# Transformer-based Language Models for Factoid Question Answering at BioASQ9b

Urvashi Khanna, Diego Mollá

*Macquarie University, Australia*

**Abstract**

In this work, we describe our experiments and participating systems in the BioASQ Task 9b Phase B challenge of biomedical question answering. We have focused on finding the ideal answers and investigated multi-task fine-tuning and gradual unfreezing techniques on transformer-based language models. For factoid questions, our ALBERT-based systems ranked first in test batch 1 and fourth in test batch 2. Our DistilBERT systems outperformed the ALBERT variants in test batches 4 and 5 despite having 81% fewer parameters than ALBERT. However, we observed that gradual unfreezing had no significant impact on the model's accuracy compared to standard fine-tuning.

**Keywords**

Transfer learning, DistilBERT, ALBERT, Question Answering, BioASQ9b

## 1. Introduction

Nowadays, the use of language models that have been pretrained on massive amounts of data are the norm [1, 2, 3]. Rather than making significant task-specific architecture improvements, these pretrained models can be fine-tuned for various tasks by making minor changes to the language model architecture, such as adding an output layer on top. Fine-tuning approaches are critical for learning the distributions of the target task and improving the language model's adaptability. However, fine-tuning a language model on small datasets like BioASQ can lead to catastrophic forgetting and overfitting. Furthermore, training all layers simultaneously on data of different target tasks may result in poor performance and an unstable model [4]. A schedule for updating the pretrained weights may be critical for preventing catastrophic forgetting of the source task's knowledge. Scheduling techniques like chain thaw [5] and gradual unfreezing [6] have improved the performance of multiple Natural Language Processing (NLP) tasks. Gradual unfreezing involves gradually fine-tuning model layers rather than fine-tuning all layers at once.

Pretrained language models are usually trained on general language and then adapted to downstream tasks of varied domains. Many domain-specific tasks, however, face the problem of the scarcity of labelled datasets. Auxiliary signal through multi-task fine-tuning helps the language model to adapt on smaller datasets better [7, 8, 9]. Multi-task fine-tuning (also referred

*CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania*

Urvashi.Khanna@mq.edu.au (U. Khanna); Diego.Molla-Aliod@mq.edu.au (D. Mollá)

https://researchers.mq.edu.au/en/persons/diego-molla-aliod (D. Mollá)

0000-0003-2345-5596 (U. Khanna); 0000-0003-4973-0963 (D. Mollá)

to as sequential adaptation in some literature [4]) is the intermediate fine-tuning stage in which the model is fine-tuned on a larger dataset before fine-tuning on a low-resource dataset. In this paper, we describe the experiments of our participating systems[1] at the BioASQ9b challenge[2]. We discuss two of our systems, mainly focusing on factoid questions. Both systems adapt the multi-task fine-tuning technique of fine-tuning on a larger dataset before fine-tuning on the BioASQ9b dataset. Our first system fine-tunes the pre-trained model ALBERT on SQuAD2.0 and then on the BioASQ9b dataset. This system performed exceedingly well on BioASQ9b Test batches 1 and 2. Our second system investigates the effect of the gradual unfreezing technique on the smaller, compact transformer-based model, DistilBERT. We assess this system via two of our submissions at the BioASQ9b Challenge. One of our submissions of DistilBERT ranked sixth in the BioASQ9b leaderboard[3]. From our results, we conclude that gradually unfreezing DistilBERT had no significant improvement in the accuracy of the BioASQ9b test data in comparison to standard fine-tuning.

The rest of this paper is structured as follows. In Section 2, we briefly discuss related work for background. Section 3 describes the BioASQ dataset and the processing steps involved. Section 4 details our experimental setup for both our systems. Section 5 discusses the results of our systems on the BioASQ public leaderboard. Finally, Section 6 provides a conclusion to our work.

## 2. Related Work

Transfer learning has been widely used to transfer knowledge across multiple domains. The scarcity of sizable domain-specific datasets and the cost associated with manually annotating them are driving this trend. In this section, we discuss previous works that used transfer learning for the BioASQ biomedical question answering task [10].

In the 5th BioASQ challenge, Wiese et al. [11] explored domain adaptation to transfer knowledge from an already existing neural Question Answering (QA) system named FastQA [12] that was trained on SQuAD [13]. They initialised their model with the pretrained FastQA models' parameters during the fine-tuning phase. Using a combination of fine-tuning and biomedical Word2vec embeddings, their model achieved state-of-the-art results. They also used optimisation approaches such as L2 weight regularisation and forgetting cost term to minimise catastrophic forgetting.

Lee et al. [14] discovered the potential to adapt the general domain language model BERT for the biomedical domain. They presented BioBERT, the first biomedical language model. In the pretraining step, BioBERT was initialised with BERT weights and then pretrained on biomedical domain corpora. BioBERT produced benchmark results on a wide range of biomedical text mining tasks, including question answering, relation extraction, and named entity recognition. Yoon et al.'s [15] submission for task 7b topped the leaderboard in the 7th BioASQ challenge. They used a sequential adaptation technique in which pretrained BioBERT was fine-tuned first on the SQuAD dataset and then on the BioASQ dataset.

[1] Code associated with this paper is available at https://github.com/urvashikhanna/bioasq9b

[2] http://bioasq.org/

[3] http://participants-area.bioasq.org/results/9b/phaseB/

Similarly, BioELMo [16] is a biomedical version of ELMo that outperforms BioBERT on the authors' probing tasks when used as a feature extractor. However, the fine-tuned BioBERT outperforms BioELMo on named entity recognition and Natural Language Inference (NLI) tasks.

Hosein et al. [17] studied domain portability and error propagation of BERT-based QA models through their BioASQ7b submissions. Their results concluded that general domain language models could generalise and give good results for domain-specific tasks. They also observed that pretraining is more critical than fine-tuning when improving the domain portability of BERT QA models. For yes/no questions in the BioASQ7 Phase B challenge, Resta et al. [18] used an ensemble of classifiers with input from various transformer-based language models. They employed contextual embeddings from multiple pretrained language models, such as BERT and ELMO, as features to capture long-term dependencies.

Jeong et al. [9] expanded the prior work on BioBERT models [14, 15] in the 8th BioASQ challenge. They adapted multiple stages of fine-tuning by first fine-tuning BioBERT on the NLI dataset [19], then on the SQuAD dataset [13], and finally on the downstream BioASQ dataset. Their results established that tasks like NLI that capture the relationships between sentence pairs improve the accuracy of the QA systems. Additionally, they analysed and reported the number of unanswerable questions from the BioASQ7b dataset in the QA setting. Kazaryan et al. [20] used ALBERT [2] as their base language model which was fine-tuned first on SQuAD v2.0 [21], and subsequently on the BioASQ8b data.

## 3. BioASQ Data Processing

BioASQ [10] is an international biomedical challenge that comprises annual tasks on semantic indexing and biomedical question answering. The ninth BioASQ challenge consists of two shared tasks. Task 9a is a semantic indexing task that aims to annotate new PubMed articles automatically [22] with Medical Subject Headings (MeSH). Task 9b is a question answering task devised for systems to answer four types of biomedical questions: factoid, summary, list, and yes/no. The participants are provided with questions along with relevant snippets. The output generated by their systems is either an exact answer (for yes/no, factoid, and list questions) or ideal answers (for summary questions), or both. The tasks are released in five batches over two months, with 24 hours to submit the answers after the release of each test batch.

We primarily concentrate on factoid questions from the BioASQ9b dataset. The dataset contains a total of 3743 questions, 1092 of which are factoid questions. An example of a factoid question is shown in Figure 1. Our system returns exact answers for factoid-type questions that can either be a single entity or a list of entities. We regard the BioASQ challenge task as an extractive QA task because the answer to the query is extracted from the relevant snippet. The metrics used for evaluating the systems on the BioASQ leaderboard are: Strict Accuracy (SAcc), Lenient Accuracy (LAcc), and Mean Reciprocal Rank (MRR). However, MRR is the official metric used by the BioASQ organisers for factoid questions since it is often used to evaluate other factoid QA tasks and challenges [10].

The BioASQ dataset is transformed into the SQuAD format and vice versa using pre-processing and post-processing steps. In a typical span-extractive question answering task, the system is provided with a passage P and a question Q, and it must identify an answer span A ($a_{start}, a_{end}$)

**Question:** Which is the third subunit of the TSC1-TSC2 complex upstream of mTORC1?
**Ideal Answer:** TBC1D7 was identified as a stably associated and ubiquitous third core subunit of the TSC1-TSC2 complex...
**Exact Answer:** TBC1D7
**Type:** Factoid
**Snippet 1:** The tuberous sclerosis complex (TSC) tumor suppressors form the TSC1-TSC2 complex...
**Snippet 2:** Here, we identify and biochemically characterize **TBC1D7** as a stably associated and ubiquitous third core subunit of the TSC1-TSC2 complex...

**Figure 1:** Sample factoid question [23]. The answer to the question is in **bold** and is extracted from snippet 2.

in P. The SQuAD dataset is an example of a span prediction QA task containing many question-answer pairs and a passage that answers the given question. In contrast, the training dataset of BioASQ includes a question, an answer, and multiple relevant snippets. Therefore, we begin by pairing each snippet with its question and transforming it into multiple question-snippet pairs. Also, based on the exact answer provided, we locate the answer's position in the snippet and populate it as the start position of the answer span in the dataset. After performing these pre-processing steps, the BioASQ9b training data samples increased five-fold from 1092 to 5447. Table 1 shows the number of questions in the training and test batches before and after pre-processing.

**Table 1**
Summary of BioASQ9b Training and Test data before and after pre-processing.

| Dataset | Number of Factoid Questions Before Pre-processing | Number of Factoid Questions After Pre-processing |
|---|---|---|
| Training | 1092 | 5447 |
| Batch 1 | 29 | 139 |
| Batch 2 | 34 | 151 |
| Batch 4 | 28 | 132 |
| Batch 5 | 36 | 148 |

Our system returns the prediction span for each question. Because we divided the snippets into several question-snippet pairs during the pre-processing stage, we now have predictions of multiple answer spans and their probabilities for each question. Each system must submit a list of up to five responses for the official BioASQ evaluation. As a result, we select the top five answers for each question in decreasing order of probability as our submission. Thus, for each factoid question, our system returns a list of up to five responses sorted by their likelihood.

## 4. Systems Overview

This section describes our systems and the experimental setup of our submissions at the BioASQ9b challenge. Our submissions in the BioASQ9b challenge are based on two pretrained models: "DistilBERT" and "ALBERT". As mentioned above, we focus mainly on factoid questions. We submitted ALBERT variants for all the BioASQ9b test batches except test batch 3. DistilBERT-

based systems were submitted in test batches 2, 4, and 5. In this section, we detail the models, the methodology used, and the experimental setup.

### 4.1. ALBERT

For the system using ALBERT, we follow a staged fine-tuning approach by fine-tuning on a large dataset before fine-tuning on the smaller dataset. This preliminary stage of fine-tuning on a large QA task is ideal due to the small size of the BioASQ dataset. However, large-scale bio-medical QA datasets are not readily available that could be used for the first stage of fine-tuning. Therefore, we use the SQuAD dataset, a widely used extractive QA dataset. Thus, we first fine-tune ALBERT on SQuAD2.0 and later on our downstream BioASQ task. This approach is illustrated in Figure 2.

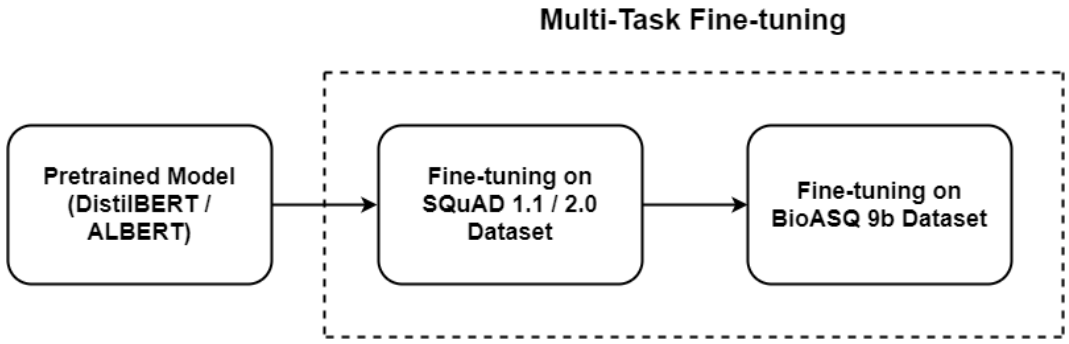


**Figure 2:** Diagram depicting our system's fine-tuning strategy.

ALBERT is a lighter version of BERT with considerably fewer parameters. Lan et al. [2] used two parameter-reduction strategies to lower the memory usage and increase the training speed of BERT. Since ALBERT models scale better than BERT, we have used the xxlarge version of ALBERT for our experiments. The BioASQ task was set up as a span-extraction QA task in which the model predicts the start and end span of answers for a given context and question. In both stages of fine-tuning, the input to the model is the concatenation of passage and question with a special token [SEP] separating them. This input is tokenized using WordPiece embeddings [24] to handle the out-of-vocabulary issues. After WordPiece tokenization, the maximum allowable input sequence length is 512 for both the ALBERT and DistilBERT models. The input has three embeddings: token, position, and sentence. In order to differentiate between the sentences, sentence embedding is appended to each sentence, and a special position token is added to identify the position of each token. The model returns the start and end scores for each word. The output of the model is the candidate span with the highest score and where the end position is greater than or equal to the start position.

We employed "ALBERT-xxlarge" version 2 as our pretrained language model along with its tokenizer, which are publicly available from the Huggingface Transformers Library [25]. This model has an additional task-specific linear question answering layer on top to output the start and end spans. Unless otherwise specified, the hyperparameters for both fine-tuning stages

were set to the default values used by the ALBERT developers. The systems were validated on the BioASQ7b test batches 1 and 2.

All the three ALBERT-based submissions use the same fine-tuning approach discussed above with slight changes to the fine-tuning hyper-parameters. The systems along with hyperparameters are listed in Table 2 and their results are listed in Table 3.

**Table 2**
ALBERT-based systems along with the hyperparameters.

| System Name | Learning Rate | Batch Size | Sequence Length | Epochs |
|---|---|---|---|---|
| ALBERT 1 | 3e-5 | 4 | 512 | 3 |
| ALBERT 2 | 2e-5 | 4 | 512 | 4 |
| ALBERT 3 | 1e-5 | 4 | 512 | 3 |

## 4.2. Gradual Unfreezing DistilBERT

In recent years, the pretrained language models are getting bigger and deeper with millions, sometimes billions of parameters [2, 26]. The success of these models on NLP tasks has fueled the race to scale up the models further. However, deploying these massive models on mobile and edge devices has implications such as environmental impact and computational cost [27], making them unsuitable for use in real-world applications. Sanh et al. [28] applied knowledge distillation [29] and proposed a smaller language model, DistilBERT, that achieves performance comparable to BERT on various NLP tasks. DistilBERT, a distilled, compact version of BERT, has 60% fewer parameters than BERT.

The focus for our second system was to study the effect of gradual unfreezing on the transformer-based language models. We used DistilBERT as our pretrained model to conduct the experiments of gradually unfreezing the transformer layers. The reason for this choice was the small size of DistilBERT and its ability to achieve close to 95% of all the NLP task benchmarks when compared to BERT.

The process of fine-tuning allows the model to learn the distribution of the downstream task. In standard fine-tuning, all the layers of the model are trained on the target task simultaneously. Howard et al. [6] introduced a fine-tuning approach of gradually unfreezing one layer at a time, starting from the top layer. They used a standard Long Short-Term Memory (LSTM) network without any attention mechanism for their experiments. Our work investigates the gradual unfreezing approach on DistilBERT using BioASQ9b as our target dataset.

DistilBERT has three blocks of layers: one embedding layer, six transformer layers, and a top task-specific layer. In our approach shown in Figure 3, we begin by fine-tuning only the top task-specific layer for one epoch while keeping all other layers frozen. Then we unfreeze the transformer layers consecutively in groups of three, fine-tune all the unfrozen layers for one epoch, and repeat until all layers are fine-tuned except the embedding layer. The decision to keep the embedding layer always frozen was based on the preliminary experiments in our previous work [23]. As a result, DistilBERT's trainable parameters have been reduced from 65 million to 42 million.

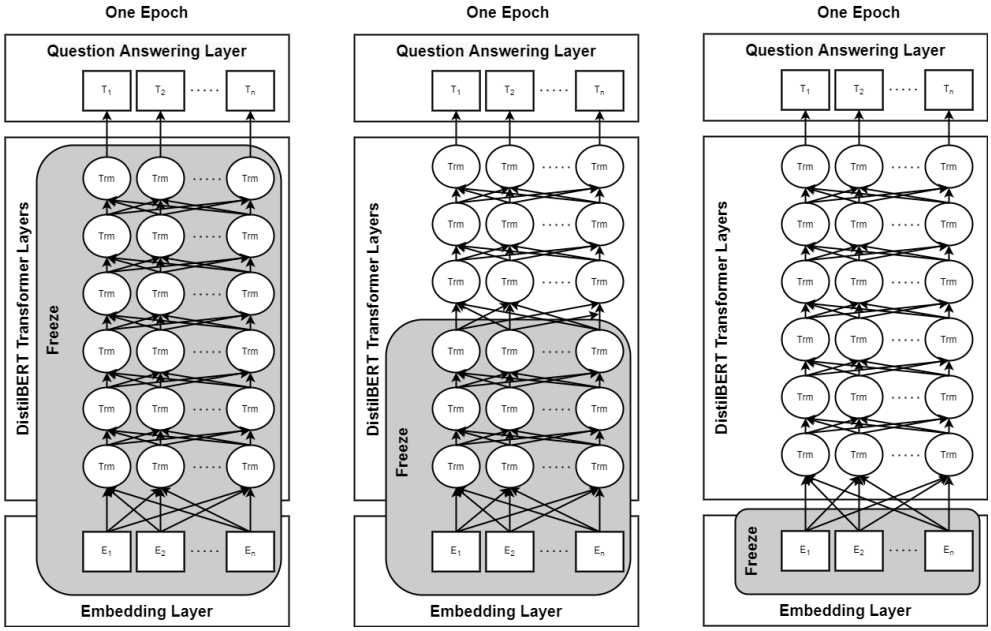


**Figure 3:** Diagram showing our unfreezing approach.

In this system, "distilbert-base-cased" [25] is first fine-tuned on SQuAD1.1 data and then on BioASQ9b task. Our gradual unfreezing approach is only applied during the second stage of fine-tuning. In the second phase of fine-tuning, we fine-tune the model at a constant learning rate of 3e-5, sequence length of 512, and for three epochs. We evaluate the unfreezing approach through two submissions at the BiOASQ challenge. The system "DistilBERT" is our baseline system. In this system, all the layers of DistilBERT are fine-tuned simultaneously. The system "Unfreezing DistilBERT" is the model that was fine-tuned using our unfreezing approach. Both systems are fine-tuned with the same hyperparameters for a fair comparison. Table 3 lists our systems with the results, along with the top-ranked system in the BioaASQ9b leaderboard. We have reported the MRR in the results table since it is the main metric used by the BioASQ organisers.

## 5. Results

The results of our submissions to the BioASQ9b Phase B challenge are shown in Table 3. From the results, we observe that "ALBERT 2" system was the best system for batch 1, and the "ALBERT 3" system was ranked fourth on the public leaderboard of the BioASQ9b challenge. Overall, the systems using the pretrained ALBERT weights have performed exceedingly well on test batches 1 and 2. However, our ALBERT variants received poor results for test batches 4 and 5. It is worth noting that all the systems will be evaluated by humans experts after the competition. However, because this data was not accessible at the time of writing this study, we rely on automatic evaluations available on the BioASQ leaderboard.

**Table 3**
Results of our five submissions along with the top-ranked system from the BioASQ9b leaderboard. The first column of the table lists the unique submission identifier along with the system names as displayed on the public leaderboard. The highest score for each batch is in **bold**.

| Submission ( Display name) | System | Factoid - Mean Reciprocal Rank (MRR) | | | |
|---|---|---|---|---|---|
| | | Batch 1 | Batch 2 | Batch 4 | Batch 5 |
| MQ TL1 (ALBERT) | ALBERT 1 | 0.4379 | 0.4667 | 0.369 | 0.4468 |
| MQ TL2 (Ensemble) | ALBERT 2 | **0.4632** | 0.501 | 0.4167 | 0.4731 |
| MQ TL-3 (Another ALBERT) | ALBERT 3 | 0.4621 | 0.5319 | 0.4375 | 0.4778 |
| MQ TL4 (Final BERT) | DistilBERT | - | 0.5059 | 0.5399 | 0.5171 |
| MQ Transfer Learning (MRes) | Unfreezing DistilBERT | - | 0.4887 | 0.5893 | 0.4917 |
| Top Ranked System | - | **0.4632** | **0.5539** | **0.6929** | **0.588** |

**Table 4**
Results from our previous work [23] on the BioASQ7b dataset. The system 'KU DMIS Team' [30, 15] is BioBERT based system that was top of the leaderboard in the BioASQ7b challenge.

| Systems | Mean Reciprocal Rank |
|---|---|
| KU-DMIS Team [30, 15] | 0.5235 |
| DistilBERT-fine-tuned | 0.4844 |
| DistilBERT-unfreeze-3 | 0.4841 |

The most noticeable difference between our DistilBERT and ALBERT variants, apart from their sizes, is the initial fine-tuning stage. In our systems, ALBERT was fine-tuned on SQuAD2.0, whereas DistilBERT was fine-tuned on SQuAD1.1. The SQuAD2.0 dataset is a reading comprehension dataset that, in addition to the SQuAD1.1 dataset, contains approximately 50,000 unanswerable questions. We need to look into whether test batches 1 and 2 had more unanswered questions after the organisers release the golden answers, and if so, how it has affected the results.

From the results of Table 3, we observe that both "DistilBERT" and "Unfreezing DistilBERT" outperformed the ALBERT variants for the test batches 4 and 5. Our system "Unfreezing DistilBERT" is ranked sixth in the BioASQ9b public leaderboard. The average MRR score of test batches 2, 4 and 5 for systems "DistilBERT" and "Unfreezing DistilBERT" is 0.5209 and 0.5232 respectively, and the difference is not statistically significant[4]. Thus, we can conclude that gradually unfreezing the transformer-based models has no significant impact on the model's accuracy compared to typical fine-tuning. These results further support the findings of our previous work [23] on gradually unfreezing DistilBERT with the BioASQ7b dataset, the results of which are shown in Table 4. The results show that gradually unfrozen models produce promising results for a few test batches, but have no overall significant impact across all the test batches.

[4]Paired t-tests were used to compute the statistical significance since the MRR can be considered as a normal distribution as it is an average of samples. We find no statistically significant difference between the gradually unfrozen model and the baseline.

## 6. Conclusion

Our participation in BioASQ9b was primarily focused on generating the ideal answers for factoid questions. We participated in four test batches, with our systems employing pretrained ALBERT and DistilBERT language models. The results were mixed, with ALBERT-based systems ranking amongst the top systems for test batches 1 and 2. For test batch 4, the compact DistilBERT variants, although having 81 percent fewer parameters, scored considerably better than ALBERT. This paves the way for a biomedical version of DistilBERT for mobile and edge devices for real life biomedical QA applications. In addition, we investigated the effect of gradual unfreezing on transformer-based language models using the BioASQ9b dataset. We conclude that gradually unfreezing the layers of DistilBERT had no significant impact on the model's accuracy in comparison to standard fine-tuning. We also investigated an unfreezing approach that makes use of only 66% of DistilBERT's parameters when fine-tuning. In the future, we will aim to investigate ensemble or hybrid models of DistilBERT and ALBERT.

## References


[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[2] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, arXiv preprint arXiv:1909.11942 (2019).

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[4] S. Ruder, M. E. Peters, S. Swayamdipta, T. Wolf, Transfer learning in natural language processing, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, 2019, pp. 15–18.

[5] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, S. Lehmann, Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 1615–1625. URL: https://www.aclweb.org/anthology/D17-1169. doi:10.18653/v1/D17-1169.

[6] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 328–339. URL: https://www.aclweb.org/anthology/P18-1031. doi:10.18653/v1/P18-1031.

[7] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification?, in: China National Conference on Chinese Computational Linguistics, Springer, 2019, pp. 194–206.
[8] S. Garg, T. Vu, A. Moschitti, Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 7780–7788.
[9] J. Kang, Transferability of natural language inference to biomedical question answering, arXiv preprint arXiv:2007.00217 (2020).
[10] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al., An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, BMC bioinformatics 16 (2015) 1–28. doi:10.1186/s12859-015-0564-6.
[11] G. Wiese, D. Weissenborn, M. Neves, Neural domain adaptation for biomedical question answering, in: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 281–289. URL: https://www.aclweb.org/anthology/K17-1029. doi:10.18653/v1/K17-1029.
[12] D. Weissenborn, G. Wiese, L. Seiffe, Making neural QA as simple as possible but not simpler, in: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 271–280. URL: https://www.aclweb.org/anthology/K17-1028. doi:10.18653/v1/K17-1028.
[13] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2383–2392. URL: https://www.aclweb.org/anthology/D16-1264. doi:10.18653/v1/D16-1264.
[14] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: pre-trained biomedical language representation model for biomedical text mining, arXiv preprint arXiv:1901.08746 (2019).
[15] W. Yoon, J. Lee, D. Kim, M. Jeong, J. Kang, Pre-trained language model for biomedical question answering, arXiv preprint arXiv:1909.08229 (2019).
[16] Q. Jin, B. Dhingra, W. Cohen, X. Lu, Probing biomedical embeddings from language models, in: Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP, Association for Computational Linguistics, Minneapolis, USA, 2019, pp. 82–89. URL: https://www.aclweb.org/anthology/W19-2011. doi:10.18653/v1/W19-2011.
[17] S. Hosein, D. Andor, R. McDonal, Measuring domain portability and error propagation in biomedical qa, arXiv preprint arXiv:1909.09704 (2019).
[18] M. Resta, D. Arioli, A. Fagnani, G. Attardi, Transformer models for question answering at bioasq 2019, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2019, pp. 711–726.
[19] A. Williams, N. Nangia, S. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, 2018,

pp. 1112–1122. URL: http://aclweb.org/anthology/N18-1101.

[20] A. Kazaryan, U. Sazanovich, V. Belyaev, Transformer-based open domain biomedical question answering at bioasq8 challenge (2020).

[21] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: Unanswerable questions for SQuAD, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 784–789. URL: https://www.aclweb.org/anthology/P18-2124. doi:10.18653/v1/P18-2124.

[22] Pubmed, Pubmed® comprises more than 30 million citations for biomedical literature from medline, life science journals, and online books., 2020. URL: https://pubmed.ncbi.nlm.nih.gov, [Online; accessed 1-December-2020].

[23] U. Khanna, Gradual unfreezing transformer-based language models for biomedical question answering, http://hdl.handle.net/1959.14/1280832, 2021. [Macquarie University, Sydney, Australia Online; accessed 03-June-2021].

[24] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's neural machine translation system: Bridging the gap between human and machine translation, arXiv preprint arXiv:1609.08144 (2016).

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. 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, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[27] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3645–3650. URL: https://www.aclweb.org/anthology/P19-1355. doi:10.18653/v1/P19-1355.

[28] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[29] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).

[30] Tsatsaronis et al, Bioasq participants area task 7b: Test results of phase b, http://participants-area.bioasq.org/results/7b/phaseB/, 2019. [Online; accessed 17-January-2021].