Self-trained Pretrained Language Models for Evidence Detection

Mohamed Elaraby
University of Pittsburgh
mse30@pitt.edu

Diane Litman
University of Pittsburgh
dlitman@pitt.edu

Abstract

Argument role labeling is a fundamental task in Argument Mining research. However, such research often suffers from a lack of large-scale datasets labeled for argument roles such as evidence, which is crucial for neural model training. While large pretrained language models have somewhat alleviated the need for massive manually labeled datasets, how much these models can further benefit from self-training techniques hasn't been widely explored in the literature in general and in Argument Mining specifically. In this work, we focus on self-trained language models (particularly BERT) for evidence detection. We provide a thorough investigation on how to utilize pseudo labels effectively in the self-training scheme. We also assess whether adding pseudo labels from an out-of-domain source can be beneficial. Experiments on sentence level evidence detection show that self-training can complement pretrained language models to provide performance improvements.

1 Introduction

In the area of Argument Mining, obtaining high-quality manually labeled data is often a complicated and expensive task (Habernal et al., 2018). Therefore, utilizing unlabeled data can help to achieve further improvements over standard supervised models. Recently, pretrained language models have achieved significant improvements in a wide variety of downstream NLP tasks (Kenton and Toutanova, 2019; Liu et al., 2019; Yang et al., 2019). These models utilize large unlabeled datasets to learn meaningful representations that are transferable across several tasks. In this work, we extend the utilization of unlabelled data in Argument Mining by proposing to use *pretrained language models in a self-training manner*. We focus on evidence detection which is an essential component in building natural language systems that are capable of arguing and debating.

Self-training is a semi-supervised technique that employs unlabeled data by using a teacher model, trained on labeled data, to generate pseudo labels out of unlabeled examples (Yarowsky, 1995; Scudder, 1965). The pseudo labeled data are then blended with manually labeled data to train a student model of a similar or larger size of the teacher model. Another way to use pseudo labels is to blend them with labeled data and train a smaller student model. This method is referred to as knowledge distillation (Hinton et al., 2015).

In this paper, we seek to answer the following research questions:

Q1. Under which conditions can self-training improve finetuned pretrained language models? To answer this question, we experiment with three different techniques to utilize the automatically generated pseudo labels to assess whether large pretrained models can benefit from self-training or not. *(a) Bootstrapped self-training:* where we select automatically annotated instances of high confidence and add them to the training data during finetuning. *(b) Pretrain on pseudo labels:* where we use the selected samples in pretraining the model before finetuning on the manually labeled set. *(c) Masked language model pretraining:* where we use the selected samples to pretrain the model using a masked language model objective before finetuning the model on the manually labeled set.

Q2. Is there a constraint on the domain of unlabeled data? In our experiments, we employ both in-domain and out-of-domain unlabeled corpora.

Q3. Does increasing the similarity between labeled and unlabeled data improve self-training? We provide a retrieval step to filter unlabeled data by considering the most similar N samples to each example in the training data . Our aim is to increase the similarity between labeled and unlabeled data by filtration.

Our contributions are as follows. *(a)* We propose a thorough investigation of self-training meth-

Proceedings of The 8th Workshop on Argument Mining, pages 142–147
Punta Cana, Dominican Republic, November 10–11, 2021.

ods for *evidence detection* over large pretrained language models. *(b)* We empirically show that with proper adjustments, self-training can indeed achieve improvements over a pretrained baseline.

2 Related Work

Argument Mining spans various lines of research work. Stab and Gurevych (2014), Habernal and Gurevych (2017), and Persing and Ng (2015) focus on identifying and classifying argument roles in text. Another direction is to mine argument units that are relevant to specific claims or topics (Shnarch et al., 2018; Biran and Rambow, 2011; Levy et al., 2017). Our work directly extends the work in this area by Shnarch et al. (2018). *Evidence detection*, as viewed in this work, aims at classifying relevant sentences to a certain topic. Shnarch et al. (2018) benefits from large-scale weakly labeled data described in Levy et al. (2017) blended with manually annotated data to train a BiLSTM with GLoVe embeddings and integrate the topic with an attention mechanism. Our work also aims at making use of weakly labeled data. However, instead of using the retrieved data described in Levy et al. (2017) directly, we employ a teacher model trained on the manually annotated data to generate the pseudo labels for the unlabeled set.

Our work falls under the umbrella of semi-supervised learning. In natural language processing, the recent use of unlabeled data has focused on pretraining large language models (Kenton and Toutanova, 2019; Howard and Ruder, 2018; Radford et al., 2018), which has led to remarkable improvements across a wide variety of NLP downstream tasks. In Argument Mining, Chakrabarty et al. (2019) retrieved sentences with seeds as *In My Opinion (IMO)* and *In My Humble Opinion (IMHO)* and used them to finetune a language model before finetuning on a claims dataset. Reimers et al. (2019) used contextualized embeddings (*ELMo* and *BERT*) to classify and cluster argument components. In their work, they reported an 80% accuracy when finetuning BERT over the evidence data described in Shnarch et al. (2018) and 81% when concatenating the topic with the input evidence. Following the same line of work, we first finetune BERT over the same evidence dataset and use it as our baseline model. We then extend finetuned BERT by utilizing it as a teacher in a self-training manner, which hasn't been explored in the literature before.

Self-training has proven to be beneficial in a wide variety of tasks (Yalniz et al., 2019; Xie et al., 2020; Zoph et al., 2020; Kahn et al., 2020; Pino et al., 2020; Wang et al., 2021). In natural language processing, Ruder and Plank (2018) evaluated several semi-supervised baselines on sentiment analysis and part of speech tagging. They built their experiment over a BiLSTM baseline. While their work established solid baselines for semi-supervised learning in neural models, their methods focused on recurrent neural models. On the other hand, our experiments study self-training in the context of large pretrained language models.

Self-trained pretrained language models haven't been studied extensively in the literature. Du et al. (2021) studied self-training as another way of leveraging unlabeled data on top of large pretrained language models. They achieved 2.6% improvements in standard classification tasks. Khalifa et al. (2021) used simple bootstrapped self-training over *BERT* to improve zero-shot and few-shot classification of Arabic dialects. We instead use self-training on top of pretrained language models for sentence-level evidence classification. Our main incentive is to explore the utility of self-training techniques in *argument mining* where acquiring manually labeled data is usually hard to get in large quantities.

3 Datasets

3.1 Manually labeled dataset

We make use of the *IBM debater* evidence dataset[1] created by Shnarch et al. (2018), which is composed of 118 topics chosen from various debate portals. The dataset consists of 5785 topic-dependent sentences in total split into two sets: 4066 instances for training and 1719 instances for testing. For each topic, Shnarch et al. (2018) retrieved sentences from *Wikipedia* which are then manually annotated by 10 workers per topic. The crowd-annotators either select whether a sentence is *evidence* or *non-evidence* for a given topic. Table 1 shows examples from the *evidence* and *non-evidence* sentences in the *IBM debater* evidence dataset.

3.2 Unlabeled datasets

In our experiments, we rely on two sets of unlabeled (for evidence) data. The first one is an in-domain argumentative corpus from *Wikipedia*. The second one is a slightly out of domain (not

[1] https://www.research.ibm.com/artificial-intelligence/project-debater/

sentences	
Topic: "We should limit executive compensation"	
evidence	A February 2009 report, published by the Institute for Policy Studies notes the impact excessive executive compensation has on taxpayers: U.S. taxpayers subsidize excessive executive compensation - by more than $20 billion per year - via a variety of tax and accounting loopholes.
non-evidence	A say on pay - a non-binding vote of the general meeting to approve director pay packages, is practised in a growing number of countries.

Table 1: Labeled *Wikipedia* examples in the *IBM Debater* evidence dataset

Wikipedia) argumentative corpus called Webis-Debate-16, that is unlike Wikipedia, constructed from online debates.

Wikipedia unlabeled data. Following the method described in Levy et al. (2017), we use the data retrieved by querying *Wikipedia* with a query composed of *"that"* + *topic_concept*. The work done by (Shnarch et al., 2018) suggests that the previous query can yield argumentative sentences in general, not just claims . The resultant corpus is composed of $29k$ candidate sentences. Table 2 shows an example unlabeled retrieved sentence out of *Wikipedia*.

sentences	
Topic: "Doping in sport"	
example	In October, the International Volleyball Federation closed the doping charges after concluding that "there is no evidence of an anti-doping rule violation" [REF].

Table 2: Unlabeled example retrieved from *Wikipedia*

Webis-Debate-16 dataset. In order to assess whether out of domain unlabeled data can improve performance, we experiment with the Webis-Debate-16 corpus[2] (created from debates extracted from *idebate.org*) as an unlabeled argumentative source for evidence detection. The dataset is labeled on the sentence level with argumentative versus non-argumentative labels. The dataset contains 10846 argumentative phrases and 5556 non-argumentative phrases, therefore, we utilize it as an unlabeled (for evidence) argumentative source. Table 3 shows examples of non-argumentative and argumentative sentences from the Webis-Debate-16 corpus.

[2]https://webis.de/data/webis-debate-16.html

sentences	
Debate Topic: "Economy"	
non-argumentative	the price tag was set as being £32.7billion.
argumentative	"high speed two will help to solve this inequality by increasing connections between north and south ."

Table 3: (Non-)Argumentative examples (unlabeled for evidence) in the Webis-Debate-16 dataset

4 Approach

Our primary goal is to investigate different methods of self-training on top of large pretrained language models for evidence extraction.

4.1 Finetune-BERT baseline

We start with finetune-BERT as our baseline, which achieved the best results on *IBM Debater* evidence dataset in Reimers et al. (2019). In our experiments, we refer to this baseline as *evidenceBERT*.

4.2 Bootstrapped self-training

Our first self-training setting is bootstrapped self-training, where we employ *evidenceBERT* to annotate unlabeled data. Every epoch, we make predictions over unlabeled data U. For each instance x in U, we extract the probability assigned to the most likely class $p(x) = argmaxM(x)$ where $x \epsilon U$ and M is our *evidenceBERT*. The examples are then ranked based on the probabilities and the *top N* examples selected. In our experiments, we determine N by choosing the percentile of the top examples. Due to limited computational resources, we vary the percentiles from 10% to maximum 50%.

4.3 Pretrain on pseudo labels

Weakly labeled data has been used in pre-training neural models in information retrieval (Dehghani et al., 2017) and sentiment analysis (Severyn and Moschitti, 2015). We employ the generated pseudo labels from *evidenceBERT* to initially finetune BERT before finetuning over the manually labeled set.

4.4 Masked language model pretraining

In these experiments, we use the *top N* examples from the automatically labeled data to train on a masked language model objective. We finetune the masked language model using Wolf et al. (2019).

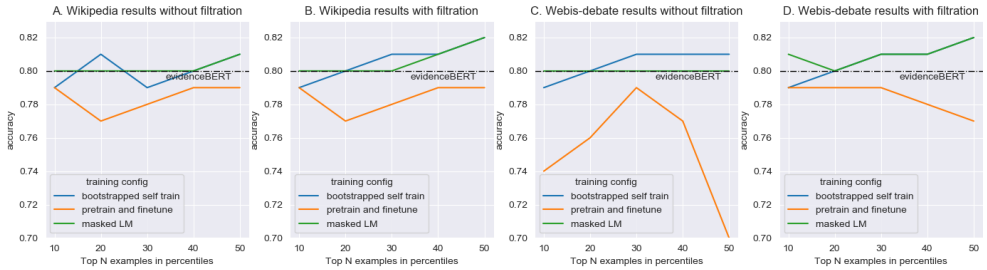

Figure 1: Self-training techniques results

4.5 K-nearest neighbors based filtration

Following the selection from unlabeled data method in Du et al. (2021), we encode manually labeled and unlabeled datasets using *XMLR RoBERTa* (Conneau et al., 2020) which achieves state-of-the-art results on the Semantic Textual Similarity benchmark (Cer et al., 2017). Then, for each instance in the manually labeled set, we select the top 5 nearest neighbors from the unlabeled set. We hypothesize that this process will yield more evidence like data from the whole argumentative set. We employ the new retrieved set in the three configurations of self-training we experiment with.

5 Results

After presenting the baseline pretrained finetuned language model results, we discuss how adding the different self-training approaches sheds light on the three research questions introduced in Section 1.

evidenceBERT. We start by replicating the results in Reimers et al. (2019), using the *BERT* implementation from Wolf et al. (2019). For training hyperparameters, we finetune *BERT* for 6 epochs of training. We employ *Adam* (Kingma and Ba, 2014) optimizer for training and an initial learning rate of $4e-5$. We chose our training parameters based on a manual search optimized on 5% of the training data. We achieve an accuracy of 80% on the test set, which is almost the same result reported in Reimers et al. (2019).

***Q1.* Under which conditions can self-training improve finetuned pretrained language models?** Results suggest that by adding appropriate N pseudo examples, bootstrapped self-training and masked language model pretraining can improve accuracy over *evidenceBERT*. Figure 1 shows that both bootstrapped self-training and masked language model pretraining can improve accuracy by 1% to 2% over finetuned BERT at optimal N.[3] While both masked language model pretraining and bootstrapped self-training can improve performance, the masked language model pretraining is more robust to the selected N pseudo examples (e.g., masked language modeling never degrades baseline performance). On the other hand, Figure 1 implies that utilizing pseudo examples in regular pretraining always performs poorly when compared to the *evidenceBERT* baseline.

***Q2.* Is there a constraint on the domain of unlabeled data?** Comparing Figure 1 C and D to Figure 1 A and B, respectively, suggests that using an unlabeled corpus from a different distribution like *Webis-Debate-16* largely can achieve similar improvements as using unlabeled *Wikipedia*. This is true for masked language model pretraining and bootstrapped self-training, both when comparing results to *evidenceBERT* at optimal N pseudo examples (except for masked language modeling in C versus A) and with respect to robustness over N.

Q3. Does increasing the similarity between labeled and unlabeled data improve self-training? Comparing Figure 1 B and D to Figure 1 A and C, respectively, shows that increasing the similarity between labeled and unlabeled data yields improvements in terms of accuracy at optimal N, particularly when using out of domain unlabeled data as both masked language modeling and bootstrapped self-training improve. Robustness also improves.

6 Conclusion and Future Work

We explored a variety of self-training configurations for evidence detection on top of BERT. Results show that 1) self-training with bootstrapped self-training and masked language model pretraining (but not with pretrain on pseudo labels) can improve finetuned large pretrained language models such as BERT; 2) unlabeled data can be utilized from both in-domain (*Wikipedia*) or out-of-domain

[3] A similar level of improvement was found by Du et al. (2021) in classification tasks outside of argument mining.

(*Webis-Debate-16*) sources; and 3) filtration of unlabeled data via selecting nearest neighbors with semantic similarity improves results. Future plans include covering more Argument Mining tasks and domains, optimizing selection of nearest neighbors instead of using fixed k, and using various blending techniques of pseudo labeled data during training.

7 Acknowledgment

The authors would like to thank Ahmed Magooda, Nhat Tran, and Mahmoud Azab for their fruitful comments and corrections.

References

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 162–168. IEEE.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Tuhin Chakrabarty, Christopher Hidey, and Kathleen McKeown. 2019. 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.

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

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74.

Jingfei Du, Édouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Jacob Kahn, Ann Lee, and Awni Hannun. 2020. Self-training for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7084–7088. IEEE.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Muhammad Khalifa, Muhammad Abdul-Mageed, and Khaled Shaalan. 2021. Self-training pre-trained language models for zero-and few-shot multi-dialectal arabic sequence labeling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 769–782.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, pages 79–84.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552.

Juan Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-training for end-to-end speech translation.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054.

Henry Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371.

Aliaksei Severyn and Alessandro Moschitti. 2015. Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 464–469.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Zhepei Wang, Ritwik Giri, Umut Isik, Jean-Marc Valin, and Arvindh Krishnaswamy. 2021. Semi-supervised singing voice separation with noisy self-training. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. 2020. Rethinking pre-training and self-training. *Advances in Neural Information Processing Systems*, 33.