Structured Fine-Tuning of Contextual Embeddings for Effective Biomedical Retrieval

Alberto Ueda, Rodrygo L. T. Santos
UFMG, Brazil
{ueda,rodrygo}@dcc.ufmg.br

Craig Macdonald, Iadh Ounis
University of Glasgow, UK
{craig.macdonald,iadh.ounis}@glasgow.ac.uk

ABSTRACT

Biomedical literature retrieval has greatly benefited from recent advances in neural language modeling. In particular, fine-tuning pretrained contextual language models has shown impressive results in recent biomedical retrieval evaluation campaigns. Nevertheless, current approaches neglect the inherent structure available from biomedical abstracts, which are (often explicitly) organised into semantically coherent sections such as background, methods, results, and conclusions. In this paper, we investigate the suitability of leveraging biomedical abstract sections for fine-tuning pretrained contextual language models at a finer granularity. Our results on two TREC biomedical test collections demonstrate the effectiveness of the proposed structured fine-tuning regime in contrast to a standard fine-tuning that does not leverage structure. Through an ablation study, we show that models fine-tuned on individual sections are able to capture potentially useful word contexts that may be otherwise ignored by structure-agnostic models.

CCS CONCEPTS

• **Applied computing** → **Bioinformatics**; • **Information systems** → *Retrieval models and ranking*.

KEYWORDS

biomedical retrieval; contextual embeddings; fine-tuned models

ACM Reference Format:
Alberto Ueda, Rodrygo L. T. Santos and Craig Macdonald, Iadh Ounis. 2021. Structured Fine-Tuning of Contextual Embeddings for Effective Biomedical Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3463075

1 INTRODUCTION

An increasing number of biomedical articles are being published each year. The latest snapshot of MEDLINE/PubMed — a well-known biomedical article database — has more than 30 million articles. The use of information retrieval (IR) models in this domain has become of ever-increasing importance to find the most relevant content to a given biomedical information need. Consider the question: *"how long does coronavirus survive on surfaces?"* Even since 2020, there has been a huge number of articles published on coronavirus [33]. Meanwhile, it is reported that experienced biomedical search professionals can take between 30 seconds and several minutes to screen a single article [29], which demands a high effectiveness from the available biomedical search tools. Indeed, evidence-based medicine necessarily relies on scientific literature retrieval [16].

For this reason, there have been several evaluation campaigns aimed at promoting the research and development of effective biomedical search techniques. Among recent campaigns, we note the TREC Clinical Decision Support/Precision Medicine tracks [26, 27], as well as the CLEF eHealth Lab for systematic reviews [13]. More recently, the TREC Covid campaign was designed around an evolving corpus of coronavirus-related biomedical literature [33]. One finding of note from the TREC Covid participants [32] is the high effectiveness of state-of-the-art neural ranking models based on pretrained contextual embeddings, such as BERT [23] and SciBERT [4], when applied to the raw text of biomedical abstracts. Such models are fine-tuned using a large number of queries. However, these models neglect the inherent structure available from a considerable percentage of abstracts in biomedical articles [3, 25, 31], which clearly identify section headings such as "background" and "conclusions".[1] In this paper, we investigate how neural rankers based on SciBERT models can be fine-tuned to effectively take into account the section structure underlying biomedical abstracts.

The contributions of this work are two-fold: (i) we propose an alternative fine-tuning regime that produces multiple ranking models focused on different sections of the input text; and (ii) we thoroughly evaluate the proposed structured fine-tuning regime across two standard test collections for biomedical retrieval. Our results demonstrate the effectiveness of the proposed structured fine-tuning regime compared to a state-of-the-art contextual embedding approach fine-tuned at the abstract level.

2 RELATED WORK

The use of advanced IR techniques to improve scientific literature search has become an essential part of modern biomedical search engines [10, 13, 27, 32]. Of particular interest to this work, the BERT contextual language model (LM) [8] and many of its extensions and variants such as ALBERT [17], SciBERT [4], or RoBERTa [20] have shown significant improvements in several natural language processing (NLP) tasks, such as question answering, named entity recognition, and passage retrieval [21, 23]. More recently, some variants of the BERT language model have been proposed for the scientific literature domain, such as SciBERT [4] and BioBERT [18]. In practice, each LM variant is focused on absorbing the specific context of texts in the domain to better represent the semantics of words in a vocabulary. In this paper, we take a step further and

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

[1] https://www.nlm.nih.gov/bsd/policy/structured_abstracts.html

propose to fine-tune such models to exploit the context of individual abstract sections. To the best of our knowledge, our framework is the first to leverage the abstract structure for fine-tuning contextual embedding models for biomedical search.

The effective use of document structure has been the subject of intense research in IR over the years. For instance, Wilkinson [34] investigated strategies to combine document-level evidence and evidence from individual document fields such as "purpose", "summary", and "title". This included taking the maximum or the weighted average score across fields, potentially combined with a document-level score. Later work into field-based IR models [15, 28] showed that it may be beneficial to leverage the entire document structure jointly as opposed to estimating field-specific scores to be later aggregated at the document level. More recently, extensions of neural ranking models have also been applied to effectively leverage document structure [2, 7, 19, 30, 35]. Field-based models are particularly suited to arbitrarily structured domains with varying-length and potentially missing fields. In contrast and by construction, biomedical article abstracts show a remarkably consistent structure typically centred around five roughly evenly-sized fields: background, objective, methods, results, and conclusions [25, 31]. To take advantage of such a regularity, we revisit the hypothesis of field-level aggregation and investigate the usefulness of the available abstract structure for fine-tuning contextual embedding models as features for effectively ranking biomedical articles.

3 STRUCTURED FINE-TUNING

Fine-tuning pretrained contextual language models has shown impressive results in several downstream retrieval tasks including biomedical retrieval [27, 32]. Our approach builds upon a standard cascading architecture for fine-tuning contextual language models [23]. For each query q, we retrieve an initial ranking of k biomedical abstracts $a \in A$ to be re-ranked. Each $\langle q, a \rangle$ pair is concatenated with a [SEP] token in between and fed into a pretrained SciBERT model [4]. On top of SciBERT's [CLS] token embedding, we stack a single linear combination layer to obtain a prediction $h_A(q, a)$ that abstract a is relevant for query q. Fine-tuning is performed using the pairwise cross-entropy loss [5], according to:

$$\mathcal{L} = \log\left(1 + e^{-h_A(q,a^+,a^-)}\right), \tag{1}$$

where a^+ and a^- are respectively relevant and non-relevant abstracts sampled from A, and $h_A(q, a^+, a^-) = h_A(q, a^+) - h_A(q, a^-)$. Note the subscript A in the h_A model to emphasise the fact that the model is fine-tuned on a collection of unstructured abstracts.

Structured abstract information in the form of explicit section headings (e.g. "background", "methods", "results", "conclusions") has been strongly recommended in most medical databases to facilitate readability, organisation and retrieval [3, 25, 31]. We hypothesise that leveraging such information can also help improve representation learning on biomedical abstracts. However, given the diversity of publication venues, submission guidelines, and even different authors' writing preferences, it is estimated that only around 25% of published biomedical articles are structured and the used structure scheme varies [11, 12, 22]. As a result, several approaches have been proposed throughout the years to segment article abstracts into coherent sections [1, 22, 24, 36]. In this paper, we adopt a standard classification approach [24] for identifying article sections — whenever section information is missing — and instead focus on how to leverage this structure for effective retrieval.

Table 1: Salient statistics of the used test collections.

Collection	#Docs	#Train. Queries	#Test Queries	#Rel. per Train. Query	#Rel. per Test Query
TREC-PM	29,138,919	80	40	119.8	142.0
TREC-COVID	191,175	0	50	-	218.2

To effectively leverage abstract structure, we investigate the suitability of fine-tuning multiple pretrained SciBERT models concurrently, with each model focusing on a different abstract section. More precisely, assuming structured abstracts of the form $a = (a_s \mid s \in \{B, M, R, C\})$,[2] we learn one model h_{A_s} per section, which is fine-tuned on a projection A_s of the original collection A comprising only the target section (i.e. $A_s = \{a_s \mid a \in A\}$). Analogously to Equation (1), we learn h_{A_s} that minimises:

$$\mathcal{L} = \log\left(1 + e^{-h_{A_s}(q,a_s^+,a_s^-)}\right), \tag{2}$$

where a_s^+ (a_s^-) is analogous to a^+ (a^-) in Equation (1). In the next sections, we will assess the suitability of such a structured fine-tuning for effective biomedical retrieval.

4 EXPERIMENTAL SETUP

We aim to answer the following research questions:

RQ1. How effective is the proposed structured fine-tuning regime for biomedical literature retrieval?

RQ2. How do different biomedical abstract sections contribute to retrieval effectiveness?

Test Collections. We build upon the evaluation paradigm provided by two recent TREC biomedical search campaigns. The TREC Precision Medicine (PM) track [27] test collection comprises over 29M abstracts from MEDLINE/PubMed, one of the largest and most important scientific literature databases available in the biomedical domain. Each of its 120 queries describes a cancer patient case with a certain type of cancer and the relevant gene variants. The relevance judgements of TREC-PM are made by physicians trained in medical informatics, who assign a relevance score to an article abstract ranging from 0 (non-relevant) to 2 (highly relevant).

Our second test collection corresponds to the 5th and final round of TREC-COVID, an evaluation campaign built around the CORD-19 Open Research Dataset,[3] a joint effort to collect COVID-related scientific literature. Each of the 50 available queries comes in three variants with increasing lengths. We use a concatenation of all three query formats as input, which performed effectively in previous rounds of TREC-COVID. Relevance judgements follow the same procedure of TREC-PM. Salient statistics of the test collections used in both campaigns are shown in Table 1.

[2] B, M, R, C are short for "background", "methods", "results", and "conclusions".
[3] https://pages.semanticscholar.org/coronavirus-research

Indexing and Retrieval. We use Terrier 5.2[4] for indexing the title and abstract of each article after applying Porter's stemmer and removing standard English stopwords. For the initial retrieval in our cascading architecture, we use DFRee for TREC-PM and DFRee with Bo1 query expansion for TREC-COVID with the default parameters in Terrier and retrieve the top 1,000 abstracts per query.

Unstructured Fine-Tuning. As the base neural architecture for our experiments, we use SciBERT (SciVOCAB, Uncased) [4], which outperformed other contextual embedding models such as vanilla BERT and ColBERT in our preliminary experiments for both TREC-PM and TREC-COVID. As a strong biomedical reranking baseline, we fine-tune a pretrained SciBERT model [4] using all queries from TREC-PM 2017 and 2018 and the abstracts from the top 1,000 articles retrieved by DFRee for each query. Fine-tuning is performed using CEDR[5] with its recommended hyperparameters [21]. For ease of reference, we denote this baseline model as "global".

Structured Fine-Tuning. To test the benefits of our proposed structured fine-tuning regime, we use the same setup previously described for the unstructured fine-tuning, however with individual abstract sections rather than entire abstracts used as input, as described in Equation (2). To infer the underlying structure for abstracts not already manually structured into sections, we adopt a standard multi-class sentence classification approach, with a fixed set of four abstract sections as classes: "background", "methods", "results", and "conclusions".[6] As training instances for classification, we use 9,194 sentences extracted from the MEDLINE/PubMed corpus with their corresponding section label. To represent each sentence, we use a sparse 19,746-dimensional TF-IDF representation augmented with a numeric feature indicating the sentence position in the abstract. As our sentence classifier, we use logistic regression, which showed the highest accuracy in our preliminary investigation compared to other classifiers implemented in Scikit-learn 0.23.2.[7] Given structured abstracts, in addition to the global model h_A, we learn "local" models for all four abstract sections: h_{A_B} (background), h_{A_M} (methods), h_{A_R} (results), and h_{A_C} (conclusions).

Evaluation procedure. To assess the effectiveness of structured fine-tuning, we aggregate the predictions of all five models previously described (the global model h_A as well as the local models h_{A_B}, h_{A_M}, h_{A_R}, and h_{A_C}) to produce the final score for each query-abstract pair. As aggregation methods, we use both the sum as well as the maximum of all predictions.[8] For testing, we use two held-out query sets: PM19, which includes all 40 queries from TREC-PM 2019, and COVID, which includes all 50 queries of TREC-COVID round 5. As evaluation metrics, we report normalised discounted cumulative gain (nDCG), mean average precision (MAP), and precision@10 (P@10). All results are statistically validated using Student's paired t-test with significance levels $\alpha = 0.05$ and $\alpha = 0.01$.

5 EXPERIMENTAL RESULTS

In this section, we present the results of our experiments aimed to answer the research questions introduced in Section 4.

[4] http://terrier.org [5] https://github.com/Georgetown-IR-Lab/cedr [6] These sections correspond to official NLM category tags used in PubMed and are the most frequent among the structured abstracts available in our test collections. [7] https://scikit-learn.org/ [8] More complex aggregation methods such as LambdaMART [6] did not provide further significant improvements in our preliminary investigations.

Table 2: Retrieval effectiveness of different rankings leveraging globally (h_A) and locally ($h_{A_s} \mid s \in S = \{B, M, R, C\}$) fine-tuned SciBERT models as features. The symbols † and ‡ denote significant increases over the global model in the first row for $p < 0.05$ and $p < 0.01$, respectively.

	#	Features	Agg	NDCG	MAP	P@10
PM19	1	$\{h_A(q,a)\}$	–	0.530	0.206	0.456
	2	$\{h_A(q,a)\} \cup \{h_A(q,a_s) \mid s \in S\}$	sum	0.526	0.209	0.462
	3	$\{h_A(q,a)\} \cup \{h_{A_s}(q,a_s) \mid s \in S\}$	sum	**0.569**‡	**0.262**‡	**0.562**‡
	4	$\{h_A(q,a)\} \cup \{h_A(q,a_s) \mid s \in S\}$	max	0.517	0.198	0.428
	5	$\{h_A(q,a)\} \cup \{h_{A_s}(q,a_s) \mid s \in S\}$	max	0.533	0.217	0.465
COVID	1	$\{h_A(q,a)\}$	–	0.467	0.160	0.402
	2	$\{h_A(q,a)\} \cup \{h_A(q,a_s) \mid s \in S\}$	sum	0.486†	0.180‡	0.484‡
	3	$\{h_A(q,a)\} \cup \{h_{A_s}(q,a_s) \mid s \in S\}$	sum	**0.492**‡	**0.192**‡	**0.516**‡
	4	$\{h_A(q,a)\} \cup \{h_A(q,a_s) \mid s \in S\}$	max	0.475	0.169	0.454
	5	$\{h_A(q,a)\} \cup \{h_{A_s}(q,a_s) \mid s \in S\}$	max	0.482†	0.175‡	0.438

5.1 (*RQ1*) Retrieval Effectiveness

To address RQ1, we assess the effectiveness of our proposed structured fine-tuning regime for biomedical retrieval. To this end, we leverage both globally (h_A) as well as locally ($h_{A_s} \mid s \in S = \{B, M, R, C\}$) fine-tuned SciBERT models as features for ranking biomedical abstracts, aggregated with either a sum or max operator. Table 2 shows the results of this investigation for both the PM19 and COVID test sets in terms of nDCG, MAP, and P@10. As our first baseline, we consider a single-feature ranking leveraging the globally fine-tuned h_A model (row #1), which represents the current practice in fine-tuning BERT-based models for ranking [21, 23]. To control for the effect of simply leveraging multiple features, as our second baseline, we consider multi-feature rankings generated by applying the globally fine-tuned, section-agnostic h_A model to each individual abstract section (rows #2 and #4 for the sum and max feature aggregation, respectively). Such multi-feature baselines differ from our proposed approach (rows #3 and #5 for the sum and max aggregation, respectively), which leverages as a feature for a given section a SciBERT model locally fine-tuned on that same section.

From the top half of Table 2, we first note that, for PM19, leveraging individual sections with a globally fine-tuned model (the multi-feature baseline in rows #2 and #4) does not significantly improve upon the global model applied only once to the entire abstract (the single-feature baseline in row #1). In contrast, when locally fine-tuned features are considered (our approach in rows #3 and #5), consistent and significant improvements are attained for the sum aggregator (row #3), but not for max (row #5). This suggests different sections contribute complementary information, which in turn is only fully exploited via local fine-tuning. For COVID, in the bottom half of Table 2, even though the multi-feature baseline is effective (rows #2 and #4), our proposed fine-tuning regime once again delivers further consistent and significant improvements, this time for both the sum and max aggregators (rows #3 and #5). Recalling RQ1, these results demonstrate the effectiveness of fine-tuning multiple SciBERT models focused on different sections of the input biomedical abstracts, as an alternative to the standard practice of performing one global fine-tuning on the entire input.

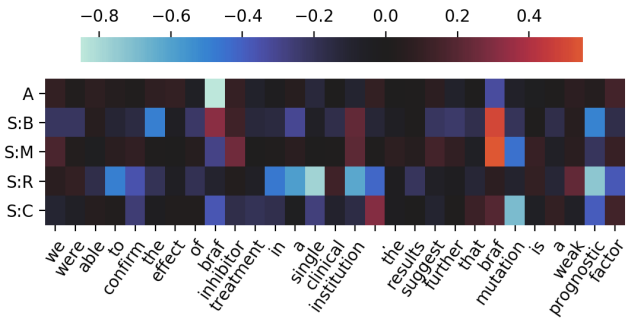

Figure 1: Term ablation study of TREC-PM topic #25 ("melanoma NRAS Q61H") for a passage (x-axis) in a highly relevant abstract (#23855428) scored by different models (y-axis). Each cell quantifies the emphasis a given model places on a given token as the observed difference in the model score for the abstract caused by removing the token (the higher the difference, the more emphasis is placed).

5.2 (*RQ2*) Feature Importance

The results in Section 5.1 suggest that multiple sections contribute to the effectiveness of our approach, as evidenced by a performance drop observed when scores from only one section are leveraged through the max aggregator in the PM19 query set. To complete our investigation, we address RQ2 by further assessing the impact of section-specific scores on the effectiveness of our proposed approach. To this end, Table 3 shows an ablation study on PM19, highlighting the importance of fine-tuned model scores for different abstract sections as well as the entire abstract as ranking features. As expected, the original abstract score plays a major role in our aggregated model, causing a significant drop in performance when this feature is removed. However, we also note that all the section-specific scores contribute significantly to the overall effectiveness of the model according to at least one of the evaluation metrics. In particular, when we discard evidence from the "conclusions" section, effectiveness drops significantly according to all metrics, being the only individual section to cause a significant decrease in all three metrics. In contrast, "background" seems somewhat redundant, incurring the least penalty in effectiveness when discarded.

Zooming into an illustrative example, Figure 1 shows how different models (y-axis) emphasise different tokens (x-axis) in a passage from a selected biomedical abstract judged relevant by medical experts and boosted by our structured fine-tuning approach from rank 20 to rank 1 in the ranking for query "melanoma NRAS Q61H". In the figure, each cell quantifies the contribution a token has on the final score assigned by each model to the selected abstract (for a given row, red indicates a positive contribution whereas blue indicates a negative contribution). From Figure 1, we first note that different models emphasise different tokens (or even the same token in different contexts). For instance, we observe that the second occurrence of token "braf" (just before "mutation") has a different impact on the overall abstract scoring depending on the model considered. In particular, the score given by the global model h_A correlates negatively (blue) with this token, while the score given by most local models h_{A_s}, most notably "background" and "methods", is positively correlated (red). Mutations in the NRAS and BRAF genes are present in 30% and 43% of metastatic melanomas, respectively [9]. Often, studies discussing NRAS-target therapies also include relevant information (such as clinical benefits) about BRAF-target therapies. Although the global, abstract-level model considered that the presence of "braf" should decrease the abstract score, this analysis suggests that, at a section-level, the models are able to detect the relevance of this specific token, which contributes to boosting the abstract in the ranking. Recalling RQ2, these observations support our hypothesis that section-based fine-tuned models can emphasize differently the tokens of an abstract, ultimately leading to more effective biomedical ranking models.

Table 3: Impact in retrieval performance on PM19 when removing individual features from our summed feature aggregation. Symbols † and ‡ denote significant decreases from the baseline (first row) for $p < 0.05$ and $p < 0.01$, respectively.

Set of features	NDCG	MAP	P@10
Using all sections	**0.569**	**0.262**	**0.562**
(-) Original Abstract	0.561 (-0.008)‡	0.249 (-0.013)‡	0.532 (-0.030)‡
(-) Background	0.565 (-0.004)	0.254 (-0.008)†	0.547 (-0.015)
(-) Methods	0.564 (-0.005)†	0.252 (-0.010)‡	0.557 (-0.005)
(-) Results	0.563 (-0.006)‡	0.253 (-0.009)‡	0.551 (-0.011)
(-) Conclusions	0.564 (-0.005)‡	0.251 (-0.011)‡	0.540 (-0.022)†

6 CONCLUSIONS

In this work, we presented a simple but effective framework for biomedical literature retrieval, which leverages the articles' abstract structure to extract and combine multiple section-based features — e.g. features from abstract sections such as "background", "methods", "results", and "conclusions" — instead of using only the original abstract text as in common approaches for biomedical IR. We show that these improvements can be achieved by using recently proposed natural language processing approaches, such as pretrained BERT-based models fine-tuned to individual abstract sections. We validated our framework using two standard and recent test collections for biomedical literature search, namely, the TREC Precision Medicine (PM) and the TREC-COVID collections.

Several directions remain open to future research. For instance, we aim to gauge if the results of our framework can be generalized to other domains beyond the biomedical field. Moreover, we intend to study the performance of combinations of this approach with other retrieval models, including other recently proposed contextual embedding models such as ColBERT [14]. In the same vein, our results revisit the discussion on modelling field-based evidence separately, in contrast to the joint modelling advocated by field-based approaches (both non-neural and neural). In particular, our results suggest that fine-tuning contextual embeddings may be a suitable direction for capturing field-based evidence that otherwise had to be modelled jointly.

ACKNOWLEDGMENTS

This work was partially funded by the authors' individual grants from CNPq, CAPES, and FAPEMIG.

REFERENCES

[1] Asan Agibetov, Kathrin Blagec, Hong Xu, and Matthias Samwald. 2018. Fast and scalable neural embedding models for biomedical sentence classification. *BMC Bioinformatics* (2018).

[2] Saeid Balaneshinkordan, Alexander Kotov, and Fedor Nikolaev. 2018. Attentive Neural Architecture for Ad-hoc Structured Document Retrieval. In *Proceedings of CIKM*.

[3] Gitanjali Batmanabane. 2018. The IMRAD Structure. In *Reporting and Publishing Research in the Biomedical Sciences*.

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).

[5] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to Rank Using Gradient Descent. In *Proceedings of ICML*.

[6] Christopher J.C. Burges. 2010. *From RankNet to LambdaRank to LambdaMART: An Overview*. Technical Report MSR-TR-2010-82. Microsoft Research.

[7] Jason Choi, Surya Kallumadi, Bhaskar Mitra, Eugene Agichtein, and Faizan Javed. 2020. Semantic Product Search for Matching Structured Product Catalogs in E-Commerce. Pre-print.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

[9] H. Ekedahl, H. Cirenajwis, K. Harbst, A. Carneiro, K. Nielsen, H. Olsson, L. Lundgren, C. Ingvar, and G. Jönsson. 2013. The clinical significance of BRAF and NRAS mutations in a clinic-based metastatic melanoma cohort. *The British Journal of Dermatology* (2013).

[10] Erik Faessler, Michel Oleynik, and Udo Hahn. 2020. What Makes a Top-Performing Precision Medicine Search Engine? Tracing Main System Features in a Systematic Way. In *Proceedings of SIGIR*.

[11] Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins Karolinska, Lin Sun, and Ulla Stenius. 2010. Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes. In *Proceedings of BioNLP*.

[12] James Hartley. 2004. Current findings from research on structured abstracts: an update. *Journal of the Medical Library Association* (2004).

[13] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and René Spijker. 2019. CLEF 2019 Technology Assisted Reviews in Empirical Medicine Overview. In *Proceedings of CLEF*.

[14] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of SIGIR*.

[15] Jin Young Kim and W. Bruce Croft. 2012. A Field Relevance Model for Structured Document Retrieval. In *Proceedings of ECIR*.

[16] Bevan Koopman, Jack Russell, and Guido Zuccon. 2017. Task-oriented search for evidence-based medicine. *International Journal on Digital Libraries* (2017).

[17] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint 1909.11942* (2020).

[18] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (2020).

[19] Binsheng Liu, Xiaolu Lu, Oren Kurland, and J. Culpepper. 2018. Improving Search Effectiveness with Field-based Relevance Modeling. In *Proceedings of ADCS*.

[20] 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 1907.11692* (2019).

[21] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized embeddings for document ranking. In *Proceedings of SIGIR*.

[22] Sejin Nam, Senator Jeong, Sang-Kyun Kim, Hong-Gee Kim, Victoria Ngo, and Nansu Zong. 2016. Structuralizing biomedical abstracts with discriminative linguistic features. *Computers in Biology and Medicine* (2016).

[23] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint 1901.04085* (2019).

[24] Sergio Ribeiro, Jing Yao, and Denis Rezende. 2018. Discovering IMRaD Structure with Different Classifiers. In *Proceedings of ICBK*.

[25] Anna M Ripple, James G Mork, John M Rozier, and Lou S Knecht. 2012. Structured abstracts in medline: Twenty-five years later. *National Library of Medicine* (2012).

[26] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, and William R. Hersh. 2016. Overview of the TREC 2016 Clinical Decision Support Track. In *Proceedings of TREC*.

[27] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, Alexander J. Lazar, and Shubham Pant. 2019. Overview of the TREC 2019 Precision Medicine Track. In *Proceedings of TREC*.

[28] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* (2009).

[29] Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Leif Azzopardi, and Shlomo Geva. 2017. A Test Collection for Evaluating Retrieval of Studies for Inclusion in Systematic Reviews. In *Proceedings of SIGIR*.

[30] Xuan Shan, Chuanjie Liu, Yiqian Xia, Qi Chen, Yusi Zhang, Angen Luo, and Yuxiang Luo. 2020. BISON:BM25-weighted Self-Attention Framework for Multi-Fields Document Search. *arXiv preprint 2007.05186* (2020).

[31] Nakayama Takeo, Nobuko Hirai, Shigeaki Yamazaki, and Mariko Naito. 2005. Adoption of structured abstracts by general medical journals and format for a structured abstract. *Journal of the Medical Library Association* (2005).

[32] Ellen M. Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection. *arXiv preprint 2005.04474* (2020).

[33] Lucy Lu Wang, Kyle Lo, Y. Chandrasekhar, Russell Reas, J. Yang, Darrin Eide, K. Funk, Rodney Michael Kinney, Ziyang Liu, W. Merrill, P. Mooney, D. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, B. Stilson, A. Wade, K. Wang, Christopher Wilhelm, Boya Xie, D. Raymond, Daniel S. Weld, O. Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 Open Research Dataset. *ArXiv preprint 2004.10706* (2020).

[34] Ross Wilkinson. 1994. Effective Retrieval of Structured Documents. In *Proceedings of SIGIR*.

[35] Hamed Zamani, Bhaskar Mitra, Xia Song, Nick Craswell, and Saurabh Tiwary. 2018. Neural Ranking Models with Multiple Document Fields. In *Proceedings of WSDM*.

[36] Sijia Zhou and Xin Li. 2020. Feature engineering vs. deep learning for paper section identification: Toward applications in Chinese medical literature. *Information Processing & Management* (2020).