. On the one side, the approach show to be capable of discriminating argumentative stances of support and opposition related to some famous argumentative patterns (Argumentation Schemes) such as Argument from Expert Opinion, and Argument from negative consequences, showing better results compared to previous studies. On the other side, the approach show that it is possible clustering Argumentation Schemes according to the criteria of the pragmatical dimension, which is a crucial aspect described in the most recent literature about Argumentation Scheme classification [10, 6]. In summary, the approach show an ability to classify argumentative evidences not only at fine-grained levels (e.g., different instances of Argument from Expert Opinion) but also at the level of large clusters (like the Argumentation Schemes coming from an external source, a class which according to some classification approaches can be used as first dichotomic criterion of discrimination among schemes [10, 6]).

Section 2 will describe the Transfer Learning methodology and the two main settings for the experiments. Section 3 will describe the datasets used for the experiments in the two scenarios. Sections 4 and 5 will show the experimental results on the two scenarios. Section 6 will describe the related works. In Section 7, some final considerations will conclude the paper.

2 Methodology

Transfer Learning methods are generally divided in two approaches: the first approach is called fine-tuning and it consists of using a pre-trained neural architecture (i.e., a Transformer architecture trained on thousands of inputs) as a starting point to perform further training steps on a downstream task (training, thus, the neural architecture on downstream data). The second approach, instead, is that of using a pre-trained neural architecture just to extract the outputs that the Neural Architecture generate for a given input at a specific stage of the neural architecture. For example, a sentence can be used as input and the output generated by the neural architecture can be extracted and used as sentence embeddings, that can represent our sentence in other downstream tasks (noticeably, the extraction of the generated output to be used as embedding can be performed at different stages of the neural architecture, not necessarily at the final layer). In this paper, the second approach will be employed: a famous pre-trained architecture will be selected, some sentences will be used as inputs for this neural architecture, and the output coming from the neural architecture will be employed as sentence embeddings to represent our data in a series of downstream classification tasks.

For the pre-trained embeddings we will employ three pre-trained models: the first one is the famous neural transformer called BERT [3] (specifically, we will use the uncased *base* version). The second and third models are two recent models which are derived from BERT, namely: distilBERT[16] and RoBERTa[9] (uncased). While BERT base consists of 12 layers, 768 hidden dimensions, 12 self-attention heads and nearly 110M parameters, RoBERTa base consists of 12 layers, 768 hidden dimensions, 12 self-attention heads and 125M parameters. Finally, distilBERT consists of 6 layers, 768 hidden dimensions, 12 self-attention heads and 66M parameters.

To extract the embeddings from the neural models, each input sentence must be firstly tokenized according to the requirements of the given model. Typically, with BERT, a [CLS] and a [SEP] special tokens are inserted at the beginning and at the end of the input (we are interested in the first one which is the token holding the classification output we are interested to extract from the input sentence). Moreover, the length of each input sentence is set to a max length: all sentences longer than that limit are shortened, while all sentences shorter than that limit are padded with the special [PAD] token. This process makes sure that all inputs have the same length before entering the neural architecture. After the tokenization, inputs are passed into the neural architecture of a BERT transformer, while deactivating the calculation of gradients.

After having transformed each input sentence of the test sets into tokens and having used these tokens as inputs for the BERT neural architecture, the resulting extracted embeddings have been used, in turn, as input of a classification using two classification procedure: a Support Vector Machine (SVM) classifier and a Logistic Regression classifier (LRC). Notice that for the experiment on D3 our SVM employed a Linear Support Vector Classifier (Linear SVC), while in all other experiments we employed a standard Support Vector Classifier (SVC).

The classification method is One vs All. Which means that the classification has been performed per each class, considering one class against all the other classes, a typical approach in multiclassification and multilabel scenarios. Finally, all classifiers have been evaluated on the relative testing set.

The experiments have been divided into different scenarios:

1. Baseline scenario: in this scenario, the classification was performed on the same setting of two previous works, taken as baselines for comparison.
2. Extended scenario: in this scenario, the classification was performed on new settings, using an extended version of two datasets from the baseline scenario.

3 Data

The experiments of this work have been applied to the datasets listed in Table 1, reporting reports also the number of instances for each dataset. These datasets have been selected because their annotations describe classes of argumentative evidence directly related to specific Argumentation Schemes. Importantly, during the experiments, all datasets have been split into train and test sets, following a standard 80/20 ratio.

Regarding the baseline scenario, D1 and D2 are a portion of Al Khatib et al. 2016 and Aharoni et al. 2014 respectively, two important dataset designed by IBM. Only two classes from the original datasets have been selected, reproducing the scenario in [7] in order to have baseline scenarios for our classifiers. D3 is a small dataset (only 638 sentences) from Liga and Palmirani 2019. It is a dataset which has different levels of granularity, depending on how many classes

Table 1. Description of all datasets used in this paper.

Dataset	Reference	Classes	Instances
	Baseline datasets:		
D1	Al Khatib et al. 2016 (only two classes extracted as in [7])	Study, Testimony	653
D2	Aharoni et al. 2014 (only two classes selected as in [7])	Study, Expert	569
D3	Liga and Palmirani 2019	Slippery Slope, Testimony, Other	638
	Extended datasets:		
D1+	Al Khatib et al. 2016 (three classes extracted following the method in [7])	Study, Testimony, Anecdotal	2253
D2+	Aharoni et al. 2014	Study, Expert, Anecdotal	1291
D2++	Rinott et al. 2015	Study, Expert, Anecdotal	4692

are considered. In this case we selected granularity three, which contains three labels.

Regarding the extended scenario, the dataset D1+ is an extension of D1: instead of extracting just two classes, it considers three classes. The inputs of the dataset from Al Khatib et al. 2016 [2] are actually structured in a very fragmented way, so we needed to rebuild the sentences following the approach suggested in [7]. Similarly, D2+ is an extension of D2 (instead of being a selection of just two classes, it considers three classes). Finally, D2++ is an extended version of the same dataset which, having many more instances, can be a useful benchmark for this kind of classifications.

Importantly, the datasets which have been employed in this work are among the few available datasets containing instances of argumentative evidences which can be related to Argumentation Schemes. Namely, the dataset in Al Khatib et al. 2016 [2] shows instances of argumentative evidences labelled as Study, Testimony and Anecdotal: these evidences support argumentative claims which refer to source-based opinions, this means that they belong to different types of source-based arguments. One of the most famous example of source-based Argumentation Scheme is the well-known Argument from Expert Opinion; another famous scheme is the Argument from witness testimony (more details about this kind of schemes can be found in [6]).

The datasets in Aharoni et al. 2014 [1] and Rinott et al. 2015 [15] present similar source-based Argumentation Schemes (however, this time the labels are Study, Expert and Anecdotal). In this case, the cluster of argumentative evidences labelled with the class Expert are likely to be compatible with the evidences of an Argumentation Scheme from Expert Opinion.

The dataset in Liga and Palmirani 2019 [8] offers instead only one class of evidences which is related to source-based arguments (Testimony) while another class is related to a cluster of evidences which can be related to the Argument from Negative Consequences and the Slippery Slope Arguments.

These three datasets can thus be used to assess whether classifiers are able to discriminate between different cluster of argumentative evidences. Since these argumentative evidences are strictly related to specific clusters of Argumentation Schemes, the ability of classifiers to discriminate different clusters of argumentative evidences is, in our opinion, a crucial step towards Argumentation Scheme discrimination.

4 Results for the Baseline Scenario

The classifications in this Section show that the proposed approach is able to outperform recent results in the Argumentation Mining literature. With this purpose, recent results on D1, D2 and D3 are reported [7, 8] and used as baseline for our classifiers.

Table 2. Results on the baseline classifiers (D1, D2, D3) considering mean F1 scores (macro) and two kinds of classifier. SVM = Support Vector Machine; LR = Logistic Regression; BS = Baseline. The columns whose mean F1 value has an asterisk refers to a Linear Support Vector Classifier. In bold are all the mean F1 scores which overcome the mean F1 of the baseline. The three grey columns represent the best classifiers for the baseline scenario.

Classes	Bert Base		DistilBERT		RoBERTa		BS
	SVM	LR	SVM	LR	SVM	LR	
D1 (Al Khatib et al. 2016)							
Study	.94	.92	.97	.97	.91	.89	.91
Testimony	.93	.91	.97	.96	.89	.86	.92
mean F1	**.94**	**.92**	**.97**	**.96**	.90*	.88	.92
D2 (Aharoni et al. 2014)							
Study	.78	.71	.72	.74	.75	.79	.69
Expert	.75	.68	.67	.72	.73	.77	.78
mean F1	**.76**	.69	.69*	**.73**	**.74***	**.78**	.73
D3 (Liga and Palmirani 2019)							
Slippery Slope	.75	.71	.79	.76	.82	.60	.70
Testimony	.90	.92	.93	.94	.93	.73	.71
Other	.85	.86	.87	.87	.87	.85	.91
mean F1	**.83***	**.82**	**.86***	**.86**	**.87***	.73	.77

In this paper, all F1 scores per class are calculated as the mean macro F1 scores, taken from each One-vs-All classification. All these scores are finally averaged and reported as mean F1 (per each classifier, i.e. SVM and LR).

As can be seen from Table 2, results outperform previous results for the same scenario, showing the ability of Transfer Learning techniques to achieve high performances. As indicated by the bold numbers in Table 4, for D1, D2 and D3 there are always at least four classifiers out of six which outperform the baseline.

5 Result for the Extended Scenario

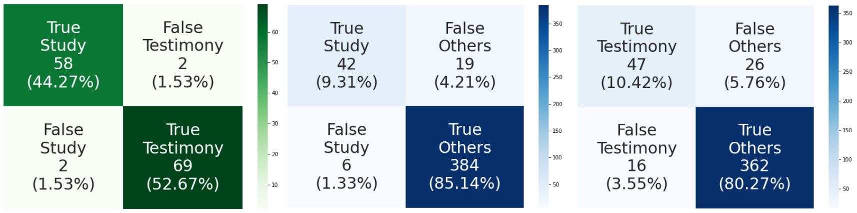
The next series of experiments have been performed on an extended version of two datasets from the baseline scenario (D1 and D2), to assess how performances change in a multiclass scenario.

Table 3. Results on D1+, D2+ and D2++ considering mean F1 scores (macro) and two kinds of classifiers. SVM = Support Vector Machine; LR = Logistic Regression; BS = Baseline. The columns whose mean F1 value has an asterisk refers to a Linear Support Vector Classifier. In bold are the top mean F1 scores. The three grey columns represent the best classifiers for the extended scenario.

Classes	Bert Base		DistilBERT		RoBERTa	
	SVM	LR	SVM	LR	SVM	LR
D1+ (Al Khatib et al. 2016)						
Study	.83	.85	.83	**.87**	.86	.77
Testimony	.77	.81	.81	**.82**	.78	.70
Anecdotal	.81	.81	.82	**.84**	.83	.77
mean F1	.80	.82	.82*	**.84**	.82*	.75
D2+ (Aharoni et al. 2014)						
Study	.89	.90	**.91**	**.91**	.90	.85
Expert	.91	.91	.92	**.93**	.90	.84
Anecdotal	.92	**.93**	.92	**.93**	.92	.92
mean F1	.91*	.91	**.92***	**.92**	.91*	.87
D2++ (Rinott et al. 2015)						
Study	.93	**.94**	**.94**	**.94**	.92	.90
Expert	.92	.92	**.93**	**.93**	.91	.88
Anecdotal	.91	**.93**	.90	.92	.87	.85
mean F1	.92*	**.93**	.92*	**.93**	.90*	.88

Table 3 shows a clear trend, with Logistic Regression on DistilBERT being the best solution for both the dataset extending D1 (i.e., D1+) and the datasets extending D2 (i.e., D2+ and D2++).

Regarding the classifications on D1+, one can see that the best performances are achieved by the Logistic Regression classifier (LR) trained on sentence embeddings extracted using DistilBERT. To have a better understanding of these results, the confusion matrix of the best classifier in this scenario (i.e., Logistic Regression on DistilBERT) are reported with the confusion matrix from the best classifier of the baseline scenario (i.e., Support Vector Machine from DistilBERT embeddings from Table 2) in Figure 1.



D1 (study vs testimony) SVM on DistilBERT

D1+ (study vs others) LR on DistilBERT

D1+ (testimony vs others) LR on DistilBERT

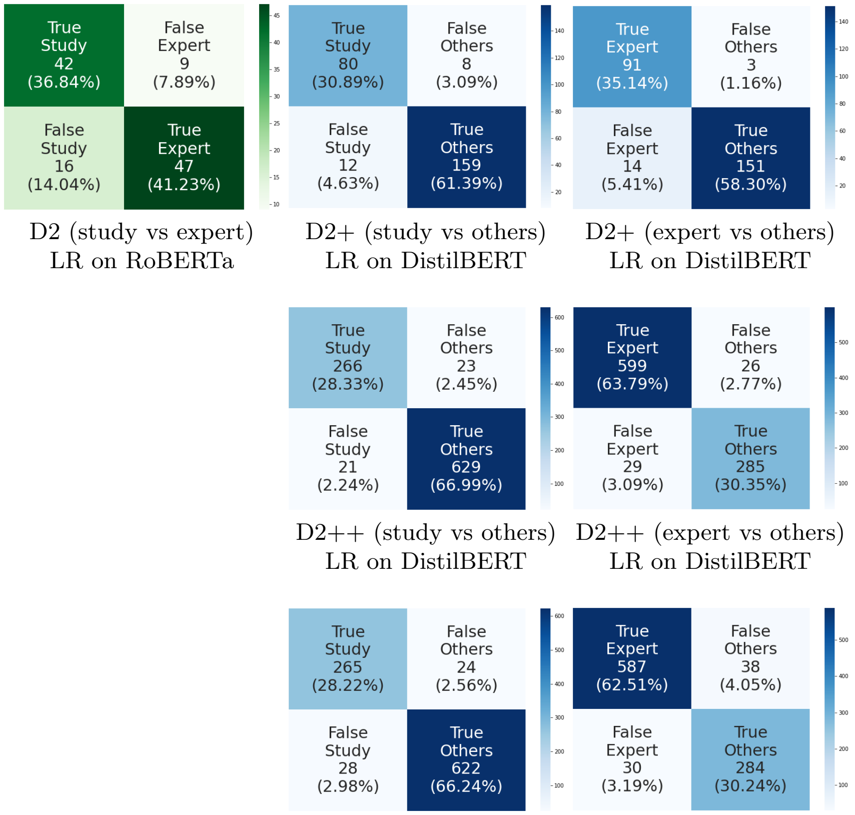
Fig. 1. Confusion matrices for D1 (in green) and D1+. The number of instances and the relative percentages are reported.

Regarding the classifications on D2+ and D2++, one can see that the best performances are achieved by the Logistic Regression classifier (LR) trained on sentence embeddings extracted using DistilBERT and Bert Base. Also in this case, to have a better understanding of the results, the confusion matrices of the best classifiers in this scenario (i.e., Logistic Regression from DistilBERT embeddings and from Bert Base) are reported with the confusion matrix from the best classifier of the baseline scenario (i.e., Logistic Regression from RoBERTa embeddings from Table 2) in Figure 2.

Notice that while confusion matrices for D1 and D2 (in green) show a binary classification, the other confusion matrices in blue (relative to D1+, D2+ and D2++) show a one-vs-all classification. These blue matrices show that classifiers are able to recognize classes also in a multiclass scenario. While Figure 1 shows an unbalance (which is probably due to the predominance of the class anecdotal), results in Figure 2 seems more balanced: the diagonal is always a 30/60 ratio, indicating the goodness of predictions.

6 Related works

Unfortunately, datasets specifically designed in a way that allow a direct link between classes and specific Argumentation Schemes are very few. A promising and growing resource, in this sense, is the corpora in AIFdb [5] thanks also to the contribute of tools like OVA+ [13] which recently added a very important component for Argumentation Scheme annotation called Argument Scheme Key [6].



D2 (study vs expert)
LR on RoBERTa

D2+ (study vs others)
LR on DistilBERT

D2+ (expert vs others)
LR on DistilBERT

D2++ (study vs others)
LR on DistilBERT

D2++ (expert vs others)
LR on DistilBERT

D2++ (study vs others)
LR on Bert Base

D2++ (expert vs others)
LR on Bert Base

Fig. 2. Confusion matrices for D2 (in green), D2+ and D2++. The number of instances and the relative percentages are reported.

Moreover, although there have been different works of text classification in Argumentation Mining, only few studies focused on classification tasks aiming at facilitating the discrimination of Argumentation Schemes.

Rinott et al. 2015 [15] achieved important results on evidence detection employing the dataset D2++. However, the approach is mostly context-dependent, while the present work is not considering the context. In Liga 2019 [7], the classification has been performed using Tree Kernels classifiers on D1 and D2, containing argumentative evidences of support among which it is possible to find evidences directly related to the Argument from Expert Opinion. The work is however limited to a binary classification. A similar approach, in a multiclass scenario, is described in Liga and Palmirani 2019 [8], where Tree Kernels are employed on D3, a small dataset which considers argumentative evidences of opposition among which one can find, for example, the Slippery Slope Argument.

Considering these two works as baselines, the approach presented in this paper seems capable of outperforming the previous achievements.

7 Conclusion

The datasets analyzed in this work are composed of argumentative evidences which are directly related to different clusters of arguments. For example, many instances which can be found in the datasets of this paper are directly related to the cluster of source-based arguments. Other instances of argumentative evidences are instead specifically related to the Argumentation Scheme from Expert Opinion, while others are related to the cluster which includes the Argument from Negative Consequences and the Slippery Slope Arguments (which do not belong to the cluster of source-based arguments).

We believe that the ability to discriminate different clusters of argumentative evidences is a crucial step in the classification of Argumentation Schemes. For example, the discrimination of clusters of Argumentation Schemes can be performed in a pipeline of binary classifications starting from source-based versus non-source-based arguments and continuing towards more specific binary classifications (similarly to the path of dichotomous choices followed by ASK, the annotation system recently elaborated in [6], which offers a valuable system of classification of Argumentation Schemes).

In general, the results presented in this paper seem encouraging, showing that pre-trained embeddings can outperform previous results in the field of Argumentation Mining related to the classification of argumentative evidences. An interesting aspect is that the proposed classifiers show encouraging results not only in the discrimination among different source-based argumentative evidences, but also in classifications involving source-based versus non-source-based argumentative evidences (i.e. with dataset D3).

However, further analysis is needed to verify if and how Transfer Learning techniques can discriminate argumentative evidences in such a way that they can facilitate Argumentation Scheme discrimination. In this regard, the present paper is just a preliminary exploration of a promising possible approach. In future works, other Transfer Learning techniques should be assessed too. For example, it could be useful to assess the performances between the two main Transfer Learning techniques: sentence embeddings and fine-tuning. Also, other pre-trained models should be employed and compared (e.g., Xlnet[18], Albert[4]).

A long-term goal is being able to connect natural language argumentative evidences to their specific Argumentation Schemes, which can be a further step in the development of an artificial Natural Argumentation Understanding.

References

1. Aharoni, E., Polnarov, A., Lavee, T., Hershcovich, D., Levy, R., Rinott, R., Gutfreund, D., Slonim, N.: A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In: Proceedings of the First Workshop on Argumentation Mining. pp. 64–68 (2014)

2. Al Khatib, K., Wachsmuth, H., Kiesel, J., Hagen, M., Stein, B.: A news editorial corpus for mining argumentation strategies. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 3433–3443 (2016)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)
5. Lawrence, J., Reed, C.: Argument mining: A survey. Computational Linguistics **45**(4), 765–818 (2020)
6. Lawrence, J., Visser, J., Reed, C.: An online annotation assistant for argument schemes. In: Proceedings of the 13th Linguistic Annotation Workshop. pp. 100–107. Association for Computational Linguistics (2019)
7. Liga, D.: Argumentative evidences classification and argument scheme detection using tree kernels. In: Proceedings of the 6th Workshop on Argument Mining. pp. 92–97 (2019)
8. Liga, D., Palmirani, M.: Detecting "slippery slope" and other argumentative stances of opposition using tree kernels in monologic discourse. In: International Joint Conference on Rules and Reasoning. pp. 180–189. Springer (2019)
9. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
10. Macagno, F., Walton, D., Reed, C.: Argumentation schemes. history, classifications, and computational applications. History, Classifications, and Computational Applications (December 23, 2017). Macagno, F., Walton, D. & Reed, C pp. 2493–2556 (2017)
11. Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B.: Camembert: a tasty french language model. arXiv preprint arXiv:1911.03894 (2019)
12. Niven, T., Kao, H.Y.: Probing neural network comprehension of natural language arguments. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4658–4664 (2019)
13. REED, M.J.J.L.C.: Ova+: An argument analysis interface. In: Computational Models of Argument: Proceedings of COMMA. vol. 266, p. 463 (2014)
14. Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., Gurevych, I.: Classification and clustering of arguments with contextualized word embeddings. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 567–578 (2019)
15. Rinott, R., Dankin, L., Alzate, C., Khapra, M.M., Aharoni, E., Slonim, N.: Show me your evidence-an automatic method for context dependent evidence detection. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 440–450 (2015)
16. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
17. Walton, D., Reed, C., Macagno, F.: Argumentation schemes. Cambridge University Press (2008)
18. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems. pp. 5754–5764 (2019)