Semantic Enrichment of Pretrained Embedding Output for Unsupervised IR

Edmund Dervakos[a,c], Giorgos Filandrianos[a,c], Konstantinos Thomas[a,c], Alexios Mandalios[a,c], Chrysoula Zerva[a,b] and Giorgos Stamou[a]

[a] *Artificial Intelligence and Learning Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece*

[b] *National Centre for Text Mining, School of Computer Science, University of Manchester, Manchester, UK*

[c] *Equal contribution authors*

Abstract

The rapid growth of scientific literature in the biomedical and clinical domain has significantly complicated the identification of information of interest by researchers as well as other practitioners. More importantly, the rapid emergence of new topics and findings, often hinders the performance of supervised approaches, due to the lack of relevant annotated data. The global COVID-19 pandemic further highlighted the need to query and navigate uncharted ground in the scientific literature in a prompt and efficient way.

In this paper we investigate the potential of semantically enhancing deep transformer architectures using SNOMED-CT in order to answer user queries in an unsupervised manner. Our proposed system attempts to filter and re-rank documents related to a query that were initially retrieved using BERT models. To achieve that, we enhance queries and documents with SNOMED-CT concepts and then impose filters on concept co-occurrence between them. We evaluate this approach on OHSUMED dataset and show competitive performance and we also present our approach for adapting such an approach to full papers, such as kaggle's CORD-19 full-text dataset challenge.

Keywords

BERT, SNOMED-CT, Semantic enrichment, scientific IR, NLP, CORD-19, Covid-19 pandemic

1. Introduction

The first weeks of the COVID-19 crisis brought together several researchers from a wide range of domains, who combined their efforts in fighting the pandemic. At the same time, a significant issue in biomedical text mining was brought to the surface; while machine learning methods keep improving, boosting the performance of supervised models in the biomedical natural language processing field (biomedical NLP or BioNLP), the domain topics change rapidly and so do the related textual resources (scientific publications, reports, clinical trials). Thus, while

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) - Stanford University, Palo Alto, California, USA, March 22-24, 2021.

eddiedervakos@islab.ntua.gr (E. Dervakos); geofila@islab.ntua.gr (G. Filandrianos); konstantinos.thomas@gmail.com (K. Thomas); amandalios@islab.ntua.gr (A. Mandalios); chrysoula.zerva@manchester.ac.uk (C. Zerva); gstam@cs.ntua.gr (G. Stamou)

https://orcid.org/0000-0002-4031-9492 (C. Zerva)

CEUR Workshop Proceedings (CEUR-WS.org)

gold standard, annotated datasets provide a solid basis for training, improving and evaluating new methods, they cannot account for emerging topics, new entities and terminology.

Indeed, navigating existing and upcoming literature, on a variety of COVID-related topics was identified as a critical task early on. The CORD-19 dataset and the *kaggle CORD-19 challenge*, reflected this need and indicated the path to addressing it. The CORD-19 dataset is an ongoing initiative (further described in Section 4.2) to collect resources that could be informative and helpful in coronavirus-related research. The kaggle challenge(s) built upon the early versions of the dataset, and invited research teams to submit systems that would address a set of key-questions spanning across domains and ranging from very information specific ones ('*What do we know about vaccines and therapeutics*?') to rather generic ones ('*What has been published about information sharing and inter-sectoral collaboration?*'). At the time no existing resources could account for COVID-specific annotations in text, calling for either unsupervised approaches or models trained on other domains. Deep neural architectures such as BERT-based models have shown great potential in information retrieval (IR) and question-answering (QA), rendering them strong vanilla models. Since COVID-19 related concepts were already incorporated in large knowledge bases such as SNOMED-CT, MeSH and UMLS, we wanted to explore the potential of using such knowledge sources in a post-processing manner in order to enhance such pre-trained models.

Since this is a preliminary study, we focused on different ways to enhance BERT-based embeddings with knowledge extracted from SNOMED-CT. BERT (Bidirectional Encoder Representations from Transformers) is a family of high performance pre-trained language models which produce state-of-the-art results in a wide variety of NLP tasks [1]. BERT's key technical innovation is applying the bidirectional training of Transformers [2] to language modelling. By using multiple attention mechanisms (multi-head attention), the model is able to capture a broader range of relationships between words than would be possible with a single attention mechanism. Moreover, BERT stacks multiple layers of attention, each of which operates on the output of the layer that came before. Through this repeated composition of word embeddings, BERT is able to form very rich representations as it gets to the deepest layers of the model. The choice of knowledge source was also dictated by performance and wide coverage of concepts; SNOMED-CT [1] is the most comprehensive clinical healthcare terminology, consisting of more than 350,000 concepts and covering clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices and specimens among others.

For evaluation we use a subset of the OHSUMED dataset (see Section 4.1.1) to assess the improvement out method can achieve for document retrieval compared to different BERT models, used as baselines. Specifically, we are interested in comparing the potential for improvement across BERT-based models trained on different domains (generic, clinical, biomedical and combinations). To this purpose, given a query we retrieve the initial set of relevant documents calculating similarity of the BERT-based representations between query and document. We then re-rank the document set using SNOMED-CT to define a new concept co-occurrence based metric. We show that SNOMED-based filtering and re-ranking can consistently boost performance across different BERT baseline models in the IR-OHSUMED task. We show that the

[1] https://www.snomed.org/snomed-ct/five-step-briefing

performance improvement is consistent across models but higher for generic domain models. Additionally, we show that we can obtain results that compare against and even outperform other semantic enhancement approaches. We provide a detailed analysis of the results and discuss how this preliminary study can set the basis for the development of further unsupervised methods, incorporating semantic knowledge in pre-trained embeddings using semantic knowledge graphs. Additionally, we demonstrate how this paradigm can be applied to answer the kaggle CORD-19 challenge questions, and present our modification to maintain robustness on large texts via summarisation.

2. Related Work

While there is a range of work that inspired and relates to the work presented in this paper, the main line line of research concerns the use of external knowledge sources (ontologies, knowledge graphs or knowledge-bases) in order to semantically enhance a natural language processing (NLP) model, in a pre-processing, feature-extraction, joint learning or a post-processing fashion. We position our work in the post-processing approaches based on that classification, but we present below an overview of the core approaches for each category, with a focus on the biomedical/clinical domain.

Early on, Zhang et al. [3] proposed a method for semantic relatedness (SR) calculation between terms, showing how we can combine information from Wikipedia and WordNet in an enhanced graph that can then be traversed to obtain a relatedness score. They then showed that the SR extracted form these graphs can improved performance in named entity disambiguation tasks. More recent work, has focused in the potential of using an external knowledge source (ontologies, graphs or knowledge-bases) to identify key concepts in text and then link textual information from different documents [4].

In many downstream tasks, and especially textual classification, significant performance boosts can be obtained by using external knowledge sources to complement the textual representations and provide more informative features [5]. In this approach the extracted features are used with an SVM to obtain the classification output. While such feature engineering was the standard method to infuse external knowledge to supervised models in traditional ML, these are static features that are integrated in a uniform manner for all instances. Deep neural networks (DNN) can better address this limitation; there have been recent attempts to directly exploit external knowledge sources during the training of DNN models, either using "knowledge focused" attention mechanisms that use the external knowledge to obtain better representations of concepts in text [6, 7], or by retrofitting information from the knowledge graph to pre-trained language models [8] in a post-processing fashion.

Focusing on the IR task, several publications use external ontologies and knowledge sources such as MESH in order to improve IR performance via semantic query expansion [9, 10, 11, 12]. Agosti et al. [13] considers the relation between text and queries and aims to reduce the semantic gap between queries and documents, by incorporating polysemy and synonymy information during the training of neural networks.

Another strand of work related to our paper concerns the unsupervised or semi-supervised IR in the biomedical domain, as well as IR approaches on the main dataset we are experiment-

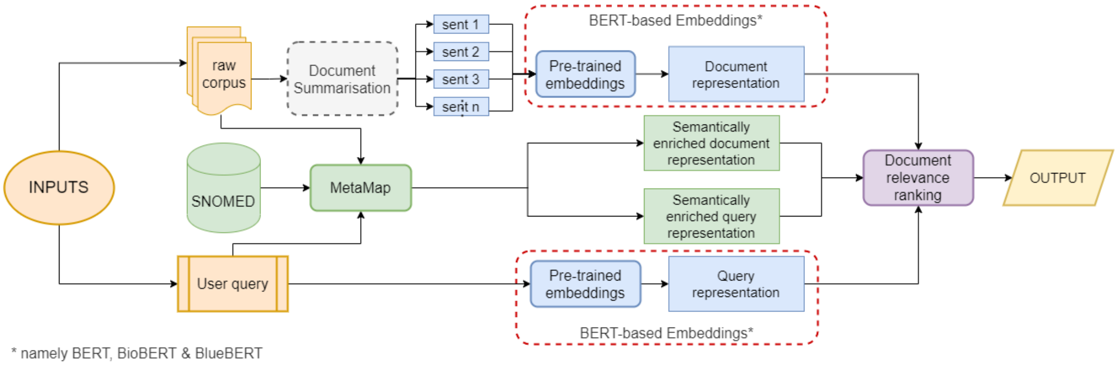


Figure 1: Summary of system architecture

ing with, namely OHSUMED. More specifically, Liu et al. [14], use UMLS to identify word relations and use this information to retrofit pre-trained word embeddings, enforcing the representations of related words to be closer together. Rais et al., compare different strategies of enriching document representations, using concepts from external knowledge sources in combinations with WSD approaches [15]. They specifically employ UMLS for the concept extraction (using a similar MetaMap enrichment to the one we describe in Section 3.4.1) and show that the replacement of terms with their respective concepts can boost IR performance on OHSUMED when used in combination with WSD approaches. We show in this work that when used in combination with pretrained DNN models (which are more robust in terms of contextualisation of terms) the conceptualisation of terms can boost performance even without the use of WSD approaches and outperform the aforementioned approach. Oh et al. [16] proposes CBEEM which exploits external dataset collections in building a feedback model to improve relevance ranking for biomedical IR. Note that instead of using a hierarchical resource in this case, Oh et al. use large external document collections to cluster documents and thus better contextualise query relevance by incorporating an additional term in the traditional feedback model related to the collection relevance estimation.

3. Research Methods

3.1. Overall system architecture

Our system assumes a large document resource that is checked against a specific user query. We use the process described in Figure 1 to obtain relevant documents that could answer the user's query. Specifically, we apply the following:

1. **Document formatting**: Summarisation of full-text documents to obtain a representative set of sentences
2. **Semantic text enrichment based on SNOMED-CT:** Identifying and aligning concepts mentioned in documents and/or queries with their respective descriptions and neighbour nodes in the knowledge-base.

3. **Vectorised text representations:** We use vectorized text representations to compare documents and queries. We employ BERT models for this purpose (see Section 3.3).

3.2. Document formatting

We distinguish two main categories of papers when searching the scientific literature, for which we apply different pre-processing before obtaining the document vectors.

1. Abstract only: Refers to articles for which only the title and abstract (and perhaps some metadata) is available to be processed, typically due to licensing. For these papers the document to be vectorised, is the concatenation of the title and abstract.
2. Full-text documents: Typically open publications where the title, the abstract and the main body and full metadata of the paper is available to any reader and/or text mining system. If the abstract is invalid (e.g., <= 3 sentences) using the full text representation can negatively influence the performance of similarity ranking against the query, due to the length discrepancy. To account for such cases (frequent in the early version of CORD-19 dataset), we opt for an extractive summarisation method, described below. The method was applied on the CORD-19 dataset only, since OHSUMED consists exclusively of abstracts.

3.2.1. Extractive summarisation for scientific publications

To reduce the gap between the abstract only and full-text documents, and be able to apply the same IR models, we opted to extracting only the most representative sentences for documents without a clearly specified abstract, in other words, we used single-document extractive summarisation for scientific documents. We thus fine-tune BERTSUM [17], which has been demonstrated to achieve high performance in scientific summarisation tasks [18, 19, 20].

The BERTSUM model provided by the authors was initially trained on summaries for news articles, whose document characteristics differ significantly compared to those of scientific articles. One of the main underlying differences, concerns the language and vocabulary used, as well as the sentence structure (longer, more complicated sentences). Additionally, the summary size in the newswire domain is significantly smaller compared to the typical scientific summary. We thus construct a new summarisation dataset based on CORD-19 dataset and fine-tune BERTSUM to this specific task, to obtain a model that can better distinguish the important sentences in a long scientific article.

The training dataset was constructed exclusively from papers which have valid and clearly distinguished abstract and main body text as they are defined above. Sentence splitting was applied on both the abstract and the body text of the papers using StandfordNlp [21]. Additional denoising was applied to remove highly frequent abbreviations. After the tokenisation and pre-processing of the texts, each paper consists of 2 parts: $abstact = [asent_1, asent_2, ..., asent_m]$ and $body = [bsent_1, bsent_2, ..., bsent_m]$ where $asent_i$, $bsent_i$ is the i^{th} sentence of the abstract and body text respectively.

The training dataset was constructed considering that the abstract contains the most important information of the paper (inductive bias). We thus score the $body$ sentences against the $abstract$ sentences using ROUGE-L score to obtain the n sentences. For the specific dataset and based on the distribution of the abstract sentences we selected $n = 3$, since in this value it appeared that there was a golden ratio between redundancy and noise. We assigned label 1 to sentences selected in the oracle summary and 0 otherwise, and fine-tuned the initial BERTSUM model on this dataset.

3.3. Vectorised embedding representations for IR

To retrieve documents relevant to a query, we employ bert-as-a-service [22] and calculate the BERT representation of the user query and each candidate document. We then rank the documents' relevance for each query by calculating the distance between two vector representations. We use cosine similarity for the distance estimation, so if we assume that d_i is the document vector and q_j is the query vector then the relevance score is calculated as:

$$relevance(d_i, q_j) = \frac{d_i \cdot q_j}{||d_i|| \cdot ||q_j||} \tag{1}$$

3.3.1. BERT-based embeddings

We opted for BERT embeddings trained on different domains, with demonstrated high performance in downstream classification tasks. Specifically we chose the following: (1) the original **BERT** model [1], trained on Wikipedia and BookCorpus, hence fine-tuned for the generic domain, (2) **BioBERT** [23] trained on Pubmed papers, hence fine-tuned on the biomedical domain and (3) **BlueBERT** [24], trained in a multi-task setting on a combination of biomedical and clinical data, hence still fine-tuned in the biomedical domain, but accounting for a wider scope of text. We expect the latter to have broader coverage and thus better performance on the queries, and generally we expect BioBERT and BlueBERT to outperform BERT due to the expected overlap between the documents they was trained on and the scientific papers in the evaluation datasets. The *base* model was used across cases (12-layer, 768-hidden, 12-head).

3.4. Semantic text enrichment

3.4.1. External knowledge source: SNOMED-CT

SNOMED-CT [25] is a collection of medical terms, and their synonyms, descriptions, etc., with an underlying description logic formal model. It contains clinical knowledge that can complement textual information, and help us process new documents. Its core components include concept hierarchy, descriptions, relations and reference sets, of which we focus on:

- **Concept hierarchy:** encoded clinical terms, organised in hierarchies. The hierarchical structure is particularly useful in the case of searching in a corpus of texts with multi-level information (high-level, general concepts vs more technical/specific information). One can move down the hierarchy in order to fetch specific results, or move up the hierarchy in the case of queries that do not match exactly any of the documents' terms.

- **Descriptions:** textual descriptions of concepts. These can be either synonyms or periphrastic definitions of the corresponding terms. Given that a term can appear with multiple surface forms, enhancing a concept with associated description (and the terms mentioned in in) we can map additional relevant text spans to a given concept.

For the purposes of our work, we make use of the concept hierarchy, where we collect the parents of medical concepts in the SNOMED CT hierarchy and the descriptions, where we incorporate alternative, equivalent ways of describing the same medical concept. One of the challenges in terms of transferring these rich SNOMED-CT concepts to raw text, is to be able to identify the relevant terms in text. For this purpose, we employ the MetaMap tool [26], which maps biomedical text to the UMLS metathesaurus. Upon identifying the text spans that correspond to UMLS concepts, we use a mapping between UMLS and SNOMED concepts in order to incorporate the SNOMED knowledge.

3.4.2. Filtering

For this preliminary study we focus on the existence of SNOMED related concepts in the text as a re-ranking and filtering method for the vector based similarity ranking. More specifically, assume a given user query q_j and an ordered list of documents D, ranked by the cosine similarity scoring described in Eq 1. Upon obtaining the list, we identify a set of text spans CD_i in each document $d_i \in D$ that correspond to SNOMED concepts, using the process described in Section 3.4.1. Similarly, we identify a set of SNOMED concepts CQ_j corresponding to the query q_j. For each identified SNOMED concept c such that $c \in CD_i$ and/or $c \in CQ_j$, we navigate the SNOMED-CT hierarchy to identify the hyper-concepts (parents) c^{hyp} as well as the terms contained in the concept c description, c^{desc} and then expand the CD_i and CQ_j sets with the respective c^{hyp} and c^{desc} concepts. We then calculate the concept intersection between the two sets as specified in Eq 2 and attribute each document with a renewed score, calculated as:

$$concept\_filter(d_i, q_j) = \|CD_i \cap CQ_j\| \tag{2}$$

$$document\_score(d_i, q_j) = relevance(d_i, q_j) \cdot concept\_filter(d_i, q_j) \tag{3}$$

We then investigate the use of different thresholds on the value of the $concept\_filter$ to filter the initial set of documents D. For the experiments presented in the following sections, when the $concept\_filter$ falls below the threshold the value is set set to 1 in Eq 3. We show that this simplified filtering technique can consistently boost performance on different BERT models, and we identify the optimal threshold on the OHSUMED dataset, based on the NDCG metric (Eq. 4).

4. Results and Discussion

4.1. Intrinsic evaluation results on OHSUMED dataset

4.1.1. OHSUMED dataset

The OHSUMED test collection is a subset of the MEDLINE database. We consider a commonly used subset [27], consisting of the first 20,000 documents from the 50,216 medical abstracts

Table 1
Query - document set size distribution for the OHSUMED-91 dataset

max #documents in OHSUMED-91	1	2	3	4	5	6	>6	Total
queries	11	8	10	5	6	5	11	56
percentage (%)	20	15	17	9	10	8	20	100

published in the year 1991. It comprises 13,929 abstracts focusing on cardiovascular disease, classified under a set of 23 Medical Subject Headings (MeSH) categories. We henceforth refer to this subset of the dataset as OHSUMED-91 dataset. The TREC-09 IR task on the OHSUMED dataset contains a total of 108 queries, each query associated with a set of relevant documents. Out of the 108 OHSUMED queries, 56 had at least one document belonging in the OHSUMED-91 dataset. Hence we use these 56 queries for the evaluation. In Table 1 we present the a more detailed breakdown of query-document sets. As we discuss in the results, the low number of related documents for some queries, has a negative impact on performance which we aim to address in the future work.

4.1.2. Evaluation

We base our evaluation on the normalised discounted cumulative gain (NDCG) metric, used to assess the model's ranking of relevant papers pertaining to a set of queries Q. It is defined for position $p \in \{0, 1, ..., N\}$ as described in Eq. 4:

$$\text{nDCG} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{IDCG_p{}^{(q)}}{DCG_p{}^{(q)}}, \quad \text{for} \quad \text{DCG}_\text{p}{}^{(\text{q})} = rel_1{}^{(q)} + \sum_{i=2}^{N}\frac{rel_i{}^{(q)}}{log_2(i)} \tag{4}$$

where $IDCG$ denotes the ideal and highest possible DCG and $rel_i{}^{(q)}$ refers to the relevance of the i^th result ranked according to query q.

While NDCG is our main metric, we also discuss the results for Precision@N and Recall@N, to provide better insights on the performance of the proposed methods (see Section 4.1.3). The Precision and Recall functions are presented below, assuming an ordered set of relevant documents $RelD = \{rel_1, ..., rel_k\}$ and an ordered set of retrieved documents $RetD = \{ret_1, ..., ret_l\}$

$$Precision@N = \frac{RelD \cap RetD_1^N}{N}, \quad Recall@N = \frac{RelD \cap RetD_1^N}{|RelD|} \tag{5}$$

4.1.3. Results on OHSUMED

In the experiments, we first identified the optimal threshold value for filtering, which is consistently $threshold = 5$ for all models. We can then see, that for the application of filtering with this threshold, NDCG performance improves consistently across models. As explained in Section 4.1.1, the OHSUMED-91 dataset, contains only a subset of documents of the original dataset, hence some queries contain only a single relevant document in the ground truth set,

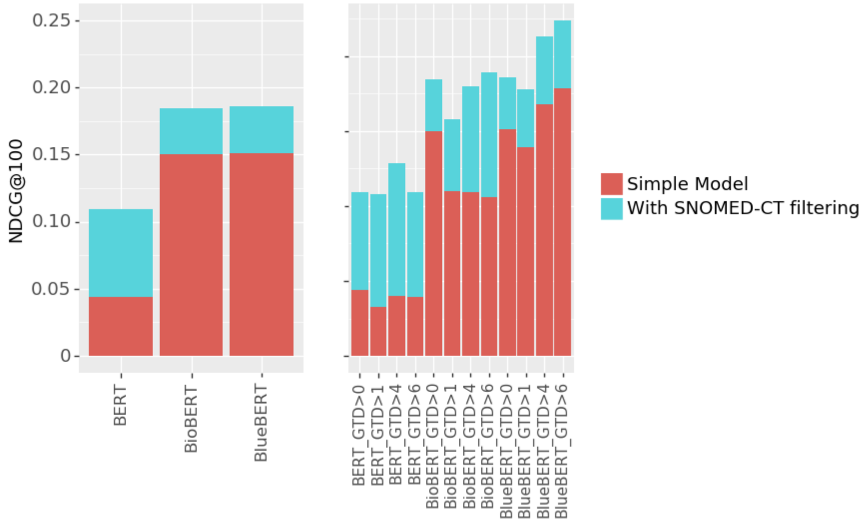
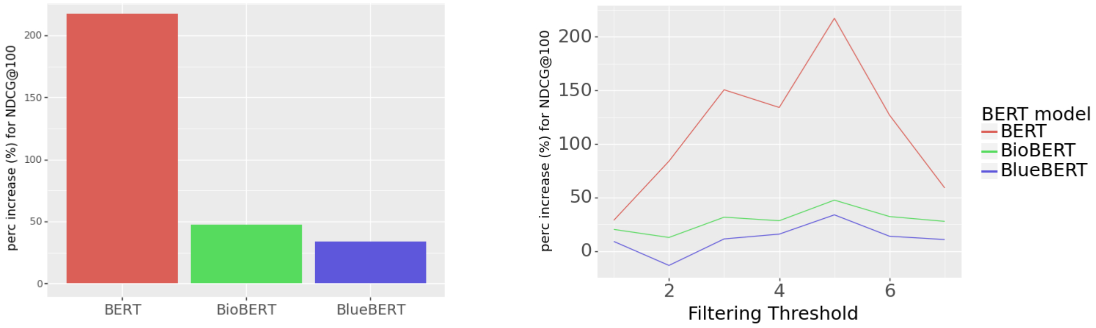


Figure 2: Performance improvement for NDCG@100 over initial BERT models with the addition of SNOMED-based filtering. We can see the improvement (light blue) for the full dataset (left) and different subsets of the dataset (right) split based on the minimum number of ground truth documents for each query (GTD).

rendering the IR problem more demanding. However, as we show in Figure 2, once we constrain the dataset to contain only queries with a higher number of relevant documents (GTD), performance improves further for all models. Thus we reach NDCG@100 performance approximating to 0.25 which is comparable with other unsupervised IR methods on the same dataset [16]. BioBERT and BlueBERT models outperform BERT in all versions, with BlueBERT reaching the best performance for most dataset subsets, confirming our initial hypothesis.



(a) Percentage increase for the optimal filtering threshold=5

(b) Percentage increase over increasing filtering thresholds

Figure 3: Analysis of (%) performance improvement for different models for the NDGC@100 metric

If we cast a closer glance at the increase over the plain model version compared to the SNOMED-CT filtering, we can see that we get a greater performance increase for the less domain specific models (see Figure 3a). Moreover, we can see that this improvement boost

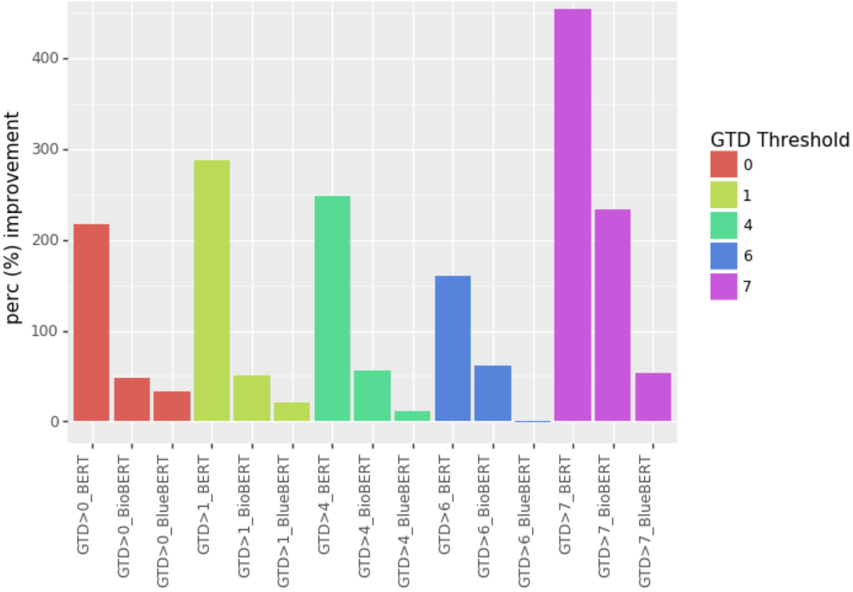


Figure 4: Percentage increase for NDCG@100 varying by the min number of ground-truth documents per query (GTD)

is consistent for all filtering thresholds (Figure 3b, and dataset subsets (Figure 4). This observation underlines the potential of semantic enhancement across domains. More importantly, it demonstrates that a variation of semantic filtering based on a knowledge graph, can successfully be applied to adapt out-of-domain models to a new domain, thus motivating our future research into transfer learning via knowledge graphs.

Regarding the impact of the filtering threshold, we can see that the initial performance of all BERT models improves consistently with the application of filtering with increasing thresholds, until we reach the optimal threshold value.The improvement holds across different dataset subsets (GTD) even if we apply the more relaxed filtering option ($threshold = 1$) which demands that there is at least one concept co-occurence between the document and the query, for the document to be considered valid. Moreover, performance increases consistently until $threshold = 5$ for all models, with a sole exception to the trend for BlueBERT $threshold = 2$. Performance drops for larger thresholds but we have to note that the optimal threshold is related to the length of the queries and documents and needs to be studied separately for different dataset setups. Additionally, the low number of related documents per query significantly impacts the Recall@N values shown in Figure 6 which start from low values for all model and threshold variations. However, we can see that still, comparing the baseline BERT models when there is no concept-based filtering and the case where we use the previously identified optimal threshold for filtering ($threshold = 5$), we get a significant improvement for recall, with emphasis on $N < 200$ (see 6a). Similar observations were seen in precision, presented in table 2, where we should note that for the BlueBERT model we obtain better performance than the one reported in [15] for P@5 and comparable for P@10. We expect that we would see considerably higher values for datasets with a larger number of related documents per query,

Table 2
Precision comparison for different models between no threshold and the optimal threshold (5) version

	Precision@5		Precision@10	
	no threshold	threshold=5	no threshold	threshold=5
BERT	0.036	0.063	0.022	0.054
BioBERT	0.054	0.100	0.036	0.090
BlueBERT	0.127	0.163	0.090	0.113

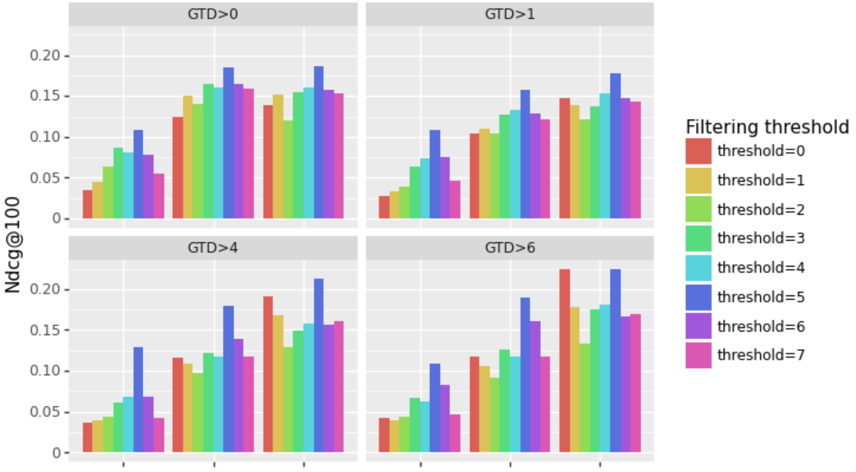
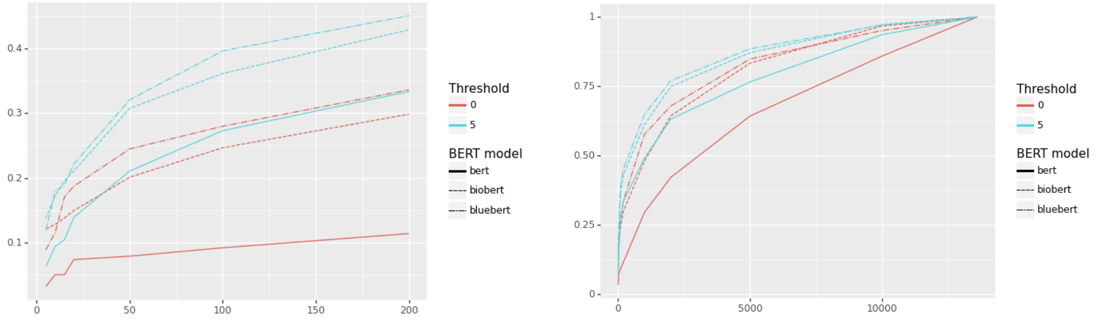


Figure 5: Performance for NDCG@100 for dataset subsets varying by the minimum number of ground truth documents per query (GTD).

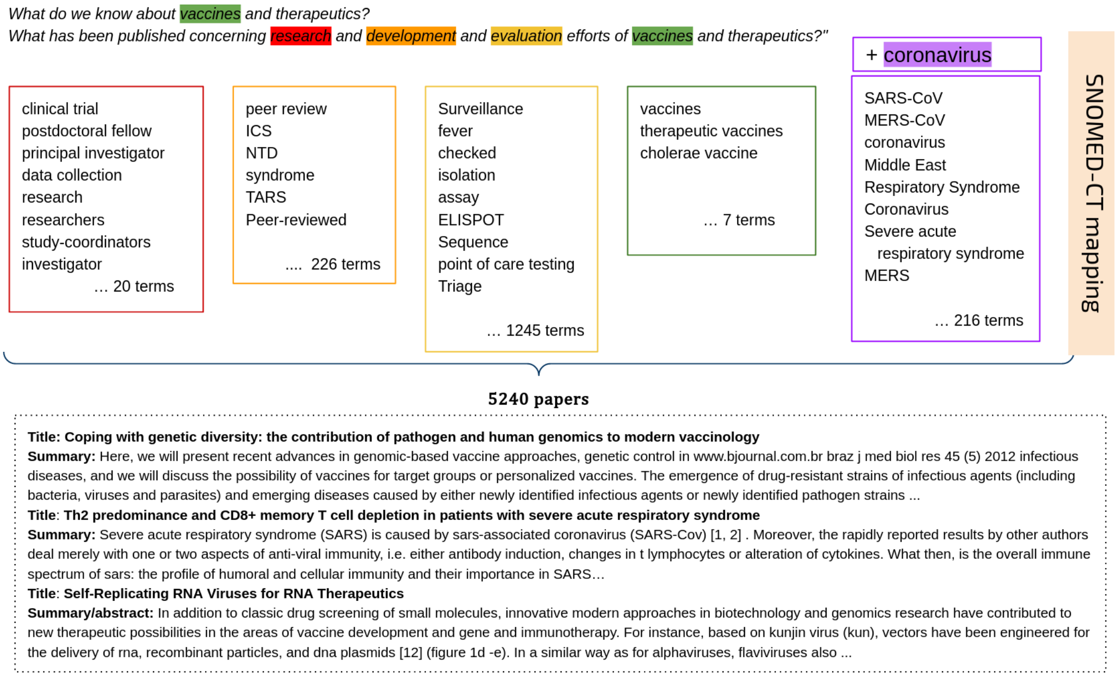


(a) Recall performance up to R@200

(b) Recall performance up to R@1300

Figure 6: Recall comparison between different models for the optimal threshold (5) and the baseline, no concept-filtering version

but we reserve such experiments for future work as explained in Section 5.


What do we know about vaccines and therapeutics?

What has been published concerning research and development and evaluation efforts of vaccines and therapeutics?"

Title: Coping with genetic diversity: the contribution of pathogen and human genomics to modern vaccinology
Summary: Here, we will present recent advances in genomic-based vaccine approaches, genetic control in www.bjournal.com.br braz j med biol res 45 (5) 2012 infectious diseases, and we will discuss the possibility of vaccines for target groups or personalized vaccines. The emergence of drug-resistant strains of infectious agents (including bacteria, viruses and parasites) and emerging diseases caused by either newly identified infectious agents or newly identified pathogen strains ...
Title: **Th2 predominance and CD8+ memory T cell depletion in patients with severe acute respiratory syndrome**
Summary: Severe acute respiratory syndrome (SARS) is caused by sars-associated coronavirus (SARS-Cov) [1, 2] . Moreover, the rapidly reported results by other authors deal merely with one or two aspects of anti-viral immunity, i.e. either antibody induction, changes in t lymphocytes or alteration of cytokines. What then, is the overall immune spectrum of sars: the profile of humoral and cellular immunity and their importance in SARS…
Title: **Self-Replicating RNA Viruses for RNA Therapeutics**
Summary/abstract: In addition to classic drug screening of small molecules, innovative modern approaches in biotechnology and genomics research have contributed to new therapeutic possibilities in the areas of vaccine development and gene and immunotherapy. For instance, based on kunjin virus (kun), vectors have been engineered for the delivery of rna, recombinant particles, and dna plasmids [12] (figure 1d -e). In a similar way as for alphaviruses, flaviviruses also ...

Figure 7: Processing sample and output for Covid-19 query related to vaccines

4.2. Exploratory results on COVID-19 queries

To demonstrate the direct applicability of the described methodology on other datasets without further fine-tuning, we also experiment with an early version of the CORD-19 dataset, as used for the respective kaggle challenge [2]. The dataset is a snapshot of the 10th of April 2020, and contains 51045 documents accompanied with their full text and metadata information. Where available, the abstract is provided as a separate metadata element, however approx. 27K articles had invalid abstracts, based on the criterion described in Section 3.2. Documents are selected based on their expected relevance to the COVID-19 pandemic, covering a wide range of biomedical, clinical and socioeconomic aspects, spanning the period from 1985 to 2020.

Since our initial motivation was the kaggle CORD-19 challenge and associated queries, we present below representative examples of query outputs showing also the difference between the highlighted queries. We show the process and output for identifying the relevant passages for one of the main kaggle questions in Figure 7. We added the *coronavirus* concept to the query when it was not explicitly stated to further adapt to the domain. We show the identified concepts and sample SNOMED-CT mappings (Figure 7), as well as the start of the produced BERTSUM summaries. We provide an interactive query UI on kaggle [3].

[2] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[3] https://www.kaggle.com/georgefila/semantic-enrichment-of-pre-trained-embeddings

5. Conclusions and Future Work

Motivated by the rapid evolution of covid-related publications and query topics, we explored options for improving unsupervised IR on emerging queries in the biomedical domain. This is preliminary work, exploring the use of SNOMED-CT to further filter the relevant documents, ranked by BERT model variations. We showed that even with a simple co-occurrence filtering method, we can significantly improve the initial results and achieve comparative performance to other unsupervised work on the same dataset. Specifically, we show that for multi-document queries and using the BlueBERT model as a basis, the filtering method reaches 0.23 for the NDCG@100 metric. Additionally, we show that we can get meaningful gains across different metrics even for models trained on generic data. Indeed, BERT-based results filtered using SNOMED-CT surpass the performance of unfiltered BioBERT results.

The aforementioned outcomes provide solid basis for further experimentation into better exploitation of knowledge graphs and concept hierarchies as a means of boosting IR on new topics in an unsupervised manner. We intend to further establish our findings by applying the described approach to the full OHSUMED corpus, as well as other biomedical datasets (CLEF, TREC CDS, TREC CORD-19, etc). Moreover, we inted to experiment with neural network architectures other than BERT, such as XLNET and ELECTRA [28]. Upon completion, our future work will be focused on two main tracks. Firstly, explore in more detail the potential of the SNOMED-CT hierarchy. More specifically, in this work we only incorporated the description and first parent node of each identified concept, without further traversing the concept graph. We hypothesise that the position of a concept in the hierarchy as well as the neighbourhood size and type we consider for each concept would impact the size of intersecting concepts between queries and documents (hence impacting the threshold value), and would potentially allow us to identify further connections between missed documents. Additionally, the incorporation of different types of neighbours for each concept would allow us to come up with more elaborate re-ranking formulas, taking into account multiple variables to produce the final document score. The second line of future work, concerns the use of external knowledge sources such as SNOMED-CT not in order to obtain a re-ranking functionality, but as a means to achieve transfer learning and distant supervision to better adapt deep neural networks and pretrained embeddings to new domains.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[3] Z. Zhang, A. L. Gentile, F. Ciravegna, Harnessing different knowledge sources to measure semantic relatedness under a uniform model, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 991–1002.

[4] B. P. C. de Castro, H. F. Rodrigues, G. R. Lopes, M. L. M. Campos, Semantic enrichment

and exploration of open dataset tags, in: Proceedings of the 25th Brazillian Symposium on Multimedia and the Web, 2019, pp. 417–424.

[5] M. CALISAN, C. O. SAKAR, Classification of short-texts by utilizing an external knowledge source, Journal of Science and Engineering 19 (2017).

[6] Z. Li, Y. Lian, X. Ma, X. Zhang, C. Li, Bio-semantic relation extraction with attention-based external knowledge reinforcement, BMC Bioinformatics 21 (2020) 1–18.

[7] J. Chen, Y. Hu, J. Liu, Y. Xiao, H. Jiang, Deep short text classification with knowledge powered attention, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 6252–6259.

[8] H. Fei, Y. Ren, Y. Zhang, D. Ji, X. Liang, Enriching contextualized language model from knowledge graph for biomedical information extraction, Briefings in Bioinformatics (2020).

[9] A. Mourão, F. Martins, J. Magalhaes, Multimodal medical information retrieval with unsupervised rank fusion, Computerized Medical Imaging and Graphics 39 (2015) 35–45.

[10] M. Song, I.-Y. Song, X. Hu, R. B. Allen, Integration of association rules and ontologies for semantic query expansion, Data & Knowledge Engineering 63 (2007) 63–75.

[11] Y. Gupta, A. Saini, A novel fuzzy-pso term weighting automatic query expansion approach using combined semantic filtering, Knowledge-Based Systems 136 (2017) 97–120.

[12] M. A. Raza, R. Mokhtar, N. Ahmad, M. Pasha, U. Pasha, A taxonomy and survey of semantic approaches for query expansion, IEEE Access 7 (2019) 17823–17833.

[13] M. Agosti, S. Marchesin, G. Silvello, Learning unsupervised knowledge-enhanced representations to reduce the semantic gap in information retrieval, ACM Transactions on Information Systems (TOIS) 38 (2020) 1–48.

[14] X. Liu, J.-Y. Nie, A. Sordoni, Constraining word embeddings by prior knowledge–application to medical information retrieval, in: Asia information retrieval symposium, Springer, 2016, pp. 155–167.

[15] M. Rais, A. Lachkar, An empirical study of word sense disambiguation for biomedical information retrieval system, in: International Conference on Bioinformatics and Biomedical Engineering, Springer, 2018, pp. 314–326.

[16] H.-S. Oh, Y. Jung, Cluster-based query expansion using external collections in medical information retrieval, Journal of biomedical informatics 58 (2015) 70–79.

[17] Y. Liu, Fine-tune bert for extractive summarization, arXiv preprint arXiv:1903.10318 (2019).

[18] R. Jia, Y. Cao, F. Fang, J. Li, Y. Liu, P. Yin, Enhancing pre-trained language representation for multi-task learning of scientific summarization, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–8.

[19] C. Zerva, M.-Q. Nghiem, N. T. Nguyen, S. Ananiadou, et al., Cited text span identification for scientific summarisation using pre-trained encoders, Scientometrics (2020) 1–29.

[20] A. Nikiforovskaya, N. Kapralov, A. Vlasova, O. Shpynov, A. Shpilman, Automatic generation of reviews of scientific papers, arXiv preprint arXiv:2010.04147 (2020).

[21] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, D. McClosky, The Stanford CoreNLP natural language processing toolkit, in: Association for Computational Linguistics (ACL) System Demonstrations, 2014, pp. 55–60.

[22] H. Xiao, bert-as-service, https://github.com/hanxiao/bert-as-service, 2018.

[23] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[24] Y. Peng, Q. Chen, Z. Lu, An empirical study of multi-task learning on bert for biomedical text mining, arXiv preprint arXiv:2005.02799 (2020).

[25] K. Donnelly, Snomed-ct: The advanced terminology and coding system for ehealth, Studies in health technology and informatics 121 (2006) 279.

[26] A. R. Aronson, F.-M. Lang, An overview of metamap: historical perspective and recent advances, Journal of the American Medical Informatics Association 17 (2010) 229–236.

[27] L. Yao, Y. Zhang, B. Wei, Z. Jin, R. Zhang, Y. Zhang, Q. Chen, Incorporating knowledge graph embeddings into topic modeling, in: 31st AAAI conference, 2017.

[28] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, arXiv preprint arXiv:2003.10555 (2020).