Students Who Study Together Learn Better: On the Importance of Collective Knowledge Distillation for Domain Transfer in Fact Verification

Mitch Paul Mithun, Sandeep Suntwal, Mihai Surdeanu
University of Arizona, Tucson, Arizona, USA
{mithunpaul,sandeepsuntwal,msurdeanu}@email.arizona.edu

Abstract

While neural networks produce state-of-the-art performance in several NLP tasks, they depend heavily on lexicalized information, which transfers poorly between domains. Previous work (Suntwal et al., 2019) proposed delexicalization as a form of knowledge distillation to reduce dependency on such lexical artifacts. However, a critical unsolved issue that remains is *how much* delexicalization should be applied? A little helps reduce over-fitting, but too much discards useful information. We propose Group Learning (GL), a knowledge and model distillation approach for fact verification. In our method, while multiple student models have access to different delexicalized data views, they are encouraged to independently learn from each other through pair-wise consistency losses. In several cross-domain experiments between the FEVER and FNC fact verification datasets, we show that our approach learns the best delexicalization strategy for the given training dataset and outperforms state-of-the-art classifiers that rely on the original data.

1 Introduction

Neural networks have achieved state-of-the-art (SOTA) performance across many natural language processing (NLP) tasks, usually in supervised settings.

However, it has been shown that there are limitations to these methods caused in part by their over-fitting on statistical and lexical nuances (or artifacts) specific to a dataset (Gururangan et al., 2018), which prevent them from transferring well across domains. A key solution to this problem is to not let these models rely on such dataset artifacts and instead encode the true underlying semantics of the dataset. Also, in recent years fact verification has emerged as a critical task with important societal implications (Vosoughi et al., 2018). Formally, the task is defined as: given a pair of claim and evidence texts, determine if the evidence supports or rejects the claim, or does not have enough information to reach a conclusion. Several fact verification datasets have been proposed recently, based on real-world news articles (Pomerleau and Rao, 2017), Wikipedia based knowledge bases (Thorne et al., 2018), fact verification websites (Wang, 2017), etc. Several transformer networks (Vaswani et al., 2017) based approaches (Liu et al., 2020) have achieved SOTA performance on these tasks.

However, as shown in (Panenghat et al., 2020; Karimi Mahabadi et al., 2020; Gururangan et al., 2018), these models are also similarly affected by syntactic and lexical artifacts seen in other NLP tasks. Specifically, (Suntwal et al., 2019) shows the presence of such artifacts in the Fact Extraction and Verification (FEVER) dataset (Thorne et al., 2018), along with demonstrating that this limits the ability of the trained models to transfer knowledge to other similar datasets such as the Fake News Challenge (FNC) dataset (Pomerleau and Rao, 2017). To mitigate this dependency on such artifacts, they proposed a data distillation (or delexicalization) approach, which replaces some lexical artifacts such as named entities with their type and a unique id to indicate the occurrence of the same artifact in claim and evidence.

A key unresolved issue in this direction is *how much* delexicalization to apply. As indicated in previous work (Suntwal et al., 2019; Mithun et al., 2021), delexicalization reduces overfitting. But too much delexicalization may discard critical information, e.g., replacing *India* with its NE label, say `LOCATION`, may remove contextual information about the country that is necessary for the correct classification of the claim-evidence pair. Our work proposes a solution for this problem, with the following contributions:

(1) Inspired by teacher-student architectures (Tarvainen and Valpola, 2017; Rasmus et al., 2015), we propose a novel architecture that combines data distillation with model distillation to improve cross-

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6968–6973
November 7–11, 2021.

domain performance of neural networks. In particular, our approach relies on multiple students that have access to different delexicalized views of the data but are encouraged to learn from each other through pair-wise consistency losses. We call our approach *Group Learning (GL)*. Once training completes, the student with the best performance is kept for evaluation purposes. Because we rely on a single model at evaluation time, our approach has the same evaluation run time cost as a single classifier. Note that our method can be seen as an inverse of an ensemble strategy, which trains individual models separately but applies them jointly. GL scales better at inference time due to its reliance on a single model at that stage.

(2) We implemented a GL architecture for fact verification using BERT (Devlin et al., 2019), as the classifier, and multiple delexicalized views of the data using FIGER (Ling and Weld, 2012) and CoreNLP (Manning et al., 2014) NER systems. We evaluated the domain transfer of the proposed method using two fact verification datasets: FEVER and FNC. Our results show that our method achieves a cross-domain accuracy of 73.06% when trained on FEVER and tested on FNC, and 74.46% in the other direction, outperforming other stand-alone trained methods that rely on the lexicalized data. Importantly, our approach chooses different students in each direction, highlighting different properties of the respective training datasets.

2 Methodology

2.1 Data Distillation

Based on the findings of (Suntwal et al., 2019) that named entities (NEs) are most likely to overfit in a fact verification task, we delexicalize our data by replacing NEs with their semantic classes (and a unique id). To detect and replace named entities with their most specific label returned by the Named Entity Recognizer (NER), we use their solution of Overlap Aware (OA-NER), which relies on CoreNLP (Manning et al., 2014) NE labels. In addition, we propose two new delexicalization approaches based on the FIGER-NER (Ling and Weld, 2012):

FIGER Abstract : Replaces NEs with the most abstract classes returned by the FIGER NER, (e.g., `LOCATION` for *Los Angeles*).

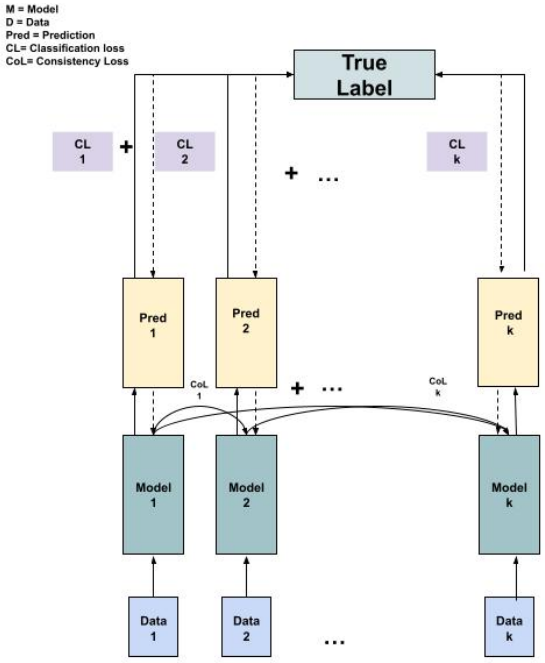

Figure 1: The multiple-student architecture for knowledge distillation.

FIGER Specific : Uses the most specific classes returned by the FIGER NER, (e.g., `CITY` for *Los Angeles*.)

2.2 Model Distillation

We propose a combined distillation strategy to mitigate the risk of distilling the data at the incorrect granularity (overly aggressive or too conservative). Specifically, we introduce a *Group Learning* architecture (shown in Figure 1), inspired from the teacher-student paradigm (Hinton et al., 2015; Tarvainen and Valpola, 2017; Laine and Aila, 2016; Sajjadi et al., 2016). In this architecture, each student method is trained on two techniques:

(a) Different versions of the same dataset, each delexicalized differently by using different data distillation techniques mentioned above.

(b) The distributions of predictions of other models.

This combined methodology of knowledge distillation encourages students to learn as much as possible from their views of the data while jointly learning with other students. Training together on the soft labels (distribution of predictions) of other student methods acts as a form of regularization between all methods. More formally, each student component includes a regular classification loss (implemented using cross-entropy) on their respective data. Additionally, each has a consistency loss

	Claim	Evidence
Plain text	J. R. R. Tolkien created Gimli .	A dwarf warrior , he is the son of Glóin -LRB- a character from Tolkien's earlier novel, The Hobbit -RRB- . Gimli is a fictional character from J. R. R. Tolkien s Middle-earth legendarium, featured in The Lord of the Rings .
OA-NER	personC1 created personC2.	A dwarf warrior, he is the son of personE1 -LRB- a character from personC1's earlier novel, The Hobbit -RRB- . personC2 is a fictional character from personC1's locationE1 legendarium, featured in The Lord of the Rings.
FIGER Specific	authorC1 created locationC1.	A dwarf warrior, he is the son of personE1 -LRB- a character from authorE1's earlier novel, The Hobbit -RRB- . locationC1 is a fictional character from authorC1's written_workE1 legendarium, featured in The Lord of the Rings.
FIGER Abstract	personC1 created locationC1.	A dwarf warrior, he is the son of personE1 -LRB- a character from personC1's earlier novel , The Hobbit -RRB- . locationC1 is a fictional character from personC1's written_workE1 legendarium, featured in The Lord of the Rings.

Table 1: The claim and evidence before and after the data distillation process, along with the distillation technique used. Note that we used actual NERs, which are imperfect tools, to generate these views.

between all other methods that minimize the difference in predicted label distributions.

The intuition behind our approach is that by providing multiple data distillation options to choose from, we encourage the student methods to 'pull' towards each other and the original underlying semantics. The part of semantic knowledge that is obscured from a student method due to the particular delexicalization technique used in the dataset version it sees is instead learned in its effort to perform on par with other methods. Thus, similar to a classroom environment where the students learn from both known labels (e.g., a textbook) and by helping another student learn, each student can thus choose the right amount of granularity needed to enhance its understanding.

2.3 Classifiers

We use BERT (Devlin et al., 2019) as the pretrained model used in all our experiments since it has achieved SOTA results in many NLP tasks, including fact verification. Specifically, we experiment with two variants, BERT-Base, and Mini BERT (Turc et al., 2019), a light-weight version of BERT, both from the Hugging Face repository (Wolf et al., 2019).

3 Experiments

Data: We use two distinct fact verification datasets for our experiments,The Fact Extraction and Verification (FEVER) dataset (Thorne et al., 2018) and the Fake News Challenge (FNC) dataset (Pomerleau and Rao, 2017).

The FEVER dataset consists of 145,449 data points, each having a claim and evidence pair. These claim-evidence pairs typically contain one or more sentences compiled from Wikipedia using an information retrieval (IR) module and are classified into three classes: *supports, refutes* and *not enough information*. The evidence for data points that had the gold label of *not enough information* were retrieved (using a task-provided IR component) either by finding the nearest neighbor to the claim or randomly. Even though the training partition of the FEVER dataset was publicly released, the gold test labels used in the final shared task were not. We therefore built our own test partition by dividing the randomized training partition into 80% (119,197 data points) and 20% (26,252 data points).

The FNC dataset comprises claim-evidence pairs that were divided into four classes, *agree, disagree, discuss* and *unrelated*. These claim-evidence pairs were created using the headlines and content section of real news articles, respectively. While the training partition of the publicly available dataset comprised 49,972 data points, the testing partition had 25,413 data points. We further divided the training partition into 40,904 data points for training and 9,068 data points for development.

In order to evaluate the proposed methods in a cross-domain setting, we modified the label space of the source domain to match that of the target domain as done in (Suntwal et al., 2019).

Setting: In all the experiments, the performance of the underlying method on the respective original, lexicalized data is considered as the baseline. In the baseline model, we use the default hyper parameters from Hugging Face. We focus our analysis on cross-domain evaluation, i.e., we train all methods on one dataset (e.g., FEVER) and evaluate their accuracy on the other dataset (e.g., FNC). At the

end of the training, the best student model from the list of all the trained models is saved to be used for evaluation.

	Configuration			
Train Eval	FEVER FEVER	FEVER FNC	FNC FNC	FNC FEVER
Decomposable Attention Lex	83.43%	48.86%	68.99%	41.16%
Decomposable Attention Delex	75.26%	46.71%	45.51%	51.77%
Mini BERT Lex	83.86%	69.50%	89.33%	54.11%
Mini BERT **GL**	83.74%	73.06%*	89.72%	74.46%*
BERT-Base Cased Lex	90.88%	66.68%	99.21%	73.78%
BERT-Base Cased **GL**	84.88%	**75.37%***	97.07%	75.51%
BERT-Base Uncased Lex	91.95%	64.21%	99.45%	76.59%
BERT-Base Uncased **GL**	86.12%	73.61%*	98.42%	**77.67%**

Table 2: In-domain and cross-domain accuracies for various methods. All scores reported are averaged across three random seeds. "Lex" is the stand alone model trained on the original lexicalized data; "GL" denotes the student in the proposed multi-student 'Group Learning' architecture. Decomposable Attention Delex refers to the best performing model in (Suntwal et al., 2019), a stand alone decomposable attention model (Parikh et al., 2016) which was trained on data that was delexicalized using the OA-NER and Super Sense tagging techniques. * indicates that the corresponding result is significantly better than its baseline ("lex" in the same column), under a bootstrap resampling test with 1,000 samples, and p-value < 0.05.

4 Results

Table 2 summarizes our results. We focus on two sets of experiments using each training method and setting: in-domain (columns 2 and 4) and cross-domain (columns 3 and 5). For comparison purposes, we also show the results from (Suntwal et al., 2019) where a decomposable attention model was used for the same experiments.As shown in Table 2, although all the baseline models perform well (83.43%–99.45%) in-domain, they transfer poorly when evaluated cross-domain where up to 35% drop in performance is observed. This verifies our findings that the signal the model learns from the un-masked text does not generalize well between domains. On the other hand, we observe a marginal in-domain drop in performance for the student models trained on the GL architecture (e.g., 6% in the FEVER/FEVER setting for BERT-Base Cased) compared to their lexical counterparts. This indicates that GL models retain most signal from lexical data. Importantly, the GL models perform considerably better than their corresponding lexical versions across domain (e.g., up to 20.35% improvement in the FNC/FEVER setting for Mini-BERT). This demonstrates that data distillation and model distillation can be successfully combined as a strategy to improve domain transfer of fact verification methods.

5 Discussion and Conclusion

Previous work (Suntwal et al., 2019) has shown that delexicalization is useful in learning domain transferable knowledge. However, the level of delexicalization suitable for each task is unclear. In this work, we provide multiple delexicalization choices to neural network models and encourage them to choose the most appropriate choice. We suspect this approach acts as regularization (through the consistency losses), as well as a form of data noise (because of the imperfect NERs), which has been shown to aid in knowledge distillation paradigms (Hinton et al., 2015; Tarvainen and Valpola, 2017).

Also, note that the BERT models perform better than the decomposable attention (DA) (Parikh et al., 2016) model in most of the cases, especially in the FNC dataset. More importantly, the cross-domain performance gain when using BERT with the proposed group learning architecture is higher than that achieved by the decomposable attention model. This is likely caused by three reasons. First, BERT has a considerably larger number of parameters than DA. Second, in applications involving text pairs, the decomposable attention model individually encodes these text pairs before using bidirectional cross attention. Instead, BERT combines these two stages using the self-attention mechanism that operates jointly over the two concatenated sentences (and separated with the `[SEP]` token). Lastly, BERT is pretrained on massive amounts of texts related to the datasets used here, whereas DA learns its parameters from scratch.

Further, analyzing the selection made by the GL framework for various random seeds, we observed that when trained on FEVER and tested on FNC, GL selects the lexicalized student, while in the other cross-domain direction, the common choice is the student delexicalized with OA-NER. We hypothesize this happens for two reasons. First, the training data in the FNC dataset is smaller

(40,904 data points) compared to the FEVER dataset (119,197 training data points). Therefore it is more prone to overfitting in the original, lexicalized form. Second, as mentioned above, since the FNC dataset is derived from real-world news articles, the number of evidence sentences in the FNC is higher than FEVER sentences. This means that delexicalized sentences in FNC preserve enough lexical signal for training, even in their delexicalized forms. The opposite observations hold in the other direction (FEVER to FNC), which caused the lexicalized students to perform better. Also, please note that even though we use only four student methods in our experiments to train with each other, this can be extended to any number of methods. However, the correct number of student models (and their corresponding delexicalized datasets) for a given task needs to be empirically determined.

Our approach demonstrates that: (a) delexicalization helps model generalization, (b) the amount of delexicalization to apply varies from dataset to dataset, and (c) it is possible to learn how much delexicalization to apply through our proposed GL architecture.

Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under the World Modelers and HABITUS programs. Mihai Surdeanu declares a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee, and is managed in accordance with its conflict of interest policies. The authors would also like to thank Steve Bethard, Becky Sharp, and Marco Valenzuela-Escárcega for all their valuable comments and reviews.

References

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.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. 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 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

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

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Samuli Laine and Timo Aila. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *AAAI*, volume 12, pages 94–100.

Xiaodong Liu, Yu Wang, Jianshu Ji, Hao Cheng, Xueyun Zhu, Emmanuel Awa, Pengcheng He, Weizhu Chen, Hoifung Poon, Guihong Cao, and Jianfeng Gao. 2020. The Microsoft toolkit of multi-task deep neural networks for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 118–126, Online. Association for Computational Linguistics.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Mitch Paul Mithun, Sandeep Suntwal, and Mihai Surdeanu. 2021. Data and model distillation as a solution for domain-transferable fact verification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4546–4552.

Mithun Paul Panenghat, Sandeep Suntwal, Faiz Rafique, Rebecca Sharp, and Mihai Surdeanu. 2020. Towards the necessity for debiasing natural language inference datasets. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6883–6888.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge.

Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. 2015. Semi-supervised learning with ladder networks. *arXiv preprint arXiv:1507.02672*.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171.

Sandeep Suntwal, Mithun Paul, Rebecca Sharp, and Mihai Surdeanu. 2019. On the importance of delexicalization for fact verification. 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 3413–3418, Hong Kong, China. Association for Computational Linguistics.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

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*, pages arXiv–1910.