Learning from Auxiliary Sources in Argumentative Revision Classification[*]

Tazin Afrin and Diane Litman

University of Pittsburgh, Pittsburgh PA 15260, USA
{taa74,dlitman}@pitt.edu

Abstract. We develop models to classify desirable reasoning revisions in argumentative writing. We explore two approaches – multi-task learning and transfer learning – to take advantage of auxiliary sources of revision data for similar tasks. Results of intrinsic and extrinsic evaluations show that both approaches can indeed improve classifier performance over baselines. While multi-task learning shows that training on different sources of data at the same time may improve performance, transfer-learning better represents the relationship between the data.

Keywords: Writing · Revision · Natural Language Processing.

1 Introduction

Our research focuses on the *automatic classification of desirable revisions of reasoning*[1] in argumentative writing. By reasoning, we refer to how evidence is explained and linked to an overall argument. Desirable revisions (e.g., reasoning supporting the evidence) are those that have hypothesized utility in improving an essay, while undesirable revisions do not have such hypothesized utility [1]. Identifying desirable revisions should be helpful for improving intelligent feedback generation in automated writing evaluation (AWE) systems [5]. In this study, we focus on improving our model learning by taking advantage of auxiliary data sources of revisions. For example, we would like to see how college-level essay data might be beneficial for elementary-school essays. We train two types of models – a *multi-task learning* (MTL) model, and a *transfer learning* (TL) model. In our MTL experiment, we allow information sharing during training using different source data. In our TL experiment, we fine-tune a model pre-trained on source data to see which type of source data might improve performance on the target data. Our results show that both MTL and TL are beneficial for the datasets written by more novice writers. However, for more expert writers (e.g., college students), it is difficult to further improve classifier performance.

Prior Natural Language Processing (NLP) research in academic writing has focused on classifying argumentative revision purposes [8] and understanding revision patterns [5]. While some have classified revisions in terms of *quality* [7],

---

[1] Such revisions of text *content* are considered more useful in revising [8].

Table 1: Comparison of essay corpora used in this study.

Corpus	#Students	Grade Level	Feedback Source	Essay Drafts Used	Essay Score Range	Improvement Score Range
E	143	5th & 6th	AWE	1 and 2	[1, 4]	[0, 3]
H1	47	12th	peer	1 and 2	[0, 5]	[-2, +3]
H2	63	12th	peer	1 and 2	[17, 44]	[-14, +12]
C	60	college	AWE	2 and 3	[15, 33]	-1, +1

[2] were the first to consider a revision's utility in improving an essay in alignment with previously received feedback. They also investigated state-of-the-art models to identify desirable revisions of evidence and reasoning on three different corpora [1]. We extend their revision annotation framework on one additional corpus, as well as *leverage the corpora as auxiliary sources for MTL and TL*. MTL in NLP is widely used to learn from a limited data source (e.g., [6] used the same primary task of argument classification for multiple low-resource datasets framed as MTL). Following [6], we explore MTL to classify desirable reasoning revisions for multiple datasets. To the best of our knowledge this is the *first exploration of MTL for revision classification*. Transfer learning in NLP is used to reduce the need for labeled data by learning representations from other models [3,4]. Unlike previous works, we first train our model using *source revision data*, then fine-tune the model for a *target revision data*.

2   Data and Annotations

Our data consists of reasoning revisions from four corpora of paired drafts of argumentative essays used in previous revision classification tasks [8,1]. All essays were written by students in response to a prompt, revised in response to feedback, and graded with respect to a rubric. A corpus comparison is shown in Table 1. Their diversity along multiple dimensions makes it challenging to train one model for all. However, since our target is to classify revisions following one annotation framework, it is compelling to investigate how these datasets might be related. In the *elementary* (E) school corpus, students wrote Draft1 about a project in Kenya, then received feedback from an automated writing evaluation (AWE) system. All essay pairs were later graded on a scale from 0 to 3 to indicate improvement from Draft1 to Draft2 in line with the feedback [2]. Two corpora contain essays written by *high-school* students and revised in response to peer feedback – H1 and H2. Drafts 1 and 2 of each high-school corpus were graded using separate rubrics by expert graders. We create an improvement score for each essay pair, calculated as the difference of the holistic score between drafts. The *college* (C) essays unlike the other corpora involving proprietary data, were downloaded from the web [8]. Students received general feedback after Draft1, revised to create Draft2, then revised again to create Draft3 after receiving essay-specific feedback from an AWE system. We create a binary improvement score, calculated as 1 if Draft3 improved compared to Draft2, -1 otherwise.

Table 2: Average number of revisions over 10-fold cross-validation is shown before and after data augmentation (D = Desirable, U = Undesirable).

Corpus	N	Before Augmentation			Augmented for MTL			Augmented for TL		
		Total	D	U	Total	D	U	Total	D	U
E	143	389	186	203	5120	2376	2744	7725	3881	3844
H1	47	387	202	185	5120	2750	2370	5780	2963	2817
H2	63	329	169	160	5120	2770	2350	10986	5997	4989
C	60	207	114	93	5120	2894	2226	5515	3186	2329

For all essays in each corpus, sentences from the two drafts were aligned manually based on semantic similarity. Aligned sentences represent one of four operations between drafts – no change, modification, sentence deleted from Draft1, sentence added to Draft2. Each pair of changed aligned sentences was then extracted as a *revision* and annotated for its *purpose* (e.g., revise reasoning), using the scheme introduced for the college corpus [8]. From among the full set of annotations, we only use reasoning revisions for the current study because they are the most frequent across the four corpora. The reasoning revisions were then annotated for its desirability [1]. We leverage the annotated E, H1, and C data from the previous study [1] and extend the annotation of a new data, H2.

Deep learning requires more than our limited amount of data for training. We use the synonym replacement data augmentation strategy from our prior work [1] to generate more training examples. Since MTL is trained batch by batch for each data in a round robin fashion, we selected a fixed number of instances from each dataset to stay consistent. TL used all available data. Table 2 shows the number of desirable and undesirable revisions for each corpus.

3 Models

Single-task Learning Model (STL). Our STL model is a neural network model used in previous desirable revision classification task [1]. The input to the model is the revision sentence pair. The model uses the pre-trained BERT ('bert-base-uncased') embedding with a BiLSTM and a Dense layer. The output is a sigmoid activation function for the binary classification task. Classifying desirable reasoning in each corpus is considered an individual task due to the difference in corpora summarized in Table 1. Following previous work, we also select the learning rate $1e^{-3}$ and batch size 16, and apply the same to all data.

Multi-task Learning Model (MTL). The individual tasks in STL are combined in MTL with a *shared BiLSTM layer*. After encoding the revision using the off-the-shelf BERT encoder, we send this to the BiLSTM layer. The BiLSTM layer learns shared information between different tasks. Each task has an individual Dense layer and a Sigmoid output layer to learn task-specific information. During training, we use the same settings as in STL. In MTL, we train the model in the sequence of C, H1, H2, and E data in a round robin fashion for each batch. This sequence is repeated for all the batches for 10 epochs. Since our

batch size is very small, we believe the training will not be affected substantially by the order of selecting the data. We apply 10-fold cross-validation. During testing, we use the respective task for the respective data.

Union Model. Unlike STL (where we train a separate model for each of the four corpora/tasks), for the Union Model we train only one STL model. We use the union of all task data as input. The training is performed following the MTL (e.g., batch by batch) training process. Compared to STL, the Union model will help us understand if using extra data as a source of information is beneficial. Comparison of the Union and MTL models will help us validate that any MTL improvement is not just due to more training data.

Transfer Learning Models (TL). In transfer learning, we learn from a source dataset and apply it to a target dataset to understand the relationship between our data, which may not be obvious from MTL. The source and target data are taken from all possible combinations of our datasets (Table 1). TL also adopts the STL model to first train with the source data, then fine-tune with the target data using all the augmented data available (shown in Table 2).

4   Results

In our *intrinsic* evaluation, classification performance was compared to baseline models in terms of average unweighted F1-score over 10-folds of cross-validation. We compare MTL with two baselines (STL and Union), while TL was compared against one baseline (STL trained on target data). *Extrinsic* evaluation checked how often desirable and undesirable revisions (gold annotations) are related to improvement score using Pearson correlation. We then replicate the process for the predicted revisions to see if they are also correlated in the same way.

MTL evaluation. Intrinsic evaluation in Table 3a shows that in-general MTL has higher average f1-scores. However, MTL and baseline results are close with no significant difference. Further investigation showed MTL outperformed baselines in identifying undesirable revisions. MTL also showed improvement over Union baseline, indicating that MTL's success over STL is not just due to more training data. Union performed worse than STL, emphasizing the importance of data usage. In extrinsic evaluation, MTL showed significant positive correlations for predicted desirable reasoning for the E data (Table 3b), which is consistent with the Gold correlation. In other cases, either MTL is not consistent with Gold, or the correlation is not to be significant. In contrast, both STL and Union often showed significant correlation to essay improvement. Our results suggest treating our datasets as individual tasks to better relate to student writing improvement (extrinsic evaluation). However, we found sharing features (via MTL) useful for identifying desirable reasoning (intrinsic evaluation).

TL evaluation. Elementary-school students can be considered as the least experienced writers in our datasets considering the age group. Hence the result in Table 3c may indicate that model for elementary-school students needed to learn the structure from better-written essays. Unlike how the H2 data as a source helped the H1 data as the target, the reverse is not entirely true. More-

Table 3: MTL and TL evaluation. Best results are bolded. ↑ indicates TL improved over STL. Extrinsic evaluation shows Desirable results only. Significant predicted correlations consistent with using gold labels are bolded. * p< .05.

(a) MTL Intrinsic Evaluation

Corpus	STL	Union	MTL
E	0.597	0.583	**0.607**
H1	**0.649**	0.631	0.627
H2	0.633	0.622	**0.658**
C	0.613	0.539	**0.619**

(b) MTL Extrinsic Evaluation

	Gold	STL	Union	MTL
E	**0.450***	**0.339***	**0.347***	**0.317***
H1	**0.351***	0.249	0.266	0.222
H2	**0.301***	**0.274***	**0.300***	0.232
C	0.029	0.039	-0.057	0.003

(c) TL Intrinsic Evaluation.

		Target			
		E	H1	H2	C
	STL	0.597	0.649	0.633	0.613
Source	E		0.661↑	**0.652↑**	0.606
Source	H1	0.607↑		0.636↑	**0.644↑**
Source	H2	0.606↑	**0.678↑**		**0.644↑**
Source	C	**0.641↑**	0.638	0.598	

(d) TL Extrinsic Evaluation.

		Target			
		E	H1	H2	C
	Gold	**0.450***	**0.351***	**0.301***	0.029
	STL	**0.339***	0.249	**0.274***	0.039
Source	E		0.262	**0.262***	0.033
Source	H1	**0.337***		**0.308***	0.008
Source	H2	**0.360***	**0.376***		-0.060
Source	C	**0.350***	**0.292***	**0.250***	

over, transfer from C decreased performance for H2. Finally, for C data, transfer learning the weights from both high-school datasets helped improve performance over the baseline (trained only on the target data). Although E as the target domain improved most when learning from the C data, the reverse is not true. College-level students were comparatively experienced writers in our corpora, so inexperienced student writing may have not helped.

Table 3d shows that when our target is the E data, TL results are consistent with Gold annotation results for desirable reasoning. Undesirable revisions are not significant. When H1 is the target, H2 yields the highest correlation, which might be because it is also high-school data. C data also showed significant correlation in this case. While all models are consistent with Gold annotation for H2 as the target, H1 showed the highest correlation of desirable revision to essay improvement score. This is surprising because H1 did not improve in intrinsic evaluation. Finally, C as target did not see any significant correlations.

Overall, transfer learning shows that the availability of more data or more information is not enough. Rather, which data is used to pre-train or how it is being used to train the model (e.g., MTL) is also important. Extrinsic evaluation also supports the fact that more data does not mean improvement. For example, transfer from other high-school data yield stronger results for H1 or H2 data compared to transfer from E or C data. Overall, our results from the transfer learning experiments show that for each target data there were one or more source corpora that improved the classifier performance.

5 Discussion of Limitations and Conclusion

Our corpora were originally annotated using a detailed revision scheme [2], then used a simplified scheme [1] to create binary revision classification task. In a real-world scenario, an end-to-end AWE system deploying our model would have errors propagated from alignment and revision purpose classification and perform lower than the presented model. Moreover, we need to examine our additional but less frequent revision purposes too. We also plan to explore other data augmentation techniques to experiment with more complex models. Although we do not have demographic information, the students in the college corpus include both native English and non-native speakers [8]. Another limitation is that the MTL model training process is slow. The current methods also require GPU resources. Moreover, we only investigated one sequence of training the MTL.

We explored the utility of predicting the desirability of reasoning revisions using auxiliary sources of student essay data using multi-task learning and transfer learing. Both experiments indicate that there is common information between datasets that may help improve classifier performance. Specifically, the results of our intrinsic and extrinsic evaluations show that while desirable revision classification using auxiliary sources of training data can improve performance, the data from different argumentative writing tasks needs to be utilized wisely.

References

1. Afrin, T., Litman, D.: Predicting desirable revisions of evidence and reasoning in argumentative writing. In: The 17th Conference of EACL, May 2023
2. Afrin, T., Wang, E.L., Litman, D., Matsumura, L.C., Correnti, R.: Annotation and classification of evidence and reasoning revisions in argumentative writing. In: Workshop on Innovative Use of NLP for Building Educational Applications (2020)
3. Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., Hwang, A.: AMPERSAND: Argument mining for PERSuAsive oNline discussions. In: Proceedings of the 2019 EMNLP-IJCNLP. ACL (November 2019)
4. Ghosh, D., Beigman Klebanov, B., Song, Y.: An exploratory study of argumentative writing by young students: A transformer-based approach. In: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. ACL (July 2020)
5. Roscoe, R.D., Snow, E.L., Allen, L.K., McNamara, D.S.: Automated detection of essay revising patterns: Applications for intelligent feedback in a writing tutor. Technology, Instruction, Cognition and Learning **10**(1), 59–79 (2015)
6. Schulz, C., Eger, S., Daxenberger, J., Kahse, T., Gurevych, I.: Multi-task learning for argumentation mining in low-resource settings. In: Proceedings of NAACL - HLT, Volume 2 (Short Papers). ACL (June 2018)
7. Tan, C., Lee, L.: A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication. In: Proceedings of the 52nd ACL. vol. 2: Short Papers (June 2014)
8. Zhang, F., Hashemi, H., Hwa, R., Litman, D.: A corpus of annotated revisions for studying argumentative writing. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2017)