Multi-task Learning in Argument Mining for Persuasive Online Discussions

Nhat Tran
University of Pittsburgh
nlt26@pitt.edu

Diane Litman
University of Pittsburgh
dlitman@pitt.edu

Abstract

We utilize multi-task learning to improve argument mining in persuasive online discussions, in which both micro-level and macro-level argumentation must be taken into consideration. Our models learn to identify argument components and the relations between them at the same time. We also tackle the low-precision which arises from imbalanced relation data by experimenting with SMOTE and XGBoost. Our approaches improve over baselines that use the same pre-trained language model but process the argument component task and two relation tasks separately. Furthermore, our results suggest that the tasks to be incorporated into multi-task learning should be taken into consideration as using all relevant tasks does not always lead to the best performance.

1 Introduction

Argument mining (AM) focuses on automatically identifying argumentative structures in text, and utilizing these structures in applications. AM tasks include identifying argument components (e.g., "claim") and relations between them (e.g., "support"). However, most AM studies have focused on monologues or micro-level models of arguments (Peldszus and Stede, 2015; Persing and Ng, 2016; Stab and Gurevych, 2017). AM in dialogues and macro-level models have received less attention (Bentahar et al., 2010; Chakrabarty et al., 2019b).

In this study, we extend the work of Chakrabarty et al. (2019b) in AM for persuasive online discussions. Particularly, we take advantage of a multi-task learning (MTL) approach to automatically identify the argument structures in persuasive dialogues that contain both micro-level and macro-level argumentation. We identify argument components (claim, components, non-argumentative) and two types of relations: intra-turn relations within one post and inter-turn relations across posts. Our results demonstrate that using MTL improves the performance of both argument component and intra-turn/inter-turn relation classification. However, further analysis shows that the tasks in the MTL configuration should be chosen carefully depending on the focused task. We then try several techniques to increase the innate low precision of the relation classification tasks due to the highly imbalanced data, specifically SMOTE (Chawla et al., 2002) and XGBoost (Chen and Guestrin, 2016). Our results demonstrate that SMOTE is not very helpful but XGBoost, when used with the representations learnt from MTL, can increase the precision and F-scores of the relation identification tasks.

2 Related Work

Our work is closely related to Chakrabarty et al. (2019b). Their system, called AMPERSAND, tackles three AM tasks on a dataset created from the Change My View (CMV) subreddit[1] (Hidey et al., 2017) and focuses on transfer learning approaches with BERT (Devlin et al., 2019) that take advantage of discourse and dialogue context. Specifically, they define three separate tasks: argument component classification and intra/inter relation identification. For the first task, the requirement is to classify a given sentence into either Claim, Premise or Non-argumentative. For the intra-relation identification task, given a pair of argumentative sentences from the same post, we need to answer if an argumentative relation between these two sentences exists. The inter-relation identification task is similar, except that the two sentences are from different posts. However, they treated the tasks of argument component classification and relation prediction separately and had independent BERT models for the tasks. Our approach works on the assumption that the three tasks are related to each other.

Many studies have shown that jointly learning several tasks during training usually leads to better

[1] https://www.reddit.com/r/changemyview

Proceedings of The 8th Workshop on Argument Mining, pages 148–153
Punta Cana, Dominican Republic, November 10–11, 2021.

performance in NLP problems (Søgaard and Goldberg, 2016; Yang et al., 2016; Liu et al., 2019; Peng et al., 2020). Focusing on one single domain and dataset, Eger et al. (2017) treats AM as a sequence tagging problem and uses sub-tasks such as component identification and relation classification as auxiliaries in MTL to improve performances. Schulz et al. (2018) also formalizes argument component identification as a sequence tagging problem but utilizes multiple datasets from different domains in their MTL setup. They observe that the results on a small AM dataset can be improved when other AM datasets are leveraged as auxiliary tasks. These approaches, however, work on monologues where each data instance is from one person and therefore ignore the macro-structure of arguments. Our work tackles AM at dialogical level, specifically on online discussion forums. We hypothesize that MTL can help represent both micro and macro structure and use BERT with a MTL setup to classify argument components and relations at the same time.

3 Data

We use the same data from Chakrabarty et al. (2019b). They reuse the CMV corpus (Hidey et al., 2017), where each sentence in a thread of the CMV subreddit is annotated as claim, premise or non-argumentative. Additionally, they annotate the argument relation among these propositions (inter-turn/intra-turn) and expand the corpus by annotating additional argument components using the same guidelines.

The final dataset consists of 112 threads with 2756 sentences. The proportions of claims, premises and non-argumentative components are 34%, 43% and 23% respectively. Although several types of relations are annotated, the relation identification task only uses a binary label to represent if a relation exists between two components. The dataset is highly imbalanced in terms of relations, with only 4.6% of 27254 possible pairs having intra-turn and only 3.2% of 26695 having inter-turn relations, making low precision a major modeling challenge.

Below is an example of a discussion. User A makes a claim and supports it with a premise (intra-turn relation). User B, however, disagrees with the reasoning made by user A (inter-turn relation).

> A: [I think the biggest threat to global stability comes from the political fringes.]$_{0:\text{CLAIM}}$ [It has been like that in the past.]$_{1:\text{PREMISE:SUPPORT:0}}$
>
> B: [What happened in the past has nothing to do with the present]$_{2:\text{ATTACK:1}}$

Realizing that the data size is small, Chakrabarty et al. (2019b) utilizes distant-labeled data and uses transfer learning for fine-tuning BERT depending on the context. The IMHO+context dataset (Chakrabarty et al., 2019a) is used as micro-level context data. This is a corpus of opinionated claims in the form of sentences containing the internet acronyms IMO (in my opinion) or IMHO (in my humble opinion) from Reddit. The assumtion is that a relation exists between a sentence containing IMHO and the following one. For macro-level context data, they use the Reddit quote feature and construct the QR dataset containing quote-response pairs. In Reddit, when responding to a post, a user can quote another user's response and this feature is used to highlight what part of someone's argument a particular user is targeting in the CMV corpus. Specifically, the QR dataset treats the quoted text and the following sentence as a positive inter-turn relation example. For a fair comparison, we also fine-tune BERT using both distant datasets.

4 Methods

We use AMPERSAND's (Chakrabarty et al., 2019b) two relation classification constraints. For intra-turn relations, the source has to be a premise and the target can be a premise or a claim. For inter-turn relations, the source must be a claim.

4.1 Multi-task Learning on BERT

We follow the architecture of Liu et al. (2019) for MTL. It has lower BERT encoder layers shared across all tasks with task-specific classification layers on top of them. In this procedure, each task can be either single-sentence classification or sentence pair classification, which fits our tasks of component classification and relation identification. The latter task can be further divided into intra-turn and inter-turn relations, resulting in three tasks in total.

We have three MTL configurations, each one represents a different combination of tasks incorporated in the MTL process. First, all three tasks are used for MTL (**MTL_ALL**). Second, only argument component and intra-turn relation tasks are used (**MTL_intra**). Third, in **MTL_inter**, the intra-turn relation classification task is excluded.

Our reason is that intra-turn and inter-turn relations can be different in nature and including inter-turn prediction could possibly degrade intra-turn prediction, or vice versa. The argument component classification task is essential for both relation identification tasks since it helps filter out pairs of sentences that do not follow the constraints. Thus, it is kept in all MTL configurations.

4.2 Low Precision in Relation Prediction

Due to imbalanced data with less than 5% of pairs having relations, low precision is expected. AMPERSAND (Chakrabarty et al., 2019b) attempts to increase intra-turn relation precision with window clipping. Specifically, the best F-scores are reported when limiting the prediction of an intra-turn relation to be within a window of 1. Since this approach only works for intra-turn relations and is dependent on the data, we instead try two universal approaches which are corpus-independent to raise model precision.

SMOTE (Chawla et al., 2002) is an oversampling technique where synthetic samples are generated for the minority class. It focuses on the feature space to generate new instances by using interpolation between positive instances that lie together.

Gradient boosting is also useful when data is highly skewed (Brown and Mues, 2012; Teramoto, 2009). We experiment using XGBoost (Chen and Guestrin, 2016), a decision-tree-based boosting algorithm, as the classifier on top of the BERT representation instead of the normal softmax layer.

4.3 Implementation Details

In the **MTL_ALL** setup, we first fine-tune the BERT model on the IMHO+context and QR datasets using both the masked language modeling and next sentence prediction objectives. We then fine-tune BERT using MTL by learning the three tasks jointly. For the **MTL_intra** configurations, only the IMHO+context data is used for the first fine-tuning step and only the argument component classification and intra-turn identification tasks are used in the MTL procedures. The same settings are applied for **MTL_inter**, but the QR data is used for the first fine-tuning step instead.

Peng et al. (2020) observe that additional fine-tuning after the training process can increase performance. They remove the last layer, which is basically a linear and a softmax layer on top of the BERT representation to make the final classification, of the trained model and replace it with a new untrained one. Then they use a smaller learning rate to continue training all layers on each specific task. We call this step *refinement*.

AMPERSAND uses an additional RST classifier and ensembles its result with the prediction from the BERT classifier to predict the existence of a relation. Rhetorical Structure Theory (RST) provides an explanation for the coherence of text, in the form of a tree where leaves represent elementary discourse units and other nodes represent discourse relations. Specifically, they create a RST parse tree for the concatenated two argumentative components and take the predicted discourse relation at the root of the parse tree as a categorical feature in a binary classifier. They also use a candidate target selection procedure built from extractive summarization for inter-turn relation identification. Since these two are not involved in training and only work as additional filters, we keep them unchanged.

For XGBoost, since the 768-dimension vector from BERT is too large, we reduce the dimension to 128 using a two-layer neural network (512 and 128 neurons, respectively). We did an experiment and see that this reduction only affects performance with XGBoost, so we keep the 128 dimensions for all models to make the comparisons fair.

5 Results

Using the same train/test split from **Ampersand** (10% of the data for testing) (Chakrabarty et al., 2019b), we compare our results with AMPERSAND. In Tables 1, 2 and 3, *XG* stands for XGBoost, *SMO* for SMOTE, and *refine* for the refinement step from Sec. 4.3. Best results are in bold.

To make the comparisons more consistent with our models, we applied refinement and XGBoost on top of the final BERT representation of AMPERSAND. The reported numbers from the second row of the tables are from our own rerun of Ampersand and therefore they are slightly different from ones in the original paper. Although we used the AMPERSAND published code, the difference in Pytorch version could be the cause for this discrepancy. The best results of AMPERSAND from the original paper (Chakrabarty et al., 2019b) are reported in the first rows of the tables as Ampersand*.

5.1 Argumentative Component Classification

Table 1 shows that compared to Ampersand, in two MTL configurations MTL_ALL and MTL_intra,

Method	C	P	NA
Ampersand*	67.1	72.5	75.7
Ampersand	67.1	72.3	75.3
+ *refine*	67.3	72.9	76.1
MTL_ALL	67.8	75.2	77.8
+ *refine*	**70.0**	**76.6**	**78.1**
MTL_intra	68.3	74.8	76.5
+ *refine*	69.5	75.1	76.9
MTL_inter	66.2	72.1	73.9
+ *refine*	66.8	73.4	75.3

Table 1: F-scores for 3-way Classification: Claim (C), Premise (P), Non-Argument (NA). The best results from Chakrabarty et al. (2019b) are reported as Ampersand*

F-scores are improved for all three classes. The MTL_ALL model achieves 67.8, 75.2 and 77.8 for Claim, Premise and Non-argumentative respectively. When the inter-turn relation identification task is removed from the MTL configuration, MTL_intra model observes a slight drop in Premise (0.4%) and NA (1.3%) but a small increase in Claim (0.5%). On the other hand, taking out the intra-turn relation identification task (MTL_inter) degrades the F-scores in all categories. This implies that in our setting, intra-turn relation identification plays a crucial role in classifying components. Furthermore, including only inter-turn relation identification can hurt component classification as MTL_inter is inferior to AMPERSAND.

The additional fine-tuning step on each separate task also helps boost the F-scores. We witness slight increases in all of the three classes for all of the MTL configurations and the Ampersand model with this refinement step. Our best results are obtained with the MTL_ALL model with refinement.

5.2 Relation Prediction

For each metric in Tables 2 and 3, we report results with both gold-standard (G) and predicted (P) components from the argument component classifier.

5.2.1 Intra-turn Relations

The results from Table 2 demonstrate that MTL is still helpful in this task. Both MTL_ALL and MTL_intra have higher F-scores in comparison with the equivalent version of AMPERSAND.

MTL_intra models outperform the equivalent MTL_ALL models in terms of precision and F_1 scores. This suggests that we should eliminate the inter-turn relation task from MTL specifically for the intra-turn relation task. Our reasoning is that the inter-turn relations have some special characteristics and are harder to identify, which leads to the decrease in performance for the intra-turn task when it is included in the MTL configuration.

Method	Precision		Recall		F-score	
	G	P	G	P	G	P
Ampersand*	16.7	15.5	73.0	70.2	27.2	25.4
Ampersand	16.7	15.5	73.0	70.0	27.2	25.4
+ *refine*	16.7	15.7	73.0	70.0	27.2	25.6
/w XG	17.0	16.5	73.1	68.9	27.6	26.6
MTL_ALL	17.4	15.9	72.1	69.6	28.0	25.9
+ *refine*	18.1	16.3	72.2	68.4	28.9	26.3
/w SMO	17.7	16.2	**73.1**	71.0	28.5	26.4
/w XG	20.8	19.5	72.4	**73.0**	32.3	30.8
MTL_intra	19.3	17.3	71.1	69.5	30.4	27.7
+ *refine*	19.9	18.4	71.8	70.2	31.2	29.2
/w SMO	20.0	18.4	69.3	70.1	31.0	29.1
/w XG	**23.6**	**22.5**	**73.1**	69.8	**35.7**	**34.0**

Table 2: Results for Intra-turn Relation Prediction

Method	Precision		Recall		F-score	
	G	P	G	P	G	P
Ampersand*	18.9	17.5	**79.4**	**75.6**	30.5	28.3
Ampersand	18.7	17.1	**79.4**	**75.1**	30.3	27.9
+ *refine*	19.3	18.1	77.8	**75.1**	30.9	29.2
/w XG	17.0	16.5	73.1	68.9	27.6	26.6
MTL_ALL	20.3	18.2	79.1	74.5	32.3	29.3
+ *refine*	20.3	18.3	79.1	74.5	32.5	29.4
/w SMO	20.5	18.0	78.8	74.9	32.5	29.0
/w XG	21.2	**19.5**	75.7	65.2	33.1	**29.7**
MTL_inter	20.1	17.9	79.1	74.0	32.1	28.8
+ *refine*	20.2	18.3	79.0	74.2	32.2	29.4
/w SMO	20.0	18.2	**79.4**	74.9	32.0	29.3
/w XG	**21.5**	18.8	77.5	67.4	**33.7**	29.4

Table 3: Results for Inter-turn Relation Prediction

The refinement step generally helps improve the performance, but the gain is not very remarkable, especially in the case of MTL_ALL when the increases in F-score are less than 1 point for both gold and predicted components.

The XGBoost classifier raises the already low precision scores noticeably for MTL. For both MTL configurations, precision scores are increased by at least 2.7 points while recall scores are not decreased by more than 0.4 points. This leads to an improvement in F-scores based on predicted components of 4.5 points for MTL_ALL and 4.8 points for MTL_intra. Our best results are obtained by using XGBoost on the features of MTL_intra. In contrast, SMOTE does not help much.

5.2.2 Inter-turn Relations

For this task, the results from Table 3 demonstrate that MTL models still generally outperform the comparable baselines, but the gap is marginal com-

pared to the previous two tasks.

In contrast to intra-turn relation prediction, removing the intra-turn task does not always improve the result of the inter-turn task. In other words, MTL_inter models do not always outperform the equivalent MTL_ALL model. Also, the gain with XGBoost is now smaller, with less than 2 points in F-scores for both MTL configurations, regardless of gold-standard or predicted components. The reason is due to a now large recall drop (e.g., 9.3% and 6.8% drop for MTL_ALL and MTL_inter respectively, with predicted components).

For predicted components, MTL_ALL with XGBoost achieves the best F-score, while for gold components, MTL_inter with XGBoost is best.

6 Qualitative Analysis of Intra-turn Degradation using MTL_ALL

One noticeable observation from Section 5.2.1 and Table 2 is that the incorporation of the inter-turn prediction task into the MTL process indeed hurts the performance of intra-turn. To further analyze this phenomenon, we retrieve examples which were predicted correctly by MTL_intra but incorrectly by MTL_ALL. In many of these examples, there is a wrong "inference" that if A has an **inter-turn** relation with C and B has an **inter-turn** relation with C, then A has an **intra-turn** relation with B.

Below is a concrete example of this error. C_0 and C_1 are two argumentative components from post P_1, while C_2 and C_3 are two consecutive argumentative components from another post P_2 replying to P_1. The MTL_ALL model predicts there exists an intra-turn relation between C_2 and C_3, which is incorrect. In this example, C_0 presents a claim that "There have been many dark animated movies that become famous" and premise C_1 supports this claim with two examples of "The Iron Giant" and "Land Before Time". Although both C_2 and C_3 challenge the connection from one of the two mentioned movies to the claim of C_0, there should not be an intra-turn relation between them. C_2 and C_3 may both support a claim attacking C_0, but they do not support or attack each other.

> P_1: [There have been a great many "dark" animated movies and shows that grew to become extremely famous.]$_{C_0:\text{CLAIM}}$ [If we're using a level of "dark" of the level of Brave Little Toaster then why did things like **The Iron Giant** and **Land Before Time** get a ton of love associated with them.]$_{C_1:\text{PREMISE:SUPPORT:0}}$
>
> P_2: [**Land before Time** has about 15 other movies in the franchise which make it popular, much like toy story,]$_{C_2:\text{PREMISE:ATTACK:}C_1}$ and [**Iron Giant** doesn't really deal with much that's terribly dark or controversial.]$_{C_3:\text{PREMISE:ATTACK:}C_1}$

This type of error raises the number of false positive cases in intra-turn relation identification. As a result, the precision scores of MTL_ALL are inferior to MTL_intra.

7 Conclusion

We show that using multi-task learning with micro and macro structures represented improves the performance of argumentative component classification and two relation prediction tasks, both with and without *refinement*. Also, we observe that combining all tasks may not always be beneficial since we can have conflicts between some of them. Further, our results demonstrate that using the XGBoost model as the final classifier on top of the representation from BERT, while not affecting the recall much, raises the precision scores for the intra-turn and inter-turn relation tasks. In sum, we achieve better results with MTL compared to the single-task training of Chakrabarty et al. (2019b), with and without our refinement and XGBoost enhancements. Future plans include leveraging contextual information to further improve performance and conducting further analyses on the incompatibility between intra-turn and inter-turn relation identification task in MTL.

Acknowledgements

We would like to thank Ahmed Magooda, Mingzhi Yu, Muhammad Salem and Ravneet Singh for their constructive feedback on the initial draft of the paper and the anonymous reviewers for their helpful comments.

References

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259.

Iain Brown and Christophe Mues. 2012. An experimental comparison of classification algorithms for

imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3):3446–3453.

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019a. IMHO fine-tuning improves claim detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 558–563, Minneapolis, Minnesota. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019b. AMPERSAND: Argument mining for PERSuAsive oNline discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An empirical study of multi-task learning on BERT for biomedical text mining. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 205–214, Online. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394, San Diego, California. Association for Computational Linguistics.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41, New Orleans, Louisiana. Association for Computational Linguistics.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany. Association for Computational Linguistics.

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

Reiji Teramoto. 2009. Balanced gradient boosting from imbalanced data for clinical outcome prediction. *Statistical applications in genetics and molecular biology*, 8(1).

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *CoRR*, abs/1603.06270.