Boosting Low-Resource Biomedical QA via Entity-Aware Masking Strategies

Gabriele Pergola[1], **Elena Kochkina**[1,3], **Lin Gui**[1], **Maria Liakata**[1,2,3], **Yulan He**[1]
[1]University of Warwick, UK [2]Queen Mary University of London, UK
[3]The Alan Turing Institute, UK
{gabriele.pergola,e.kochkina,lin.gui,yulan.he}@warwick.ac.uk
m.liakata@qmul.ac.uk

Abstract

Biomedical question-answering (QA) has gained increased attention for its capability to provide users with high-quality information from a vast scientific literature. Although an increasing number of biomedical QA datasets has been recently made available, those resources are still rather limited and expensive to produce. Transfer learning via pre-trained language models (LMs) has been shown as a promising approach to leverage existing general-purpose knowledge. However, fine-tuning these large models can be costly and time consuming, often yielding limited benefits when adapting to specific themes of specialised domains, such as the COVID-19 literature. To bootstrap further their domain adaptation, we propose a simple yet unexplored approach, which we call *biomedical entity-aware masking* (BEM). We encourage masked language models to learn entity-centric knowledge based on the pivotal entities characterizing the domain at hand, and employ those entities to drive the LM fine-tuning. The resulting strategy is a downstream process applicable to a wide variety of masked LMs, not requiring additional memory or components in the neural architectures. Experimental results show performance on par with state-of-the-art models on several biomedical QA datasets.

1 Introduction

Biomedical question-answering (QA) aims to provide users with succinct answers given their queries by analysing a large-scale scientific literature. It enables clinicians, public health officials and end-users to quickly access the rapid flow of specialised knowledge continuously produced. This has led the research community's effort towards developing specialised models and tools for biomedical QA and assessing their performance on benchmark datasets such as BioASQ (Tsatsaronis et al., 2015). Producing such data is time-consuming and requires involving domain experts, making it an expensive process. As a result, high-quality biomedical QA datasets are a scarce resource. The recently released CovidQA collection (Tang et al., 2020), the first manually curated dataset about COVID-19 related issues, provides only 127 question-answer pairs. Even one of the largest available biomedical QA datasets, BioASQ, only contains a few thousand questions.

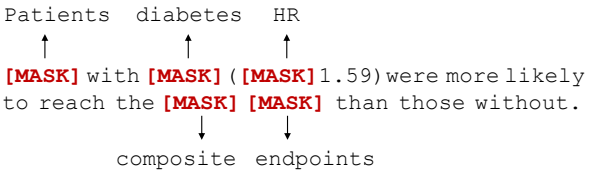

Figure 1: An excerpt of a sentence masked via the BEM strategy, where the masked words were chosen through a biomedical named entity recognizer. In contrast, BERT (Devlin et al., 2019) would randomly select the words to be masked, without attention to the relevant concepts characterizing a technical domain.

There have been attempts to fine-tune pre-trained large-scale language models for general-purpose QA tasks (Rajpurkar et al., 2016; Liu et al., 2019; Raffel et al., 2020) and then use them directly for biomedical QA. Furthermore, there has also been increasing interest in developing domain-specific language models, such as BioBERT (Lee et al., 2019) or RoBERTa-Biomed (Gururangan et al., 2020), leveraging the vast medical literature available. While achieving state-of-the-art results on the QA task, these models come with a high computational cost: BioBERT needs ten days on eight GPUs to train (Lee et al., 2019), making it prohibitive for researchers with no access to massive computing resources.

An alternative approach to incorporating external knowledge into pre-trained language models is to drive the LM to focus on pivotal entities characterising the domain at hand during the fine-

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1977–1985
April 19 - 23, 2021.

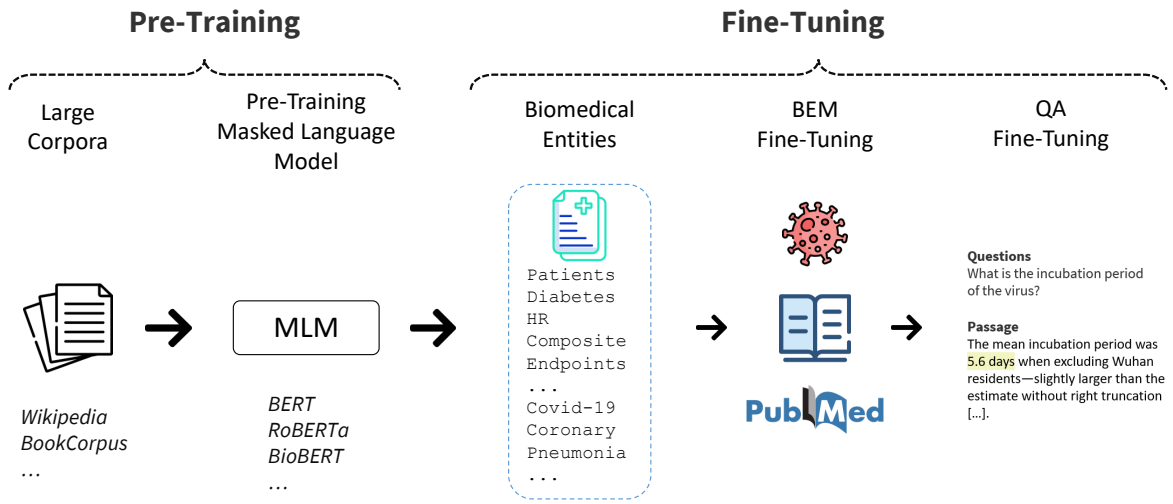

Figure 2: A schematic representation of the main steps involved in fine-tuning masked language models for the QA task through the biomedical entity-aware masking (BEM) strategy.

tuning stage. Similar ideas were explored in works by Zhang et al. (2019), Sun et al. (2020), which proposed the ERNIE model. However, their adaptation strategy was designed to generally improve the LM representations rather than adapting it to a particular domain, requiring additional objective functions and memory. In this work we aim to enrich existing general-purpose LM models (e.g. BERT (Devlin et al., 2019)) with the knowledge related to key medical concepts. In addition, we want domain-specific LMs (e.g. BioBERT) to re-encode the already acquired information around the medical entities of interests for a particular topic or theme (e.g. literature relating to COVID-19).

Therefore, to facilitate further domain adaptation, we propose a simple yet unexplored approach based on a novel masking strategy to fine-tune a LM. Our approach introduces a *biomedical entity-aware masking* (BEM) strategy encouraging masked language models (MLMs) to learn entity-centric knowledge (§2). We first identify a set of entities characterising the domain at hand using a domain-specific entity recogniser (SciSpacy (Neumann et al., 2019)), and then employ a subset of those entities to drive the masking strategy while fine-tuning (Figure 1). The resulting BEM strategy is applicable to a vast variety of MLMs and does not require additional memory or components in the neural architectures. Experimental results show performance on a par with the state-of-the-art models for biomedical QA tasks (§4) on several biomedical QA datasets. A further qualitative assessment provides an insight into how QA pairs benefit from the proposed approach.

2 BEM: A Biomedical Entity-Aware Masking Strategy

The fundamental principle of a masked language model (MLM) is to generate word representations that can be used to predict the missing tokens of an input text. While this general principle is adopted in the vast majority of MLMs, the particular way in which the tokens to be masked are chosen can vary considerably. We thus proceed analysing the random masking strategy adopted in BERT (Devlin et al., 2019) which has inspired most of the existing approaches, and we then introduce the biomedical entity-aware masking strategy used to fine-tune MLMs in the biomedical domain.

BERT Masking strategy. The masking strategy adopted in BERT randomly replaces a predefined proportion of words with a special `[MASK]` token and the model is required to predict them. In BERT, 15% of tokens are chosen uniformly at random, 10% of them are swapped into random tokens (thus, resulting in an overall 1.5% of the tokens randomly swapped). This introduces a rather limited amount of noise with the aim of making the predictions more robust to trivial associations between the masked tokens and the context. While another 10% of the selected tokens are kept without modifications, the remaining 80% of them are replaced with the `[MASK]` token.

Biomedical Entity-Aware Masking Strategy We describe an entity-aware masking strategy which only masks biomedical entities detected by a domain-specific named entity recogniser (SciS-

#	Model	CovidQA			BioASQ 7b		
		P@1	R@3	MRR	SAcc	LAcc	MRR
1	**BERT**	0.081*	0.117*	0.159*	0.012	0.032	0.027
2	+ BioASQ	0.125	0.177	0.206	0.226	0.317	0.262
3	+ STM + BioASQ	0.132	0.195	0.218	0.233	0.325	0.265
4	+ BEM + BioASQ	0.145	0.278	0.269	0.241	0.341	0.288
5	**RoBERTa**	0.068	0.115	0.122	0.023	0.041	0.036
6	+ BioASQ	0.106	0.155	0.178	0.278	0.324	0.294
7	+ STM + BioASQ	0.112	0.167	0.194	0.282	0.333	0.300
8	+ BEM + BioASQ	0.125	0.198	0.236	0.323	0.374	0.325
9	**RoBERTa-Biomed**	0.104	0.163	0.192	0.028	0.044	0.037
10	+ BioASQ	0.128	0.355	0.315	0.415	0.398	0.376
11	+ STM + BioASQ	0.136	0.364	0.321	0.423	0.410	0.397
12	+ BEM + BioASQ	0.143	0.386	0.347	**0.435**	0.443	0.398
13	**BioBERT**	0.097*	0.142*	0.170*	0.031	0.046	0.039
14	+ BioASQ	0.166	0.419	0.348	0.410†	0.474†	0.409†
15	+ STM + BioASQ	0.172	0.432	0.385	0.418	0.482	0.416
16	+ BEM + BioASQ	*0.179*	**0.458**	*0.391*	0.421	**0.497**	**0.434**
17	**T5 LM**						
18	+ MS-MARCO	**0.282***	0.404*	**0.415***	—	—	—

Table 1: Performance of language models on the CovidQA and BioASQ 7b1 dataset. Values referenced with * come from the Tang et al. (2020) work and with † from Yoon et al. (2020).

pacy[1]). Compared to the random masking strategy described above, which is used to pre-train the masked language models, the introduced entity-aware masking strategy is adopted to boost the fine-tuning process for biomedical documents. In this phase, rather than randomly choosing the tokens to be masked, we inform the model of the relevant tokens to pay attention to, and encourage the model to refine its representations using the new surrounding context.

Replacing strategy We decompose the BEM strategy into two steps: (1) *recognition* and (2) *sub-sampling and substitution*. During the *recognition phase*, a set of biomedical entities $\mathcal{E}$ is identified in advance over a training corpus.

Then, at the *sub-sampling and substitution* stage, we first sample a proportion ρ of biomedical entities $\mathcal{E}_f \in \mathcal{E}$. The resulting entity subsets $\mathcal{E}_f$ is thus dynamically computed at batch time, in order to introduce a diverse and flexible spectrum of masked entities during training. For consistency, we use the same tokeniser for the documents d_i in the batch and the entities $e_j \in \mathcal{E}$. Then, we substitute all the k entity mentions $w^k_{e_j}$ in d_i with the special token `[MASK]`, making sure that no consecutive entities are replaced. The substitution takes place at batch time, so that the substitution is a downstream process suitable for a wide typology of MLMs. A diagram synthesizing the involved steps is reported in Figure 2.

[1]https://scispacy.apps.allenai.org/

3 Evaluation Design

Biomedical Reading Comprehension. We represent a document as $d_i := (s^i_0, \ldots, s^i_{j-1})$, a sequence of sentences, in turn defined as $s_j := (w^j_0, \ldots, w^j_{k-1})$, with w_k a word occurring in s_j. Given a question q, the task is to retrieve the span $w^j_s, \ldots, w^j_{s+t}$ from a document d_j that can answer the question. We assume the extractive QA setting where the answer span to be extracted lies entirely within one, or more than one document d_i.

In addition, for consistency with the CovidQA dataset and to compare with results in Tang et al. (2020), we consider a further and sightly modified setting in which the task consists of retrieving the sentence s^i_j that most likely contains the exact answer. This sentence level QA task mitigates the non-trivial ambiguities intrinsic to the definition of the exact span for an answer, an issue particularly relevant in the medical domain and well-know in the literature (Voorhees and Tice, 1999)[2].

Datasets. We assess the performance of the proposed masking strategies on two biomedical datasets: CovidQA and BioASQ.

[2]Consider, for instance, the following QA pair: *"What is the incubation period of the virus?"*, *"6.4 days (95% 175 CI 5.3 to 7.6)"*, where a model returning just *"6.4 days"* would be considered wrong.

BERT with STM	BERT with BEM
*What is the **OR** for severe infection in COVID-19 patients with hypertension?*	
- There were significant correlations between COVID-19 severity and [..], diabetes [OR=2.67], coronary heart disease [OR=2.85]. - Compared with the non-severe patient, the pooled odds ratio of hypertension, respiratory system disease, cardiovascular disease in severe patients were (OR 2.36, ..), (OR 2.46, ..) and (OR 3.42, ..).	- There were significant correlations between COVID-19 severity and [..], diabetes [OR=2.67], coronary heart disease [OR=2.85]. - Compared with the non-severe patient, the pooled odds ratio of hypertension, respiratory system disease, cardiovascular disease in severe patients were (OR 2.36, ..), (OR 2.46, ..) and (OR 3.42, ..).
*What is the **HR** for severe infection in COVID-19 patients with hypertension?*	
- - - -	- After adjusting for age and smoking status, patients with COPD (HR 2.681), diabetes (HR 1.59), and malignancy (HR 3.50) were more likely to reach to the composite endpoints than those without.
*What is the **RR** for severe infection in COVID-19 patients with hypertension?*	
- - - -	- In univariate analyses, factors significantly associated with severe COVID-19 were male sex (14 studies; pooled RR=1.70, ...), hypertension (10 studies 2.74 ...),diabetes (11 studies ...), and CVD (..).

Table 2: Examples of questions and retrieved answers using BERT fine-tuned either with its original masking approach or with the biomedical entity-aware masking (BEM) strategy.

CovidQA (Tang et al., 2020) is a manually curated dataset based on the AI2's COVID-19 Open Research Dataset (Wang et al., 2020). It consists of 127 question-answer pairs with 27 questions and 85 unique related articles. This dataset is too small for supervised training, but is a valuable resource for zero-shot evaluation to assess the unsupervised and transfer capability of models.

BioASQ (Tsatsaronis et al., 2015) is one of the larger biomedical QA datasets available with over 2000 question-answer pairs. To use it within the extractive questions answering framework, we convert the questions into the SQuAD dataset format (Rajpurkar et al., 2016), consisting of question-answer pairs and the corresponding *passages,* medical articles containing the answers or clues with a length varying from a sentence to a paragraph. When multiple passages are available for a single question, we form additional question-context pairs combined subsequently in a postprocessing step to choose the answer with highest probability, similarly to Yoon et al. (2020). For consistency with the CovidQA dataset, we report our evaluation exclusively on the factoid questions of the BioASQ 7b Phase B1.

Baselines. We use the following unsupervised neural models as baselines: the out-of-the-box BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), as well as their variants BioBERT (Lee et al., 2019) and RoBERTa-Biomed (Gururangan et al., 2020) fine-tuned on medical and scientific corpora.

To highlight the impact of different fine-tuning strategies, we examine several configurations depending on the data and the masking strategy adopted. We experiment using the BioASQ QA training pairs during the fine-tuning stage and denote the models using them with `+BioASQ`. When we fine-tune the models on the corpus consisting of PubMed articles referred within the BioASQ and AI2's COVID-19 Open Research dataset, we compare two masking strategies denoted as `+STM` and `+BEM`, where `+STM` indicates the standard masking strategy of the model at hand and `+BEM` is our proposed strategy. We additionally report the T5 (Raffel et al., 2020) performance over CovidQA, which constitutes the current state-of-the-art (Tang et al., 2020)[3].

Metrics. To facilitate comparisons, we adopt the same evaluation scores used in Tang et al. (2020) to assess the models on the CovidQA dataset, i.e. mean reciprocal rank (MRR), precision at rank one (P@1), and recall at rank three (R@3); similarly, for the BioASQ dataset, we use the strict accuracy (SAcc), lenient accuracy (LAcc) and MRR, the BioASQ challenge's official metrics.

4 Experimental Results and Discussion

We report the results on the QA tasks in Table 1.

Among the unsupervised models, BERT achieves slightly better performance than RoBERTa on CovidQA, yet the situation is reversed on BioASQ (rows 1,5). The low precision of the two models (especially on the BioASQ dataset) confirms the difficulties in generalising to the biomedical domain. Specialised language

[3]We attach supplementary results in Appx. A on SQuAD (Tab. A1) and the *perplexity* of MLMs when fine-tuned on the medical collection with different masking strategies (Fig. A1)

models such as RoBERTa-Biomed and BioBERT show a significant improvement on the CovidQA dataset, but a rather limited one on BioASQ (rows 9,13), highlighting the importance of having larger medical corpora to assess the model's effectiveness. A general boost in performance is shared across models fine-tuned on the QA tasks, with a large benefit from the BioASQ QA. The performance gains obtained by the specialised models (BioBERT and RoBERTa-Biomed) suggest the importance of transferring not only the domain knowledge but also the ability to perform the QA task itself (rows 9,10; 13,14).

A further fine-tuning step before the training over the QA pairs has been proven beneficial for all of the models. The BEM masking strategy has significantly amplified the model's generalisability, with an increased adaptation to the biomedical themes shown by the notable improvement in R@3 and MRR; with the R@3 outperforming the state-of-the-art results of T5 fine-tuned on MS-MARCO (Bajaj et al., 2018) and proving the effectiveness of the BEM strategy.

Table 2 reports questions from the CovidQA related to three statistical indices (i.e. Odds Ratio, Hazard Ratio and Relative Risk) to assess the risk of an event occurring in a group (e.g. infections or death). We notice that even though the indices are mentioned as abbreviations, BERT fine-tuned with the STM is able to retrieve sentences with the exact answer for just one of three questions. By contrast, BERT fine-tuned with the BEM strategy succeeds in retrieving at least one correct sentence for each question. This example suggests the importance of placing the emphasis on the entities, which might be overlooked by LMs during the training process despite being available.

5 Related Work

Our work is closely related to two lines of research: the design of masking strategies for LMs and the development of specialized models for the biomedical domain.

Masking strategies. Building on top of the BERT's masking strategy (Devlin et al., 2019), a wide variety of approaches has been proposed (Liu et al., 2019; Yang et al., 2019; Jiang et al., 2020).

A family of masking approaches aimed at leveraging entity and phrase occurrences in text. SpanBERT, Joshi et al. (2020) proposed to mask and predict whole spans rather than standalone tokens and to make use of an auxiliary objective function. ERNIE (Zhang et al., 2019) is instead developed to mask well-known named entities and phrases to improve the external knowledge encoded. Similarly, KnowBERT (Peters et al., 2019) explicitly model entity spans and use an entity linker to an external knowledge base to form knowledge enhanced entity-span representations. However, despite the analogies with the BEM approach, the above masking strategies were designed to generally improve the LM representations rather than adapting them to particular domains, requiring additional objective functions and memory.

Biomedical LMs. Particular attention has been devoted to the adaptation of LMs to the medical domain, with different corpora and tasks requiring tailored methodologies. BioBERT (Lee et al., 2019) is a biomedical language model based on BERT-*Base* with additional pre-training on biomedical documents from the PubMed and PMC collections using the same training settings adopted in BERT. BioMed-RoBERTa (Gururangan et al., 2020) is instead based on RoBERTa-*Base* (Liu et al., 2019) using a corpus of 2.27M articles from the Semantic Scholar dataset (Ammar et al., 2018). SciBERT (Beltagy et al., 2019) follows the BERT's masking strategy to pre-train the model from scratch using a scientific corpus composed of papers from Semantic Scholar (Ammar et al., 2018). Out of the 1.14M papers used, more than 80% belong to the biomedical domain.

6 Conclusion

We presented BEM, a biomedical entity-aware masking strategy to boost LM adaptation to low-resource biomedical QA. It uses an entity-driven masking strategy to fine-tune LMs and effectively lead them in learning entity-centric knowledge based on the pivotal entities characterizing the domain at hand. Experimental results have shown the benefits of such an approach on several metrics for biomedical QA tasks.

Acknowledgements

This work is funded by the EPSRC (grant no. EP/T017112/1, EP/V048597/1). YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (UKRI) (grant no. EP/V020579/1).

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A human generated machine reading comprehension dataset.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

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), NAACL*.

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, ACL20*.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics (TACL)*, 8.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics 2019*.

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.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, Florence, Italy.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. 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 (JMLR)*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*.

Raphael Tang, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly bootstrapping a question answering dataset for COVID-19.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1).

Ellen M. Voorhees and Dawn M. Tice. 1999. The trec-8 question answering track evaluation. In *In Text Retrieval Conference TREC-8*, pages 83–105.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni,

and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*.

Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. 2020. Pre-trained language model for biomedical question answering. In *Machine Learning and Knowledge Discovery in Databases*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL19*, Florence, Italy.

A Appendix

We further examined whether the fine-tuning of the QA pairs affects not only the model adaptation to the QA task but it further helps realign the repression for the domain at hand. The report scores point out that the vanilla LMs are the ones gaining the most when using in-domain QA pairs, such as BioASQ, compared to the SQuAD (rows 2,3; 9,10). The advantage tends to be reduced on already specialised LMs (rows 16,17; 23;24).

#	Model	CovidQA			BioASQ 7b		
		P@1	R@3	MRR	SAcc	LAcc	MRR
1	**BERT**	0.081*	0.117*	0.159*	0.012	0.032	0.027
2	+ SQuAD	0.110	0.131	0.158	0.292	0.343	0.318
3	+ BioASQ	0.125	0.177	0.206	0.226	0.317	0.262
4	+ STM + SQuAD	0.114	0.146	0.173	0.305	0.355	0.336
5	+ STM + BioASQ	0.132	0.195	0.218	0.233	0.325	0.265
6	+ BEM + SQuAD	0.126	0.173	0.191	0.317	0.371	0.349
7	+ BEM + BioASQ	0.145	0.278	0.269	0.241	0.341	0.288
8	**RoBERTa**	0.068	0.115	0.122	0.023	0.041	0.036
9	+ SQuAD	0.098	0.134	0.160	0.353	0.365	0.328
10	+ BioASQ	0.106	0.155	0.178	0.278	0.324	0.294
11	+ STM + SQuAD	0.107	0.148	0.175	0.361	0.388	0.347
12	+ STM + BioASQ	0.112	0.167	0.194	0.282	0.333	0.300
13	+ BEM + SQuAD	0.114	0.162	0.185	0.368	0.391	0.353
14	+ BEM + BioASQ	0.125	0.198	0.236	0.323	0.374	0.325
15	**RoBERTa-Biomed**	0.104	0.163	0.192	0.028	0.044	0.037
16	+ SQuAD	0.111	0.308	0.288	0.376	0.382	0.358
17	+ BioASQ	0.128	0.355	0.315	0.415	0.398	0.376
18	+ STM + SQuAD	0.118	0.314	0.297	0.381	0.390	0.367
19	+ STM + BioASQ	0.136	0.364	0.321	0.423	0.410	0.397
20	+ BEM + SQuAD	0.121	0.331	0.323	0.385	0.397	0.378
21	+ BEM + BioASQ	0.143	0.386	0.347	**0.435**	0.443	0.398
22	**BioBERT**	0.097*	0.142*	0.170*	0.031	0.046	0.039
23	+ SQuAD	0.161*	0.403*	0.336*	0.381	0.445	0.397
24	+ BioASQ	0.166	0.419	0.348	0.410†	0.474†	0.409†
25	+ STM + SQuAD	0.161	0.411	0.339	0.387	0.447	0.401
26	+ STM + BioASQ	0.172	0.432	0.385	0.418	0.482	0.416
27	+ BEM + SQuAD	0.168	0.427	0.354	0.391	0.458	0.423
28	+ BEM + BioASQ	*0.179*	**0.458**	*0.391*	0.421	**0.497**	**0.434**
29	**T5 LM**						
30	+ MS-MARCO	**0.282***	0.404*	**0.415***	—	—	—

Table A1: Performance of language models on the CovidQA and BioASQ 7b1 dataset. Values referenced with * comes from the Tang et al. (2020) work and with † from Yoon et al. (2020).

In Figure A1, we report the LM perplexity obtained when fine-tuning the model with the standard masking strategy versus the BEM strategy with different proportion of medical entities. Vanilla LMs experienced a huge gain with just a small fraction of entities, while already specialised LMs has a lower but still significant improvement. This could be expected as the specialised LMs has already encoded a large domain knowledge with representations that need to be realigned to the new ones.

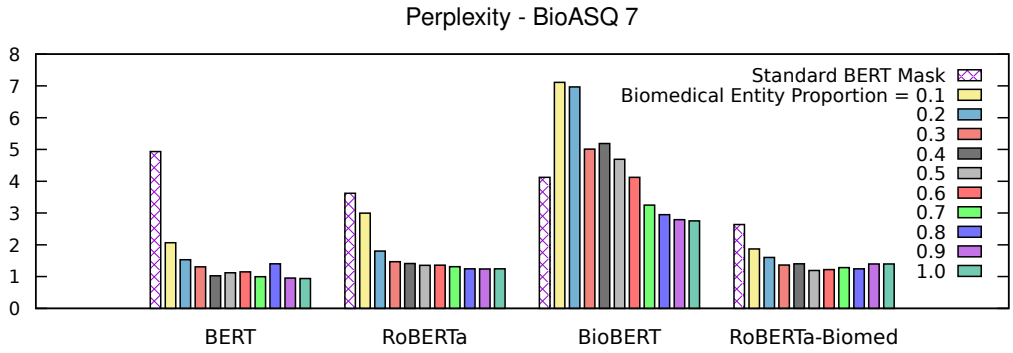

Figure A1: Perplexity of MLMs using different masking strategies on the collection of medical articles.