

BioXtract: Learning Biomedical Knowledge From General and Random Data

Stanford CS224N Custom Project

Mark Endo

Department of Computer Science
Stanford University
markendo@stanford.edu

Jimmy Le

Department of Computer Science
Stanford University
jimmyle@stanford.edu

Gilbert Rosal

Department of Computer Science
Stanford University
rosalg@stanford.edu

Abstract

The privacy of medical documents and protected healthcare information can often-times limit the accessibility of accurate biomedical natural language processing models. Distillation can be used to transfer knowledge from these models, but it typically relies on having related data to distill on. In this work, we investigate the distillation of BERT-based biomedical models using transfer datasets from varying domains, including general data, randomized general data, and biomedical data. We find that general data can be used to learn task-specific biomedical knowledge, especially when we can initialize student models with similar weights to the teacher. We observe that randomized general data can also be used to transfer knowledge, but it is not as effective as general data. We hope that our findings bring attention to both the benefits and potential dangers of the widespread use of mixed-domain pretraining in NLP, particularly relating to models that continue their pretraining process on private data.

1 Key Information to include

- Mentors: Akshay Smit, Pranav Rajpurkar

2 Introduction

Model distillation has proven to be a valuable technique in natural language processing for transferring the knowledge of one model to another [1]. Typically, distillation is done using a transfer dataset relevant to the task-at-hand [2, 3]. However, this can limit the ability to create distilled models, particularly when teacher models are trained using private data. This issue is especially apparent in the biomedical setting, where patient data is oftentimes private. Recent work shows that BERT-based APIs can be extracted using solely general or random data [4]. At the same time, the concept of distilling (or extracting) biomedical BERT-based models using general or random data remains unexplored.

In this work, we distill biomedical BERT-based models using transfer datasets containing either general, randomized general, or biomedical data. First, we find that student models can learn biomedical task-specific knowledge using teacher model outputs on general and random data. We discover that student models better learn this knowledge when the teacher uses mixed-domain pretraining over pretraining from scratch on biomedical data. For question answering, distilling with

general data reduces the AUC gap from BERT initialization to the teacher model by 48.4%, and using random data reduces this gap by 23.2%. Second, we observe that fine-tuning student models on general data without using teacher model outputs results in inferior performance. This shows that the biomedical task-specific knowledge learned during distillation with general data is caused by biomedical knowledge transfer from the teacher instead of the student solely learning general task knowledge that can be applied during final evaluation. Lastly, we find that distilling with more relevant data to the evaluation task does not necessarily greatly improve results. For example, students trained on question answering biomedical data only perform 4.3%-6.8% better than students trained on general data. We hope that our findings bring attention to both the benefits and potential dangers of the widespread use of mixed-domain pretraining in NLP.

3 Related Work

3.1 Model Distillation and Extraction

With the increasing popularity of large mixed-domain pretraining models such as BERT, GPT-2, and GPT-3, distillation of NLP transformers became popular as a way to reduce their size and make them more deployable on resource-restricted machines like mobile devices. Unlike some model compression techniques, model distillation does not require student and teacher models to share the same architecture [5]. This same approach of training a model on predicted outputs of another is also used in model extraction. In this context, the goal of extraction is to duplicate the functionality of a model without prior knowledge of its parameters or training data [6]. Most related to our work, Krishna et al. [4] demonstrated that general and even nonsensical data from adversaries can be used for successful model extraction. The authors explore implications from successfully extracting models with general and random data. First, model extraction becomes significantly cheaper because adversaries no longer have to use good data—which means we can have cheaper models. However, second, cheaper extraction means higher susceptibility to security leaks represented in the data collected. Lastly, cheaper extraction also means intellectual property theft becomes more likely. We extend this work to the biomedical setting, exploring whether BERT-based biomedical models can be distilled using general or random data. Currently, distillation of biomedical models is done with relevant task-specific biomedical data [2, 3]. However, this limits the distillation ability of models which are trained on private data.

3.2 Biomedical NLP Models

With the rise of transfer learning in natural language processing and the use of large, pretrained transformer models, many have investigated how to adapt this pretraining setup to the biomedical setting [7, 8, 9]. Such methods initialize models with weights tuned on general domain texts (e.g. Wikipedia and BookCorpus) and continue pretraining on large-scale biomedical corpora. These methods assume that domain-specific pretraining can benefit by starting from general language models. Recent work challenges this assumption by pretraining from scratch on an abundant amount of biomedical text [10]. The authors reach state-of-the-art performance with this pretraining from scratch method, showing that with enough compute power the mixed-domain pretraining setup can be outperformed. In our work, we analyze how effectively the two biomedical model setups can be distilled.

4 Approach

4.1 Distillation

We distill various BERT-based biomedical models using transfer datasets from differing domains. In distillation, a student model is trained on a transfer dataset with the objective of matching its output target distribution to that of a teacher model [5]. Specifically, we minimize the following loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N q^T \log q^S + (1 - q^T) \log(1 - q^S) \quad (1)$$

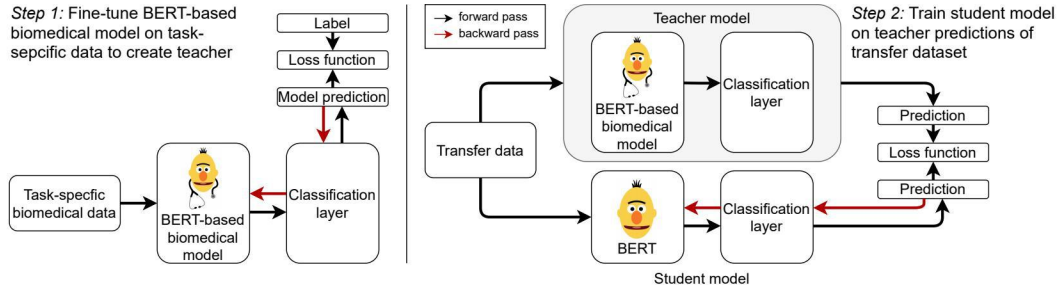


Figure 1: Overview of our model distillation setup. Teacher models are obtained by fine-tuning BERT-based biomedical models on task-specific data. From there, student models are trained on the teacher predictions of a transfer dataset, which is comprised of either general data, randomized general data, or task-specific biomedical data.

where T and S represent the teacher and student respectively, and q is the probability of the positive class of the form $1/(1 + \exp(-z/\tau))$ (z represents the logits and τ is the sigmoid temperature parameter). We don't include ground truth labels as part of the loss function since we aren't focused on the student model's performance on the transfer dataset. Instead, we are using the transfer dataset to learn knowledge from the teacher model that can be applied to the task-specific biomedical evaluation dataset. Excluding ground truth labels in our loss also allows us to use a large amount of unlabeled data.

Our transfer datasets are comprised of either general data, randomized general data, or task-specific biomedical data. We go into more detail on the specific datasets used for each task in the experiments section.

4.1.1 Teacher Model Pretraining

We experiment with various teacher models to analyze the relative difficulty of distilling a BERT-based biomedical model pretrained with mixed-domain data versus one pretrained from scratch on domain-specific data. Mixed-domain pretraining refers to the approach of initializing a model with weights learned on general data and then continuing the pretraining process on domain-specific data (often referred to as continual pretraining). One such example is BioBERT [7]. This approach is initialized with BERT weights [11], which are tuned on Wikipedia and BookCorpus data. Then, BioBERT continues pretraining on PubMed data. Domain-specific pretraining from scratch refers to the process of training solely on biomedical data without any initialization from weights tuned general data. In our experiments, the domain-specific pretraining model we use is PubMedBERT [10], which is pretrained from scratch on PubMed abstracts and full texts. Each of our teacher models is fine-tuned on the final evaluation task.

4.1.2 Student Model Architecture

In all of our experiments, student models are initialized with BERT-Base weights [11]. We leave the experimentation of smaller student model architectures to future work.

4.2 Fine-Tuning

As a comparison to our distillation method, we use the transfer datasets to directly fine-tune BERT-initialized models. In this case, teacher model output predictions are replaced by ground truth labels during training. We use this method to examine whether a student model's performance on the downstream task can be attributed to learning biomedical knowledge from the teacher's outputs of the transfer dataset, or if it is solely caused by training on data that has some relevance to the evaluation dataset. This is particularly relevant when the transfer datasets are comprised of either general data or randomized general data.

4.3 Implementation details

Our training and evaluation code is modified from the BioBERT repository, which uses Hugging Face Transformers [12]. We built the distillation workflow, which saves teacher models’ target outputs on the transfer dataset then trains student models on these outputs. We also developed the testing architecture, as well as the data-preprocessing steps to get all the datasets in the same format for each task.

5 Experiments

5.1 Data

5.1.1 Question Answering

We use the BioASQ dataset to evaluate model performance on biomedical question answering [13]. The dataset contains yes/no, factoid, list, and summary questions. Similar to Gu et al. [10], we focus on the yes/no questions (Task 8b) and leave the other tasks to future work. We use a modified version of the original test set, balanced to include equal numbers of yes and no answers. Also, we split around 900 examples off of the train set to create a dev set.

We use BoolQ, BoolQ-Random, PubMedQA, and BioASQ as transfer datasets. BoolQ is a yes/no QA dataset containing questions that are “naturally occurring” [14]. We use this as our general transfer dataset. To create BoolQ-Random, similar to Krishna et al. [4], we gather all passages within the BoolQ dataset to form a new vocabulary and employ a unigram frequency-based sampling on this vocabulary to form new passages. Then, we uniformly sample words from the passages and prepend starter question words (e.g. "is", "what", "how") to the beginning to create questions. PubMedQA is a biomedical-related yes/no QA dataset collected from PubMed abstracts¹ [15]. Compared to BoolQ and BoolQ-Random, PubMedQA and BioASQ are more conventionally used transfer datasets in this setting. This is because they are very related to the downstream task (BioASQ even is the downstream task). As such, we use these biomedical transfer datasets for baselines to compare against using general and random data.

5.1.2 Relation Extraction

For domain-specific teacher to student training, we used the Exploring and Understanding Adverse Drug Reactions (EUADR) dataset as the transfer set [16]. The dataset has approximately 500 data points. For general teacher to student training, we used the TAC Relation Extraction Dataset (TACRED) as the transfer set [17]. Although the dataset has approximately 100,000 data points, we used 60,000 for the sake of time (and the selection of 60,000 was arbitrary). We also preprocessed both of these datasets to coincide with the original format of BioBERT Relation Extraction computation.

We use the Genetic Association Dataset (GAD) as the test set to evaluate model performance on biomedical relation extraction [18]. The GAD dataset consists of approximately 5,000 examples that had a sentence classifier structure mirroring the format used by the original BERT: a [CLS] token representing the words to consider for relation surrounded by context. For example, "Results do not support any overall association of the Ala-9Val @GENE\$ polymorphism to the development of @DISEASE\$", where the tokens represent the words we are checking for relation. Because of GAD’s smaller size, we used a five-fold cross validation as inspired by previous work [19].

5.2 Evaluation method

For question answering and relation extraction, we use AUROC (AUC) as the evaluation method on the BioASQ test dataset. For relation extraction, we use the averaged AUC scores on the 5 crossfold-validation of the GAD test dataset.

5.3 Experimental details

For training, we use an AdamW optimizer with a learning rate of $8e-6$ and a maximum gradient norm of 1. For distillation, we use a sigmoid temperature of 1. Each model is trained on the transfer dataset

¹PubMedQA answers are of the form yes/no/maybe, but we remove the questions with maybe answers

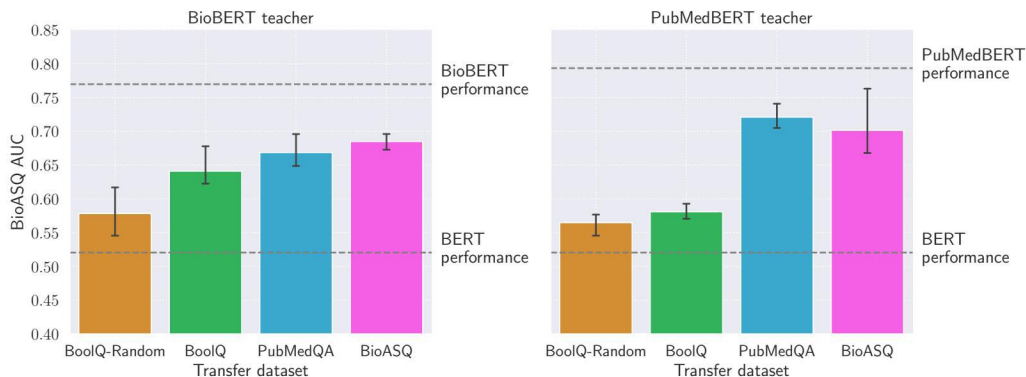


Figure 2: Biomedical question answering AUC scores for distillation using either BioBERT or PubMedBERT as teachers and various transfer datasets. Dashed lines denote teacher model and BERT performance. Using a BioBERT teacher model results in the best distillation performance on general data.

for approximately 1000 to 1200 steps with a batch size of 12. We run each model 3 times using different random seeds and report the average score on the evaluation task.

5.4 Results

5.4.1 Learning Biomedical Knowledge From General and Random Data

When using a teacher model that employs mixed-domain pretraining (BioBERT), we find that distillation with general data and even randomized general data improves upon the BERT initialization biomedical-task performance. On question answering, distilling with teacher outputs of the BoolQ dataset results in a BioASQ AUC of 0.641, reducing the AUC gap from BERT initialization to the teacher model by 48.4%. When distilling with randomized general data, we still find that student models improve upon BERT initialization. For question answering, using BoolQ-Random as the transfer dataset reduces the AUC gap from BERT initialization to the teacher model by 23.2%.

With this teacher model setup, we find that using more relevant data as the transfer dataset does not radically improve student model performance. For example, student models trained on the PubMedQA dataset only perform 4.3% better than students trained on general data. Similarly, students trained on BioASQ perform 6.8% better than students trained on general data.

When using a teacher model pretrained from scratch on domain-specific data (PubMedBERT), we find that student models are unable to learn as much knowledge from the teacher’s outputs of general data. When using PubMedBERT’s outputs of the BoolQ dataset, student models obtain a 0.581 AUC score on the BioASQ test set, which is .060 lower than using a BioBERT teacher. At the same time, distilling the PubMedBERT teacher results in improved performance when using more relevant data as the transfer dataset. For example, student models trained on PubMedBERT outputs of PubMedQA have a .052 higher AUC on BioASQ than models trained on BioBERT outputs of PubMedQA. Detailed results are shown in Figure 2.

5.4.2 Fine-tuning on Transfer Datasets

We find that fine-tuning on general data does improve upon BERT initialization performance, but the best student model trained on teacher model outputs of general data outperforms this approach. Fine-tuning BERT on BoolQ data achieves an AUC of 0.590 on BioASQ, compared to an AUC of 0.641 when training on BioBERT outputs of BoolQ. We see a similar result when working with the PubMedQA dataset. Fine-tuning BERT on PubMedQA data achieves an AUC of 0.680, compared to an AUC of 0.721 for the best performing student model. The only time we see fine-tuning on the transfer dataset outperform distillation is when the transfer dataset is the same used to evaluate performance. Detailed comparisons between distillation and fine-tuning methods are shown in Figure 3.

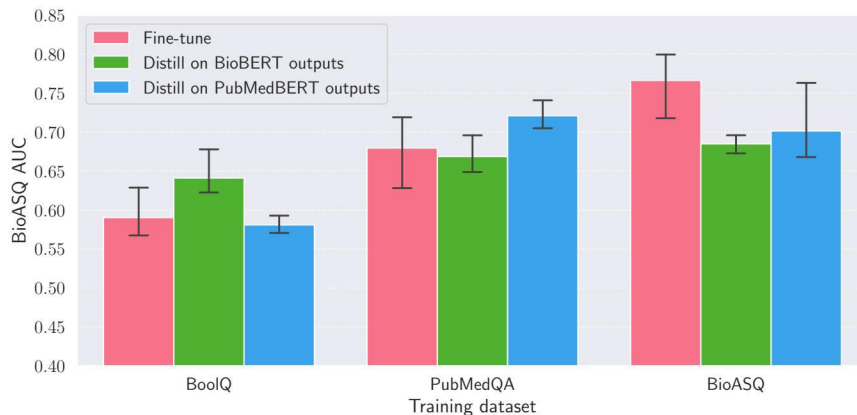


Figure 3: Biomedical question answering AUC scores for fine-tuning versus distilling on training datasets. Distillation outperforms fine-tuning when using BoolQ and PubMedQA as the training dataset, whereas fine-tuning performs the best when training on the BioASQ dataset.

6 Analysis

6.1 Similarity Between Student Initialization and Teacher Matters

As shown in the results section, we find that student models better learn biomedical task-specific knowledge from general data when the teacher uses mixed-domain pretraining instead of pretraining from scratch on biomedical data, even when the teacher model pretrained from scratch has a better performance. This can be caused by the fact that student initialization (BERT) more closely matches a teacher that uses mixed-domain pretraining. This claim is supported by previous work which determined that model extraction works better when the attacker (student) uses the same model architecture as the victim (parent) [4]. More generally, this finding illustrates both the benefits and potential dangers of the widespread use of transfer learning in NLP. On one hand, the availability of large pretrained language models allows us to effectively distill models that employ this pretraining approach, even without access to the fine-tuned training data. On the other hand, this ease of distillation may not be desired, especially when training datasets are private and costly to acquire.

When using biomedical task-specific data, we no longer see a correlation between student model performance and its similarity to the teacher model. One potential reason for this could be that once the transfer dataset is similar enough to the downstream task, the student model is learning valuable knowledge from the training dataset itself. This is supported by our finding that fine-tuning on BioASQ outperforms the distillation of teacher models on BioASQ.

6.2 Fine-tuning Requires Data Relevant to the Evaluation Task

In our results section, we find that fine-tuning on general data is outperformed by training on BioBERT’s outputs of general data. This shows that the biomedical task-specific knowledge learned during distillation with general data is caused by biomedical knowledge transfer from the teacher instead of the student solely learning general task knowledge that can be applied during final evaluation. It also suggests that merely fine-tuning on general data can cause the model to overfit on the general dataset whereas training on a teacher model’s probability outputs does not run into this issue. At the same time, we find that fine-tuning on the same dataset that is used for final evaluation outperforms a distillation setup. This result is somewhat expected since we are much less likely to overfit to the train set.

6.3 Relation Extraction Limitations

When attempting our second task of relation extraction, although we made significant progress, we ran into numerous problems that prevented us from getting significant data. We successfully fine-tuned our baseline teacher models on biomedical task-specific data. We also used those teacher

models to write predictions on the domain specific dataset (GAD) and a general dataset (TACRED). When creating the transfer datasets, we had to handle converting TACRED to a suitable format for the models and obtaining probabilities instead of classifications (which are what BioBERT and BERT normally provided) for the transfer dataset that would be used for training by the student. We were able to get students to run preliminary tests, successfully running them on the train and test sets, but realized the datasets needed adjustment such that the student model could successfully learn. Ultimately, we were held back as a result of time constraints and cache issues with the transformers package. The biggest challenges were learning the interface and outputs of various packages used by the student and teacher models, and manipulating those outputs. Relation extraction results are in Appendix 1.

7 Conclusion

In this work, we distill biomedical BERT-based models using transfer datasets from various domains. We find that general data can be used to learn task-specific biomedical knowledge, particularly when we initialize student models with similar weights to the teacher. In addition, we find that fine-tuning student models on general data without using teacher model outputs does not replicate our results. This shows that the student models are truly learning biomedical task-specific knowledge from the teacher’s outputs on general data. We also observe that randomized general data can be used to distill biomedical BERT-based models, but its performance is inferior to using general data. When distilling with more relevant biomedical data, we find that the student initialization similarity to the teacher no longer plays as large of a role on student performance. The primary limitation of our work is that we have not shown that our results generalize to tasks beyond question answering. This can be explored in future work to better prove that general data can universally be used to effectively distill task-specific biomedical transformer models. Other future work includes using smaller student models, testing the ability to learn with general and randomized data, and also improve upon the usability of teacher models. Also, future work can look at different methods for randomizing data, such as using question generators for the question answering task. Another major limitation involved relation extraction: the size of our domain specific data was incredibly small, totalling to around 5500 pieces of data. Especially for model distillation, smaller datasets prove incredibly ineffective, so future work could invest in better data for the teacher models to distill to the student model. In closing, we hope that our findings bring attention to both the benefits and potential dangers of the widespread use of mixed-domain pretraining in NLP.

References

- [1] Subhabrata Mukherjee and Ahmed Hassan Awadallah. Distilling bert into simple neural networks with unlabeled transfer data, 2020.
- [2] Hongbin George Du and Yanke Hu. Squeezebiobert: Biobert distillation for healthcare natural language processing. In Sriram Chellappan, Kim-Kwang Raymond Choo, and NhatHai Phan, editors, *Computational Data and Social Networks*, pages 193–201, Cham, 2020. Springer International Publishing.
- [3] Macarious Abadeer. Assessment of DistilBERT performance on named entity recognition task for the detection of protected health information and medical concepts. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 158–167, Online, November 2020. Association for Computational Linguistics.
- [4] Kalpesh Krishna, Gaurav Singh Tomar, Ankur Parikh, Nicolas Papernot, and Mohit Iyyer. Thieves of sesame street: Model extraction on bert-based apis. 2020.
- [5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.
- [6] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis, 2016.
- [7] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

- [8] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets, 2019.
- [9] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2020.
- [10] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing, 2021.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- [12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.
- [13] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel Ngonga, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138, 2015.
- [14] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions, 2019.
- [15] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering, 2019.
- [16] Erik M. van Mulligen, Annie Fourier-Reglat, David Gurwitz, Mariam Molokhia, Ainhua Nieto, Gianluca Trifiro, Jan A. Kors, and Laura I. Furlong. The eu-adr corpus: Annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, 45(5):879–884, 2012. Text Mining and Natural Language Processing in Pharmacogenomics.
- [17] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45, 2017.
- [18] Àlex Bravo, J. P. González, N. Queralt-Rosinach, Michael Rautschka, and L. Furlong. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16, 2015.
- [19] Balu Bhasuran and Jeyakumar Natarajan. Automatic extraction of gene-disease associations from literature using joint ensemble learning. *PLOS ONE*, 13(7):1–22, 07 2018.

A Appendix

Baseline Models	GAD AUC
BioBERT	0.808
BERT	0.701
Distill On Biomedical Data	GAD AUC
BERT + BioBERT outputs of EUADR	0.471
Distill On General Data	GAD AUC
BERT + BioBERT outputs of TACRED	0.503

Table 1: Biomedical relation extraction AUC scores for distillation using BioBERT.