Effective Knowledge Distillation Generalization for Language Models

Taha Shabani Mirzaei
Department of Electrical and Computer Engineering
The University of British Columbia
Vancouver, BC, Canada
taha.shabani@ece.ubc.ca

Mobina Shahbandeh
Department of Electrical and Computer Engineering
The University of British Columbia
Vancouver, BC, Canada
mobinashb@ece.ubc.ca

Abstract

With the increasing prevalence of large-scale machine learning and deep learning models, deploying such models on real-world hardware becomes harder. Knowledge distillation can transfer knowledge from such cumbersome models to smaller, shallow models. These distilled models aim to capture the larger models' knowledge into a smaller single model, which is easier to deploy without significant loss in performance. In this work, we propose a new diverse dataset along with a student-teacher architecture for our distillation tasks, utilizing generative models for dataset augmentation. Then, this dataset is used to transfer knowledge from the teacher model to the student model. The proposed student model with knowledge distillation and augmented transfer set has reached over 88% accuracy of the teacher model in multiple tasks while having a small and simple architecture.

1 Introduction

With the pervasiveness of large-scale data and complex tasks for machine learning models, these models are getting excessively larger every day. Deploying such unwieldy models, especially on less powerful devices, is often complicated. Although such cumbersome models, e.g., very deep neural networks, language models, and ensembles of many models, have a high knowledge capacity, this capacity may not be effectively used in environments with limited resources. In addition, these models' performance is often optimized on the training data, whereas the target is to generalize to unseen data. Therefore, these models may fail to meet performance and latency expectations on real-world data.

Knowledge distillation helps overcome these impediments by transferring knowledge from a large model to a smaller model without loss of validity. In the natural language processing (NLP) domain, deep learning models such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. [2018]), trained on a very large dataset of text, have obtained notable results on

*Author contributions: Taha proposed the research problem, took care of the teacher part, prepared some of the slides, and wrote part of the paper. Mobina implemented the student models, prepared some of the slides, and wrote part of the paper.

Preprint. Under review.

NLP tasks such as sentiment analysis and text similarity detection. Knowledge distillation (Ba and Caruana [2014]) can be used to transfer knowledge from these large and deep models to shallow word embedding-based neural networks. In this approach, the smaller model tries to mimic the larger model in order to achieve similar or higher accuracy.

The transfer dataset is constructed using sampling on pre-trained language models. The smaller model, called the "student", considers the output of the larger model, called the "teacher", the ground truth. This work's aim is to improve the student model in terms of generalization and performance in different language understanding tasks using a high-quality dataset.

In this regard, first, we offer a new comprehensive dataset using social media and standard benchmarks. Then, we propose a practical approach combining generative models with language models, along with RNN-based student models, for a variety of language processing tasks. Our technique could provide powerful models even in environments with limited resources.

2 Related Work

Considering the elaboration of employing large language models (LM), there is assorted research conducted to transfer the knowledge of these huge models into more shallow architectures. Ba and Caruana [2014] propose knowledge distillation as an approach to enhance the results of a smaller model, i.e., the student model, using larger teacher model outputs as the desired outcome. Assuming we have h_S as an untrained student model and h_T as a fine-tuned teacher model, on classification tasks, the loss of the student models is as below:

$$\mathcal{L}_{\mathrm{KD}} = \frac{1}{N} \sum_{i=1}^{N} \left\| h_S\left(x_i\right) - h_T\left(x_i\right) \right\|_2^2$$

Which N is the size of the dataset, and KD represents the term knowledge distillation.

Student models generally have fewer parameters, so they naturally tend to imitate the teacher models in a specific task. For example, Zhu et al. [2021] introduce a combination of curriculum learning and knowledge distillation for efficient dialogue generation models. They first cluster the dataset according to their complexity. Then, they propose a learnable distillation model to distill knowledge by hierarchical curriculum learning.

In addition to text generation studies, distillation techniques are widely used for classification tasks. For instance, Melas-Kyriazi et al. [2020] leverage large LMs in classification using unlabeled training examples. This technique uses a generative model alongside a LM to generate new training examples across only three low-resource single-sentence classification datasets, then utilizes the main training sets combined with the generated ones to train the student model. Also, Hahn and Choi [2019] have proposed a knowledge distillation method for natural language translation using two soft target probabilities that are obtained based on the word embedding space.

Furthermore, knowledge distillation is used in other areas, such as computer vision and robotics. An example in computer vision is a paper in which the authors have used this technique to recognize human actions from images (Chapariniya et al. [2020]). Additionally, Kwon et al. [2021] have proposed a method of using knowledge distillation for collision detection in robotics.

Aside from the previous work, Tang et al. [2019a] use a bidirectional long short-term memory network (BiLSTM) as the student model. Next, the model is trained with a fabricated training dataset, utilizing rule-based generative techniques for classification tasks such as sentiment analysis and sentence similarity. This paper shows that their final model outperforms OpenAI GPT on some datasets. However, this model only relies on limited classification topics. Moreover, this approach doesn't research the best option for the teacher model. Likewise, Tang et al. [2019b] explore distilling the knowledge from BERT Devlin et al. [2018] as the teacher into a simple BiLSTM-based model with much fewer parameters than BERT. In addition, in Sanh et al. [2019], a distilled version of BERT is introduced with the same general architecture as BERT with removing some layers and optimizing some operations. Nevertheless, these approaches also lack data domain generalization and teacher analysis.

On the other hand, we employ a technique similar to Tang et al. [2019a] but for various tasks in single sentence classification, sentence-pair classification, and sentence-pair similarity. In this regard, we

generate a new diverse dataset using social media data alongside GLUE Benchmark Wang et al. [2019] and analyze the student model with it. Furthermore, we propose optimized student architectures for different tasks, showing comparable results with large transformer-based machine-learning models.

3 Method

3.1 Overview

In order to optimize the pre-defined distillation methods for classification tasks, we proposed a practical approach using a combination of generative and transformer-based language models. As student models are ordinarily small and shallow, they are generally task-specific, e.g., sentiment analysis. We extend the previously provided dataset offered by GLUE Benchmark Wang et al. [2019] and try to provide a generalized dataset for our students.

To achieve this goal, First, we define our initial dataset, which consists of three different tasks as described in Section 4.1. Next, for each assignment, we use the initial dataset to fine-tune GPT-2 (345M parameter version, Radford et al. [2019]). Using the fine-tuned GPT-2, we augment the initial dataset with new unlabeled data. Further analysis for data augmentation is provided in Section 3.2.

Afterward, we utilize the collection of the initial and augmented datasets to, respectively, fine-tune and make predictions with our BERT-based model (teacher). The teacher's prediction would be our ground truth for further steps. We refer to the set of initial and labeled augmented datasets as the transfer dataset. Finally, this transfer dataset is used to train the student.

Additional details about the research methodology are shown in Figure 1

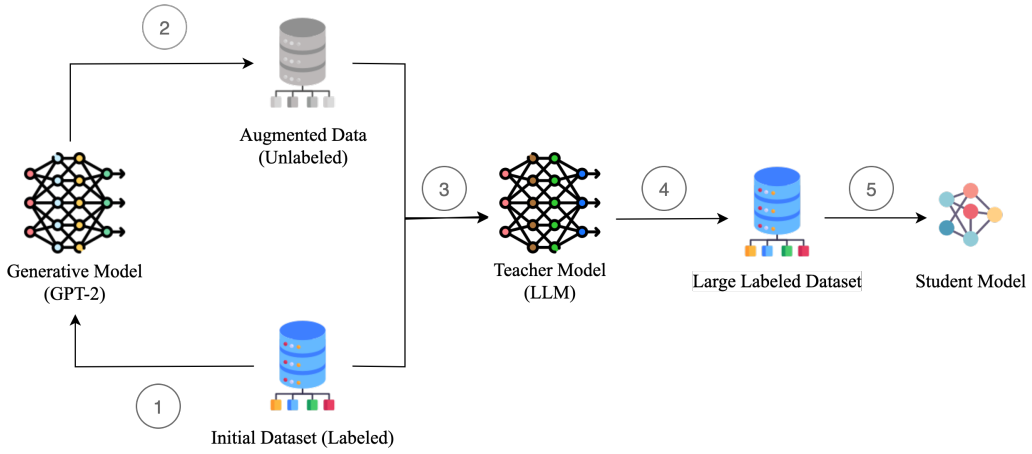

Figure 1: Our proposed distillation procedure with generative models, large language models (LLMs), and BiLSTM. The labels are as follows: (1) Fine-tuning a generative model (GPT-2) using the initial dataset. (2) Generation of the augmented dataset using fine-tuned GPT-2. (3) Fine-tuning LLM (BERT, RoBERTA) with the initial dataset and prediction for augmented data using fine-tuned LLM. (4) Combining the initial dataset with labeled augmented data (transfer dataset). (5) Training the student model (BiLSTM) with the transfer dataset.

3.2 Data Augmentation Using the Generative Model

Although the basis of our approach is to extend the initial dataset using a generative model, we use different techniques for each type of task of our initial dataset. More detail of the initial dataset is provided in Section 4.1.

For single sentence classification tasks, after fine-tuning the generative model, we use a conceptually empty prompt for Text generation. This Prompt contains only the start-of-text (SOT) token. On the other hand, as sentence-pair tasks inquire about the semantic relationship between sentences, we use a generated prompt as the first sentence and generate the second sentence accordingly. Augmenting

datasets for sentence-pair tasks is more time-consuming; consequently, we created less data for these tasks than single sentence classification. Table 1 depicts the size of the original dataset and augmented dataset after repetitive removal.

Table 1: Glue-Benchmark Subset Tasks

Task	Initial size (Train)	Augmented size
CoLA Warstadt et al. [2018]	8.5k	84.4k
SST-2 Socher et al. [2013]	67k	142.2k
MRPC Dolan and Brockett [2005]	3.7k	82.3k
WNLI Levesque et al. [2012]	634	12.4k
RTE Dagan et al. [2010]	2.5k	10.1k
STS-B Cer et al. [2017]	7k	15.2k

3.3 Teacher Model

For the teacher model, we have used pre-trained BERT Devlin et al. [2018], and RoBERTA Liu et al. [2019] and compared the results. The accuracy of both models for our tasks were almost the same. However, we chose BERT, which is a smaller model with better time consumption for fine-tuning. for each task and fine-tuned it on each of the datasets.

3.3.1 Single-sentence Classification

In this task, the target is to classify a single sentence as the input into two or multiple classes. The datasets that we have used to train the models in this task are the Stanford Sentiment Treebank V2 (Socher et al. [2013]) and CoLA (Warstadt et al. [2018]).

3.3.2 Semantic Textual Similarity

In this task, the goal is to classify a single sentence as the input into two or multiple classes. The dataset that we have used to train the models in this task is the Semantic Textual Similarity Benchmark (Cer et al. [2017]). This dataset consists of 5749 samples in the training set, 1500 samples in the validation set, and 1379 samples in the test set.

3.3.3 Sentence-pair Classification

In this task, we aim to determine the similarity or entailment of two sentences as the inputs. The datasets that we have used to train the models in this task are the Microsoft Research Paraphrase Corpus (Dolan and Brockett [2005]), WNLI (Levesque et al. [2012]), and RTE (Dagan et al. [2010]). Our main dataset for these tasks is MRPC, which consists of 3668 samples in the training set, 408 samples in the validation set, and 1725 samples in the test set.

3.4 Student Model

For each of the three tasks, we have considered a proper technique. For the single-sentence classification task, we have constructed a BiLSTM neural network. For the semantic textual similarity task, we have measured the cosine similarity. Finally, for the sentence-pair classification task, we have used an RNN architecture. We have implemented these models on top of Tensorflow (Abadi et al. [2015]).

3.4.1 Single-sentence Classification

In this task, we construct a BiLSTM network with two dense and two dropout layers. The dropout layers were added later as a remedy to the overfitting of the model. Each of these layers has a frequency of 0.1. The input is first tokenized and then fed to the network. We tried both the BiLSTM and CNN architecture, but the former yielded better results in terms of accuracy and loss. Accordingly, we concluded that BiLSTM could model time collection and word dependencies better.

Moreover, we use the Root Mean Squared Propagation(RMSProp) optimizer along with the binary cross entropy loss since this task is a binary classification, and the Adam optimizer seemed to lead to earlier overfitting. The output layer of the model uses the sigmoid function as the activation function

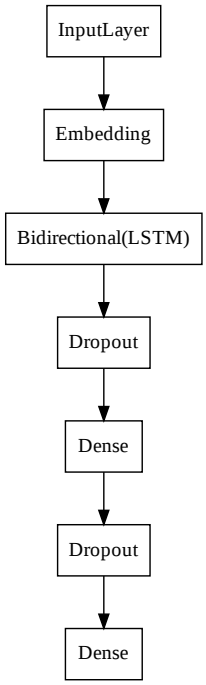

Figure 2: A graph illustrating the architecture of the student model for the single-sentence classification task

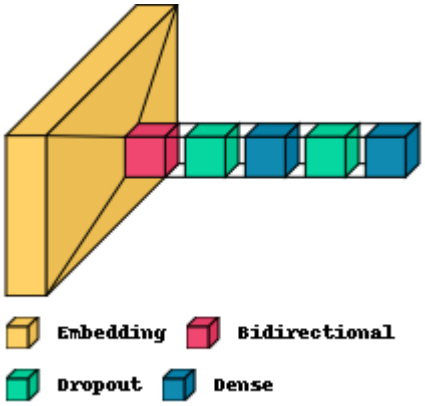

Figure 3: Architecture of the student model for the single-sentence classification task in 3D

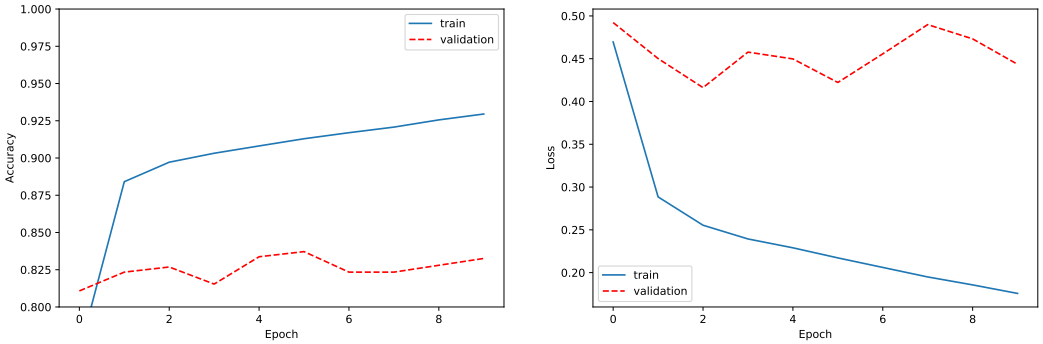

Figure 4: Training and validation accuracy and loss curve of the student model for the single-sentence classification task on the base dataset

and is of size 1. The training and validation loss and accuracy curves for 10 epochs on the base STS-B dataset and without distillation are illustrated in Figure 4. The model tends to quickly overfit on the training set, and we have reduced this effect by adding dropouts, changing the optimizer, and reducing the complexity of the model by reducing the number of neurons in each layer.

3.4.2 Semantic Textual Similarity

To calculate semantic similarity between two sentences, we first preprocess the sentences by tokenizing them into words and removing the stopwords. Then, we calculate the TF-IDF (term frequency-inverse document frequency) of the training set and compute the embeddings of the sen-

tences using the TF-IDF (Sammut and Webb [2010]) of the words of the sentences using equation 1 in which tf represents the term frequency, i.e., the number of times a specific term is present in the document. The idf is calculated as in equation 2.

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D) \tag{1}$$

$$\frac{idf(t, D) = \log |D|}{1 + |\{d \in D : t \in d\}|} \tag{2}$$

We get the embeddings from Word2Vec model (Mikolov et al. [2013]). Finally, we determine the sentence similarity by calculating the cosine similarity of the two embedding vectors of the two sentences as in equation 3. To evaluate the approach, we use the Pearson-Spearman correlation.

$$\cos(\mathbf{t}, \mathbf{e}) = \frac{\mathbf{te}}{\|\mathbf{t}\|\|\mathbf{e}\|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}} \tag{3}$$

3.4.3 Sentence-pair Classification

In this task, we leverage an RNN architecture with fully connected time-distributed recurrent layers with droupouts. As the input, we feed the tokenized data to the model. The weights of the first layer of the model (the embedding layer) are initialized using the GloVe (Pennington et al. [2014]) word embeddings. We use the binary cross entropy loss and the Adam optimizer for this model.

3.5 Distillation Phase

For the distillation step, we feed the dataset labeled by the teacher model to the student model. The student model tries to match its output to the one generated by the teacher model. This results in a smaller model that has learned to mimic the bigger model. We trained the student models with and without the generated sets for each of the tasks.

4 Experiments

4.1 Initial Dataset

As the quality of the initial dataset is crucial for further steps of our research, we need to utilize reliable, diverse, and unbiased textual sources. For this purpose, we formed our initial data using GLUE benchmark Wang et al. [2019] along with Twitter datasets such as Twitter sentiment analysis corpus Naji [2012] and TweetQA Xiong et al. [2019], which consists of question-answer pairs.

Social media platforms play a significant role in our lives. People communicate, express their feelings and passions, and inform or get informed about various news via these platforms. In recent years, Twitter has been a key source of information dissemination as one of the most powerful social networks.

Moreover, GLUE is a manifold benchmark with nine different tasks in three different categories, i.e., single sentence classification, sentence-pair classification, and sentence-pair similarity. A standard language understanding benchmark like GLUE with various tasks could help achieve a more generalized student model. For our analysis, we used a subset of 6 datasets shown in Table 2.

4.2 Evaluation

To evaluate our proposed approach, we consider various tasks. The student model is compared against two baseline models, including BERT itself and the distilled model of BERT proposed in Tang et al. [2019a]. We have also compared our model with a student without distillation.

In this experiment, we compare the student model against the teacher model in the tasks noted in Table2. We have considered the student model without distillation, and the student model with distillation and with the extended transfer set. Table 3 shows the results. As can be inferred from the results, the $\text{Student}_{\text{KD}}$ + $\text{TS}_{\text{GPT-2}}$ model (student model with knowledge distillation and augmented

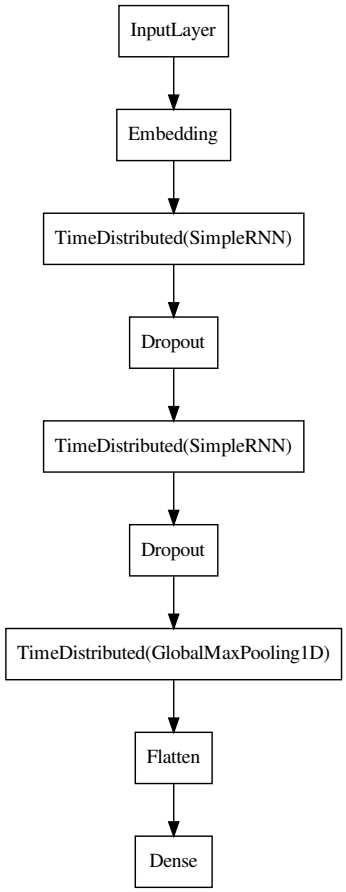

Figure 5: A graph illustrating the architecture of the student model for the sentence-pair classification task

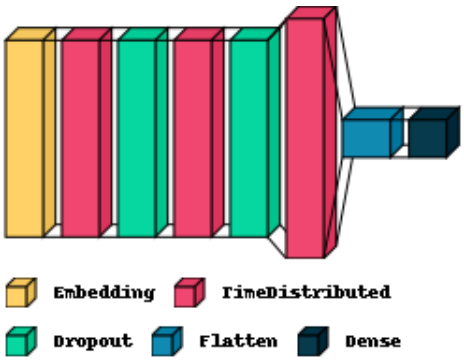

Figure 6: Architecture of the student model for the sentence-pair classification task in 3D

Table 2: Glue-Benchmark Subset Tasks

Task	Type	Description
CoLA Warstadt et al. [2018]	Single sentence classification	Linguistic acceptability
SST-2 Socher et al. [2013]	Single sentence classification	Sentiment Analysis
MRPC Dolan and Brockett [2005]	Sentence-pair classification	Research Paragraph equivalency
WNLI Levesque et al. [2012]	Sentence-pair classification	Entailment and ambiguity
RTE Dagan et al. [2010]	Sentence-pair classification	Textual entailment
STS-B Cer et al. [2017]	Sentence-pair Similarity	Similarity of two statements

transfer set) yields better results in all tasks, although showing very little improvement in some tasks. The $\text{Student}_{\text{KD}}$ + $\text{TS}_{\text{GPT-2}}$ model yields over 90% of the teacher model's accuracy for the SST-2 and WNLI tasks, and over 88% and 76% of the teacher model's accuracy for the RTE and MRPC tasks, respectively.

Table 3: Test results

#	Model	SST-2 Acc.	CoLA MCC	STS-B r/ρ	MRPC F_1/Acc.	WNLI Acc.	RTE Acc.
1	Student	83.3	10.4	64.0/64.3	70.2/58.3	59.6	57.6
2	BiLSTM (GLUE Scratch)	82.8	11.6	70.3/67.8	81.8/74.3	65.1	57.4
3	$\text{Student}_{\text{KD}}$ + $\text{TS}_{\text{GPT-2}}$	**85.1**	**20.1**	64.4/64.4	78.6/65.4	60.9	**61.7**
4	Teacher (BERT)	93.5	52.1	86.5/87.6	88.9/85.4	65.1	70.1

5 Conclusion and Future Work

In this work, we proposed a generalization approach for knowledge distillation in the language understanding domain. We constructed an augmented dataset using the GLUE benchmark, Twitter as a source of social media data, and generative models (GPT-2). Then, we fed the set to the student model for training. We considered multiple tasks, and in all of the tasks, knowledge distillation led to some improvements. The proposed approach yields an accuracy of over 88% of the larger teacher model's accuracy for the smaller student model in most of the tasks, which helps in providing a solution for deploying large language models to less powerful hardware.

To improve the quality of our model, one direction of future work is to explore more complicated student models. Furthermore, examining our approach using a larger dataset and more epochs might help. Similar to the reason for the outperformance of RoBERTa over BERT. Moreover, prompt engineering for creating better initial statements could also be beneficial for the quality of the augmented dataset.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL `https://proceedings.neurips.cc/paper/2014/file/ea8fcd92d59581717e06eb187f10666d-Paper.pdf`.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, 2017. doi: 10.18653/v1/s17-2001. URL `https://doi.org/10.18653%2Fv1%2Fs17-2001`.

Masoumeh Chapariniya, Seyed Sajad Ashrafi, and Shahriar Baradaran Shokouhi. Knowledge distillation framework for action recognition in still images. *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 274–277, 2020.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105, 2010.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002.

Sangchul Hahn and Heeyoul Choi. Self-knowledge distillation in natural language processing. *CoRR*, abs/1908.01851, 2019. URL http://arxiv.org/abs/1908.01851.

Wookyong Kwon, Yongsik Jin, and Sang Jun Lee. Uncertainty-aware knowledge distillation for collision identification of collaborative robots. *Sensors*, 21(19), 2021. ISSN 1424-8220. doi: 10.3390/s21196674. URL https://www.mdpi.com/1424-8220/21/19/6674.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Luke Melas-Kyriazi, George Han, and Celine Liang. Generation-distillation for efficient natural language understanding in low-data settings. *CoRR*, abs/2002.00733, 2020. URL https://arxiv.org/abs/2002.00733.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

Ibrahim Naji. TSATC: Twitter Sentiment Analysis Training Corpus. In *thinknook*, 2012.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Claude Sammut and Geoffrey I. Webb, editors. *TF–IDF*, pages 986–987. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_832. URL https://doi.org/10.1007/978-0-387-30164-8_832.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

Raphael Tang, Yao Lu, and Jimmy Lin. Natural language generation for effective knowledge distillation. In *Workshop on Deep Learning Approaches for Low-Resource Natural Language Processing (DeepLo)*, 2019a. doi: 10.18653/v1/D19-6122.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. Distilling task-specific knowledge from BERT into simple neural networks. *CoRR*, abs/1903.12136, 2019b. URL http://arxiv.org/abs/1903.12136.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. 2019. In the Proceedings of ICLR.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*, 2018.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. Tweetqa: A social media focused question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Qingqing Zhu, Xiuying Chen, Pengfei Wu, JunFei Liu, and Dongyan Zhao. Combining curriculum learning and knowledge distillation for dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1284–1295, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.111. URL https://aclanthology.org/2021.findings-emnlp.111.