Rethinking Task-Specific Knowledge Distillation: Contextualized Corpus as Better Textbook

Chang Liu[1,2], **Chongyang Tao**[1], **Jianxin Liang**[1,3], **Tao Shen**[4], **Jiazhan Feng**[1,3], **Quzhe Huang**[1,3], **Dongyan Zhao**[1,2,5,6*]

[1] Wangxuan Institute of Computer Technology, Peking University
[2] Center for Data Science, Peking University
[3] School of Intelligence Science and Technology, Peking University
[4] AAII, University of Technology Sydney
[5] Institute for Artificial Intelligence, Peking University
[6] The MOE Key Laboratory of Computational Linguistics, Peking University
{liuchang97,chongyangtao,jsleung,fengjiazhan,huangquzhe,zhaody}@pku.edu.cn
tao.shen@uts.edu.au

Abstract

Knowledge distillation has been proven effective when customizing small language models for specific tasks. Here, a corpus as 'textbook' plays an indispensable role, only through which the teacher can teach the student. Prevailing methods adopt a two-stage distillation paradigm: general distillation first with task-agnostic general corpus and task-specific distillation next with augmented task-specific corpus. We argue that such a paradigm may not be optimal. In general distillation, it's extravagant to let the diverse but desultory general knowledge overwhelms the limited model capacity of the student. While in task-specific distillation, the task corpus is usually limited and narrow, preventing the student from learning enough knowledge. To mitigate the issues in the two gapped corpora, we present a better textbook for the student to learn: contextualized corpus that contextualizes task corpus with large-scale general corpus through relevance-based text retrieval. Experimental results on GLUE benchmark demonstrate that contextualized corpus is the better textbook compared with jointly using general corpus and augmented task-specific corpus. Surprisingly, it enables task-specific distillation from scratch without general distillation while maintaining comparable performance, making it more flexible to customize the student model with desired model size under various computation constraints.

1 Introduction

Pre-trained language models (PLMs) have achieved remarkable success in a wide range of tasks (Devlin et al., 2019; Liu et al., 2019; Song et al., 2019; Raffel et al., 2020; Brown et al., 2020). However, their satisfactory performance comes at the expense of high computation cost, which makes them not applicable in resource-scarce scenarios. To ease the burden, substantial efforts have been made to compress large PLMs into small ones with minimum performance degradation, among which we focus on knowledge distillation (Hinton et al., 2015).

In the literature of language model distillation, a major line of research is the objective functions (Sanh et al., 2019; Jiao et al., 2020; Hou et al., 2020; Sun et al., 2020; Fu et al., 2021; Wang et al., 2021; Park et al., 2021; Li et al., 2021; Liu et al., 2022; Zhou et al., 2022) that define how to teach the student. However crucial, there is another factor that plays an indispensable role in knowledge distillation: the corpus as 'textbook', only through which the student can be taught. Previous work (Jiao et al., 2020) proposed a general-then-task-specific distillation paradigm, where the student is first taught on large-scale general corpus by a task-agnostic PLM and then distilled on task corpus by an in-task fine-tuned PLM. Though verified effective, we argue that this paradigm may not be optimal for two reasons. (1) Given the downstream task, it's wise to directly transfer the task-specific knowledge instead of overwhelming the student of limited model capacity with diverse but desultory general knowledge derived from general corpus and task-agnostic teacher. (2) When transferring task-specific knowledge, the task corpus is usually limited and narrow, which is not enough for the in-task fine-tuned teacher to transfer its abundant task-specific knowledge before the student begins overfitting.

We propose that the silver bullet for customizing a small student model into specific tasks is to transfer as much task-specific knowledge as possible while preventing the student model from overfitting. To achieve this goal, an in-task fine-tuned

* Corresponding author: Dongyan Zhao.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10652–10658
December 7-11, 2022

teacher (aka. task-specific teacher) is indispensable without doubt. But what 'textbook' should we use to teach the student? A large-scale general corpus? Abundant and diverse but lacking task-relevance. A task corpus with some kind of data augmentation? Highly relevant to tasks but limited and narrow. Herein, we propose contextualized corpus, a better 'textbook' that marries up task-relevance with abundance and diversity for task-specific knowledge distillation. The contextualized corpus, as the name implies, is constructed by contextualizing the task corpus with large-scale general corpus through relevance-based text retrieval, enriching the limited task corpus with abundant data that not only holds high task-relevance but also keeps as diverse as general corpus. Therefore, it acts as the better textbook with which the in-task fine-tuned teacher can transfer enough task-specific knowledge to the student without worrying about overfitting.

We conduct experiments of task-specific knowledge distillation on GLUE benchmark (Wang et al., 2018). The results demonstrate that the utilization of the proposed contextualized corpus largely improves the performance over the previous method that jointly used general corpus and augmented task corpus. Moreover, we find that contextualized corpus enables us to get rid of general distillation as initialization, making it possible to customize small models with desired model size flexibly.

Our contributions are two folds: (1) We propose contextualized corpus, a better textbook compared with the combination of general corpus and augmentation task corpus, through which task-specific knowledge can be better transferred to student. (2) We conduct comprehensive experiments to study the utilization of contextualized corpus under different distillation settings and provide detailed analyses to demonstrate its superiority.

2 Method

2.1 Contextualized Corpus as Textbook

When transferring the knowledge from a teacher to a student, a corpus as textbook is indispensable. For different purpose, various types of corpus are adopted such as general corpus and task corpus with possible data augmentation (Jiao et al., 2020; Liang et al., 2020). Aiming at task-specific knowledge distillation, we propose a better textbook, contextualized corpus to mitigate the aforementioned issues of the two gapped corpora by marrying up task-relevance with abundance and diversity.

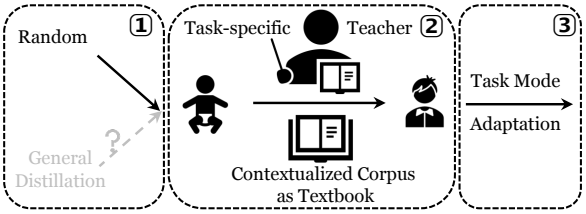

Figure 1: Our framework for task-specific distillation.

Given a task, we aim to gather the task-specific knowledge from abundant and diverse world knowledge as its 'context' to enrich the limited and narrow original task corpus. Specifically, we treat each sentence x_i in a task corpus $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ as an anchor with which we retrieve the top-k relevant sentences from a large-scale general corpus using BM25 (Robertson and Zaragoza, 2009) for its simplicity and efficiency following Yao et al. (2022). Hereby we obtain a contextualized corpus for this task $\mathcal{D}_{\text{CC}} = \{\{x_i^j\}_{j=1}^{k}\}_{i=1}^{n}$. Then we simply merge the retrieved sentences derived from all the anchor sentences and remove duplicate sentences to obtain the final version of contextualized corpus.

2.2 Distillation Framework

Our distillation framework is illustrated in Fig. 1, which can be disentangled into three parts:

(1) Student initialization where a student model is initialized in a task-agnostic manner (i.e., initialized either randomly or with general distillation). We experiment on different student initialization.

(2) Task-specific knowledge injection where an in-task fine-tuned teacher model transfers its task-specific knowledge to the student using some objective functions with some corpus as textbook. We mainly focus on the choice of textbook while adopting the off-the-shelf MiniLMv2 (Wang et al., 2021) as the objective since most studies of distillation objectives in the context of language model compression is task-agnostic. Concretely, we study three types of textbook: *general corpus* which is the combination of English Wikipedia and BookCorpus (Zhu et al., 2015) (abbr. **WikiBook**), *task corpus* with data augmentation by contextual word replacement (Jiao et al., 2020) (abbr. **TaskDA**), and our proposed *contextualized corpus* (abbr. **CC**).

(3) Task mode adaptation where the student equipped with task-specific knowledge is finally adapted to the input-output mode (e.g., classification) of the task, which can be achieved by fine-tuning it with gold labels, distilling it with teacher's

		MNLI	SST-2	RTE	MRPC	CoLA	STS-B	QQP	QNLI
Task	# Cases	392703	67350	2491	3669	8551	5750	363847	104744
TaskDA	# Cases	8017849	1107141	143018	225057	210911	319959	7573531	4229751
	File Size	1.4G	66M	46M	53M	9.2M	39M	924M	940M
CC	# Cases	3848189	3821674	3433959	3777625	3677210	3543323	3732885	3899863
	Top-k	10	150	1000	800	900	500	20	50
	File Size	1.6G	1.5G	1.7G	1.8G	1.6G	1.7G	1.7G	1.8G

Table 1: The statistics of various types of corpus.

predictions, or the combination of both. As we find that these choices of task mode adaptation don't change comparison results of (2), we choose fine-tuning on the labeled task corpus for simplicity.

Under this distillation framework, we aim to systematically study the utilization of contextualized corpus as well as to demonstrate its superiority over other corpora. Therefore, we compare the performance of using contextualized corpus as textbook against other counterparts considering different influential factors, including the student (randomly initialized v.s. with general distillation) and the teacher (task-specific v.s. task-agnostic / general).

3 Experiments

Datasets. We conduct experiments on GLUE benchmark (Wang et al., 2018). Based on original GLUE datasets, we construct two types of corpora: TaskDA and CC. For TaskDA, we follow the exact data augmentation setting as Jiao et al. (2020). For CC, we first collect the anchor set based on which the relevant sentences are retrieved. For single-sentence tasks (i.e., SST-2, CoLA), each sentence is an anchor. For sentence-pair tasks (i.e., MNLI, QQP, QNLI, RTE, MRPC, STS-B), we treat each sentence of the pair as a separate anchor. After removing duplicate anchor sentences in the anchor set, we next retrieve top-k relevant sentences from the 160G pre-training corpus of RoBERTa (Liu et al., 2019) for each anchor sentence using BM25 (Robertson and Zaragoza, 2009) with the extracted keywords as queries (Rose et al., 2010). We choose different top-k for different tasks to make sure CC for each task has similar amounts of data (around 1.6G-1.8G). We build this retrieval system following Yao et al. (2022). The statistics of various corpus are shown in Table 1.

Implementations. For distillation, we adopt $\text{RoBERTa}_{\text{large}}$ (Liu et al., 2019) as the teacher model and a transformer (Vaswani et al., 2017) with 6 layers and 384 hidden dimensions as the student model. When we use CC or WikiBook as the textbook to teach the student, the maximum input length is 128, the maximum training step is 400k, the warmup ratio (WR) is 0.01, the weight decay (WD) is 0.01, the batch size (BS) is 256, the learning rate (LR) is 6e-4, following Wang et al. (2021). As for using TaskDA, when the student model is randomly initialized, the hyperparameters remain the same for a fair comparison. When the student model has already been distilled (i.e., initialized with general distillation or task-specific distillation on CC), we change the following hyperparameters: LR is 1e-4, WR is 0.06, the maximum training steps for tasks are defined by their maximum training epoch which is 10 for MNLI, QNLI, QQP, 20 for SST-2, MRPC, RTE, STS-B and 50 for CoLA, following the suggestion by Jiao et al. (2020). When we first fine-tune $\text{RoBERTa}_{\text{large}}$ and finally fine-tune the task-specific distilled student model, the maximum input length is 128, the training epoch is 10, WR is 0.06 and WD is 0.01. For MNLI, QQP, QNLI and SST-2, BS is 32, LR is 1e-5. For MRPC, RTE, CoLA and STS-B, we choose LR from {1e-5, 2e-5} and BS from {16, 32}.

3.1 Main Results

The main results are shown in Table 2. Overall, the incorporation of contextualized corpus largely boosts the performance. Moreover, we provide three detailed findings as follows:

Contextualized Corpus is the better textbook. It can be observed that however the student is initialized, using the proposed CC as the textbook is better than using WikiBook, TaskDA, and their combination. We first compare WikiBook and TaskDA. When the student is randomly initialized and the task-specific teacher is used, using WikiBook as the textbook is significantly better than using TaskDA on low-resource tasks (e.g., RTE, MRPC, CoLA, and STS-B) and is comparable on other tasks that have moderate or large amounts of training examples. A similar trend can be found in the setting where the student is initialized from general distil-

Init	Teacher	Data	MNLI-m	SST-2	RTE	MRPC	CoLA	STS-B	QQP	QNLI	AVG
	RoBERTa$_{\text{large}}$		90.6	96.2	89.5	90.9	72.3	92.2	92.2	94.8	89.8
	RoBERTa$_{\text{base}}$		87.6	94.8	78.7	90.2	63.6	91.2	91.9	92.8	86.4
Random	General	WikiBook$_{++}$ ♦	84.1	92.0	67.9	87.5	35.6	88.5	90.5	90.5	79.6
		WikiBook	83.2	92.0	65.0	87.0	37.4	87.0	89.8	91.1	79.1
		TaskDA	83.2	92.4	57.0	80.4	40.0	84.3	90.8	89.9	77.3
		CC	84.3	93.5	73.7	87.8	44.8	88.8	91.0	90.8	81.8
	Task	WikiBook	85.8	93.4	78.7	90.9	49.2	90.5	90.9	91.9	83.9
		TaskDA	86.6	93.0	59.9	81.7	37.8	85.0	91.6	89.8	78.2
		CC	87.0	**95.2**	**82.3**	**90.9**	51.4	**91.5**	91.3	92.1	85.2
		CC&TaskDA	**87.6**	94.8	81.2	90.7	**58.4**	91.3	**91.8**	**92.4**	**86.0**
MiniLM$_{\text{v2}}$	General	WikiBook	84.0	91.9	67.3	87.0	36.4	88.1	90.3	90.2	79.4
		TaskDA	84.1	92.4	71.8	88.5	45.0	88.4	90.7	90.5	81.4
		CC	84.2	93.5	72.6	86.5	47.6	88.8	90.9	91.3	81.9
	Task	WikiBook	85.8	93.8	79.8	90.0	47.2	90.6	91.3	92.1	83.8
		TaskDA ♠	87.0	93.6	74.4	88.7	52.2	90.1	91.6	91.7	83.7
		CC	87.1	**95.2**	**82.0**	**91.2**	52.2	**91.4**	91.3	92.5	85.3
		CC&TaskDA	**87.4**	95.0	80.9	90.4	**57.5**	91.0	**91.7**	**92.4**	**85.8**

Table 2: Evaluation results on the dev set of GLUE benchmark. Model with ♦ is the generally distilled MiniLM$_{\text{v2}}$ model using ~160G general corpus, and model labeled with ♠ is the framework proposed by Jiao et al. (2020).

lation. This finding indicates that abundance and diversity are much more crucial than task-relevance when task corpus is limited and narrow, given a task-specific teacher. Moreover, the model using CC achieves superior performance than using WikiBook, TaskDA, and their combination (i.e., previous general-then-task-specific framework (♠) (Jiao et al., 2020)) by increasing task-relevance while not damaging abundance and diversity. In addition, we also explore whether an additional distillation stage using TaskDA can further improve the student distilled with CC and fail to find consistent results among tasks. We leave the strategy of joint utilization of CC and TaskDA for future work.

General distillation is dispensable when provided with contextualized corpus. Comparing the models distilled by task-specific teacher using TaskDA but with different initialization, we can find that a general distillation as initialization significantly improves the performance of low-resource tasks since models easily overfit the limited task data without general distillation as a good initialization point. But now with the introduction of contextualized corpus that is both abundant and task-relevant, the teacher can transfer its abundant task-specific knowledge to a randomly initialized student without worrying about overfitting, making it possible to distill a task-specific student model from scratch while keeping comparable performance with general distillation as initialization. This allows for more flexible customization of student model with various model sizes under different resource requirements rather than being trapped by a few released generally distilled models.

Task-specific teacher itself can transfer task-specific knowledge whatever data is used. Comparing the models using general teacher with task-specific teacher, we can find that the latter is generally much better on all types of data. The most interesting comparison is between two types of teacher on WikiBook, which is a task-agnostic general corpus. The one with task-specific teacher largely outperforms the one with general teacher on all tasks and achieves comparable performance with the previous method (♠) (Jiao et al., 2020) that strongly depends on carefully designed task data augmentation. This finding indicates that task-specific knowledge can be transferred through a general corpus by a task-specific teacher.

3.2 Discussions

Larger CC, better results? Recall that in our primary experiment we collect roughly 1.6G-1.8G data for CC of each task. Now we analyze the influence of data scale based on Figure 2. As the data scale increases (i.e., from purple to red), the curves first move up then coincide on resource-rich task (i.e., MNLI). While for other tasks with moderate or scarce data, the best performance is achieved with top-80% or even top-40% of full CC. This observation indicates that blindly enlarging CC is not worthwhile since there is a trade-off between abundance and task-relevance.

Why does contextualized corpus so helpful? We

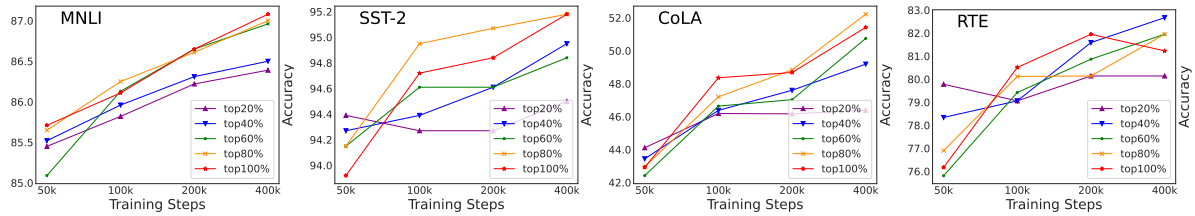

Figure 2: The accuracy curve of different training steps based on different proportions of contextualized corpus.

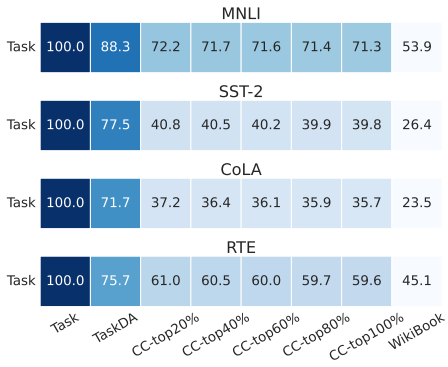

Figure 3: vocabulary overlap among different corpora.

assume the reason why contextualized corpus is such effective lies in that it keeps as abundant and diverse as general corpus while increasing task-relevance. To verify this assumption, we analyze the diversity and task-relevance of different types of corpus. For diversity, we adopt Distinct-n (Li et al., 2016; Tao et al., 2018) (abbr. D-n) which calculates the ratio of distinct n-grams in a corpus as the metric. To make this metric comparable across corpora, we randomly sample a subset for larger corpora to make sure it has approximately the same # uni-grams as the smallest corpus (i.e., Task in the upper block and TaskDA in the lower block in Table 3) in each comparison group. It can be observed from the upper block that both Task and TaskDA are much narrow and lacks diversity compared with CC and WikiBook. Moreover, when augmenting the task corpus from Task to TaskDA, the diversity gap between TaskDA and CC / WikiBook is further widened, indicating that traditional data augmentation method enlarges the task corpus at the cost of diversity degradation. While for task-relevance, we employ top-5k vocabulary overlap as the measurement following Gururangan et al. (2020). From Figure 3 we can indicate that TaskDA keeps high task-relevance with the original Task corpus as we expect, and CC shows much more task-relevance than WikiBook. Therefore by analyzing these two perspectives, we conclude that the superiority of CC lies in that it successfully combines task-relevance with abundance and diversity.

	MNLI		SST-2		CoLA		RTE	
	D-1	D-2	D-1	D-2	D-1	D-2	D-1	D-2
Task	0.6	12.3	2.5	13.5	8.5	42.8	11.9	60.1
TaskDA	0.6	19.1	3.7	37.2	10.9	53.5	11.4	56.1
CC	2.4	31.4	**8.2**	**56.2**	**18.9**	**73.8**	**14.8**	**69.1**
WikiBook	**2.5**	**31.7**	8.1	54.4	18.4	70.9	14.2	65.9
TaskDA	0.03	3.8	0.3	11.8	1.0	16.8	0.4	10.3
CC	**1.0**	14.5	**2.6**	**32.5**	4.9	**46.0**	2.8	33.3
WikiBook	0.8	**15.7**	2.5	32.0	**5.3**	45.4	**3.0**	**34.4**

Table 3: Distinct-n metric of different corpora.

4 Conclusion

We study different influential factors of task-specific knowledge distillation and propose contextualized corpus, a theoretically simple yet highly effective textbook through which the student can better learn task-specific knowledge from teacher.

Limitations. We improve the performance of task-specific knowledge distillation by proposing contextualized corpus, a better textbook for the student to learn task-specific knowledge from the teacher. Though theoretically simple, the construction of contextualized corpus is a bit more complex than traditional data augmentation, which needs a large-scale general corpus as the candidate pool as well as a text retrieval pipeline that should be accurate and efficient.

Ethical Statement. This paper studies task-specific knowledge distillation in natural language understanding and proposes contextualized corpus through which task-specific knowledge can be better transferred to the student. This research doesn't pose ethical issues. The datasets we adopted are publicly available and generally used by other researchers. The proposed method introduces no ethical/social bias.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (No. 2020AAA0106600).

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Hao Fu, Shaojun Zhou, Qihong Yang, Junjie Tang, Guiquan Liu, Kaikui Liu, and Xiaolong Li. 2021. Lrc-bert: latent-representation contrastive knowledge distillation for natural language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12830–12838.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

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

Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic bert with adaptive width and depth. *Advances in Neural Information Processing Systems*, 33:9782–9793.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Lei Li, Yankai Lin, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. 2021. Dynamic knowledge distillation for pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kevin J Liang, Weituo Hao, Dinghan Shen, Yufan Zhou, Weizhu Chen, Changyou Chen, and Lawrence Carin. 2020. Mixkd: Towards efficient distillation of large-scale language models. In *International Conference on Learning Representations*.

Chang Liu, Chongyang Tao, Jiazhan Feng, and Dongyan Zhao. 2022. Multi-granularity structural knowledge distillation for language model compression. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1001–1011, Dublin, Ireland. Association for Computational Linguistics.

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*.

Geondo Park, Gyeongman Kim, and Eunho Yang. 2021. Distilling linguistic context for language model compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 364–378, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1(1-20):10–1002.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936. PMLR.

Siqi Sun, Zhe Gan, Yuwei Fang, Yu Cheng, Shuohang Wang, and Jingjing Liu. 2020. Contrastive distillation on intermediate representations for language model compression. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 498–508, Online. Association for Computational Linguistics.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *IJCAI*, pages 4418–4424.

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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Xingcheng Yao, Yanan Zheng, Xiaocong Yang, and Zhilin Yang. 2022. Nlp from scratch without large-scale pretraining: A simple and efficient framework. In *International Conference on Machine Learning*, pages 25438–25451. PMLR.

Wangchunshu Zhou, Canwen Xu, and Julian McAuley. 2022. BERT learns to teach: Knowledge distillation with meta learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7037–7049, Dublin, Ireland. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.