SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of

Informatics, Systems and Communication

PhD program in Computer Science, Cycle XXXVI

Adaptation of Neural-enhanced Retrieval Models to Domain-specific Tasks

Oscar Javier Espitia Mendoza

Registration number 865289

Supervisor: Prof.GabriellaPasi

Coordinator: Prof.LeonardoMariani

ACADEMIC YEAR 2022-2023

Adaptation of Neural-enhanced Retrieval Models to Domain-specific Tasks

Enhancing on Medical and Academic search: Domain-specific Applications of Advanced Retrieval

Oscar E. Mendoza

Supervisor: Prof. Gabriella Pasi

Department of Computer Science
University of Milano-Bicocca

This dissertation is submitted for the degree of
Doctor of Philosophy

January 2024

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgments. This dissertation contains fewer than 65,000 words, including appendices, bibliography, footnotes, tables, and equations, and has fewer than 150 figures.

Oscar E. Mendoza
January 2024

Acknowledgements

I extend my gratitude to the following individuals and institutions whose support and guidance have been instrumental in the completion of my doctoral journey:

I express my appreciation to my main supervisor, Prof. Gabriella Pasi, whose guidance and mentorship played a pivotal role in shaping my research. I am also grateful to Prof. Allan Hanbury and Prof. Andrew Yates, my supervisors during periods abroad, for their valuable insights and support.

To my local colleagues, especially Ricky, Georgos, and Pranav, I extend my appreciation for the enriching environment. A special acknowledgement goes to my colleagues during my stays abroad: Wojciech, Sophia, and Thong, who helped me get the most out of those experiences.

I acknowledge the financial support provided by the Marie Sklodowska-Curie Actions Innovative Training Networks project "Domain Specific Systems for Information Extraction and Retrieval"–DoSSIER. Thanks to all network members, particularly Florina, for their collaboration and encouragement.

I express my gratitude to the Department of Computer Science at the University of Milano-Bicocca for their institutional support.

Appreciations to my friends back in Colombia and those who shared with me during these three years. Special thanks to Luis and Riccardo, my cherished friends, who provided companionship and camaraderie during the challenging phases of my PhD. Their friendship and shared moments of respite were invaluable.

I am profoundly thankful to my family, especially Yuri Mejia, Ayde Mendoza, and Mario Espitia Mendoza. Their life experiences and invaluable insights have been a source of inspiration and support, shaping my personal and professional approach to work. Thank you. This thesis is dedicated to you.

Abstract

Information retrieval (IR) plays the role of ranking information items in search engines widely used in many scenarios. The criteria used to produce a rank of information items are matching signals between information needs, expressed as queries, and the items. These signals are related to the notion of relevance used to judge such items.

This thesis studies how to design IR models in specific scenarios, characterized by their domains, based on contextual factors from particular instances. We have considered the task contexts of Clinical Trials Retrieval (CTR) and Scholarly Document Retrieval (SDR), specifically addressing search processes in clinical trials collections and collections of academic documents, respectively. Compared to traditional ad-hoc text retrieval, CTR and SDR exhibit different challenges: the queries could be much longer and more complex than common keyword-based queries; the definition of the relevance of a document to a query is beyond general topical relevance (i.e., the semantic relationship between texts), and as such, its assessment may require expert knowledge.

Curriculum learning is an approach in machine learning that involves designing a curriculum to enable the model to learn concepts progressively from simple to complex, especially when a task can be decomposed. We proposed a curriculum learning approach to address the CTR problem, in which a model is first optimized based on topical relevance and then on eligibility classification (i.e., screening the criteria given in a trial for patient enrollment). This setting is used to establish a re-ranking pipeline. Our proposed re-ranking formula explicitly models the eligibility decisions instead of using only the topical relevance and shows additional performance improvement comparable to more expensive approaches.

In the case of SDR, classifying scholarly documents according to their research themes is an important task to improve their retrievability. To establish a benchmark

for research theme classification, we present experiments and evaluation results with traditional machine learning models and compare them to a more sophisticated ensemble with state-of-the-art models. A clear limitation is the overlap between disciplines that leads to incorrect predictions when considering mutually exclusive categories. We consider, then, a fine-grained theme distribution. We leverage the capabilities of large pre-trained Transformer models in an architecture that uses a sequence-to-sequence learning system to map text to fine-grained themes. We evaluate an approximation to Learned sparse retrieval (LSR) to directly introduce these theme annotations to a sparse model.

Constraining search with theme classification can contribute to the performance of a retrieval system based on the results of the different tasks, and LSR has shown to be a potential channel to incorporate contextual domain-specific information in the IR system.

Table of contents

List of figures **xiii**

List of tables **xvii**

Nomenclature **xix**

I Context **1**

1 Introduction **3**
1.1 Research Outline and Questions 5
1.1.1 Effective Clinical Trials Search: An Application of Neural Information Retrieval in the Medical Domain 5
1.1.2 Enhancing the Retrievability of Domain-specific Documents with Neural Models 6
1.2 Thesis Contributions 7
1.3 Definition and Scope Clarification: 9
1.4 Overview 9
1.5 Origins 11

2 Background **13**
2.1 The Retrieval Problem 14
2.2 Neural-enhanced Retrieval Models 15
2.2.1 Neural Ranking 16
2.2.2 Dense Retrieval 17
2.2.3 Learned Sparse Retrieval 18
2.2.4 Generative Retrieval 18

2.3 Neural-enhanced Retrieval Models in Domain-specific tasks 19

II Effective CT Search: An Application of NIR on the Medical Domain 21

3 Curriculum Learning for Neural Information Retrieval in Clinical Trials Retrieval 23
3.1 Introduction 23
3.2 Related Work 25
3.3 Problem Statement 26
3.4 Curriculum Learning for NIR in CT 27
3.5 Experiments 30
3.5.1 Dataset 30
3.5.2 Evaluation 30
3.5.3 Baselines 31
3.5.4 Implementation details 31
3.6 Results 32
3.7 Discussion 34

III Enhancing the Retrievability of Domain-specific Documents with Neural Models 37

4 Broad Theme Classification 39
4.1 Introduction 39
4.2 Related work 40
4.3 Ensemble Model for Theme Classification 41
4.3.1 Transformer-based Classifier 42
4.3.2 Data Enrichment 43
4.3.3 Extending Labels to Enriched Data 44
4.3.4 Aggregating Predictions from Enriched Data 44
4.4 Experiments 44
4.4.1 Training Settings 45
4.4.2 Prediction Settings 45
4.4.3 Evaluation metrics 46

4.4.4 Baseline Models 46
4.5 Results 47
4.5.1 Validation Results 47
4.5.2 Test Results 50
4.6 Discussion 50

5 Large-Scale Hierarchical Classification Analysis 53
5.1 Introduction 53
5.2 Related Work 54
5.3 Large Scale Classification as a Generation Task 55
5.4 Experiments 57
5.4.1 Datasets 57
5.4.2 Baselines 58
5.4.3 Implementation Details 58
5.5 Results 58
5.6 Discussion 59

6 An Approximation to Learned Sparse Retrieval for Domain-specific Documents 61
6.1 Introduction 61
6.2 Related Work 63
6.3 Problem Statement 65
6.4 Contextual Expansions for Retrieval 65
6.4.1 Overview 66
6.4.2 Document representation and indexing 67
6.4.3 Query representation 67
6.4.4 Document scoring 68
6.5 Experiments 68
6.5.1 Datasets 68
6.5.2 Implementation Details 69
6.5.3 Baselines 69
6.6 Results 69
6.7 Discussion 73

7 Discussion and Conclusions 75
7.1 Main Findings and Results 75

7.2 Upsides Compared to Earlier Investigations 76
7.3 Prospective Applications 77
7.4 Limitations 77
7.5 Future Directions 78

References **79**

Appendix A Datasets Descriptions **93**
A.1 CTR Datasets 93
A.2 Theme Classification Dataset 94
A.3 SDR Dataset 96

Appendix B Supplementary Results **97**
B.1 CTR Supplementary Results 97
B.2 CTR TREC Records 99
B.3 Supplementary Results on Broad Theme Classification 101
B.4 Qualitative analysis of context-enhanced SDR 102
B.4.1 Expert Assessment 102
B.4.2 Query Analysis 102
B.5 Supplementary results on SDR 106

List of figures

2.1 Representation and interaction-based scoring. (a) Representation-based models focus on optimizing text representations, while the score is usually computed as a simple similarity operation. (b) Interaction-based models focus on studying interactions between queries and documents that can bring relevance signals that contribute to the final score while the representations remain static. . . . 16
2.2 Bert-based scoring. Analogous to interaction-based models, BERT's attention layers capture interactions between text inputs. The Last hidden state of the [CLS] token is used as input for a traditional classifier to perform relevance classification. 17
2.3 Learned sparse scoring. An example of a model that replaces TF values with predicted weights for indexing documents. 18

3.1 Patient to CT matching problem: (left) Patient record: free text similar to an EHR with demographics and medical history of the patient ; (right) CT example: Structured multi-fielded document from ClinicalTrials.gov database with trial description in different levels and criteria for eligibility. 27
3.2 Neural re-ranking setup TCRR. Training system for CT retrieval based on two objectives: topical relevance and eligibility. The model is first trained for discriminating relevant and irrelevant trials given a patient and then for classifying their eligibility for the relevant trial. 29

4.1 Ensemble for research theme classification. CLS stands for classifier. For a publication, various sections are evaluated as independent samples with the classifier; the final prediction is achieved by aggregating sections' predictions. 45

4.2 Confusion Matrix for validation results for a sample of classes. For Clinical Medicine, most examples where the model's incorrect prediction are classified as Allied Health Professions, Nursing and Pharmacy, and Biological Sciences. Similar behavior can be observed with related fields of study; see Appendix B for an extended analysis. 49

5.1 Large Scale Classification as a Generation Task. During training, the seq2seq model takes documents and target labels to build an embedding space for a given set of categories. During inference time, given a text snippet, it is supposed to generate a set of labels matching the input. 56

6.1 Vocabulary expansion: Example of representing a query with the expanded vocabulary. Here, the original vocabulary is extended with members of an ontology. The example corresponds to the computer science field. 66

A.1 Breakdown of the research theme classification dataset by theme. About 60,000 records from the REF dataset were selected along 36 categories and enriched for this dataset. 94

B.1 Averaged per patient count of relevant (top) and excluded (bottom) trials depending on a cut-off of K trials retrieved (x-axis) for TREC CT 2022 collection. Both versions of Bm25 show a positive impact of the field selection in finding more eligible trials. The TCRR neural re-ranking retrieves twice as many trials for which a patient is eligible than excluded; it helps to remove ineligible until the first 15 retrieved trials. 98

B.2 Confusion Matrix for validation results for 25 most frequent classes. The remaining 11 classes are grouped in the 'others' category. Notice that for Clinical Medicine, most examples where the model's incorrect prediction are classified as Allied Health Professions, Nursing and Pharmacy, and Biological Sciences. Similar behavior can be observed with other closely related disciplines. 101

B.3 Qualitative performance report. Performance comparison (measured by P@10 and nDCG@10) between our CESR and the BM25 for each of the 20 proposed queries. Overall, there is an improvement achieved by CESR in the global comparison provided by the histogram. 103

B.4 Search result example in the qualitative analysis for CESR. Parallel between CESR and BM25 for the proposed query *T 06*. A sample of the distribution of expansions is also shown for each retrieved document (relevant expansions are highlighted). 103
B.5 Example of content selection for the SDR task. We highlight the relevant content for two examples given query examples. Relevant document expansions are also highlighted. 107
B.6 Frequency of document expansions on retrieved documents by query examples. BM25 retrieves a set of documents for a given query; since we previously expanded documents with concepts from an ontology, we can measure how frequent expansion terms are in a retrieved set of documents. 108
B.7 Expansions scores for documents retrieved with the query examples. Heat maps for the topic distribution of the first 30 ranked retrieved documents 109

List of tables

3.1 Neural re-ranking evaluation results on TREC test set. Underlined values indicate the highest score among general models. **Bold** values indicate the highest score achieved by our approach. †-marked models indicate that there is a significant improvement over the BM25 baseline using the Student's paired t-test with a 95% confidence level. 33

4.1 Micro F1-score results form the comparison using different input features for prediction. BERT_T stands for BERT model trained on titles only, BERT_{T+A} means model trained on both titles and abstracts. 48

4.2 Three experiments testing the utility of individual sections on BERT_{T+A}. The augmentation is evaluated by independent sections combined with titles. Samples are selected such that corresponding sections are available for all documents. 48

4.3 Validation results using different fields for BERT_{T+A}. The experiments vary in the prediction and aggregation settings. The aggregations we use are simply weighted sums with uniform weights and assigned arbitrarily according to Section 4.4.2. 49

4.4 Test results with different experimental (Run) settings. The experiments vary in the training (T), prediction (P), and aggregation (Agg.) settings. The aggregations we use are simply weighted sums with uniform weights (U) and compensation weights (C) assigned according to section 4.4.2. 50

5.1 Large scale classification benchmarks. Statistics of WOS, EURLEX-57K, and CSO benchmarks for multi-label classification. 57

5.2 Large scale classification results (R and F1). We show results for the generative annotation compared with specific baselines on WOS, EURLEX-57K, and CSO benchmarks. Underlined values are the best results. 59

6.1 Relevant expansions for a sample of queries. Strike-through items correspond to examples of apparent bad expansion terms. Expansion terms are the result of PRF on expanded documents. 70
6.2 Theme concentration analysis. Report on average Diversity and Entropy, measured over the document expansions for retrieved documents. Underlined values correspond to the best results. 71
6.3 Retrieval evaluation. Comparative results reported on retrieval metrics. Underlined values correspond to the best results. 71
6.4 Retrieval evaluation at different rank levels. Comparative results reported on retrieval metrics at different levels of the rank. Underlined values correspond to the best results, and the †-mark indicates statistically significant improvement. 72

A.1 Statistics of TREC CT datasets from 2021 and 2022. The train set is from the 2021 edition, and the test set is from the 2022 edition. . . . 93
A.2 Statistics for the research theme classification dataset. Percentages of information available for each record on both train and test sets. . 95
A.3 SDR Dataset statistics. Queries and judgments available from the dataset for SDR by SanJuan et al. [95]. 96

B.1 Official TREC CT 2022 evaluation results for TCRR. Comparative results reported on retrieval metrics. Underlined values correspond to the best results. 100

Nomenclature

Acronyms / Abbreviations

AZ	Argumentative Zoning
CT	Clinical Trial
DPR	Dense Passage Retrieval
EHRs	Electronic Health Records
GenIR	Generative Retrieval
IDF	Inverse Document Frequency
IR	Information Retrieval
LLM	Large Language Model
LM	Language Model
LSR	Learned Sparse Retrieval
MLM	Masked Language Model
nDCG	Normalized discounted cumulative gain
NIR	Neural Information Retrieval
NN	Neural Networks
P	Precision

PRF Pseudo Relevance Feedback

R Recall

RR Reciprocal rank

SDR Scholarly Document Retrieval

Seq2seq Sequence-to-sequence

SR Sparse Retrieval

TF Term Frequency

Part I

Context

1

Introduction

In many scenarios, users often turn to search engines for help. Such systems point them in the direction of possible relevant information items. Information retrieval (IR) plays the role of ranking these information items to generate an ordered list retrieved from a corpus in response to an information need a user might have. The criteria used to produce a rank are matching signals between information needs, expressed as queries, and the information items. These signals are related to the notion of relevance used to judge the result.

Different volumes of data, document sizes, the structure of the documents, jargon, and the way information needs are defined, among others, are features that justify that a retrieval model should not handle all information equally when it comes to domain-specific retrieval tasks, e.g., in healthcare, and academia.

Some of those features can be considered contextual elements, i.e., factors influencing how an IR system is used and how its performance should be evaluated. The concept of context in IR has been extensively studied within the contextual IR paradigm. This paradigm in itself aims to optimize the retrieval performance by defining the search context to be considered in the information selection process and in assessing the search outcome [65].

According to different interpretations, context can be considered from different perspectives based on the sources of evidence available to the search system, such as

content (queries, content of the collection, supplementary content, etc.), user behavior (visited content, explicit/implicit feedback), and environment (location, time, device, etc.) [31]. Thus, there are approaches that have used content as context, domain as context, environment as context, user preferences as context, and search task or type of information as context.

In this thesis, we study how to design IR models in specific domains based on a set of particular instances where we first consider the task contexts:

> *Patient-clinical trials matching*. Clinical trials (CT) are experiments to develop new medical treatments, drugs, or devices. Recruiting candidates for a trial motivates the IR task of matching eligible patients to CT [51] (we will refer to as CT Retrieval or CTR).
>
> *Searching for academic literature*. Broadly, queries are usually given under an informational intent [22], i.e., by submitting them to the search system, we want to obtain some related information that is assumed to be available. Additionally, the user might be aiming to solve a task of information gathering, which involves collecting information from multiple sources [47] (we will refer to this task as Scholarly Document Retrieval or SDR).

Exhibiting different challenges, these tasks involve queries that can be much longer and more complex than common keyword-based queries. For example, in CTR, queries often arise from patient descriptions in unstructured snippets. Furthermore, determining the relevance of a document to a query goes beyond general topical relevance. In CTR, a document is considered relevant not only based on its semantic relationship to the query but also, for instance, to the eligibility criteria. This complexity in assessment may require expert knowledge. These tasks also exhibit domain-specific properties (e.g., jargon, multi-fielded structure), allowing the exploration of domain-based information sources, particularly domain-specific knowledge. In the context of search, it has the potential to constrain the information space to find relevant documents more effectively than in open-domain applications.

Neural networks (NN) represent a tool for learning text representations, defining ranking models capable of handling different kinds of documents, and even introducing additional elements in the retrieval process. Models using NN offer the flexibility for adapting models based on domain-specific features, including those related to contextual factors.

NN in IR has produced a family of Neural-enhanced retrieval models currently shaping state-of-the-art in multiple retrieval tasks: neural ranking models use neural methods for relevance classification, dense retrieval models generate optimized embeddings for scoring, learned sparse retrieval models employ learned weights instead of bag-of-words, and generative retrieval embeds documents into sequence-to-sequence model parameters (see Section 2.2 for details).

We explore concepts from neural-enhanced retrieval models such as neural ranking for tailoring the relevance estimation to our study cases; generative retrieval for embedding domain-knowledge, and Learned sparse retrieval incorporating domain-specific information into the document scoring process.

1.1 Research Outline and Questions

As mentioned above, this thesis has explored the properties of recent NN models to define IR models in specific domains and for specific tasks. We focus on two research objectives: (1) Effective Clinical Trials Search: An Application of Neural Information Retrieval in the Medical Domain, and (2) Enhancing the Retrievability of Domain-specific Documents with Neural models at different levels.

Below, we describe the main research questions for each chapter. In each chapter, we describe more fine-grained subquestions that we ask to answer each main research questions.

1.1.1 Effective Clinical Trials Search: An Application of Neural Information Retrieval in the Medical Domain

The general retrieval model design is focused on general signals of relevance. For domain-specific search tasks, the effectiveness-efficiency trade-off in performance can be sorted with a task-oriented model design, which considers more specific signals. A particularly challenging task in the medical domain is CT recruitment. From the IR perspective, it has been framed as a search task under the patients-to-trials evaluation paradigm. In our first study, we approach the problem of finding eligible trials given patient description using a neural-enhanced IR model:

RQ1 Given the clinical trials IR problem, can we optimize a model upon different signals of relevance?

We propose splitting the problem based on two objectives considered for defining the relevance of CTR: (i) the semantic relationship between texts and (ii) the eligibility criteria. We optimize a model based on these two objectives and evaluate the contribution to the performance of selected trials. However, splitting the problem into granular problems does not help to increase the performance of a general-purpose model. To address this limitation, in our study, we consider an integrated approach:

RQ2 Given the multitask nature of CTR, can we design a task-oriented model under a reasonable effectiveness-efficiency trade-off?

Curriculum learning is the approach to design a curriculum for learning concepts from simple to complex when a task can be decomposed. We proposed a curriculum learning approach to address the CTR problem in which the model is first optimized based on topical relevance and then on eligibility classification.

1.1.2 Enhancing the Retrievability of Domain-specific Documents with Neural Models

Classifying research papers according to their research themes is an important task to improve their retrievability and, in general, to support approaches for analyzing and making sense of the research environment. Identifying research themes for a given scholarly document poses a challenging task due to the lack of large, high-quality labeled data. This made it difficult both to train high-performance classification models as well as to compare models' performance across studies.

RQ3 Can academic publications be effectively discriminated into broad themes when high-quality data is provided?

To establish a benchmark for research theme classification, we present experiments and evaluation results with traditional machine learning models and compare them to a more sophisticated ensemble with state-of-the-art models. A clear limitation is the overlap between disciplines that leads to incorrect predictions when considering mutually exclusive categories. We consider, then, a fine-grained Themes distribution:

RQ4 Can fine-grained themes be embedded into model parameters for annotating documents?

We study how distributions of more fine-grained themes can be embedded into model parameters for performing the classification of different types of hierarchically arranged categories, as usually defined for large-scale classification. We leverage the capabilities of large pre-trained Transformer models in an architecture that uses a sequence-to-sequence (seq2seq) learning system [101] to directly map text to labels. Drawing on the recent advances in generative language models. By harnessing the context understanding of the Transformer architecture, our proposed method enables robust integration of semantic information into the text-to-label mapping process. Although it was previously stated that topic modeling has potential use in retrieval, a clear pipeline needs to be defined to implement an IR model that integrates these two tasks:

RQ5 How can fine-grained document annotations be used in a neural-enhanced retrieval scenario?

We evaluate an analogous scenario to Learned Sparse Retrieval as a way of directly introducing theme annotations to the retrieval model. We propose a hybrid model that expands documents using a generative model that predicts fine-grained annotations and expands queries using expansions of the feedback from retrieved documents. Then, modify the sparse retrieval scoring formula to allow expansions and their importance to tailor results.

1.2 Thesis Contributions

In this section, we summarize the main contributions of this thesis into three categories: methodological contributions regarding theoretical and algorithmic insights, empirical contributions regarding experimentation and analysis, and resources we released throughout developing this thesis.

Methodological contributions

1. Effective training strategy for CTR, a method that applies the idea of curriculum learning with diverse matching signals to estimate relevance (Chapter 3).
2. Establishing a benchmark for scholarly document classification (Chapter 4).
3. A method for generating granular themes annotations as an application of large-scale classification (Chapter 5).

4. A method to introduce domain context into the retrieval model to improve first-stage retrieval (Chapter 6).

Empirical contributions

5. Effective retrieval model for CT (Chapter 3).
    a. Empirical comparison of our proposed training strategy and other retrieval models.
    b. Empirical ablation study, breaking down our multistage ranking pipeline.
6. Classification model for scholarly documents (Chapter 4).
    a. Empirical comparison of our proposed ensemble model and multiple traditional models.
    b. Empirical evaluation of the contribution of multiple sources of information for classification performance.
    c. Error analysis of common errors made by our ensemble.
7. Annotating text with hierarchically arranged sets of labels (Chapter 5).
    a. Empirical comparison of our proposed strategy with state-of-the-art approaches.
    b. Empirical evaluation of the approach in multiple scenarios with varying domains/hierarchies.
8. Leveraging context trough LSR (Chapter 5).
    a. Empirical evaluation of the model with multiple experiments.
    b. Empirical evaluation of the approach in multiple scenarios: in comparison with alternative baselines and analysis of multiple parameters.

Resources

9. Open source implementations.[1,2]
10. Wide range of baselines for establishing a new benchmark for text classification.

[1] https://github.com/ProjectDossier/sdp2022
[2] https://github.com/ProjectDossier/patient-trial-matching

11. Neural Models survey.[3]
12. Neural Models tutorial.[4]

1.3 Definition and Scope Clarification:

Even though they are related concepts, in the context of this thesis, "themes" and "topics" hold distinct roles, each serving a specific purpose in the framework of our research.

> *Themes*, as used within the scope of research theme classification (chapters 4 and 5), denote the fundamental conceptual categories. These themes represent the fundamental ideas or concepts that recur throughout datasets, aiding in the interpretation and understanding of the underlying patterns or principles that characterize the subject matters under investigation.
>
> *Topics*, on the other hand, are used in the context of IR (chapters 3 and 6) to refer to the contextual elements provided for a query in an evaluation setting. This is also related to topical relevance, a concept discussed later in Chapter 2. These topics serve as the key points or subject areas that define the information needs of the user. They are usually given to human assessors for facilitating judgments and are relevant to the development of IR models for considering the definition of relevance inside certain contexts.

Although this thesis delineates the specific applications of these terms within their respective analytical frameworks, we find it useful for the reader to establish a clear distinction between the use of "themes" in research theme classification and "topics" in the domain of Information Retrieval.

1.4 Overview

The thesis is organized in three parts: Part I discusses background and state-of-the-art, and Parts II and III describe the contributions of this work.

[3] D. Alexander, O. E. Mendoza, Y. Ghafourian, K. Pathak, G. Peikos, L. Azzopardi, G. Pasi. State of the Art Models Survey. Deliverable for Marie Sklodowska-Curie Actions project DoSSIER.
[4] D. Alexander, O. E. Mendoza, Y. Ghafourian, K. Pathak, G. Peikos, L. Azzopardi, G. Pasi. Models Tutorial. Deliverable for Marie Sklodowska-Curie Actions project DoSSIER.

Specifically, in the first part, we discuss the context of our contributions. We introduce our motivations and how neural models are used to enhance IR (Chapter 2).

In the second part, we study the effectiveness-efficiency trade-off with NIR models for CTR (Chapter 3).

In the third part, we study how to enhance the retrievability of documents in domain-specific search by classifying documents into themes (Chapters 4 and 5) and by evaluating the effect of considering classification in the first-stage retrieval model (6)

In Chapter 7, we conclude the thesis and discuss directions for future work.

1.5 Origins

Below, we list the publications and resources that contributed to this thesis:

- W. Kusa, O. E. Mendoza, P. Knoth, G. Pasi, A. Hanbury. Effective Matching of Patients to Clinical Trials using Entity Ex-traction and Neural Re-ranking. Journal of Biomedical Informatics. 2023. (Chapter 3)

  OEM designed the Neural methods and ran the experiments. WK designed non neural methods ran the experiments. All authors conributed to the text, OEM and WK did most of the writing.

- O. E. Mendoza, W. Kusa, P. Knoth, F. Piroi, and G. Pasi, A. Hanbury. Benchmark for Research Theme Classification of Scholarly Documents. Workshop on Scholarly Document Processing. Coling. Korea. 2022. (Chapter 4)

  OEM designed the Neural methods and ran the experiments. WK helped with algorithmic design and non neural methods and ran the experiments. PK provided dataset and descriptions. All authors contributed to the text, OEM did most of the writing.

- O. E. Mendoza, G. Pasi. Domain Context-centered Retrieval for the Content Selection task in the Simplification of Scientific Literature. Conference and Labs of the Evaluation Forum - CLEF. Greece. 2023. (Appendix B.5–Chapter 6)

The thesis also benefited from insights gained from the following publications:

- W. Kusa, O. E. Mendoza, M. Samwald, P. Knoth, G. Pasi, A. Hanbury. CSMED: Bridging the Dataset Gap in Automated Citation Screening for Systematic Literature Reviews. NeurIPS. New Orleans - USA. 2023. (Not part of the thesis)

- O. Espitia, and G. Pasi. Neural IR for Domain-Specific Tasks. Italian Information Retrieval Workshop - IIR. 2021. Italy. (Not part of the thesis)

- G. Peikos, O. Espitia, and G. Pasi. UNIMIB at TREC 2021 Clinical Trials Track, TREC. 2021. (Not part of the thesis)

2

Background

IR is the process of getting information items relevant to an information need from within collections of some sort. IR has an essential role in many scenarios nowadays, for instance, navigating digital libraries of different kinds of data, finding experts, and Web search overall. As there might be a number of candidate results, the outcome of this process is typically given as a ranked list. The rank is achieved with respect to some notion or signal of relevance. Defining ranking models is one of the most studied research problems in IR.

A significant textual IR application is ad-hoc retrieval, which is the classic retrieval task where users specify their information needs through queries. Submitting a query to an information system initiates searching for likely relevant items to those users. The term ad-hoc refers to the scenario where documents in the collection remain relatively static while new queries are submitted to the system continually [7].

Relevance in many applications is situated in the user and task context and is an important consideration in the design of IR models. The concept of relevance has been extensively studied [10], and although it has many more nuances, we focus on the definition of the various criteria or features users may evaluate in the process of judging retrieved information objects. One of the main features used for designing IR systems is topical relevance, i.e., the relationship between the topic of a request and the information objects retrieved about that topic. Standardized approaches

to relevance classification in IR use statistical models to identify the presence or absence of specific topics that might make a document relevant to the searcher. These approaches have been used to predict better relevance on the basis of what the document is about [68].

Many different ranking models for ad-hoc retrieval have been proposed over the past decades, including vector space models [94], probabilistic models [89], and learning to rank (LTR) models [59], and more recently neural-enhanced models. In this chapter, we discuss the latter, which is the basis of our contributions.

2.1 The Retrieval Problem

A generalized IR problem is focused on finding the optimal scoring function $f(q, d_k)$, such that

$$f(q, d_k) = F\Big(\underbrace{\Phi(q)}_{\substack{\text{query}\\ \text{representation}}} \quad , \quad \underbrace{\Theta(d_k)}_{\substack{\text{document}\\ \text{representation}}}\Big) \quad , \tag{2.1}$$

where F is the interaction function computing interactions between query q and document d_k representations ($\Phi(q)$ and $\Theta(d_k)$, respectively). For convenience, we omit the subscript $k \in [1, \cdots, |D|]$ referring to the k-th document d in a collection D. Different retrieval models arise depending on whether the approach to the problem is defining an interaction function F or the representation functions $\Phi(\cdot)$ and $\Theta(\cdot)$. Part of the research in IR focuses on developing encoders whilst the scoring function remains as a simple dot product operation, such that

$$f(q, d) \triangleq \sum_{j=1}^{N} \Phi(q)_j \times \Theta(d)_j, \tag{2.2}$$

where N is the size of the vector representation. A well-known example of this formulation is the efficient sparse model BM25 [29],[1] in which $N = |V|$ is equivalent to the size of the vocabulary $V = \{v_1, \cdots v_N\}$, thus,

[1] BM25 is a parameterized model (with parameter values b and k_1) that uses term frequency (TF) and inverse document frequency (IDF); it also takes into account the averaged document length (avgdl).

$$
\begin{aligned}
f(q,d) &\triangleq \text{BM25}(q,d) \\
&= \sum_{i=1}^{|q\cap d|} \text{IDF}(q_i) \times \frac{\text{TF}(q_i,d)\times(k_1+1)}{\text{TF}(q_i,d)+k_1\cdot\left(1-b+b\cdot\frac{|d|}{\text{avgdl}}\right)}, \\
&= \sum_{j=1}^{N} \underbrace{\mathbf{1}_q(v_j)\text{IDF}(v_j)}_{\text{query encoder}} \times \underbrace{\mathbf{1}_d(v_j)\frac{\text{TF}(v_j,d)\times(k_1+1)}{\text{TF}(v_j,d)+k_1\cdot\left(1-b+b\cdot\frac{|d|}{\text{avgdl}}\right)}}_{\text{document encoder}},
\end{aligned}
\tag{2.3}
$$

where $q_i \in q$ is the i-th token of q, and $\mathbf{1}$ is an indicator function.[2]

2.2 Neural-enhanced Retrieval Models

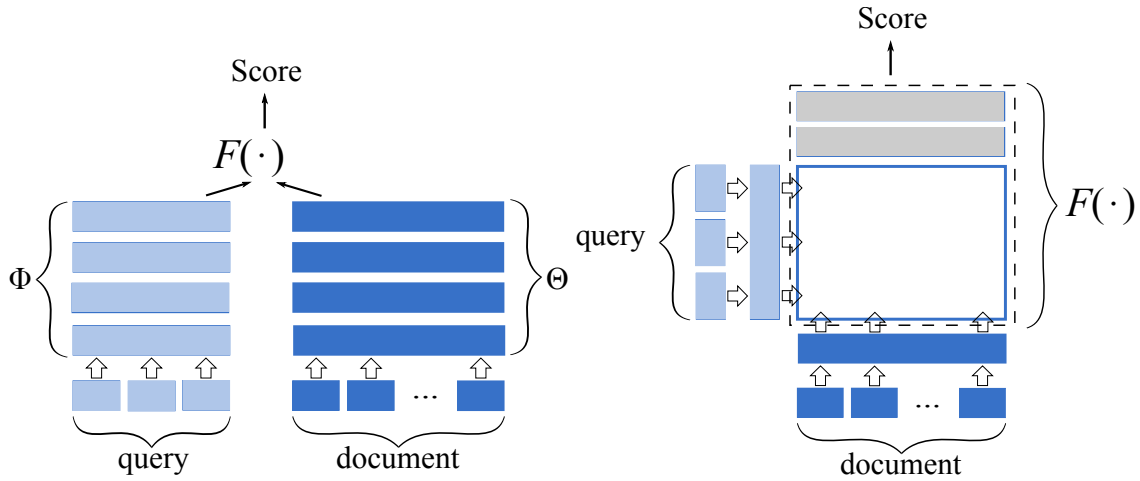
The introduction of pre-trained language models substantially impacted the effectiveness of retrieval approaches, as witnessed at the largest-scale evaluation of retrieval techniques [15]. Transformers-based approaches in IR not only enabled significant improvements [116] but also influenced the development of a wide variety of advanced models specialized in different problems in the IR ground. Before Transformers, NN were used in IR also within the main two model variants: representation-based, focused on optimizing encoders, and interactions-based models, focused on scoring functions taking into account interactions between query and documents. We illustrate both architectures at a high level in Figure 2.1.

These two branches shaped the principles of neural architectures that later would influence, for instance, the widely used Transformers-based dual-encoders[3] and cross-encoders[4]. From this point, neural approaches have been studied mainly in four categories: neural ranking models, dense retrieval models, learned sparse retrieval models, and the most recent generative retrieval models.

[2] $\mathbf{1}_A(x) := \begin{cases} 1 & \text{if } x \in A\,, \\ 0 & \text{if } x \notin A\,. \end{cases}$

[3] Dual-encoders (also known as bi-encoders) usually refer to a siamese representation-based model which optimizes embeddings from pre-trained transformers.

[4] Cross-encoders (also known as mono-encoders) usually refer to pre-trained transformers used to classify query-document pairs in terms of the binary relevance.



(a) Representation-based (b) Interaction-based

Fig. 2.1 Representation and interaction-based scoring. (a) Representation-based models focus on optimizing text representations, while the score is usually computed as a simple similarity operation. (b) Interaction-based models focus on studying interactions between queries and documents that can bring relevance signals that contribute to the final score while the representations remain static.

2.2.1 Neural Ranking

Neural retrieval models (usually referred to as Neural IR or NIR) have been widely studied in the past few years, leading to considerable improvements in retrieval effectiveness [37]. NIR models are mostly oriented to re-rank a small subset of documents ranked by more efficient models, such as sparse retrieval (SR) models. Notable models of this category include DRMM [36], Duet [67], KNRM [18], and BERT [21] used for re-ranking [17, 63]. These models fall into the pipeline-based architectures, which comprehend settings with a stack of multiple models, refining the outcome as it progresses in the pipeline; thus, the performance of these models is bounded by the quality of the early-stage retrieval models. The summarized score for a given query from such models can be formulated as follows:

$$f(q,d) \triangleq \alpha \cdot \mathrm{SR}(q,d) + (1-\alpha) \cdot f_{\mathrm{NN}}(q,d), \tag{2.4}$$

where $\alpha \in [0 \cdots 1]$, $\mathrm{SR}(q,d)$ is usually set as the normalized BM25 scores and $f_{NN}(q,d)$ is a neural ranker, e.g., BERT re-ranker (cross-encoder):

$$
\begin{aligned}
f_{\text{NN}}(q,d) &\triangleq f_{\text{BERT}}\left([\text{CLS}] \oplus q \oplus [\text{SEP}] \oplus d \oplus [\text{SEP}]\right), \\
&= \text{Softmax}\left(h_{[\text{CLS}]}W + b\right)_1,
\end{aligned}
\tag{2.5}
$$

where [CLS] and [SEP] are BERT reserved tokens[5], $\oplus$ denotes text concatenation, the representation of the [CLS] token $h_{[\text{CLS}]} \in \mathbb{R}^N$ is passed through a single-layer fully-connected NN, $W \in \mathbb{R}^{N\times 2}$ is a weight matrix and $b \in \mathbb{R}^2$ is a bias term. In this case, N is the model embedding dimension, and $\text{Softmax}(\cdot)_i$ denotes the i-th element of the Softmax output. BERT follows the Transformers architecture using stacked self-attention and point-wise, fully connected layers [21]. Analogous to interaction-based models (see Figure 2.1b), BERT's all-to-all attention, at each Transformer layer, captures interactions between and within terms from the query and the document. We illustrate this approach at a high level in Figure 2.2.

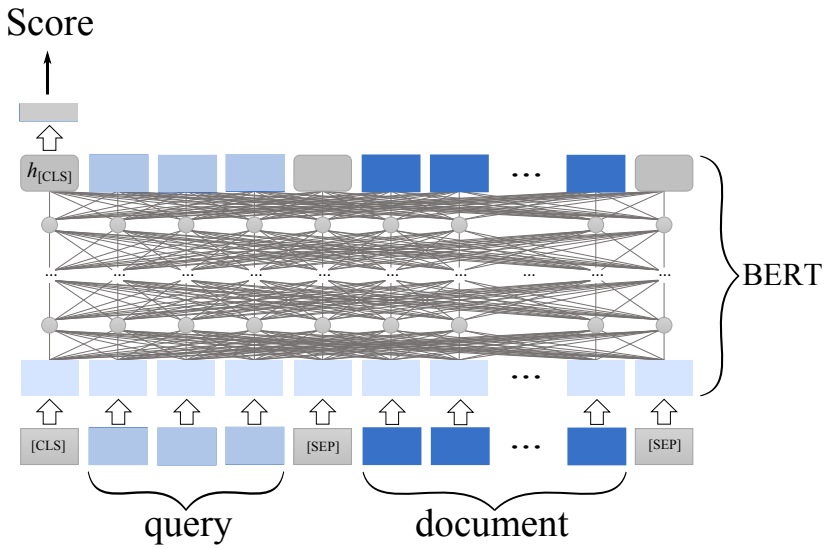


Fig. 2.2 Bert-based scoring. Analogous to interaction-based models, BERT's attention layers capture interactions between text inputs. The Last hidden state of the [CLS] token is used as input for a traditional classifier to perform relevance classification.

2.2.2 Dense Retrieval

Models involving nearest neighbor search are investigated in the lookup for alternative approaches to the traditional SR models. Aligned with representation-based models and dual-encoders, dense retrieval models (usually referred to as dense passage retrieval or DPR) learn dense vector representations as in equation (2.1), optimized

[5]CLS stands for classification and SEP stands for separator.

upon retrieval objectives, enabling dense indexing and vector search, thus, following the formulation (2.2). Models within this category have been widely studied and cited in the literature as part of the state-of-the-art in many scenarios [48, 115, 81, 46]. However, it is uncertain to what extent this paradigm can replace the established and robust SR models that use inverted indexes for efficient retrieval [39].

2.2.3 Learned Sparse Retrieval

Learned sparse retrieval (LSR) is another category of retrieval methods that use encoders to project queries and documents to sparse vectors with the size of the vocabulary for creating inverted indexes [30, 118, 64]. The way document scoring is performed follows the traditional formulation of SR methods, such as in equation (2.3). As BM25 uses TF and IDF components, LSR models use term weights predicted by neural models optimized upon retrieval objectives (see Figure 2.3 for an overview of an LSR scoring). As different models based on Transformers, LSR has the limitation of reduced input size.

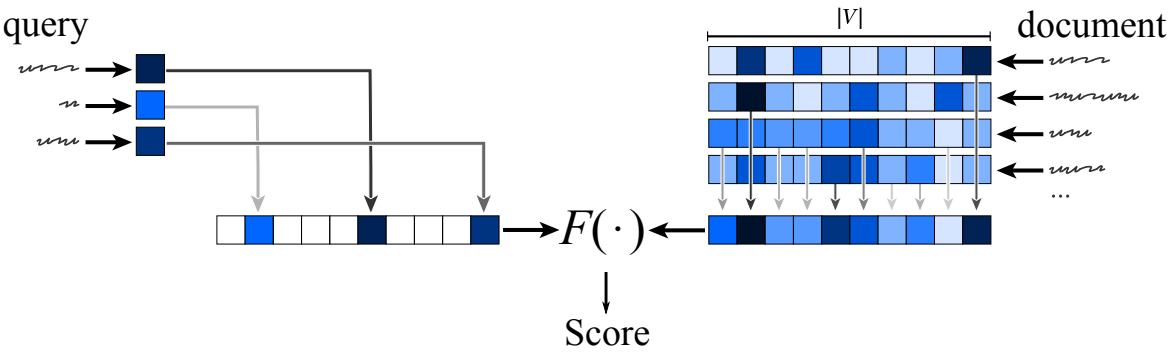


Fig. 2.3 Learned sparse scoring. An example of a model that replaces TF values with predicted weights for indexing documents.

2.2.4 Generative Retrieval

With the large language models (LLMs) hype, generative retrieval (GenIR) has emerged as the most recent retrieval paradigm [104, 113]. GenIR focuses on model-based approaches, alternatively to pipeline-based approaches. In this case, indexing and retrieval are performed with a single encoder-decoder transformer model. During training time, the model learns to index documents by memorizing the association between documents and identifiers; during inference time, it is supposed to generate relevant document identifiers in response to a query. This relatively recent family of

models has shown promising results; however, it faces challenges with updating the index and scaling to large collections [77]. For this family of models, the document ranking scores are computed as

$$f(q,d) \triangleq f(q,D) = \text{Softmax}\left(W_D^\top h_q\right), \tag{2.6}$$

where $W_D = [h_{d_1}, h_{d_2}, \cdots, h_{d_{|D|}}]^\top$ and h_{q/d_k} are the output vectors, i.e., last hidden state, of the encoder-decoder for the q or a document d_k.[6]

2.3 Neural-enhanced Retrieval Models in Domain-specific tasks

Task context is another feature that can affect relevance estimation. It has been brought up in many context taxonomies and frameworks for designing search systems [66]. We focus on the concept related to search tasks without disregarding a broader definition related to work tasks that also contribute to search behavior overall. There are different criteria to consider depending on work tasks. Searchers in different domains usually have specific patterns of search behavior and information needs, e.g., in the academic patent and legal domains, users are usually required to find very specific or gather information for the topic being searched. Users in the medical domain might have complex and unusual medical cases where they need to find specific and potentially rare information that supports decision-making.

Nested with the work task, the domain is a criterion that seems to play an important role in search scenarios. However, most of the research on IR is focused on developing ad-hoc open-domain models. There is a rising interest in domain-specific search tasks in the main IR venues and with different initiatives. For instance, TREC Clinical Trials Track[7], the CLEF SimpleText Track[8], focused on medical and academic IR problems, respectively.

CTR, specifically, aims to search for CT documents for recruiting candidates for a trial. SDR, with mainly an informational intent, aims to gather all related information about certain topic being searched.

[6] Notice that the notation here uses the subscript k because the model considers the whole set of documents D instead of individual documents.

[7] http://www.trec-cds.org/2022.html

[8] http://simpletext-project.com/2023/clef/

According to results reported for these tasks by the different initiatives, several approaches using Transformer-based architectures and pre-trained models, such as BERT, have achieved state-of-the-art effectiveness in some biomedical information processing applications.

In CT retrieval, there have been multiple attempts to use BERT embeddings in both dual-encoders and cross-encoders retrieval setups with different pre-trained models such as BioBERT or ClinicalBERT [43, 91, 90]. These results correspond to implementations of methods applied to traditional ad-hoc retrieval tasks and have not outperformed multiple experiments under traditional retrieval models [85, 86]. On the contrary, Pradeep et al. [78] applied a pipeline-based neural ranking system for the CTs matching problem, relying on T5-based models, currently with state-of-the-art results in multiple retrieval tasks, including CT.

NIR settings have also been exploited in the context of academic search, approaches well known for their effectiveness in ad-hoc scenarios, and then fine-tuned on domain-specific corpus [25, 26].

These set of Domain-specific tasks have relevant nuances that might come from the broader context of the task, as information needs or intents could have different motivations to open domain retrieval or might come from the context of the search tasks, which gives the opportunity to exploit domain-based information sources (e.g., domain knowledge) or domain adaptation. In this thesis, we aim to explore those contextual factors, how they can be incorporated better into the retrieval scenario, or how they can be taken into account for the adaptation of suitable Neural-enhanced retrieval models relative to each task.

Part II

Effective CT Search: An Application of NIR on the Medical Domain

3

Curriculum Learning for Neural Information Retrieval in Clinical Trials Retrieval

In the first part of this thesis, we study Domain-specific retrieval from the nuances of the CTR. In this chapter, we aim to answer *RQ1*: Given the clinical trials IR problem, can we optimize a model upon different signals of relevance?, and *RQ2*: Given the multitask nature of CTR, can we design a task-oriented model under a reasonable effectiveness-efficiency trade-off?

3.1 Introduction

Clinical trials are crucial to the progress of medical science, specifically in developing new treatments, drugs, or medical devices [79]. Awareness and access to these studies are still challenging both for patients and physicians, making the recruitment of patients a significant obstacle to the success of trials [70, 79].

Even if a sufficient number of participants is found, the recruitment process requires screening the patients for eligibility, which is a labor-intensive task [24]. Automated identification of eligible participants not only promises great benefits for translational

science [70] but also aids patients by allowing them to be included in specific trials [52].

In recent years, several initiatives have been proposed to build automatic systems for matching patients to CTs [99, 52, 87, 85]. The task has been defined as an IR problem under the patient-to-trials evaluation paradigm [88]. Here, the query is constituted by patient-related information, either in the form of electronic health records (EHRs) or ad-hoc queries, and the documents are the CTs [52].

This retrieval task involves the semantic complexity of matching the patients' information with heterogenous, multi-fielded CT documents [90]. The existing approaches have revealed a significant lack of an efficiency-effectiveness trade-off to date. While pipeline-based models showcase promising performance, the substantial model sizes required to achieve competitive results raise concerns regarding costly deployment and limitations on reproducibility.

This chapter presents a system for CT matching that uses a pipeline-based model with a Transformer network with a moderate size. We define a training strategy for re-ranking trials using a pre-trained language model in a two-step schema that leverages the structure of CT by considering not only the traditional topical relevance objective but also the eligibility criteria. Taking the result from our first stage retrieval process, we then match the patient's information with descriptive sections of the trials for re-ranking based on topical relevance. Later, we further train this model by matching patient data with trial eligibility criteria in an attempt to discriminate documents as eligible or excluded.

We break down *RQ1* into two research sub-questions. First, we ask how we can effectively split the CTR into multiple tasks (*RQ1.1*). Second, we consider whether we can cast CTR models in a curriculum learning framework (*RQ1.2*). We then break down the *RQ2* into two research sub-questions, where first, we examine if we can improve the performance by using cascade models instead of independent purpose models (*RQ2.1*). Second, what's the effectiveness-efficiency trade-off compared with the state-of-the-art models (*RQ2.2*).

We perform experiments using TREC CT track 2021 and 2022 collections and show that our method improves finding relevant trials under a reasonable efficiency-effectiveness trade-off.

3.2 Related Work

In this section, we discuss the background of the task we intend to tackle and related approaches proposed in the same context.

The TREC CTs track focuses on the task of matching single patients to CT documents. Related tasks to CT matching, which also involve handling CT documents, are, e.g., cohort-based retrieval [53], trial-to-trial retrieval [114] and other healthcare-related TREC tracks.

Traditional SR models may exhibit low performance in CTR as both the patient's description and the CTs contain many irrelevant terms, thereby introducing noise. Additionally, SR may not be suitable for CTR since both topics and documents contain negated key terms (e.g., the exclusion criteria), which are essential for deciding eligibility [34, 102]. Besides, the sections of queries and documents may have different importance because of their time dependency (i.e., past or present conditions) and because they can refer to either patients or patients' family medical history.

Using a simple SR model to tackle CTR may pose difficulties as both the patient's description and the CTs contain many irrelevant terms, thereby introducing noise. Moreover, both can contain negated key terms (for instance, the exclusion criteria), which are essential for deciding eligibility but may not be trivial even when using SR or NN-based models [34, 102]. Additionally, the sections of queries and documents may have different importance because of their time dependency (i.e., past or present conditions) and because they can refer to either patients or patients' family medical history.

Many systems reported in the TREC CT used variants of BM25 or the Divergence from Randomness (DFR) model [4] that has demonstrated potential in the biomedical IR field. Leveling [58] annotated a corpus with terms from medical dictionaries and with negations for retrieving trials for the TREC Precision Medicine track. In previous works, the CT matching task was also approached by using various lexical and neural models.

Several approaches using Transformer-based architectures and pre-trained models, such as BERT [21], have achieved state-of-the-art effectiveness in some biomedical information processing applications. In CT retrieval, there have been multiple attempts

to use BERT embeddings in both dual-encoder and cross-encoder retrieval setups with different pre-trained models such as BioBERT or ClinicalBERT [43, 91, 90].

These results correspond to implementations of methods applied to traditional ad-hoc retrieval tasks and have not outperformed multiple experiments under traditional SR models [85, 86]. On the other hand, Pradeep et al. [78] proposed a multi-stage neural ranking system for the CTs matching problem, relying on T5-based models, currently with state-of-the-art results in multiple retrieval tasks, including CT.

According to the findings presented in TREC CT 2021 [85], T5-based models currently outperform smaller Transformers models in CT retrieval. In this chapter, we propose an effective training strategy that takes into account various aspects of clinical trial retrieval, including topical relevance and eligibility criteria, as separate learning objectives. We evaluate its quality both on the general, pre-trained BERT model as well as biomedical domain-focused versions. Our approach results in a strong competitor to T5-based models with a much simpler architecture, as demonstrated by the official results reported in TREC CT 2022 [86]. Specifically, our model performs second-best overall, outperformed only by the model proposed by Pradeep et al. [78]. These findings suggest that BERT-based models with our proposed training strategy can provide a viable alternative to T5-based models in CTR.

3.3 Problem Statement

In the TREC CT track context, patient-related information is written as free text, whereas the document collection consists of a snapshot of the ClinicalTrials.gov database.[1] Each clinical trial contains multiple fields, including two titles (brief and official), conditions, summary, detailed description, and eligibility criteria. The content of these fields can range from structured (e.g., gender and age of eligible patients) through semi-structured (e.g., eligibility criteria section) to unstructured (e.g., description and summary). The eligibility criteria field contains inclusion and exclusion criteria, a core aspect of the CT matching task. Trials were judged using a graded relevance scale of three points: 0 if the patient is not relevant to the CT, 1 if the patient is topically relevant but excluded based on the eligibility criteria, and 2 when the patient fulfills the eligibility criteria.

[1] https://clinicaltrials.gov

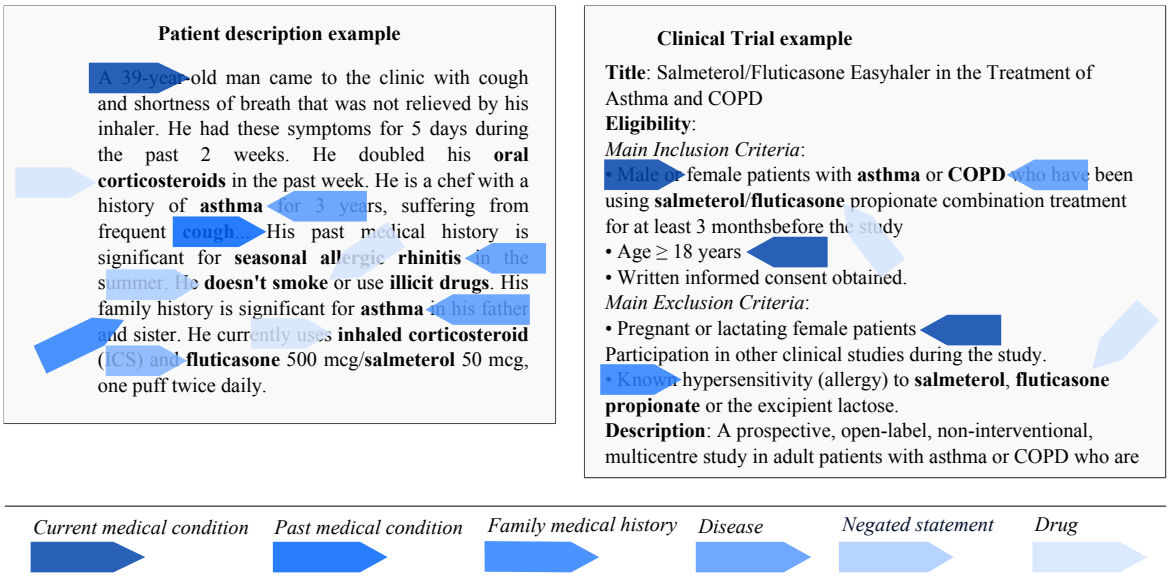


Fig. 3.1 Patient to CT matching problem: (left) Patient record: free text similar to an EHR with demographics and medical history of the patient ; (right) CT example: Structured multi-fielded document from ClinicalTrials.gov database with trial description in different levels and criteria for eligibility.

Formally, given a patient q, and a collection of trials $D = \{d_1, d_2, \cdots, d_{|D|}\}$, the task is to identify and rank the trials $D' = \{d'_l \mid d'_l \in D \wedge \text{eligible}(q, d'_l)\}$, where $\text{eligible}(q, d'_l)$ denotes that d'_l should be a trial for which the given patient q is eligible. A CT document d can be described by the set of CT components $\{d_{\text{description}}, d_{\text{conditions}}, d_{\text{official title}}, d_{\text{brief title}}, d_{\text{brief summary}}, d_{\text{detailed description}}, d_{\text{criteria}}\}$, whereas a patient q is represented by an unstructured set of descriptors, e.g., $\{q_{\text{age}}, q_{\text{gender}}, q_{\text{description}}\}$.

3.4 Curriculum Learning for NIR in CT

This section presents our approach to CTR, specifically how we optimize a model upon different signals of relevance and how we design a task-oriented model for CTR.

In identifying the ranked set D' of eligible trials, we look at two aspects: first, that the trial and the patient are actually related, which is aligned with the notion of topical relevance in IR, and second, that the patient meets the eligibility criteria. This follows the definition of the TREC CT track and the way results are evaluated.

Based on this idea, we split the task into two: retrieval and eligibility classification. Then, taking advantage of the structure of the documents, we define a training schema with two objectives. For that, we follow the notion of curriculum learning: the approach aimed at decomposing complex knowledge and designing a curriculum for learning concepts from simple to hard [117].

It follows from the way the screening of trials is performed, i.e. finding related trials/patients and then assessing eligibility criteria, eligibility classification is a harder task than retrieval. We also consider this heuristic in our approach.

With the heuristic that the CT retrieval task can be decomposed into two sub-tasks, first, we set the retrieval objective, which simply relies on discriminating topical relevance (both eligible and excluded documents are relevant). Second, we set the objective of eligibility classification (only eligible documents are relevant).

We use the pre-trained language model BERT [21] with the standard cross-encoder NIR approach. For ranking, a linear combination layer is stacked atop the Transformer network, whose parameters are tuned with a ranking loss function. We use a pairwise loss function and train the model for re-ranking outputs of the SR model.

Thus, the model is trained for these two objectives consecutively, such that there are two instances of the same model that we optimize first upon the retrieval objective, with the loss given by:

$$\mathscr{L}(q, d_{\mathrm{T}}^{+}, d_{\mathrm{T}}^{-}; W) = \max(0, 1 - s(q, d_{\mathrm{T}}^{+}; W) + s(q, d_{\mathrm{T}}^{-}; W)), \tag{3.1}$$

and then upon the criteria classification objective, with the loss:

$$\mathscr{L}(q, d_{\mathrm{C}}^{+}, d_{\mathrm{C}}^{-}; W) = \max(0, 1 - s(q, d_{\mathrm{C}}^{+}; W) + s(q, d_{\mathrm{C}}^{-}; W)), \tag{3.2}$$

where $d_{\mathrm{T}} = [\, d_{\text{description}}, d_{\text{conditions}}, d_{\text{official title}}, d_{\text{brief title}}, d_{\text{brief summary}}, d_{\text{detailed description}}]$, $d_{\mathrm{C}} = [d_{\text{criteria}}]$ the superscript $^{+/-}$ denotes the condition of relevant or non-relevant items to the query q, W represents the model's parameters with the final linear layer, and s is the predicted score.

As shown in Figure 3.2, we match patient information with descriptive sections of the trials for re-ranking based on topical relevance ((d^{+})- corresponds to the descriptive sections of relevant trials). We consider the eligibility classification a harder task

than traditional retrieval, and we hypothesize that a model for criteria classification could benefit from the knowledge that it already has from the previous on retrieval. We further train this model by matching patients' information with criteria in an attempt to discriminate documents as eligible or excluded ((d^+)- corresponds to trials categorized as eligible, and (d^-)- corresponds to trials categorized as relevant but discarded).

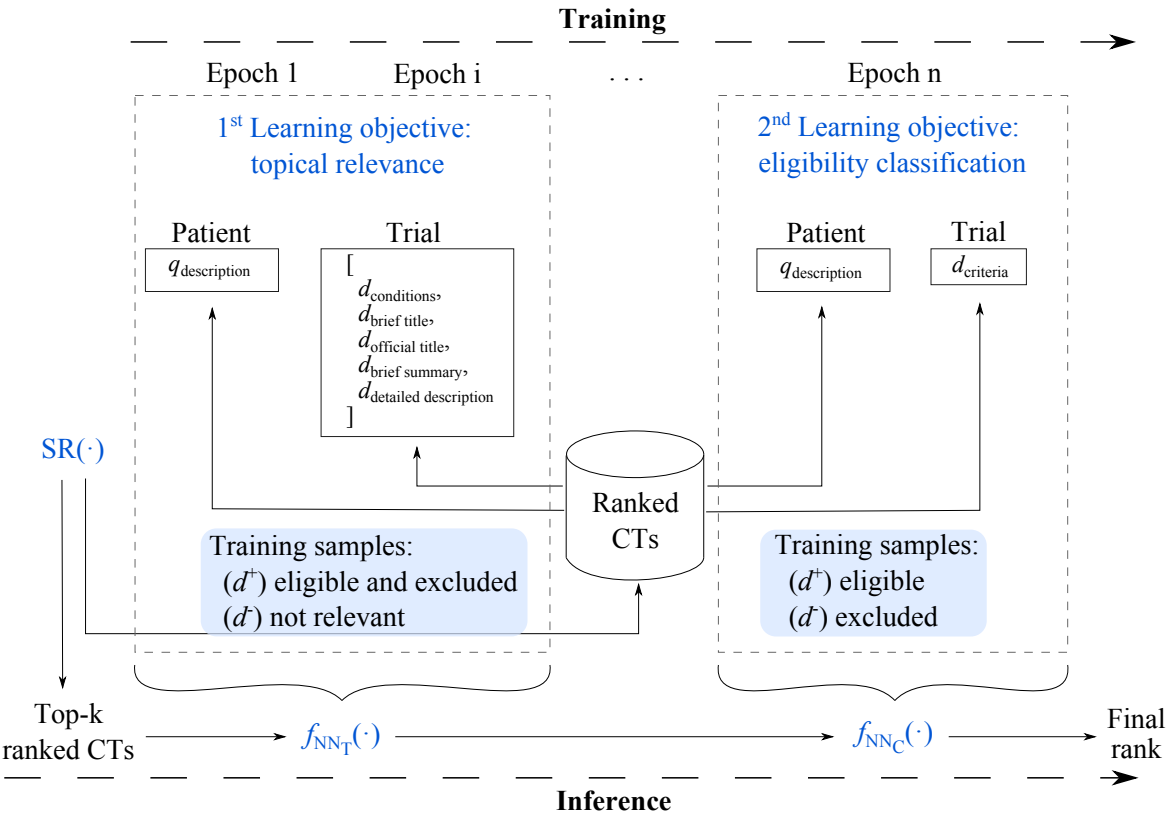


Fig. 3.2 Neural re-ranking setup TCRR. Training system for CT retrieval based on two objectives: topical relevance and eligibility. The model is first trained for discriminating relevant and irrelevant trials given a patient and then for classifying their eligibility for the relevant trial.

This process results in two different models. During inference time, we follow a similar schema: we take the SR rank and re-rank twice the top-k ranked trials using the two resulting models, respectively, such that our ranking function then follows the pipeline:

$$\begin{aligned} f(q,d) &\triangleq \beta \cdot (\alpha \cdot \mathrm{SR}(q,d) + (1-\alpha) \cdot f_{\mathrm{NN_T}}(q,d)) + (1-\beta) \cdot f_{\mathrm{NN_C}}(q,d)) \\ &= \gamma \cdot \mathrm{SR}(q,d) + (\beta-\gamma) \cdot f_{\mathrm{NN_T}}(q,d) + (1-\beta) \cdot f_{\mathrm{NN_C}}(q,d), \end{aligned} \tag{3.3}$$

where $\gamma = \alpha \cdot \beta$, α and $\beta \in [0 \cdots 1]$, $\text{SR}(q,d)$ is the normalized BM25 scores and $f_{\text{NN}_{\text{T/C}}}$ is the neural ranker. When referring to this re-ranking procedure, we refer to it as TCRR: *Topical and Criteria Re-Ranking*.

3.5 Experiments

This section details the experimental setting implemented for evaluating our approach. Specifically, we discuss the dataset, the evaluation setup, algorithm implementation details, and the baselines we compared our system with.

3.5.1 Dataset

The corpus released by TREC contains 375,580 clinical trials. The train data available (2021) consists of 75 topics (patient notes), while the test data available (2022) comprehends 50 topics. There are 35,832 relevance judgments in 2021 and 35,394 in 2022. More details of these datasets can be found in Section A.1 of A. Clinical trial documents released by TREC are in XML format and consist of several sections. In our experiments, we consider the following sections: brief title, official title, description, summary, conditions, and criteria.

For our experiments, we use the sets of topics as they were provided. For neural re-ranking, we present results on the test set of 2022. Additional splitting for training and development for neural models is described in Section 3.5.4.

3.5.2 Evaluation

We follow the evaluation procedure from the TREC Clinical Trials track, which is the standard evaluation procedure for ad-hoc retrieval tasks. As the relevance assessment is given in a graded relevance scale (eligible, excluded, or not relevant), the performance of the models is measured using normalized discounted cumulative gain (nDCG). We present results as reported by TREC, using nDCG@5 and nDCG@10, Precision (P@10), and Reciprocal Rank (RR). We treat unjudged documents as non-relevant, ensuring that our results are not biased towards models that retrieve a large number of unjudged documents. Furthermore, we focus on Precision as the primary metric for optimizing retrieval models. Our goal is to identify eligible trials, and Precision provides strict feedback to achieve this aim.

3.5.3 Baselines

As discussed in Section 3.4, we train two different models for our custom re-ranking: Topical and Criteria re-ranking. When used separately, we consider them as baselines:

TopicalRR: the model trained for re-ranking based on the topical objective is initialized with the weights of *bert-base-uncased*[2] (as well as other three domain-specific trained models: BioBERT[3], Clinical-BERT[4] Blue-BERT[5]).

CriteriaRR: the model trained for re-ranking based on the eligibility criteria classification objective is initialized with the weights of the TopicalRR. We further train this model.

Additionally, we consider the following two neural models as baselines:

TraditionalRR: the cross-encoder we use to compare our proposed training procedure with the traditional training, we train the model from the same checkpoint *bert-base-uncased*.

MonoBERT: one of the comparable models implemented from the TREC CT track. It is based on the cross-encoder architecture and trained on data drawn from the corpus in a weakly supervised fashion[6]

We also consider different BM25 implementations from [56] corresponding to enhanced query representations and document representations, specifically BM25_14 indexes all descriptive trial fields and the inclusion criteria, and BM25_14d does the same but enhances the query representations by adding current and family medical conditions to the query representation.

3.5.4 Implementation details

As a main lexical retrieval model, we use the BM25 [109] "out-of-the-box", i.e., without parameter optimization.

[2] https://huggingface.co/bert-base-uncased
[3] https://huggingface.co/seiya/oubiobert-base-uncased
[4] https://huggingface.co/Tsubasaz/clinical-pubmed-bert-base-512
[5] https://huggingface.co/bionlp/bluebert_pubmed_uncased_L-24_H-1024_A-16
[6] Notice that 'weakly supervised fashion' refers to a machine learning approach where the training data is labeled only partially or with noisy labels, as an alternative to supervised learning that uses gold standards with high-quality labeled data.

On the other hand, we use PyTorch Lightning [28] and Transformers[7] to implement the neural re-ranking pipeline. As discussed in Section 3.4, we train different models for re-ranking with different configurations (see Section 3.5.3). The TopicalRR, after splitting the datasets into train (60%), development (10%), and test (30%), is trained on 8192 samples from the training set per epoch divided into batches of 16 samples. We further train this model on 1024 samples with batches of 16 to get to the CriteriaRR. Samples for these two models were selected as described in Section 3.4 and as shown in Figure 3.2. We pick positive samples only present in BM25 rankings as well as hard negatives from ranked-irrelevant or unlabeled documents. We re-rank top-50 trials from the BM25 experiment.[8] As for the α and β parameters, we set them to .7, which was found by exhaustive search. Finally, to compare our proposed training procedure with the traditional training of a cross-encoder, we train the TraditionalRR from the same checkpoint *bert-base-uncased* on 2048 samples, where relevant documents are only those categorized as eligible.

All neural models are trained for ten epochs, with early stopping based on Precision. Our training was performed on an Nvidia Quadro RTX 8000 GPU.

3.6 Results

In this section, we present the results of our curriculum learning approach to CTR. We show comparisons with different baselines and models aimed at showing an ablation study. Supplementary results can be found in Appendix B.1 and Appendix B.2.

Table 3.1 shows the results of the re-ranking procedure discussed in Section 3.4. We used the different models for re-ranking the results of a BM25 rank. We report the evaluations on the 2022 data. Models were trained on the 2021 data. The result of the TCRR model corresponds to the official TREC CT 2022 evaluation [86].

As we hypothesize, in the context of CTs, the model benefits from the decomposition of the retrieval problem into two objectives, as it is experienced by TCRR (see Section 3.4), the model exposed to the two learning objectives and best performing. We also provide results for TopicalRR and CriteriaRR, independently, which are the models exposed only to the first (topical relevance) and second (eligibility classification) learning objectives. Additionally, we present results for the regular re-ranking setup TraditionalRR.

[7] https://github.com/huggingface/transformers

[8] We ran experiments changing the cutoff from 20 to 100 with a step of 10 to find 50 as the optimal cutoff.

Table 3.1 Neural re-ranking evaluation results on TREC test set. Underlined values indicate the highest score among general models. **Bold** values indicate the highest score achieved by our approach. †-marked models indicate that there is a significant improvement over the BM25 baseline using the Student's paired t-test with a 95% confidence level.

Model	nDCG@5	nDCG@10	P@10	RR
TREC median	—	0.392	0.258	0.411
TREC best	—	0.612	0.508	0.726
BM25_14	0.464	0.437	0.312	0.520
BM25_14d	0.502	0.460	0.328	0.521
TopicalRR	0.558	0.529	0.414	0.630
CriteriaRR	0.382	0.387	0.294	0.428
Fused_TC	0.559	0.548	0,438	0.645
TraditionalRR	0.453	0.437	0.364	0.508
MonoBERT	0.509	0.491	0.362	0.527
TCRR†	0.573	0.557	0.456	0.619
CTRR	0,562	0,545	0,426	0,628
$TCRR^{\dagger}_{Bio}$	0.627	**0.604**	**0.482**	0.672
$TCRR_{Clinical}$	0.425	0.423	0.358	0.492
$TCRR_{Blue}$	**0.631**	0.583	0.452	**0.691**

Furthermore, we present results of a dummy experiment fusing scores from the TopicalRR, CriteriaRR with the average (see Fused_TC in Table 3.1); and an experiment in which we change the order of objectives in our pipeline, then, the model is first trained for eligibility classification and then for topical discrimination (see CTRR in Table 3.1). These results are in accordance with our hypothesis about the curriculum design – The strategy that benefits the most from the performance of the model is TCRR.

For this set of experiments, we are mainly interested in the evaluation in terms of Precision since, in a real scenario, only eligible trials are considered. Given that, on average, other proposed systems perform poorly, as shown by the TREC CT median results [85, 86], precision (P@10) anywhere near 50% is regarded as a good result for this task. We report results also obtained by the best-performing model at TREC, corresponding to the system by Pradeep et al. [78]. We analyze results from the proposed approach and find a significant improvement between the performance of TCRR models (TCRR and $TCRR_{Bio}$) and BM25 at a 95% confidence level. On

average, this approach allows Bert-based models to gather more relevant documents than the selected baselines in the top 10.

We report results on different domain-specific pre-trained models we trained following our proposed approach. Again, we evaluated our best-performing model, TCRR_{Bio}, in terms of Precision and found the improvement statistically significant. For additional results on query analysis, see Section B.1 of Appendix B. Section B.2 of Appendix B shows results comparing TCRR with other baselines.

3.7 Discussion

RQ1.1 In this chapter, we revisit the pipeline-based model for patient-to-CT matching. We propose an adaptation of training a cross-encoder to the CT problem, taking advantage of the structured nature of the documents and the task considered. We find that the inclusion criteria section has a considerable impact on the retrieval score and exploit that independently from the traditional signal of relevance in our re-ranking setup.

RQ1.2 Our re-ranking formula, based on curriculum learning concerning eligibility, shows additional improvement for this task. It explicitly models the eligibility decisions instead of using only the topical relevance. This distinguishes our study from the previous works concerning clinical trial re-ranking [90].

RQ2.1 We show results for experiments on different configurations of our pipeline and compare our approach with different models previously used for the task. We focus on BERT-based models, which so far have not necessarily outperformed sparse models for the clinical trial matching task.

RQ2.2 Even though the results in Table 3.1 also show how changing the initial weights of the model can affect the overall performance (i.e., by choosing a domain-specific model like BioBERT), we show that the improvements of our proposed approach are not due to the selection of a domain-specific pre-trained model, which is the case of the TCRR. These results also provide an idea of which pre-trained model fits the task best. Overall, the TCRR initialized with BioBERT weights shows promising results, while ClinicalBERT weights were not the best choice in this scenario. Our results are comparable to the more expensive approach using the T5 architecture [78].

Although this work focused on CT retrieval, we believe the approach can also be applied to other IR tasks where first, they involve ranking documents based on topics,

and, in a second instance, the retrieval results are tailored by considering more specific criteria or constraints. One example of such a task is the selection of primary studies (citation screening) for the systematic literature reviews [55].

This study has several limitations, both related to the dataset and the models. The TREC CT collection usage implies that the patient descriptions are relatively short, i.e., EHR admission note-style documents. We acknowledge that our approaches could have problems handling longer sequences.

Additional limitations are related to the amount of data available for training and evaluating systems on the CT retrieval task. This issue, in our study, explicitly affects the curriculum learning scenario in the eligibility determination objective. It may limit the model in learning relevant patterns needed to scale to different clinical settings or patient populations.

Furthermore, the topics are written only in English. This does not concern CT, for which the ClinicalTrials.gov database is the leading international source. Nevertheless, multilingual medical retrieval may present challenges for both lexical and neural models, as the nuances and complexities of medical terminology can vary significantly across languages. Addressing these limitations and developing strategies for multilingual medical retrieval is an essential area for future research.

Part III

Enhancing the Retrievability of Domain-specific Documents with Neural Models

4

Broad Theme Classification

In the second part of this thesis, we move to the research theme of enhancing the retrievability of documents, especially scholarly documents, with neural models. In this Chapter, we aim to answer *RQ3* Can academic publications be effectively discriminated into broad themes when high-quality data is provided?

4.1 Introduction

With the recent demise of the widely used Microsoft Academic Graph (MAG) [100], the scholarly document processing community is facing a pressing need to replace MAG with an open-source community-supported service. In order to create a comprehensive scholarly graph, it is necessary to represent each paper as a node on the graph correctly. This requires condensing meta-information, such as authorship, research organizations, research themes, etc., of research papers to one node.

In general, classifying scholarly documents is important, whether for understanding scientific fields' dynamics or organizing scientific literature more effectively for retrieval purposes. So far, identifying research themes for a given scholarly document has been challenging due to the lack of large-quality labeled data. This made it difficult both to train high-performance classification models as well as to compare models' performance across studies.

To establish a benchmark for research theme classification, we present experiments and evaluation results with traditional machine learning models and compare them to a more sophisticated Transformer-based ensemble model. The data we used is based on a large human-annotated corpus of scholarly papers across 36 themes defined by the UK Research Excellence Framework, the largest overall assessment of university research outputs ever undertaken globally (the Research Excellence Framework - 2014)[1] [16]. We started with a labeled dataset containing publications and subjects to which they belong, which contains descriptions or abstracts, the first author, DOI, and year of publication.

Our ensemble model exploits all these textual fields for each scholarly document and maps these documents to CORE and Semantic Scholar [32] to gather further external information. Thus, the ensemble consists of a Transformer-based classifier used to produce multiple predictions for individual publications (split into multiple textual fields) that are aggregated to produce a single final prediction. When available, we aggregate predictions from titles, abstracts, references, citations, and related titles for every publication. Furthermore, we use abstracts, PDFs, and full texts available to identify argumentative zones [106] to use them as additional fields. We report on the results of using aggregation for different combinations of these predictions.

Given the quality of the data and the multiple sources available, we focus on the *RQ3*.

4.2 Related work

In this section, we describe the relevant background and related approaches to scholarly document classification.

In previous literature, classifying scholarly documents typically relies on textual features such as titles, author keywords, and abstracts, as well as the interrelationships between the documents (i.e., citations and co-authorship). Full texts are frequently not available, and processing a large amount of text can be computationally expensive.

A wide variety of classification features have been proposed at different levels of granularity, e.g., disciplines, themes, and subjects. A large proportion of classification methods rely on semantic similarity [112, 96, 92, 38, 11]. Others include approaches for clustering documents based on keyword co-occurrence [110, 49].

[1] https://ref.ac.uk/2014/

Further approaches leverage the relationship graph representation built from citations and co-authorship [103, 98, 41].

One promising but unexplored approach to theme classification is using information about argumentative zoning (AZ) [106]. AZ refers to the examination of the argumentative status of sentences in scientific articles and their assignment to specific argumentative zones. Its main goal is to collect sentences that belong to predefined zones, such as "claim" or "method". Annotated AZ corpora have been created by Teufel et al. [105, 106], Teufel and Moens [107], Teufel et al. [108] with approaches to AZ identification reported by Liu [60]. Given that for some cases in the dataset, full-text is available, we evaluate to what extent can the AZ signal support the classification of scholarly documents into research themes.

Classification models previously applied to this task include traditional machine learning models, such as k-Nearest Neighbours [111, 62], K-means [49] and Naïve Bayes [27]. These models have been reported to encounter performance challenges related to overly coarse classifications and low accuracy [19]. There are applications of deep NN models as well, such as convolutional NN [84, 19] and recurrent NN [96, 41]. More recent deep learning approaches take advantage of pre-trained language models [45, 38].

One of the common practices to evaluate approaches for classifying scientific text is to use classification systems from digital libraries [45, 33, 103, 35], such as the ACM Computing Classification System,[2] the Web of Science Categories[3] and Science-Metrix.[4] Other practices involve generating automatic annotations for scientific collections that can be completely synthetic [111] or curated by experts [92, 27, 19, 38, 74]. However, to date, there has been no established benchmark to evaluate these approaches. With access to a high-quality dataset used for the first time, we established the baseline experiments on evaluating research theme classification.

4.3 Ensemble Model for Theme Classification

This section depicts the approach we used to estimate probabilities of academic publications belonging to a specific theme and the heuristics we follow for classification. In general, we want to exploit all the information available for the scholarly documents

[2] ACM Computing Classification System
[3] Web of Science Categories
[4] Science-Metrix

that need to be classified. Academic publications are typically well-structured documents with multiple textual fields and metadata. We rely on open-access platforms to enrich the data with additional information (Section 4.3.2).

Currently, Transformer-based contextual language models like ELMo [76] or BERT [20] outperform most feature-based representation methods. We use a classifier based on contextual word embeddings to evaluate the utility of individual textual fields in the classification of academic publications.

4.3.1 Transformer-based Classifier

We rely on the pre-trained general language model BERT [20], which achieves outstanding performance on different NLP tasks through fine-tuning for the downstream tasks [2], in this case, multiclass classification.

It is to be noted that the BERT model was already referred to in Section 2.2.1 and in Section 3.4 and used in the architecture known as cross-encoder, which is focused on the task of next sentence prediction, we further explain the alternative way in which the model was used in this case next.

We allow all layers of BERT to be updated as we are learning the relevant context from the training data. A custom operation is added on top of the model, which takes the last hidden state tensor from the encoder and then passes it to a linear layer. At the end of the linear layer, we have a vector with a size equal to the number of classes, and each element corresponds to a category of the provided labels. Specifically, we use the following setting to build the model base:

Input layer. It builds the model's input sequence. The input sequence is segmented according to the WordPiece embeddings and the token vocabulary. The final input representations are then produced by adding each token's position embeddings, word embeddings, and segmentation embeddings.

BERT encoder. It consists of multiple Transformer blocks and multiple self-attention heads that take an input of a sequence of a limited number of tokens and output the representations of the sequence. The representation can be a specific hidden state vector or a time-step sequence of hidden state vectors.

Output layer. It consists of a simple linear layer with a Softmax classifier on top of the encoder for computing the conditional probability distributions over predefined categorical labels.

The cross-entropy loss is used to optimize the model with the Adam optimizer.

4.3.2 Data Enrichment

Taking advantage of the open-access libraries available for scientific publications, we search for complementary data for each example provided for the task. Specifically, we use the CORE [50] and the Semantic Scholar [5] APIs to map publication titles to the various fields available for each publication.

The original Theme classification dataset (A.2) includes mainly titles with metadata. Our goal with the enrichment is to collect more information related to the publication to match the themes better. After mapping the papers to results from the search using the APIs, we add a list of references and citations, full papers, abstracts, and PDFs for the cases when they are available. Moreover, we searched for five recommended papers using the title of every publication using the CORE API.

We believe that regardless of the performance of the classification model if there is enough evidence for a publication to belong to a specific theme, we should be able to classify it with enough certainty. For instance, given a publication title, which can be ambiguous, we hypothesize that considering the multiple references or citations leads to disambiguation and deciding effectively to which theme this publication should belong. The list of references or citations can be classified the same way as single inputs, and the classification result can consider the multiple corresponding outputs for the final decision.

Since this data is not guaranteed to be available for all the original samples, we exploit all available sections, including the full text and PDFs. However, since processing such an amount of text is expensive, we use AZ [106]. Here, we define four zones that cover the main components of scientific articles, namely: *Claim*, *Method*, *Result*, and *Conclusion*.

In order to extract sentences that cover the four zones from the available PDF scientific articles, we follow an approach similar to a previously proposed approach by El-Ebshihy et al. [23], which generates an article summary by expanding the article abstract. To sum up, the sentence selection and labeling with zones process goes as follows: (1) we convert the PDF papers to an XML format using the GROBID PDF parser [61], which identifies the paragraphs of the article, (2) the paragraphs are fed

into a Solr[5] index, (3) the sentences in the article's abstract are passed as queries to the Solr index in order to find the top most similar paragraphs to the abstract sentences, (4) sentences of the retrieved paragraphs, as well as the sentences of the abstract, are labeled to zones using a pre-trained BERT model based on the approach proposed by Accuosto et al. [1], and (5) we use the labeled sentences to extend our training data with four extra text fields that represent the *Claim*, the *Method*, the *Result* and the *Conclusion* — we refer to these extra fields as Argumentative Zones. In case we cannot find the PDF source of the article, we use the article abstract, if found, to generate these fields.

4.3.3 Extending Labels to Enriched Data

During training, the model takes text examples and the associated labels. Since examples for this task are academic publications, and we want to use different sections independently, we rebuild the dataset considering each section as a single sample but associated with the same publication, and we use the same label for all samples of the same publication.

In this way, we end up with an extended version of the initial dataset, in which new samples are created for titles, abstracts, citations, references, and recommendations.

4.3.4 Aggregating Predictions from Enriched Data

During inference time, we compute multiple predictions associated with the same publication. These predictions can either agree or disagree, so we formulate the final prediction as the aggregation of the different predictions. Figure 4.1 illustrates the prediction procedure used to obtain the final theme prediction for a publication in which various sections are evaluated as independent samples with the classifier. Section 4.4 describes how this aggregation is parameterized for the experiments.

4.4 Experiments

This section describes the implementation details of our study, including evaluation metrics and baseline models. As for the benchmark used for evaluation, we dedicate Appendix A.2 to it.

[5]https://lucene.apache.org/solr/

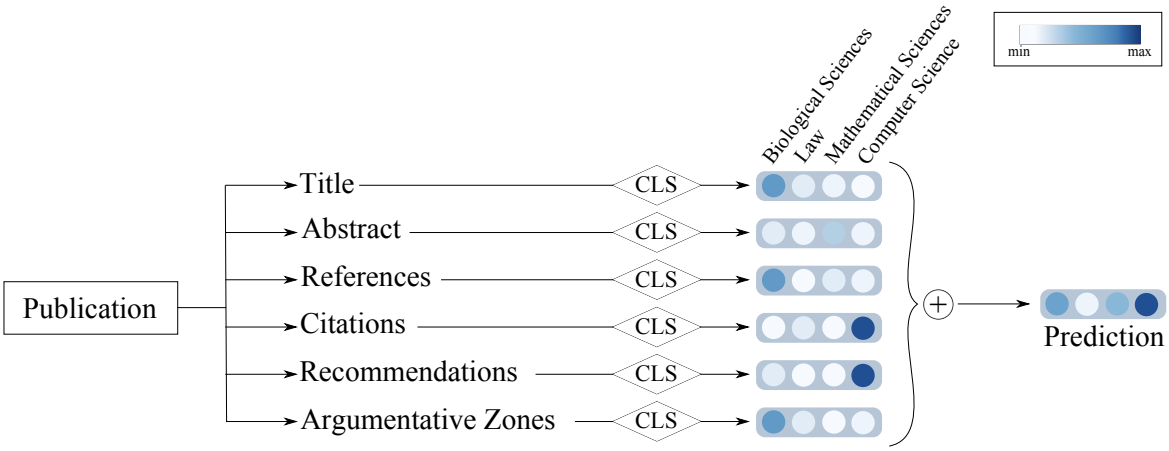


Fig. 4.1 Ensemble for research theme classification. CLS stands for classifier. For a publication, various sections are evaluated as independent samples with the classifier; the final prediction is achieved by aggregating sections' predictions.

4.4.1 Training Settings

Given the labeled training samples, we train the model using two different sets. The first training set consists of the list of titles, while the second takes both titles and available abstracts. We argue that although more information can be available per publication, the labels provided match only titles and abstracts, and further assumptions can hurt the model's performance. However, we define an additional training set under our data enrichment procedure. We refer to the first model as BERT_T and to the second one as BERT_{T+A}.

We train the model for ten epochs, with early stopping based on the performance measured using the evaluation metric (see Section 4.4.3) and patience of 3 epochs. The training samples are picked randomly, searching for a uniform distribution over the classes per batch. To prevent overfitting in the case of unbalanced batches, we use the weighted cross-entropy loss and assign the weights dynamically according to the result of the random selection of samples in the batch. We use 16,384 samples from the training set per epoch divided into batches of 64 samples and train the models on an Nvidia Quadro RTX 8000 GPU.

4.4.2 Prediction Settings

As well as the training strategy, we evaluate the utility of having multiple predictions per publication in the test set compared to a single prediction. To do so, we prepare

different evaluation sets, following the same training set schema. Thus, we evaluate the model using only titles, then using titles and abstracts, and finally, using the set created under our data enrichment procedure.

Since we have to produce a single prediction per publication, and the sets are not uniform, in the sense that certain publications may not have extra fields (see Table A.2, for instance, abstracts are available for only 32% of publications), we parameterize the prediction aggregation based on the different sets of fields. We consider the aggregation to be a weighted sum. The motivation for selecting a weighted sum, instead of just summing up the outputs, is that we can introduce offsetting through the weights. Thus, we give an advantage to the labeled fields in the original dataset over the extended data.

For our experiments, in the case of the set with titles and abstracts, we use uniform weighting. In the case of the extended set, we assign weights such that 0.5 is distributed uniformly between title and abstract, and 0.5 is uniformly distributed between all the additional fields available per publication. This setting is compared experimentally to a uniform weighting across all the fields.

4.4.3 Evaluation metrics

The evaluation metric used for evaluating classification results is micro F1-Score. The F1 score, commonly used in machine learning, measures accuracy using precision, and recall.

The F1 metric weighs recall and precision equally, and a good classification algorithm will maximize both precision and recall simultaneously. Thus, moderately good performance on both will be favored over extremely good performance on one and poor performance on the other.

4.4.4 Baseline Models

We implement several baseline models for comparison to the ensemble described in Section 4.3:

K-nearest neighbours classifier with TF-IDF representation

Logistic Regression classifier with TF-IDF representation

Naïve Bayes classifier with TF-IDF representation

Support Vector Machine classifier with TF-IDF representation

fastText classifier [44] with word vectors pretrained on wikipedia[6]

We also present scores using two dummy classifiers: selecting the most frequent category and sampling from a multinomial distribution parameterized by prior probabilities. All classifiers except for fastText are implemented using scikit-learn [75].

4.5 Results

This section presents the results and comparisons for our approach. We offer results on a validation set we sample from the training set as well as on the original test set.

4.5.1 Validation Results

Given the provided training data, we create balanced splits such that 60% is used for train, 10% for early stopping, and 30% for validation. All the sets are enriched following the process described earlier. Table 4.1 shows some preliminary results for experiments we performed to select the model and the training setup. We compare the two different BERT models with traditional models. The performance of the model trained using titles and abstracts is slightly better, and we use it for further experiments.

Furthermore, we evaluated the utility of enriching the dataset by comparing predictions from titles only with aggregated predictions using titles and additional available fields. Table 4.2 shows that adding information improves the classification for all three experiments. Notice that the experiments are not comparable to each other because the dataset samples are different. Subsamples are selected such that corresponding sections are available for all documents.

Table 4.3 shows the results obtained for the validation set using different variants of the ensemble. In general, we can improve the classification performance while adding more data, although the difference between the experiments is small. The best score reached is 0.526, using titles, abstracts, citations, references, and argumentative zones.

For the best configuration, Figure 4.2 shows the confusion matrix for a sample of classes (see section B.3 of Appendix B for a complete picture of the results). It should

[6]https://dl.fbaipublicfiles.com/fasttext

Table 4.1 Micro F1-score results form the comparison using different input features for prediction. BERT_T stands for BERT model trained on titles only, BERT_{T+A} means model trained on both titles and abstracts.

Model name	Titles	Titles and abstracts
Dummy: most frequent	— 0.095 —	
Dummy: stratified random	— 0.048 —	
K-nearest Neighbours	0.132	0.468
Logistic Regression	0.457	0.498
Naïve Bayes	0.460	0.493
Support Vector Machine	0.474	0.506
fastText	0.454	0.473
BERT_T	0.498	–
BERT_{T+A}	<u>0.500</u>	<u>0.512</u>

Table 4.2 Three experiments testing the utility of individual sections on BERT_{T+A}. The augmentation is evaluated by independent sections combined with titles. Samples are selected such that corresponding sections are available for all documents.

Sections	Sample size	F1-score (title)	F1-score (all sections)
Title + Abs.	31.3%	0.503	0.539
Title + Cit. + Refs	25.4%	0.492	0.541
Title + AZ	1.6%	0.548	0.552

be noted that for Clinical Medicine, most examples where the model's prediction is incorrect are classified as Allied Health Professions, Nursing and Pharmacy, and Biological Sciences. Similar behavior can be observed with related fields of study. Further analysis must be done to evaluate overlapping between disciplines.

Table 4.3 Validation results using different fields for BERT_{T+A}. The experiments vary in the prediction and aggregation settings. The aggregations we use are simply weighted sums with uniform weights and assigned arbitrarily according to Section 4.4.2.

Title	Abs.	Cit.	Refs	AZ	Recs.	F1
×	–	–	–	–	–	0.500
×	×	–	–	–	–	0.512
×	×	×	×	–	–	0.523
×	×	×	×	×	–	<u>0.526</u>
×	×	×	×	×	×	0.525

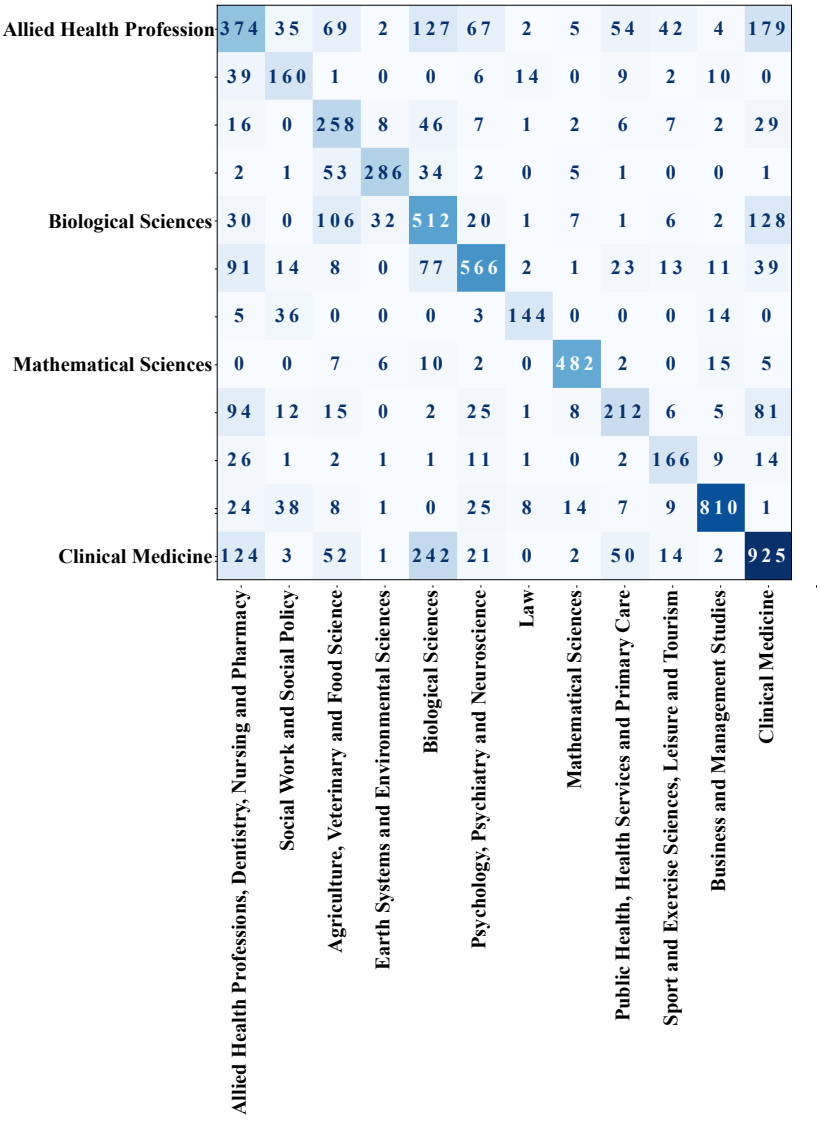


Fig. 4.2 Confusion Matrix for validation results for a sample of classes. For Clinical Medicine, most examples where the model's incorrect prediction are classified as Allied Health Professions, Nursing and Pharmacy, and Biological Sciences. Similar behavior can be observed with related fields of study; see Appendix B for an extended analysis.

4.5.2 Test Results

In this section, we show the results for the test set (see Table 4.4). In general, our approach has a positive impact, considering we could not get additional information for all the items in the original dataset.

Table 4.4 Test results with different experimental (Run) settings. The experiments vary in the training (T), prediction (P), and aggregation (Agg.) settings. The aggregations we use are simply weighted sums with uniform weights (U) and compensation weights (C) assigned according to section 4.4.2.

Experiment	Title	Abs.	Cit.	Refs	AZ	Recs.	Agg.	F1
1	T+P	T+P	–	–	–	–	U	0.569
2	T+P	T+P	P	P	P	–	C	0.575
3	T+P	T+P	P	P	P	P	C	0.571
4	T+P	T+P	P	P	P	P	U	0.577
5	T+P	T+P	T+P	T+P	–	T+P	C	0.556

In this set of experiments, we evaluate a different aggregation setting: uniform weighting through all the fields (experiment 4), and the result is the best score for the set of runs. Furthermore, we also evaluate an additional model trained with all the fields available (experiment 5), and we see no improvements.

4.6 Discussion

In this chapter, we use for the first time the new gold-standard human-annotated dataset of over 60k papers complete with paper metadata, research themes, and additional textual information, including the papers' abstract and full-text where available (see Appendix A.2). To our knowledge, our work was the first to utilize REF research evaluation for the purposes of building machine learning models for theme classification and highlighted the significant potential of this dataset for developing state-of-the-art models.

We use a high-quality dataset to establish a new benchmark for research theme classification, testing a range of classic machine-learning models under the same laboratory conditions. Unsurprisingly, our results confirm that models trained with both titles and abstracts as input features consistently achieve higher results than when using titles alone. These results hold for baseline models and our newly introduced

ensemble BERT model. While the results confirm that the BERT-based ensemble model outperforms traditional models, the performance of SVMs is only marginally worse.

Interestingly, using all available features for training (experiment 5 in Table 4.4) decreases the score compared to the model trained on titles and abstracts only. We hypothesize that a large proportion of false negatives can be attributed to noise introduced by reference sections within the full texts, especially for closely aligned domains. The confusion matrix (Figure B.2) shows that many of the incorrect classifications happened in closely related disciplines (Clinical Medicine / Biological Science, for example).

Regarding *RQ3*, this behavior is indicative of the difficulty of this task, mainly when presented with closely matched or overlapping domains. Indeed, one limitation of our approach may be the classification of each paper into a single research field. In real-world examples, a document could often be classified into multiple domains. Another limitation is that our ensemble model requires both title and abstract availability, which are necessary for the AZ approach, which we have seen contributes to the performance.

Assigning research themes to scholarly documents has wide-ranging applications. These include enhanced domain-specific search; for instance, search in Chemistry is a complex task due to the need to index chemical compounds and identify emerging research trends. Further, a significant problem with current bibliometric methodologies is accounting for cross-disciplinary differences in both publishing and citation practices. Identifying the research theme enables accounting for disciplinary differences by, for instance, calculating normalized citation counts.

5

Large-Scale Hierarchical Classification Analysis

In the previous chapter, we studied how to classify scholarly documents into a set of broad research themes. Advances in classification contribute to handling digital libraries. However, broad themes provide limited insights to establish how classification can contribute to the definition of relevance in the context of an IR model. Therefore, in this chapter, we aim to answer *RQ4* Can fine-grained themes be embedded into model parameters for annotating documents?

5.1 Introduction

Large-scale classification is a multi-label text classification problem that can include hierarchically related categories. Multiple real-life applications can be framed as hierarchical classification: labeling scientific publications with concepts from ontologies, annotating documents for legal document management systems, and assigning labels to medical records.

From task to task, it varies how relationships between categories need to be considered to some extent. This is due to different factors; one, for instance, is related to ambiguity between overlapping or closely related categories. It can lead to the

propagation of errors. Another example is related to the granularity of the hierarchy, which may affect how the texts' semantics interact with the categories' semantics.

In general, classification systems assign a set of labels to a given text snipped. The creation of such systems has been driven by the need to process increasingly large document collections. Novel approaches provide specialized understanding at each level of the hierarchies through stacks of deep learning architectures [54]. This shows a particular emphasis on hierarchy-aware models [14]. However, models focused on text semantics are still explored, especially with Transformers models, which also have gained prominence in this field [13].

In terms of training strategies, unsupervised approaches, and multi-task setups have been explored, demonstrating improved performance compared to alternative methods [93]. These advancements highlight the growing focus on leveraging hierarchical structures and semantic relationships to address the complexities of hierarchical text classification tasks.

This chapter presents an architecture leveraging the capabilities of large pre-trained Transformer models. It employs a seq2seq learning system [101] to directly map text to categories. Taking advantage of the recent advancements in generative LM, we investigate the possibility of implicitly learning hierarchy information with the model, resulting in an effective approach. By harnessing the context understanding of the architecture, we enable a robust integration of the categories' and texts' semantics into the text-to-category mapping process.

During inference time, the trained model takes as input a text snipped and outputs categories from a hierarchically arranged set of categories using beam search. As we show, this process can work consistently well on different classification tasks. In our experiments, it was tested on different types of hierarchies.

In this chapter, we focus on the *RQ4* and evaluate another approach to theme classification in a rather more fine-grained distribution of themes for scholarly documents and other domains.

5.2 Related Work

In this section, we describe recent methods used to tackle the large-scale classification task with hierarchically arranged categories.

The research on large-scale hierarchical text classification has been focused on increasing the precision of predictions at lower levels of the hierarchies [54, 93], understanding hierarchical relationships [80], semantic alignment [14], and error propagation [83].

Previously proposed solutions rely on contextual embeddings at different levels, mainly for understanding text semantics. In contrast, the hierarchical information is then processed through explicit hierarchical architectures [54], making predictions along hierarchical structures [80, 93], optimizing multiple objectives [80, 14] or tailoring the output space to models' label dependencies [14, 83].

Enforcing a hierarchical structure may not benefit some applications where documents do not naturally belong to only one hierarchical organization or have a clear hierarchical relationship between all categories assigned. The effectiveness of previous approaches may be limited in this case. It is a clear challenge to balance the importance of text and label semantics.

As Risch et al. [83], we want to focus on the task as a seq2seq generation and benefit from the strengths of Transformer models. However, we want to study how a pretrained model understands training samples given in a specific hierarchy and evaluate the impact of training without specifically enforcing a particular structure.

5.3 Large Scale Classification as a Generation Task

In this section, we describe our formulation for large-scale classification as a generation task.

Figure 5.1 shows the core idea behind the approach we follow, which is to encode the structure of categories within a neural model that will produce annotations to given documents. The model should be trained to associate the content of documents with a specific set of categories. Similar to Risch et al. [83], we model this as a seq2seq approach. We expect then that given an input document, the model is able to produce a list of labels, achieved with autoregressive generation.[1]

We train the model to predict categories given a sequence of document tokens so that it learns the labels to be assigned to a document. We may consider different ways of matching documents to labels, but since we want the model to predict directly those

[1] Autoregressive generation refers to generating sequential data by modeling the conditional probability distribution of each item generated given the previous one.

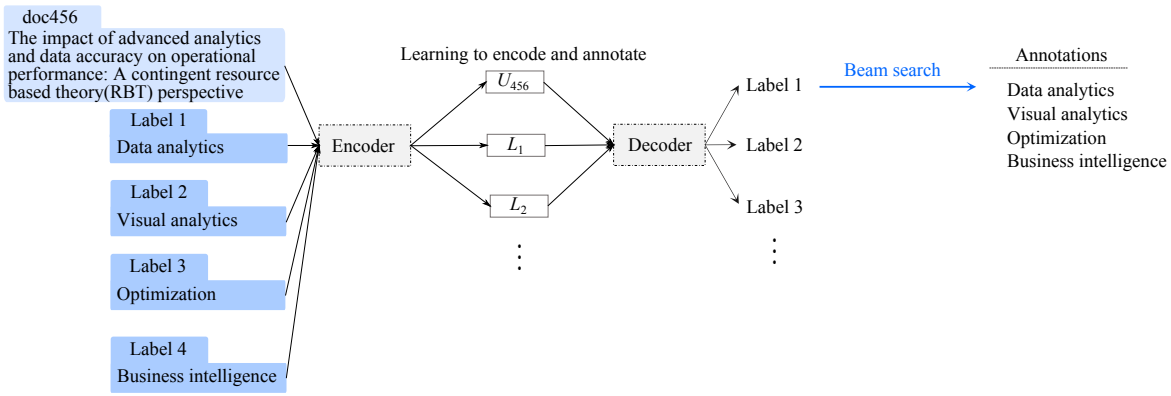


Fig. 5.1 Large Scale Classification as a Generation Task. During training, the seq2seq model takes documents and target labels to build an embedding space for a given set of categories. During inference time, given a text snippet, it is supposed to generate a set of labels matching the input.

categories, we used a straightforward inputs-to-targets strategy. In this way, we frame the task such that the seq2seq is actually of *doc_tokens-to-category*. This directly binds the categories to the document tokens by putting them in closer proximity, according to the loss function.

For the document representation, we directly input the document to the model, taking the first M tokens of a document with sequential order preserved and associating them with the categories. We also treat categories directly as tokenizable strings so that the annotation is accomplished by decoding a label string sequentially, one token at a time. When decoding, beam search is used to obtain the predicted best annotations. We use the partial beam search tree to construct the set of annotations. Formally, the output probability over categories follows (6.1), i.e.,

$$\mathscr{L}(d) = \text{Softmax}\left(W_L^\top h_d\right), \tag{5.1}$$

where W_L is the embedding matrix, where each row corresponds to the embeddings of a category. Models we train are optimized for seq2seq cross-entropy loss and trained with teacher forcing. We follow a straightforward strategy that consists of training a model to learn the different categories. Here, we observe that our setup is not looking at the specific hierarchy but trying to learn multiple relationships inside a text snippet, which can easily mix multiple categories, as learned on our results from Chapter 4.

5.4 Experiments

This section presents details on the experimental implementation, datasets, and baselines.

5.4.1 Datasets

To evaluate the effectiveness of our model, we conduct experiments on three widely-studied datasets for hierarchical multi-label text classification (WOS [54] and EURLEX [13]) and because we want to expand to more complex hierarchies, we also include experiments on a small benchmark created to evaluate classification on a large scale ontology (CSO [93]). Table 5.1 presents statistics on these datasets. Next, we describe them in detail.

WOS is a collection of public papers available from the Web Of Science. It contains the abstract, domain, and keywords of this set of published papers. The text in the abstract is the input for classification, while the domain name provides the label for the top level of the hierarchy. The keywords provide the descriptors for the next level in the classification hierarchy.

EURLEX is a large hierarchical multi-label text classification dataset containing English EU legislative documents tagged with European Vocabulary labels.

CSO provided a benchmark for evaluating text classification using members of the CSO. It contains research abstracts and keywords annotated by experts. We consider research abstracts from the Aminer collection annotated with the CSO for training.

Table 5.1 Large scale classification benchmarks. Statistics of WOS, EURLEX-57K, and CSO benchmarks for multi-label classification.

Dataset	n-labels	Train	Dev.	Test
WOS	141	30,070	7,518	9,397
EURLEX-57K	4,271	45,000	6,000	6,000
CSO	>10k	–	–	70

5.4.2 Baselines

We compared the performance of the generative model with state-of-the-art approaches for each benchmark we used:

WOS. We use results from [14], which proposed a model where both textual and label semantics are projected to the same embedding space using BERT.

EURLEX. We use results from [13], which uses a similar approach to the one used in Chapter 4 with an extra dense layer on top of BERT, with sigmoids, that produces a logit per label, which is assumed to be correlated with class probabilities.

CSO. We use results from [93], which shows results on a model using static embeddings and lexical matches.

5.4.3 Implementation Details

All models are initialized using standard pre-trained T5-small [82] model configurations. We use the Hugging Face implementation for our experiments. Models are trained for a maximum of 200,000 steps using a batch size of 32. We pick the best checkpoint based on the classification validation performance. Our training hardware consists of an Nvidia Quadro RTX 8000 GPU.

Models are trained on the collections described in Section 5.4.2. The target vocabularies are constrained to the different structures of categories for generating annotations. We train using the sets reported on Table 5.1; for the CSO benchmark, we trained the model in a weakly supervised fashion.

5.5 Results

In this section, we discuss the results and comparisons of our approach to the large-scale classification task.

Table 5.2 reports classification results for the different benchmarks in terms of recall (R) and f1 measure (F1). We compared our approach to other models that specialize in the different tasks evaluated. Our model achieves slightly better results among the various settings, especially regarding F1. We evaluate only results on labeling documents rather than focus on different levels of hierarchies, as it is shown for all the baselines.

Table 5.2 Large scale classification results (R and F1). We show results for the generative annotation compared with specific baselines on WOS, EURLEX-57K, and CSO benchmarks. Underlined values are the best results.

Benchmark	Model	R	F1
WOS	GenLSC	0.8639	0.870
	Chen et al. [14]	-	0.867
EURLEX57K	GenLSC	0.722	0.738
	Chalkidis et al. [13]	0.796	0.732
CSO	GenLSC	0.599	0.669
	Salatino et al. [93]	0.58	0.65

Overall, the results are comparable to previous results. It suggests that the generative setting is a good alternative for large-scale classification.

5.6 Discussion

In this chapter, we discussed the alternative of large-scale classification as a generation task. We specifically set the research question (*RQ4*) following the previous chapter, where we evaluated different classification settings and found overlapping disciplines the most concerning issue for increasing performance on classification.

We study how distributions of more fine-grained themes can be embedded into model parameters for performing the classification of different types of hierarchically arranged categories, as usually defined for large-scale classification. Our experiments are indicative that a strong model such as T5 can be used in this task with competitive performance, and the proposed setting could be used for annotating documents.

As discussed in Chapter 4, classification plays an important role in handling digital libraries, and in the context of scholarly documents, it represents a tool for improving the retrievability of documents. We extended our work on scientific document classification to a more granular distribution of themes.

A clear limitation of this approach compared with other models is the specialization on different levels of hierarchies in the case of strictly nested distributions of annotations. We focus on annotating documents that can be classified as multiple overlapping

themes but also possibly unrelated ones. Conversely, the work of Risch et al. [83] focuses more on the nested annotation of documents.

We specifically aim to use annotations as a way of introducing context to the search system, and this outcome provides a tool we further exploit in Chapter 6.

6

An Approximation to Learned Sparse Retrieval for Domain-specific Documents

In this chapter, we study *RQ5* How can fine-grained document annotations be used in a neural-enhanced retrieval scenario? We specifically explore an analogous scenario to Learned Sparse Retrieval to directly introduce theme annotations to the retrieval model and evaluate the use of scores assigned from a decoder model to annotations for scoring documents. Moreover, we use the model discussed in Chapter 5 for bridging context and sparse retrieval.

6.1 Introduction

Contextualization is suitable in some specific scenarios where context-free IR systems have limitations. The constraints for providing models with some level of customization have been discussed extensively during the past years [31, 73, 66]. More importantly, research interest in this direction has focused on the problem of how to use context models to benefit from it when searching.

Overall, contextualization aims to improve the quality of search results by ranking the candidate document list based on knowledge of contextual factors (as discussed in Chapter 1). To this aim, two main research questions are usually investigated: how

to model the information used as context and how to exploit such a model in the search process [73].

Query reformulation techniques, such as query expansion, are ways of approaching the issue of effectively exploiting context knowledge in search. In the field of contextual search, methods for enhancing the query with contextual information have been studied, besides the ones focused on combining scores from the contextual and global features and re-ranking [73].

On the one hand, the expansion represents a way of mitigating the vocabulary mismatch problem in IR tasks [42], which traditionally consists of reformulating the original query with additional meaningful terms. Ideally, the terms for expanding the query have a higher probability of matching relevant documents, leading to improvements in the IR effectiveness.

Expansion methods are usually classified as local (based on documents retrieved by the query) or global (using resources independent of the query) [6]. A key problem in both cases is the identification of expansions to be added to a query. The selection of possible expansions can be either automated or guided by the user through explicit interactions, which involve real users, making it expensive in terms of financial and time resources and less popular than the automated methods.

Global expansion methods rely on query and collection independent resources such as thesauri [9], controlled vocabularies or ontologies [8], among others. Automated global methods can increase query effectiveness as well as be harmed by adding irrelevant terms to the query, which creates the need for methods to discriminate expansions [12, 42].

On the other hand, expansion methods have also demonstrated improvements by enriching the document representation prior to indexing [71]. These methods aim to model the importance of terms in the documents and propagate it through expansions, i.e., identifying the most important terms to bias expansions towards them. A way of predicting the importance of terms is by using LMs. Pipelines based on LMs are not necessarily interpretable, and to achieve that, they usually require additional studies. However, expansion models achieve interpretability by grounding the representations in the original vocabulary (e.g., [64, 30]), i.e., the output of the expansions will be terms that are part of the corpus, even though the channel for getting them involves non-interpretable systems.

In this chapter, we focus on *RQ5* specifically by investigating the potential of an automated global expansion method for introducing contextual information through these expansion mechanisms. We closely follow the line of LSR model research, which allows enriching sparse representations for first-stage retrieval. Specifically, we used a hybrid setup where the vocabulary is extended by using the vocabulary of a domain ontology. We consider concepts of the ontology as terms for which we will measure importance using the model discussed in Chapter 5. When documents are expanded, we consider then a local expansion method to extend queries based on documents retrieved by the query. Instead of considering the document's original terms, we consider the document's expansion terms.

6.2 Related Work

Context models are often used to influence the relevance estimation and tailor the ranked list of documents. A way of exploiting additional information in the retrieval process is through expansion, which, besides alleviating lexical mismatch between a query and the documents, is a potential way of adding contextual information from different levels of context (as the ones outlined in Section 2.3). There are expansion models based on pseudo-relevance feedback (PRF) [119, 69], and others based on external knowledge resources [72, 9], We discuss some examples below that are also instances of NIR.

Zheng et al. [119] expand queries dynamically using chunks of text from the top feedback documents in a ranked list produced by a lexical-neural re-ranking model. Most relevant chunks are selected and combined with the original queries to compute a new document score and create a refined list of ranked results.

Padaki et al. [72] explore methods of expanding original keyword queries in an LM-based re-ranker. First, by adding structural words that help to create coherent natural language sentences (synthetic questions). Second, by adding terms with new concepts to the original queries through the classic PRF. It finds related concepts out of the list of feedback documents from an initial ranking. As a signal for supervision, the reformulated query is compared to Google's query suggestions. If it is present in the suggestions, the original query is replaced by the reformulation for the final ranking. The combination of the two mechanisms benefits a re-ranking model.

Naseri et al. [69] develop a feedback model for query expansion based on LM representations. It builds the distribution of terms' importance derived from the representations of feedback documents. The model explicitly incorporates the query focus based on its similarity to these documents, so the final representation depends on the original query. The result is an updated query LM that can be used independently or combined with other representations for re-ranking.

Blloshmi et al. [9] proposed a query expansion mechanism based on word sense disambiguation that provides sense definitions as additional semantic information for the query. The query representation combines the original keyword-based query and the definitions, which add both structure and new concepts to the query formulation. The enhanced version of the queries is used for re-ranking documents using a neural model.

Overall, expansion methods can suffer from introducing non-relevant information when expanding the query, regardless of the source of the expansion. Imani et al. [42] discussed recent approaches that can have the same limitation.

Without the explicit aim of contextualizing search, document enhancing has also been explored as a mechanism for improving ranking performance. One method that has demonstrated effectiveness along different tasks is doc-to-query [71]. It predicts a set of queries for which each document will be relevant. Given a dataset of (query, relevant document) pairs, it uses a seq2seq Transformer model that takes as an input the document terms and produces a query. Once the model is trained, it predicts a number of queries using top-k random sampling and appends them to each document in the corpus. The expanded documents are indexed, and the model retrieves a ranked list of documents for each query using BM25.

Another line of research focused on expansion mechanisms is LSR. MacAvaney et al. [64], which proposed the first method leveraging the masked language model (MLM) encoder to aggregate term weights over the logits produced by BERT, and expansions are selected as the highest-scored terms. In this case, the model is trained to do document expansion and term scoring end-to-end at once. Formal et al. [30] proposed the use of a shared MLM architecture on both the query and document sides. The MLM enables end-to-end weighting and expansion for both query and document. Instead of selecting top-k terms, a sparse regularizer is introduced during training to guarantee sparse representations.

In this Chapter, we focus on both (query and document) expansion mechanisms to contextualize search with an external source of expansions. We explore PRF as a way of expanding queries and *document topic modeling* to expand documents.

6.3 Problem Statement

We focus on an ad-hoc IR task for scholarly documents. Formally, given a query q under a given narrative[1] Q, and the collection of documents $D = \{d_1, d_2, \cdots, d_{|D|}\}$, the task is to identify and rank the documents $D' = \{d'_l \mid d'_l \in D \wedge \text{relevant}(Q, d'_l)\}$, where relevant$(Q, d'_l)$ denotes that d'_l should have content relevant to the narrative Q.

Academic search tasks have relevant nuances that we aim to take into account. In tasks involving exploratory search, queries are often given under an informational intent [22], i.e., by submitting them to the search system, we want to obtain some related information that is assumed to be available in the collection. Search tasks, such as systematic literature reviews, are information-gathering tasks that involve collecting information from multiple sources [47].

On a more global level, academic search tasks also have domain-specific properties, and it gives the opportunity to explore domain-based information sources. In particular, domain-specific knowledge is often encoded in ontologies that can be used for document annotation, and in the context of search, it has the potential to constrain the information space to find relevant documents more effectively than in open-domain applications.

6.4 Contextual Expansions for Retrieval

Because of the nature of the task, we focus on broadly analyzing document themes. We introduce contextual knowledge to the task through a large-scale ontology of research themes. This allows us to have feedback on how the information available and presented as an answer to a query is distributed according to the themes and then to constrain the results based on these distributions. In this section, we describe how we use LSR to introduce contextual information to a first-stage retrieval model.

[1] In evaluation frameworks for IR, a narrative provides a complete description of the documents the searcher would consider relevant. We make this difference because evaluation reports often include queries enriched by using narratives or another context given to the queries.

6.4.1 Overview

Similar to Formal et al. [30], we investigate how to take advantage of sparse mechanisms, such as query or document expansion. However, we explore these mechanisms specifically to bring context from knowledge resources. We aim to provide stronger-interpretable first-stage retrieval with hybrid sparse representations.

Assume that queries and documents are composed of sequences of terms taken from a vocabulary V. We also aim to represent any sequence of terms, either a query or a document, by members of the vocabulary in the ontology V_{ont}. More formally, let $f : V^n \rightarrow V_{\text{ont}}^m$ denotes such a function associating an input sequence d of n terms $t_1, \cdots, t_n$ to its m labels $L \in \mathbb{R}^m$. Overall, given all possible labels, we first consider expanding the original vocabulary V with the vocabulary V_{ont}. The way we represent an input query or document is illustrated in Figure 6.1. We use a hybrid representation where a sequence is represented by terms of the original vocabulary, and the terms form the ontology. The model mapping sequences to the ontology perform term weighting for that representation.

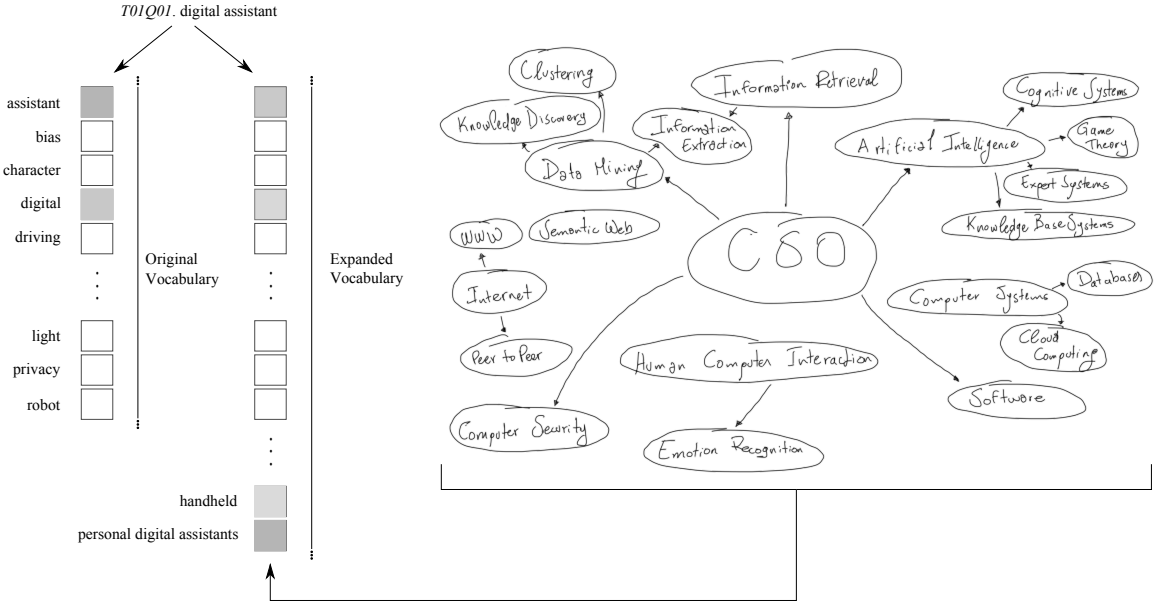


Fig. 6.1 Vocabulary expansion: Example of representing a query with the expanded vocabulary. Here, the original vocabulary is extended with members of an ontology. The example corresponds to the computer science field.

As shown in Figure 6.1 our approach also operates directly on sparse high-dimensional vectors (in the vocabulary space) in two ways. First, adding new terms to the document and estimating their term importance (term weighting), analogous to TF-IDF.

We use a large-scale domain ontology to represent the retrieved documents and study the concepts they are centered on. We specifically used the model discussed in Chapter 5 for the computer science ontology. There, we trained a seq2seq model to generate themes from the ontology in a weakly supervised fashion and then used beam search to generate multiple concepts so that we could consider the distribution themes in a retrieval scenario.

We now illustrate the process for constructing document representations, query representations, and the final query-document similarity score.

6.4.2 Document representation and indexing

A document d is represented firstly as a sparse vector. To perform document expansion, the document is projected into a $|V_{\text{ont}}| \times e$-dimensional space, i.e., $g(d) : d \rightarrow W_O^\top h_d$, where $W_O = [h_{c_1}, h_{c_2}, \cdots, h_{c_{|O|}}]^\top \in \mathbb{R}^{|V_{\text{ont}}| \times e}$ and h_{d/c_k} represents the ontology embedding matrix, i.e., last hidden states of the encoder-decoder for the document d or a concept c_k.

Once the model learns model parameters, it can make estimates for any document. The importance of a label is assumed as the generation probability. Formally, the output probability over labels is generated as follows:

$$\text{TF}_E = \text{Softmax}\left(W_O^\top h_d\right). \tag{6.1}$$

For indexing, we use ontology members as additional terms to the documents, and instead of the TF values, we store these scores.

6.4.3 Query representation

We focus on keyword queries. When documents are expanded, we consider then a local expansion method to extend queries based on documents retrieved by the query.

This is an instance of PRF, which traditionally aims to leverage the most relevant documents retrieved by the initial query to improve the subsequent query. Here, the difference is that instead of considering the document's original terms, we consider the document expansion terms.

6.4.4 Document scoring

The extended index will have postings of term t and extension E of the form [$docid(d)$, $TF(t,d)$] and [$docid(d)$, $TF_E(d)$]. It can be used with the mainstream sparse model BM25. In order to use the same formula as equation 2.3 we scale and round TF_E, using a factor M which scales the predicted weights into the range of TF values. The expansions are expected to bias the retrieval to central to more general themes and improve the performance of the first stage retrieval model. We refer to this model as contextual-enhanced SR (or CESR).

6.5 Experiments

This section presents experimental details on datasets, baselines, and implementation. We first conduct a dummy experiment reported in Appendix B.4 as a qualitative analysis of our approach.

6.5.1 Datasets

In particular, we evaluate our approach to the task proposed by SanJuan et al. [95]. It consists of retrieving all documents from a large corpus of scholarly documents and bibliographic metadata relevant to given queries. Relevant abstracts should be related to specific topics of the narrative. The narratives for this task are described as finding content related to a selection of press articles from a major international newspaper for a general audience and from Tech Xplore. Queries are keyword-based, manually extracted from articles based on the fact they allow retrieving some relevant passages from the given collection that could be inserted as citations in the press article.

Statistics for the dataset are reported in Appendix A.3. It is based on the corpus provided as the Citation Network Dataset: DBLP+Citation, ACM Citation network (12th version), which is a source of scientific documents that can be used as references. It contains more than four million documents.

Queries for this benchmark are manually extracted from a selection of 40 press articles: 20 from The Guardian and 20 from Tech Xplore6, a website taking part in the Science X Network to provide comprehensive coverage of engineering and technology advances. Each article was selected in the computer science field to be in accordance with the provided corpus. Queries were selected to provide an indication of the essential technical concepts covered by the articles.

6.5.2 Implementation Details

In this section, we present the experimental setups of the experiments described previously. We are mainly interested in analyzing the distribution of themes in the collection and understanding retrieval results and how to make them useful for the aim of content selection. Some State-of-the-art retrieval models still rely on SR models to limit the number of documents to be processed in later stages (pipeline-based models). Thus, it makes sense to study the appropriateness of the set of documents retrieved first to be processed and re-ranked in the later stages.

We use BM25 for retrieval, the CS ontology as domain knowledge, and the different strategies described in section 6.4 for expansions. We found the value of the parameter $M = 10$ after experiments ran with values in the interval $[10, 100]$ using a step of 10.

6.5.3 Baselines

We report results on different SR settings. We use BM25 as the base model with multiple variations:

BM25: we use this model in the by-default setting, i.e., no parameter optimization.

PRF: we refer as PRF to results form BM25 filtered by using annotations. In this setting, the query expansions are used to exclude documents from BM25 retrieved documents, which are not annotated with those expansions of the queries.

Dual-encoder-ST corresponds to the pre-trained Bert model in the dual-encoder setup. Tuned on the train set.

PRF_doc2query we enhanced documents with the doc2query model and applied PRF to enhance queries with expansions from the same model.

6.6 Results

This section presents the main results of various experiments we ran to test our LSR approximation to SDR. Supplementary results are provided in Appendices B.4 and B.5.

In order to study how the information from the collection can meet the information needs, we first retrieved the documents using BM25. We retrieve a number of documents from the collection using the given queries. Each set of retrieved documents allows us to study the distribution of topics that the collection gathers, specifically those selected as a response to a given query. As described in Section 6.4.3, we also benefit from PRF to expand original queries. Table 6.1 shows a selection of examples of expansions for the SDR task; strike-through items correspond to examples of apparent bad expansion terms. We further discuss these results later in this chapter. Specific extended examples are shown in Figure B.6 of Appendix B.5.

Table 6.1 Relevant expansions for a sample of queries. Strike-through items correspond to examples of apparent bad expansion terms. Expansion terms are the result of PRF on expanded documents.

Query	**Relevant themes**
Digital assistant	personal digital assistants, handheld
Biases	correlation analysis, ~~sensors~~
self driving	vehicles, autonomous driving
humanoid robots	robots, humanoid robot
online safety for children	internet, education
cookies	privacy, web content
light positioning	positioning system, indoor positioning
intelligent parking	vehicles, sensors
emotional robot	robots, emotional expressions
empathy	emotional expressions, affective state
text classification	text classification, classification models
character relationship	character recognition, computer games
gene editing	~~http~~, ~~database systems~~
conspiracy theories	signature scheme, facebook
healthcare	~~communication~~, ~~information systems~~

We analyze retrieved sets of documents in terms of the diversity of themes and the entropy of their distribution. Intuitively, a higher entropy, which directly indicates more uncertainty, implies a broader range of possible themes [3]. From the size of the ontology, we derived that the entropy of the distribution of themes in the collection

is around 13. We limit the experiments to the top 200 retrieved documents, and the average entropy of the distribution of themes is 6.83 on these sets.

As discussed before, in general, we are interested in concentrated results, implying lower entropy values and lower diversity.

Table 6.2 Theme concentration analysis. Report on average Diversity and Entropy, measured over the document expansions for retrieved documents. Underlined values correspond to the best results.

Experiment	**Diversity**		**Entropy**	
	@20	@50	@20	@50
BM25	0.4740	0.3222	5.022	5.5152
Filtering PRF	0.4213	0.2812	4.7444	5.1556
CESR	0.4145	0.2791	4.6496	5.0942

Identifying the relevant themes seems to have a positive effect. There is a clear difference between retrieving documents with the original queries (Retrieval in Table 6.2) and using extensions from relevant themes (PRF in Table 6.2). We can perceive only slight change but positive when trying to constrain results by penalizing diversity on documents (CESR in Table 6.2).

In terms of retrieval performance, analyzing themes also exhibits a positive impact. Table 6.3 shows the retrieval evaluation in terms of multiple metrics. Specifically, we report results on the test set for this experiment.

Table 6.3 Retrieval evaluation. Comparative results reported on retrieval metrics. Underlined values correspond to the best results.

Experiment	**MRR**	**P@10**	**NDCG**	**Bpref**	**MAP**
BM25	0.4536	0.1912	0.2192	0.1384	0.0515
CESR	0.5201	0.2853	0.2980	0.1898	0.1141
Dual-encoder-ST^2 (kwd)	0.3505	0.2000	0.2019	0.1956	0.0667
Dual-encoder-ST (kwd + narrative)	0.3655	0.1765	0.1912	0.2043	0.0591

Re-ranking with NIR or DPR models usually helps to refine the retrieval result, but in this case, the training set is a fairly small set, and the quality of the predictions of

the tuned models is limited. As wetness on the official evaluation for this task [95], models trained on external data perform well in the re-ranking task.

We also show in Table 6.4 results at different levels of the rank to show how the enriched model can perform as a first-stage retrieval. In this case, we present results merging the train and test sets of the collection since they share the same queries.[3]

Table 6.4 Retrieval evaluation at different rank levels. Comparative results reported on retrieval metrics at different levels of the rank. Underlined values correspond to the best results, and the †-mark indicates statistically significant improvement.

Experiment	NDCG@			P@		
	50	40	20	50	40	20
CSR	0.2331	0.2261	0.2095	0.1670	0.1778	0.2063
BM25	0.6114	0.5814	0.4744	0.2993	0.3310	0.3879
(enhan. query) BM25	0,4395	0,4163	0,3531	0,3152	0,3298	0,3500
(enhan. doc.) BM25	0,4719	0,4479	0,4011	0,3114	0,3246	0,3603
PRF_doc2query	0,3176	0,2967	0,2502	0,2314	0,2381	0,2548
CESR	0.5280	0.5018	0.4442	0.3565†	0.3710†	0.4056†

We compare the results from BM25 with retrieving documents only using query and document expansions (row 1–CSR (contextual SR)–of Table 6.4) and CESR. Interestingly, using only expansions in CSR, we are still able to retrieve relevant documents, which is indicative of the positive effect they may have. These results show that while precision is increased when using the enhanced model while nDCG drops, it implies that the model is favoring the retrieval of more relevant documents (thus increasing precision) but possibly at the expense of the quality of the ordering of these relevant documents in the retrieval list (resulting in a drop in nDCG). Which overall is a desired outcome for a first-stage retrieval model.

Moreover, we enhanced documents and queries without custom weights and used them independently to measure the effect. Expanding the vocabulary and using it in an unbalanced way hurts the retrieval performance, as experienced with the results in Table 6.4–(enhan. query) BM25 and (enhan. doc.) BM25. Our goal is mainly to increase the probability of the query to match documents. Consequently, our enhancement procedures do not work independently in this keyword query scenario. On the other hand, It was expected that Doc2query without specific training to the task

[3]Traditional evaluation in IR uses disjoint sets of queries. Since we do not use the Train set, we simply merge it to the Test set to extend our evaluation.

would add miss-leading content to documents and queries, and this is experienced by the lower performance of this model (PRF_doc2query).

6.7 Discussion

In this chapter, we focus on expansion mechanisms to introduce explicit domain context by considering contextual expansions in the document scoring process. We explore a hybrid global expansion approach with PRF as a way of expanding queries and the model discussed in Chapter 5 for expanding documents.

A text collection can gather a wide variety of themes. It is clear that for specific queries, retrieval can result in a very focused set of documents (see Figure B.7a in Appendix B.5); whereas for other topics, the distribution does not exhibit a specific theme concentration (see Figure B.7b in Appendix B.5). We also show sample queries matching the themes found through the PRF. A qualitative evaluation shows that, in general, the feedback from the collection is closely related to the queries with some exceptions, such as "gene editing" and "healthcare" (see Table 6.1). This may be due to the fact that these queries are out of the domain of the ontology.

As we hypothesized, constraining themes can contribute to the performance of the retrieval system, as shown in Table 6.2, which is one of the effects that can be achieved with the query's expansion. Table 6.3 shows results on retrieval performance, which is an instance of the interaction between query and document expansions.

Even though we closely follow the line of LSR model research, for answering *RQ5*, we proposed a hybrid approach that allows bringing context to the search and measures the importance of expansions to explicitly take them into account in the document scoring process.

According to the results, our approach has the potential to be tailored to different information needs; documents gather a wide variety of themes individually, and then a more granular evaluation should be performed to decide how to incorporate the information from the distribution of themes into a content selection process. Another unexplored alternative would be passage retrieval, considering documents also in a more granular way. We illustrate this idea in Figure B.5 of Appendix B.1.

A clear limitation of this work to extend to other domains is the need for domain-specific resources. High-quality domain knowledge has limited availability overall. Thus, replacements for such resources are part of our future research.

7

Discussion and Conclusions

7.1 Main Findings and Results

In this thesis, we studied multiple IR scenarios for adapting neural-enhanced IR models to domain-specific tasks. We adopted a contextual approach, considering tasks and domain contexts. Our main findings indicate that neural-enhanced models for IR can be adapted to specific tasks considering different aspects of relevance that users in that context may have and considering domain knowledge resources to enhance content.

In Chapter 3, we revisit the pipeline-based model for patient-to-CT matching. We propose an adaptation of training a cross-encoder to the CTR problem, taking advantage of the structured nature of the considered documents and the task. We find that the inclusion criteria section considerably impacts the retrieval score and exploits that independently from the traditional signal of relevance in our re-ranking setup. Our re-ranking formula, based on curriculum learning concerning eligibility, shows additional improvement for this task. It explicitly models the eligibility decisions instead of using only the topical relevance.

In Chapter 4, we use a high-quality dataset to establish a new benchmark for research theme classification, testing a range of classic machine-learning models under the same laboratory conditions. Unsurprisingly, our results confirm that models trained

with lengthy sections of a document as input features consistently achieve higher results than when using short document fields alone. These results hold for baseline models and our newly introduced ensemble model. At the same time, the results confirm that the BERT-based ensemble model outperforms traditional models. Our error analysis shows the difficulty of this task, mainly when presented with closely matched or overlapping domains.

Alternatively, in Chapters 5 and 6, we study how distributions of more fine-grained themes can be embedded into model parameters for performing the classification to different types of hierarchically arranged categories, as usually defined for large-scale classification. We then used the model to perform annotation of documents in a hybrid sparse retrieval approach. It allows us to bring context to the search and explicitly consider it in the document scoring process for the first-stage retrieval.

7.2 Upsides Compared to Earlier Investigations

Our work differs from previous research in multiple instances we discuss separately. Overall, we focus on designing document scoring functions from a contextual search perspective with Neural-Enhanced models. Both context and neural models set a different scenario from previous research on contextual search, neural-enhanced IR, or ranking systems.

In Chapter 3, we present our re-ranking formula based on curriculum learning to consider the traditional concept of relevance and eligibility. It explicitly models the eligibility decisions instead of using only the topical relevance, as in previous works.

In Chapter 4, we use for the first time the new gold-standard human-annotated dataset for text classification (Appendix A.1). We proposed an ensemble that allows multiple sources of information to be exploited, including an approach to handling extensive content. These factors distinguish our study from previous works.

Even though we closely follow the line of LSR model research in chapters 5 and 6, we proposed a hybrid approach that allows bringing context to the search and to some extent measure the importance of context for the scoring process, distinguishing our study from previous works.

7.3 Prospective Applications

Different outcomes of this research can be extended in different ways. We highlight some insights we could take from our study.

Although our work in Chapter 3 is focused on CTR, we believe the approach can also be applied to other tasks exhibiting multiple relevance constraints. One example of such a task is citation screening for systematic literature reviews, which also involves eligibility criteria after retrieving documents based on only topical relevance.

Besides presenting baselines for a high-quality text classification benchmark, our work in Chapter 4 can be a platform for developing aggregation strategies for analogous tasks or tasks involving text classification. We give specific insights on how to extend and enrich the data for multiple predictions and how to use inputs that are considered lengthy (such as full scholarly documents).

The work presented in chapters 5 and 6 gives insights on introducing meaningful information to the search systems from independent sources. We believe other fields can also benefit from this tool. For example, data is often categorized using hierarchies in the medical or patent domains.

7.4 Limitations

This study has several limitations, some related to scalability and some related to models constraints.

CT data available may limit the model in learning relevant patterns needed to scale to different clinical settings or patient populations. Specifically, the curriculum learning scenario in the eligibility determination objective lacks fine-grained criteria, and labels and decisions are not interpretable.

Broad themes classification, in general, has the limitation of strictly having to classify documents into a single research field. A scholarly document actually could often be classified into multiple domains. Even though we give a wide variety of sources of information in this context, another limitation is the requirement of multiple document fields, which, for example, are necessary for the AZ approach, which we have seen contributes to the performance.

Although we propose an alternative model for theme classification, scaling to fine-grained themes, limitations arise, given the need for domain-specific resources. In this way, an important limitation of this work to extend to other domains is the need for domain-specific resources. High-quality domain knowledge has limited availability overall.

7.5 Future Directions

Besides approaching the specific limitations of this work, we specifically draw insights for future research.

On the one hand, in the short term, our curriculum learning approach for CTR is aimed to be applied in the task of screening documents for systematic reviews. In [57], we presented a high-quality benchmark for this task, which is a potential application for exploring our contribution.

In terms of the theme classification tasks, we would like to measure the importance of weight assignments for augmented predictions and consider the overlap between disciplines to evaluate ways of disambiguating predictions falling into related themes.

As for the LSR variant, we proposed in Chapter 6, we would like to test it on some other scenarios. In particular, in Chapter 5, a classification task in the legal domain was evaluated. We intend to approach the legal case retrieval task as we did SDR.

On the other hand, in the long term, we aim to address limitations regarding resources needed to scale our methods. We pointed out that a clear limitation of this work to extend to other domains is the need for domain-specific resources. High-quality domain knowledge has limited availability overall. Thus, replacements for such resources are part of our future research.

We believe the setting we proposed for SDR will allow us to tailor search results online. We refer specifically to Figure B.4 in Appendix B.4, where we can see annotations ranked by their importance scores. We intend to propose a demo application to evaluate tuning concepts' importance interactively when searching.

References

[1] Accuosto, P., Neves, M., and Saggion, H. (2021). Argumentation mining in scientific literature: from computational linguistics to biomedicine. In *Frommholz I, Mayr P, Cabanac G, Verberne S, editors. BIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval; 2021 Apr 1; Lucca, Italy. Aachen: CEUR; 2021. p. 20-36*. CEUR Workshop Proceedings.

[2] Acheampong, F. A., Nunoo-Mensah, H., and Chen, W. (2021). Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8).

[3] Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Göker, A., Kompatsiaris, I., and Jaimes, A. (2013). Sensing trending topics in twitter. *IEEE Transactions on multimedia*, 15(6):1268–1282.

[4] Amati, G. and Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389.

[5] Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., et al. (2018). Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*.

[6] Azad, H. K. and Deepak, A. (2019). Query expansion techniques for information retrieval: a survey. *Information Processing & Management*, 56(5):1698–1735.

[7] Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

[8] Bhogal, J., MacFarlane, A., and Smith, P. (2007). A review of ontology based query expansion. *Information processing & management*, 43(4):866–886.

[9] Blloshmi, R., Pasini, T., Campolungo, N., Banerjee, S., Navigli, R., and Pasi, G. (2021). IR like a SIR: Sense-enhanced Information Retrieval for Multiple Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

[10] Borlund, P. (2003). The concept of relevance in ir. *Journal of the American Society for information Science and Technology*, 54(10):913–925.

[11] Boyack, K. W. and Klavans, R. (2018). Accurately identifying topics using text: Mapping pubmed. In *STI 2018 Conference Proceedings*, pages 107–115. Centre for Science and Technology Studies (CWTS).

[12] Cao, G., Nie, J.-Y., Gao, J., and Robertson, S. (2008). Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 243–250.

[13] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., and Androutsopoulos, I. (2019). Large-scale multi-label text classification on eu legislation. *arXiv preprint arXiv:1906.02192*.

[14] Chen, H., Ma, Q., Lin, Z., and Yan, J. (2021). Hierarchy-aware label semantics matching network for hierarchical text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4370–4379.

[15] Craswell, N., Mitra, B., Yilmaz, E., Campos, D., and Voorhees, E. M. (2020). Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820*.

[16] Cressey, D. and Gibney, E. (2014). Uk releases world's largest university assessment. *Nature*.

[17] Dai, Z. and Callan, J. (2019). Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 985–988.

[18] Dai, Z., Xiong, C., Callan, J., and Liu, Z. (2018). Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 126–134.

[19] Daradkeh, M., Abualigah, L., Atalla, S., and Mansoor, W. (2022). Scientometric analysis and classification of research using convolutional neural networks: A case study in data science and analytics. *Electronics*, 11(13):2066.

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

[21] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[22] Dou, Z. and Guo, J. (2020). *Query Intent Understanding*, pages 69–101. Springer International Publishing, Cham.

[23] El-Ebshihy, A., Ningtyas, A. M., Andersson, L., Piroi, F., and Rauber, A. (2020). ARTU / TU Wien and artificial researcher@ LongSumm 20. In *Proceedings of the First Workshop on Scholarly Document Processing*, Online. Association for Computational Linguistics.

[24] Embi, P. J., Jain, A., and Harris, C. M. (2008). Physicians' perceptions of an electronic health record-based clinical trial alert approach to subject recruitment: a survey. *BMC medical informatics and decision making*, 8(1):1–8.

[25] Ermakova, L., Sanjuan, E., Huet, S., Azarbonyad, H., Augereau, O., and Kamps, J. (2023). Overview of the clef 2023 simpletext lab: Automatic simplification of scientific texts. In *CLEF 2023 Conference and Labs of the Evaluation Forum Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*.

[26] Ermakova, L., SanJuan, E., Kamps, J., Huet, S., Ovchinnikova, I., Nurbakova, D., Araújo, S., Hannachi, R., Mathurin, E., and Bellot, P. (2022). Overview of the clef 2022 simpletext lab: Automatic simplification of scientific texts. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 13390 LNCS, pages 470–494. Springer Science and Business Media Deutschland GmbH.

[27] Eykens, J., Guns, R., and Engels, T. C. (2021). Fine-grained classification of social science journal articles using textual data: A comparison of supervised machine learning approaches. *Quantitative Science Studies*, 2(1):89–110.

[28] Falcon, W. and team, T. P. L. (2019). Pytorch lightning.

[29] Fang, H., Tao, T., and Zhai, C. (2004). A formal study of information retrieval heuristics. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56.

[30] Formal, T., Lassance, C., Piwowarski, B., and Clinchant, S. (2021). Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.

[31] Freund, L. and Toms, E. G. (2005). Contextual search: from information behaviour to information retrieval. In *Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI*.

[32] Fricke, S. (2018). Semantic scholar. *Journal of the Medical Library Association: JMLA*, 106(1):145.

[33] Gialitsis, N., Kotitsas, S., and Papageorgiou, H. (2022). Scinobo: A hierarchical multi-label classifier of scientific publications. *arXiv preprint arXiv:2204.00880*.

[34] Grivas, A., Alex, B., Grover, C., Tobin, R., and Whiteley, W. (2020). Not a cute stroke: analysis of rule-and neural network-based information extraction systems for brain radiology reports. In *Proceedings of the 11th international workshop on health text mining and information analysis*, pages 24–37.

[35] Gündoğan, E. and Kaya, M. (2020). Research paper classification based on word2vec and community discovery. In *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pages 1032–1036. IEEE.

[36] Guo, J., Fan, Y., Ai, Q., and Bruce Croft, W. (2016). A deep relevance matching model for ad-hoc retrieval. In *International Conference on Information and Knowledge Management, Proceedings*, pages 55–64.

[37] Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., Zamani, H., Wu, C., Croft, W. B., and Cheng, X. (2020). A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067.

[38] Hande, A., Puranik, K., Priyadharshini, R., and Chakravarthi, B. R. (2021). Domain identification of scientific articles using transfer learning and ensembles. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 88–97. Springer.

[39] Hofstätter, S., Craswell, N., Mitra, B., Zamani, H., and Hanbury, A. (2022). Are we there yet? a decision framework for replacing term based retrieval with dense retrieval systems. *arXiv preprint arXiv:2206.12993*.

[40] Hofstätter, S., Lin, S.-C., Yang, J.-H., Lin, J., and Hanbury, A. (2021). Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proc. of SIGIR*.

[41] Hoppe, F., Dessì, D., and Sack, H. (2021). Deep learning meets knowledge graphs for scholarly data classification. In *Companion proceedings of the web conference 2021*, pages 417–421.

[42] Imani, A., Vakili, A., Montazer, A., and Shakery, A. (2019). Deep neural networks for query expansion using word embeddings. In *European Conference on Information Retrieval*, pages 203–210. Springer.

[43] Jin, Q., Tan, C., Zhao, Z., Yuan, Z., and Huang, S. (2021). Alibaba DAMO Academy at TREC Clinical Trials 2021: Exploring Embedding-based First-stage Retrieval with TrialMatcher. *TREC 2021*.

[44] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

[45] Kandimalla, B., Rohatgi, S., Wu, J., and Giles, C. L. (2021). Large scale subject category classification of scholarly papers with deep attentive neural networks. *Frontiers in research metrics and analytics*, 5:600382.

[46] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

[47] Kellar, M., Watters, C., and Shepherd, M. (2007). A field study characterizing web-based information-seeking tasks. *Journal of the American Society for information science and technology*, 58(7):999–1018.

[48] Khattab, O. and Zaharia, M. (2020). Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

[49] Kim, S.-W. and Gil, J.-M. (2019). Research paper classification systems based on tf-idf and lda schemes. *Human-centric Computing and Information Sciences*, 9(1):1–21.

[50] Knoth, P. and Zdrahal, Z. (2012). Core: three access levels to underpin open access. *D-Lib Magazine*, 18(11/12).

[51] Koopman, B. and Zuccon, G. (2016a). A test collection for matching patients to clinical trials. *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[52] Koopman, B. and Zuccon, G. (2016b). A test collection for matching patients to clinical trials. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 669–672.

[53] Koopman, B. and Zuccon, G. (2021). Cohort-based clinical trial retrieval. In *Proceedings of the 25th Australasian Document Computing Symposium*, pages 1–9.

[54] Kowsari, K., Brown, D. E., Heidarysafa, M., Meimandi, K. J., Gerber, M. S., and Barnes, L. E. (2017). Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 364–371. IEEE.

[55] Kusa, W., Hanbury, A., and Knoth, P. (2022). Automation of citation screening for systematic literature reviews using neural networks: A replicability study. In Hagen, M., Verberne, S., Macdonald, C., Seifert, C., Balog, K., Nørvåg, K., and Setty, V., editors, *Advances in Information Retrieval*, pages 584–598, Cham. Springer International Publishing.

[56] Kusa, W., Mendoza, Ó. E., Knoth, P., Pasi, G., and Hanbury, A. (2023a). Effective matching of patients to clinical trials using entity extraction and neural re-ranking. *Journal of biomedical informatics*, 144:104444.

[57] Kusa, W., Mendoza, O. E., Samwald, M., Knoth, P., and Hanbury, A. (2023b). Csmed: Bridging the dataset gap in automated citation screening for systematic literature reviews.

[58] Leveling, J. (2017). Patient selection for clinical trials based on concept-based retrieval and result filtering and ranking. In *TREC*.

[59] Li, H. (2022). *Learning to rank for information retrieval and natural language processing*. Springer Nature.

[60] Liu, H. (2017). Automatic argumentative-zoning using word2vec. *CoRR*, abs/1703.10152.

[61] Lopez, P. (2009). Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *International conference on theory and practice of digital libraries*, pages 473–474. Springer.

[62] Łukasik, M., Kuśmierczyk, T., Bolikowski, Ł., and Nguyen, H. S. (2013). Hierarchical, multi-label classification of scholarly publications: modifications of ml-knn algorithm. In *Intelligent tools for building a scientific information platform*, pages 343–363. Springer.

[63] MacAvaney, S., Cohan, A., Yates, A., and Goharian, N. (2019). CEDR: Contextualized embeddings for document ranking. In *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1101–1104.

[64] MacAvaney, S., Nardini, F. M., Perego, R., Tonellotto, N., Goharian, N., and Frieder, O. (2020). Expansion via prediction of importance with contextualization. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 1573–1576.

[65] Merrouni, Z. A., Frikh, B., and Ouhbi, B. (2019a). Toward Contextual Information Retrieval: A Review and Trends. *Procedia Computer Science*, 148:191–200.

[66] Merrouni, Z. A., Frikh, B., and Ouhbi, B. (2019b). Toward contextual information retrieval: a review and trends. *Procedia computer science*, 148:191–200.

[67] Mitra, B., Diaz, F., and Craswell, N. (2017). Learning to match using local and distributed representations of text for web search. In *26th International World Wide Web Conference, WWW 2017*, pages 1291–1299.

[68] Nagumothu, D., Eklund, P. W., Ofoghi, B., and Bouadjenek, M. R. (2021). Linked data triples enhance document relevance classification. *Applied Sciences*, 11(14):6636.

[69] Naseri, S., Dalton, J., Yates, A., and Allan, J. (2021). Ceqe: Contextualized embeddings for query expansion. In *European Conference on Information Retrieval*, pages 467–482. Springer.

[70] Ni, Y., Kennebeck, S., Dexheimer, J. W., McAneney, C. M., Tang, H., Lingren, T., Li, Q., Zhai, H., and Solti, I. (2015). Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. *Journal of the American Medical Informatics Association*, 22(1):166–178.

[71] Nogueira, R., Yang, W., Lin, J., and Cho, K. (2019). Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

[72] Padaki, R., Dai, Z., and Callan, J. (2020). Rethinking query expansion for bert reranking. In *European conference on information retrieval*, pages 297–304. Springer.

[73] Pasi, G. (2011). Contextual search: issues and challenges. In *Symposium of the Austrian HCI and Usability Engineering Group*, pages 23–30. Springer.

[74] Pech, G., Delgado, C., and Sorella, S. P. (2022). Classifying papers into subfields using abstracts, titles, keywords and keywords plus through pattern detection and optimization procedures: An application in physics. *Journal of the Association for Information Science and Technology*.

[75] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[76] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics.

[77] Pradeep, R., Hui, K., Gupta, J., Lelkes, A. D., Zhuang, H., Lin, J., Metzler, D., and Tran, V. Q. (2023). How does generative retrieval scale to millions of passages? *arXiv preprint arXiv:2305.11841*.

[78] Pradeep, R., Li, Y., Wang, Y., and Lin, J. (2022). Neural query synthesis and domain-specific ranking templates for multi-stage clinical trial matching. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2325–2330, New York, NY, USA. Association for Computing Machinery.

[79] Pressler, T. R., Yen, P.-Y., Ding, J., Liu, J., Embi, P. J., and Payne, P. R. (2012). Computational challenges and human factors influencing the design and use of clinical research participant eligibility pre-screening tools. *BMC medical informatics and decision making*, 12(1):1–11.

[80] Pujari, S. C., Friedrich, A., and Strötgen, J. (2021). A multi-task approach to neural multi-label hierarchical patent classification using transformers. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43*, pages 513–528. Springer.

[81] Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W. X., Dong, D., Wu, H., and Wang, H. (2020). Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*.

[82] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

[83] Risch, J., Garda, S., and Krestel, R. (2020). Hierarchical document classification as a sequence generation task. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 147–155.

[84] Rivest, M., Vignola-Gagné, E., and Archambault, É. (2021). level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling. *PloS one*, 16(5):e0251493.

[85] Roberts, K., Demner-Fushman, D., Voorhees, E. M., Bedrick, S., and Hersh, W. R. (2021). Overview of the TREC 2021 Clinical Trials Track. In *Proceedings of the Thirtieth Text REtrieval Conference (TREC 2021)*.

[86] Roberts, K., Demner-Fushman, D., Voorhees, E. M., Bedrick, S., and Hersh, W. R. (2022). Overview of the TREC 2022 Clinical Trials Track. In *Proceedings of the Thirtieth Text REtrieval Conference (TREC 2022)*.

[87] Roberts, K., Demner-Fushman, D., Voorhees, E. M., Hersh, W. R., Bedrick, S., Lazar, A. J., and Pant, S. (2017). Overview of the trec 2017 precision medicine track. In *TREC*.

[88] Roberts, K., Demner-Fushman, D., Voorhees, E. M., Hersh, W. R., Bedrick, S., Lazar, A. J., Pant, S., and Meric-Bernstam, F. (2019). Overview of the trec 2019 precision medicine track. In *The text REtrieval conference: TREC. Text REtrieval Conference*.

[89] Robertson, S. E. and Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146.

[90] Rybinski, M., Xu, J., and Karimi, S. (2020). Clinical trial search: Using biomedical language understanding models for re-ranking. *Journal of Biomedical Informatics*, 109:103530.

[91] Rybiński, M., Nguyen, V., and Karimi, S. (2022). CSIROmed Team Report of TREC 2021 Clinical Trials track: Experiments with BERT Reranking Methods.

[92] Salatino, A., Osborne, F., and Motta, E. (2022). Cso classifier 3.0: a scalable unsupervised method for classifying documents in terms of research topics. *International Journal on Digital Libraries*, 23(1):91–110.

[93] Salatino, A. A., Osborne, F., Thanapalasingam, T., and Motta, E. (2019). The cso classifier: Ontology-driven detection of research topics in scholarly articles. In *Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings 23*, pages 296–311. Springer.

[94] Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

[95] SanJuan, E., Huet, S., Kamps, J., and Ermakova, L. (2022). Overview of the clef 2022 simpletext task 1: Passage selection for a simplified summary.

[96] Semberecki, P. and Maciejewski, H. (2017). Deep learning methods for subject text classification of articles. In *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 357–360. IEEE.

[97] Sharma, P. and Li, Y. (2019). Self-supervised contextual keyword and keyphrase retrieval with self-labelling.

[98] Shen, Z., Ma, H., and Wang, K. (2018). A web-scale system for scientific knowledge exploration. *arXiv preprint arXiv:1805.12216*.

[99] Shivade, C., Hebert, C., Lopetegui, M., De Marneffe, M.-C., Fosler-Lussier, E., and Lai, A. M. (2015). Textual inference for eligibility criteria resolution in clinical trials. *Journal of biomedical informatics*, 58:S211–S218.

[100] Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., and Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW

'15 Companion, pages 243–246, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[101] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

[102] Sykes, D., Grivas, A., Grover, C., Tobin, R., Sudlow, C., Whiteley, W., McIntosh, A., Whalley, H., and Alex, B. (2021). Comparison of rule-based and neural network models for negation detection in radiology reports. *Natural Language Engineering*, 27(2):203–224.

[103] Taheriyan, M. (2011). Subject classification of research papers based on interrelationships analysis. In *Proceedings of the 2011 workshop on Knowledge discovery, modeling and simulation*, pages 39–44.

[104] Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., Qin, Z., Hui, K., Zhao, Z., Gupta, J., et al. (2022). Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35:21831–21843.

[105] Teufel, S., Carletta, J., and Moens, M. (1999a). An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*.

[106] Teufel, S. et al. (1999b). *Argumentative zoning: Information extraction from scientific text*. PhD thesis, Citeseer.

[107] Teufel, S. and Moens, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

[108] Teufel, S., Siddharthan, A., and Batchelor, C. (2009). Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1493–1502.

[109] Trotman, A., Puurula, A., and Burgess, B. (2014). Improvements to BM25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, pages 58–65.

[110] Van Eck, N. J. and Waltman, L. (2017). Citation-based clustering of publications using citnetexplorer and vosviewer. *Scientometrics*, 111(2):1053–1070.

[111] Waltman, L. and Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12):2378–2392.

[112] Wang, S. and Koopman, R. (2017). Clustering articles based on semantic similarity. *Scientometrics*, 111(2):1017–1031.

[113] Wang, Y., Hou, Y., Wang, H., Miao, Z., Wu, S., Chen, Q., Xia, Y., Chi, C., Zhao, G., Liu, Z., et al. (2022). A neural corpus indexer for document retrieval. *Advances in Neural Information Processing Systems*, 35:25600–25614.

[114] Wang, Z. and Sun, J. (2022). Trial2vec: Zero-shot clinical trial document similarity search using self-supervision. *arXiv preprint arXiv:2206.14719*.

[115] Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P., Ahmed, J., and Overwijk, A. (2020). Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

[116] Yates, A., Nogueira, R., and Lin, J. (2021). Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, pages 1154–1156.

[117] Zeng, H., Zamani, H., and Vinay, V. (2022). Curriculum learning for dense retrieval distillation. *arXiv preprint arXiv:2204.13679*.

[118] Zhao, T., Lu, X., and Lee, K. (2020). Sparta: Efficient open-domain question answering via sparse transformer matching retrieval. *arXiv preprint arXiv:2009.13013*.

[119] Zheng, Z., Hui, K., He, B., Han, X., Sun, L., and Yates, A. (2020). Bert-qe: contextualized query expansion for document re-ranking. *arXiv preprint arXiv:2009.07258*.

Appendix Overview

This section provides an overview of the supplementary materials. Appendix A reports detailed descriptions of the datasets and benchmarks used in this thesis. Appendix B presents supplementary results that provide important insights into the main results presented in the thesis.

A

Datasets Descriptions

A.1 CTR Datasets

A summary of the CTR datasets is presented in Table A.1.

Table A.1 Statistics of TREC CT datasets from 2021 and 2022. The train set is from the 2021 edition, and the test set is from the 2022 edition.

	Train (2021)	**Test** (2022)
Documents	375,580	375,580
Topics (patient notes)	75	50
Avg. topic length (tokens)	133.4	105.9
Avg. topic length (sentences)	11.2	9.4
Total judgements	35,832	35,394
– Eligible (2)	5,570	3,939
– Excluded (1)	6,019	3,036
– Not relevant (0)	24,243	28,419
Unique trials judged	26,162	26,585

A.2 Theme Classification Dataset

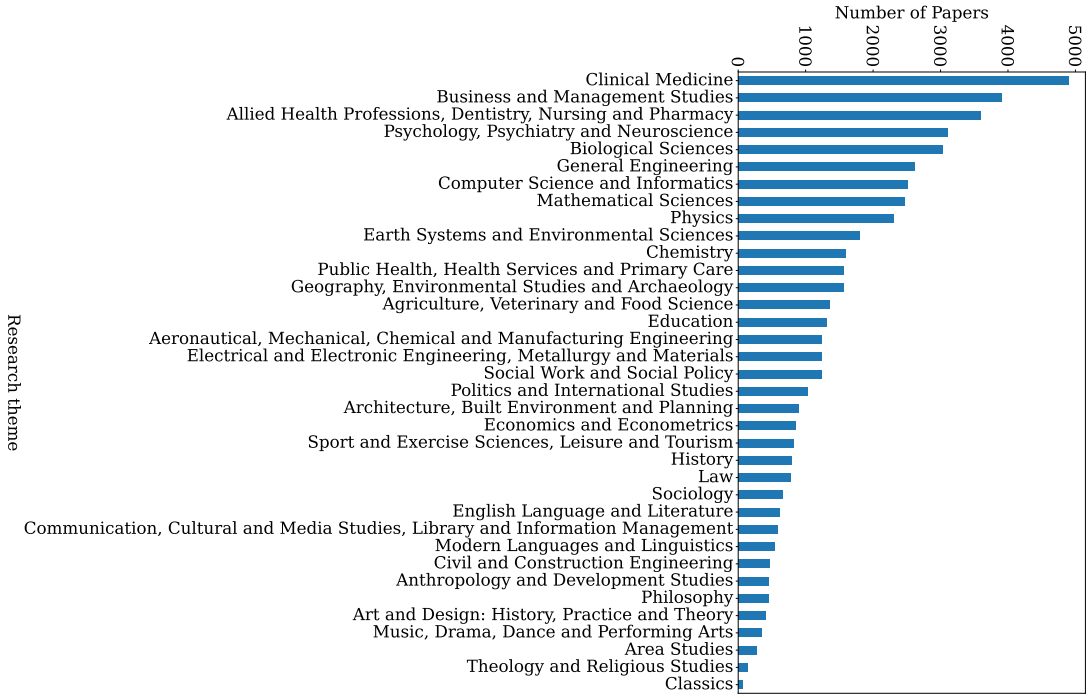


Fig. A.1 Breakdown of the research theme classification dataset by theme. About 60,000 records from the REF dataset were selected along 36 categories and enriched for this dataset.

The dataset for research theme classification was proposed by people from CORE.[1] As previously discussed, one of the significant challenges faced in the scholarly domain is the lack of large-scale labeled data for research theme classification. For the shared task, a completely new gold-standard dataset was compiled using data drawn from the U.K.'s Research Excellence Framework (REF) 2014 exercise [16]. In total, 191,000 research outputs were submitted by 154 higher education and research institutions and then peer-reviewed by experts from each domain. The REF divided research outputs into 36 'Units of Assessment' (UoA) or domain areas.

[1] https://core.ac.uk

The institutions themselves selected to which Unit of Assessment each output was submitted.

The data from the REF exercise, therefore, provides a near-perfect starting point for the task of automatically identifying research themes, as the UoA labels were manually assigned to each output by the expert academics responsible for its production.

For each output, the following were available from the REF data: publication title, publication year, publication venue, name of institution, and Unit of Assessment. These fields were fully populated for 190,628 out of 190,963 submissions to the outputs category of the REF process. We further enriched each record with the DOI, CORE ID, and abstract (where available). The CORE ID is used to identify the actual research article held by the CORE service. Not all papers in the dataset are open-access; therefore, the full-text content of all papers is not available. For non-open access papers, CORE often still has the metadata for these articles.

For the data used in this shared task, separate test and train datasets were generated. From the full REF dataset, 51,560 randomly selected records were used for the train set, and a separate 10,000 were selected for the test set. The datasets were then verified to ensure that there was no overlap between the two sets. Figure A.1 shows the cross-discipline (theme) breakdown of all records used for this task.

Statistics for the initial dataset are provided in Table A.2. Most of this dataset's publications do not contain abstracts, additional metadata, or PDFs. Theme identification algorithms should be robust to these missing features and work well when only titles are available.

Table A.2 Statistics for the research theme classification dataset. Percentages of information available for each record on both train and test sets.

	Train	**Test**
Size	51,560	10,000
% of Publications		
– available via CORE API	91.6%	92.4%
– with abstract	31.8%	31.7%
– with PDF	24.6%	25.6%
– with full text	6.3%	6.4%
– with references	8.4%	7.6%

A.3 SDR Dataset

Table A.3 shows statistics of the data available from SanJuan et al. [95] for the task of SDR. In our experiments, we merge both sets since the traditional splinting for IR comprehends disjoint sets of queries, which is not the case of the original task proposed by SanJuan et al. [95].

Table A.3 SDR Dataset statistics. Queries and judgments available from the dataset for SDR by SanJuan et al. [95].

Set	Topics	Queries	Judgments
Train	15	29	1,303
Test	15	29	3,835
Corpus		4,232,520	

B

Supplementary Results

B.1 CTR Supplementary Results

Figure B.1 presents two plots with an average per patient count of relevant and excluded trials depending on a cutoff point for the TREC CT 2022 collection. Both versions of lexical models, considering different fields from documents, show a positive impact of the field selection in finding more eligible trials. We can also see that the TCRR neural re-ranking retrieves twice as many trials for which a patient is eligible than excluded but helps in removing ineligible only for the first 15 trials. One possible explanation is that we re-ranked only the top 50 trials retrieved by the first-stage ranker.

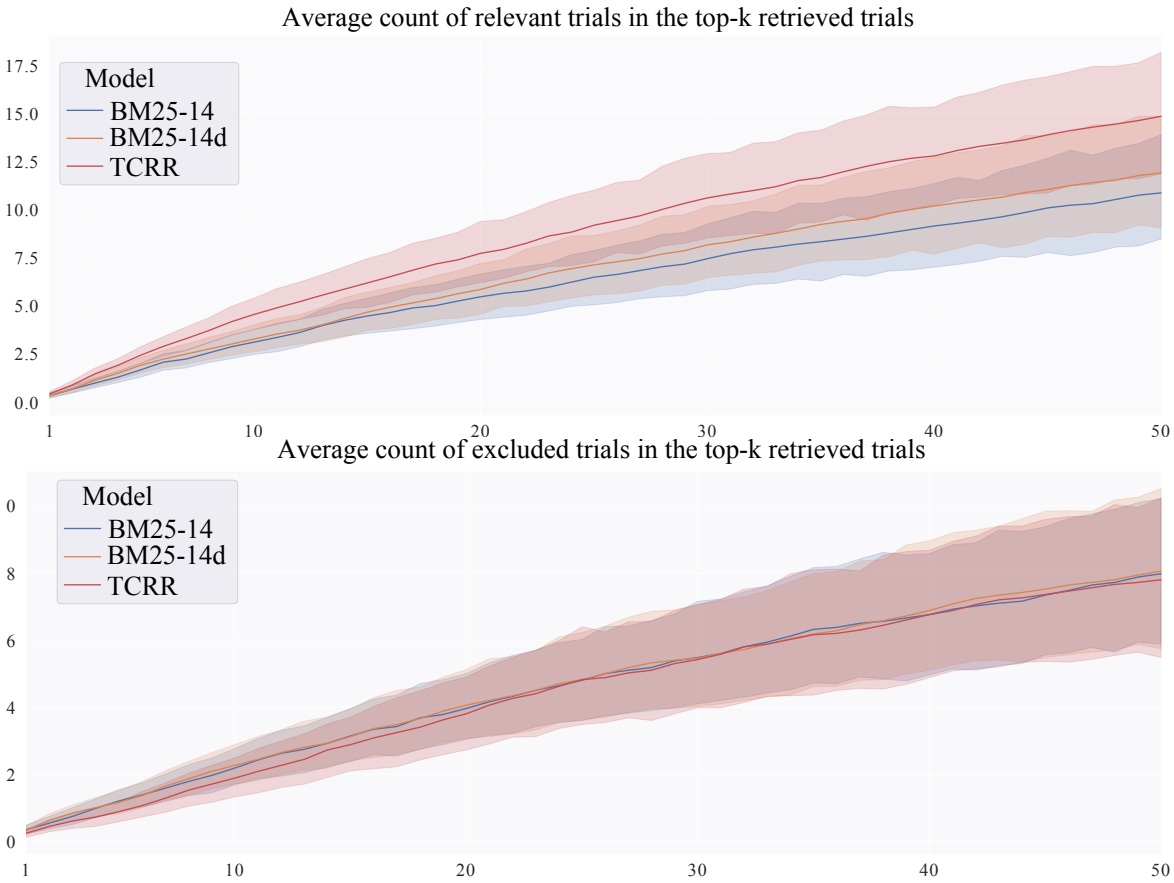


Fig. B.1 Averaged per patient count of relevant (top) and excluded (bottom) trials depending on a cut-off of K trials retrieved (x-axis) for TREC CT 2022 collection. Both versions of Bm25 show a positive impact of the field selection in finding more eligible trials. The TCRR neural re-ranking retrieves twice as many trials for which a patient is eligible than excluded; it helps to remove ineligible until the first 15 retrieved trials.

B.2 CTR TREC Records

The DoSSIER group[1] consisting of members of the IR group of TU Wien, the University of Milano-Bicocca, and the University of Vienna participated in the CT track at TREC 2022 [86]. We focus on effective and efficient approaches for retrieval and consider domain-specific characteristics of the retrieval task as well as information extraction methods.

As part of the systems used for approaching the CTR in TREC, TCRR was compared with different models. We use different ranking and neural re-ranking models, compare which parts of the text should be taken into account for ranking, and investigate how to enhance the query and the documents with extracted information in order to increase the ranking performance. For the first stage retrieval, we enrich the textual representation of the clinical trial by extracting different elements such as keywords, entities, and sections and compare BM25 retrieval based on different input texts of the clinical trial. For re-ranking of the CT track we use a dense retrieval model trained with knowledge distillation and topic-aware sampling of negatives (TASB) [40]. We further fine-tune the TASB model on the clinical trial retrieval task by using the first 50 queries of the CT 2021 test collection and their positive and negative samples.

The evaluation results show that while the TCRR exhibits performance improvements compared to the BM25 retrieval, the dense retrieval model used for re-ranking decreases the effectiveness. Furthermore, we reach the highest effectiveness in terms of nDCG@5 with BM25 by including the eligibility criteria to the query text and enriching the textual representation with keywords extracted from the text. This result is comparable to the one reached by the cross-encoder re-ranking, which, in terms of nDCG@10 and P@10, is our best-performing approach.

BM25 (keywords query) queries are sets of keywords extracted from patients' cases using [97]. The index is created using trials' summary, description, titles, conditions, and criteria sections.

BM25 (Enriched query) run with input text summary, description, titles, conditions, inclusion, lemmatized and enriched with keywords from eligibility criteria for affirmative and negative and family history entities. Query with current medical history

[1] Sophia Althammer, Sebastian Hofstätter, Oscar E. Mendoza, Wojciech Kusa, Vasiliki Kougia, DoSSIER@TREC 2022 Deep Learning and Clinical Trials Track., TREC 2022.

Table B.1 Official TREC CT 2022 evaluation results for TCRR. Comparative results reported on retrieval metrics. Underlined values correspond to the best results.

Model	nDCG@5	nDCG@10	Prec@10	RR
BM25 (keywords query)	0.4496	0.4477	0.3240	0.4482
BM25 (Enriched query)	0.5731	0.5280	0.3980	0.6607
TASB	0.4092	0.3954	0.2840	0.4076
TCRR	0.5734	0.5565	0.4560	0.6191

text enriched with all keywords and keywords from eligibility criteria for affirmative and negative and family history entities. Post-retrieval filtering.

TASB Re-ranking with the TASB model, which is trained for web search with knowledge distillation on MS Marco V1 and then domain fine-tuned on the query-document pairs of 2021 TREC Clinical Trials test collection of the first 50 queries.

B.3 Supplementary Results on Broad Theme Classification

In this section, we extend the confusion matrix of the broad theme classification. For convenience, we show the results for only the 25 most frequent classes, and we group the rest of them in a single class.

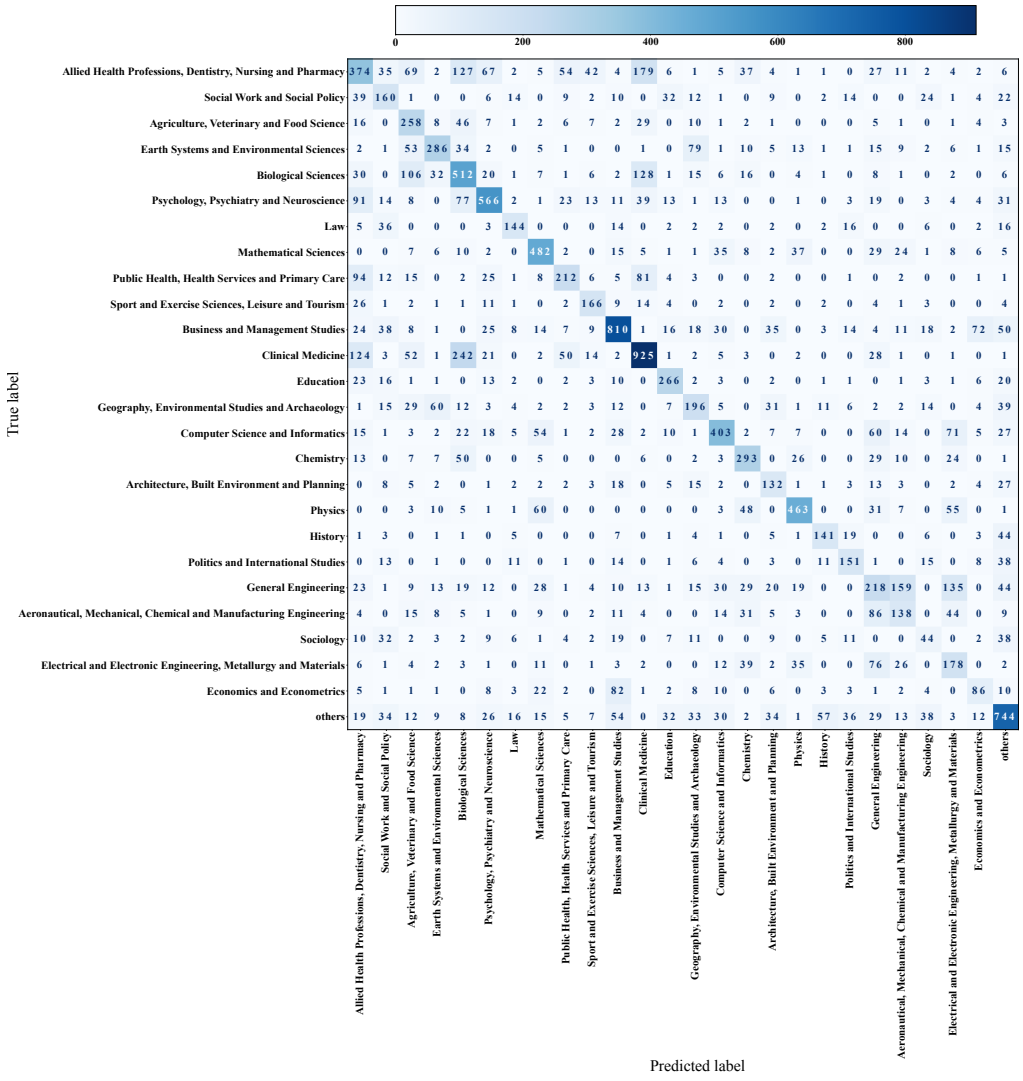


Fig. B.2 Confusion Matrix for validation results for 25 most frequent classes. The remaining 11 classes are grouped in the 'others' category. Notice that for Clinical Medicine, most examples where the model's incorrect prediction are classified as Allied Health Professions, Nursing and Pharmacy, and Biological Sciences. Similar behavior can be observed with other closely related disciplines.

B.4 Qualitative analysis of context-enhanced SDR

We used the collection of academic documents from the computer science discipline to annotate it with the CSO ontology. We evaluate the effectiveness of the approach described in Chapter 6 on a set of expert-evaluated queries.

In order to test the model with a meaningful qualitative evaluation, we manually prepare a set of 20 topics for the comparative performance measurement.

B.4.1 Expert Assessment

The topics for the experiments consist of information needs formulated by expert researchers in the computer science field. These topics were tested in a demo search engine, and the ten highest-ranked results for each topic were judged by the same experts who proposed them. The topics were judged as either relevant or not relevant. Because we evaluate the results using a binary relevance scale, the performance is measured using P@10. However, we also present results in terms of nDCG@10.

B.4.2 Query Analysis

Fig. B.3 shows the difference in performance (measured by P@10 and nDCG@10) between our approach and the baseline search for each of the 20 test queries. Overall, there is an improvement achieved by our approach in the global comparison provided by the histogram. We compute the same measures for the complete experiment. Overall, the relative improvement of our approach over the baseline is of 21.5% in terms of P@10, and of 20.7% in terms of nDCG@10. These results support our expectations with respect to the precision increment.

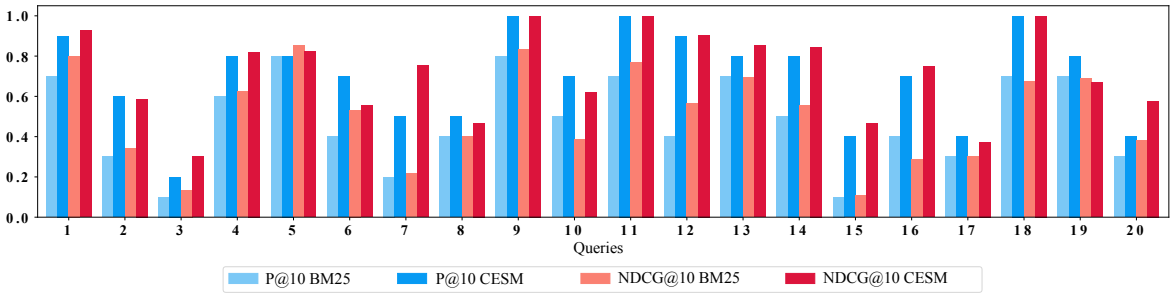


Fig. B.3 Qualitative performance report. Performance comparison (measured by P@10 and nDCG@10) between our CESR and the BM25 for each of the 20 proposed queries. Overall, there is an improvement achieved by CESR in the global comparison provided by the histogram.

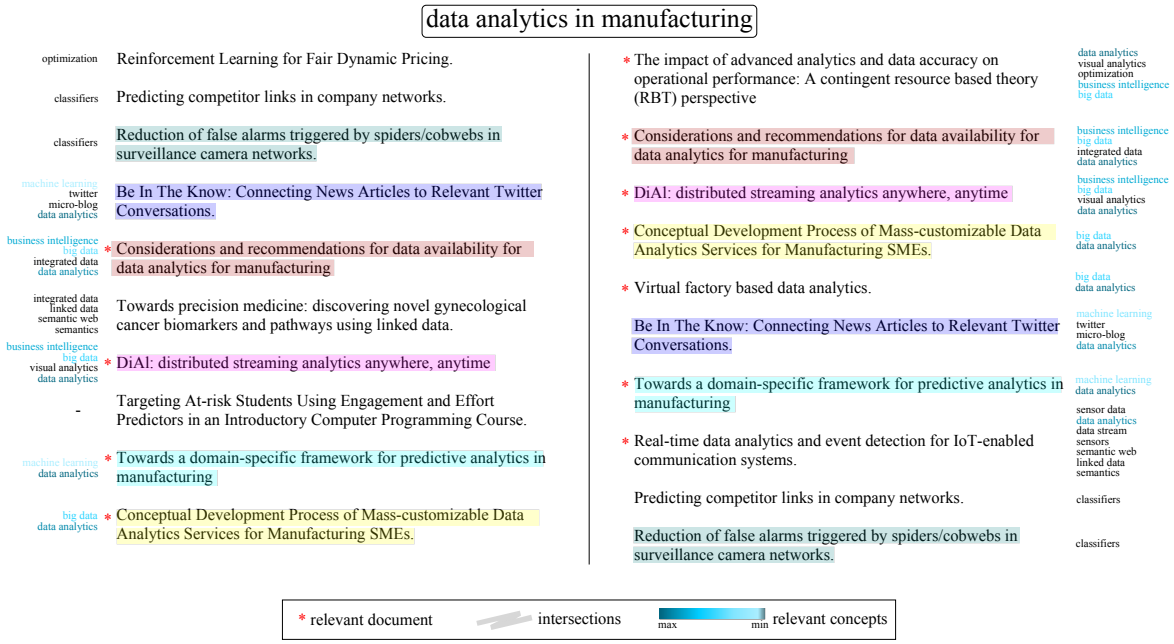


Fig. B.4 Search result example in the qualitative analysis for CESR. Parallel between CESR and BM25 for the proposed query *T 06*. A sample of the distribution of expansions is also shown for each retrieved document (relevant expansions are highlighted).

We report and discuss the observed results on three examples selected among those 20, showing different levels of performance for different characteristic cases.

T05 Frequency Recognition in steady state visually evoked potentials (SSVEPs) based on brain-computer interface signals.

Narrative: looking for the documents related to the latest techniques in frequency recognition in the steady state visually evoked potentials collected by brain computed interface. I expect statistical, optimization, and machine-learning methods to improve detection time or accuracy.

As shown in figure B.3, the performance of both models for this query is the same in terms of P@10. In general, we believe the performance, in this case, is high for both models, which can be because it is an information need easy to satisfy. However, in terms of nDCG@10, the baseline performs slightly better than the proposed method. Paying attention to this, we analyze the re-ranking in this case in terms of the diversity of the annotations and the document frequencies. We found that the standard deviation between frequencies is 5.05, which is indicative of a nonuniform distribution of frequencies and the presence of outliers. In a sample of the top 20 documents, the most frequent concept is present in 17 items. This concept is not characteristic of a specific group or cluster of documents, and its presence can lead to erratic high scores in our re-ranking method.

T06 data analytics in manufacturing.

Narrative: looking for reports that inform me about the use of data analytics methods in the manufacturing sector. I expect particular machine learning or data management tools, including predictive approaches or data storage, used in manufacturing situations.

Figure B.4 shows an example using this query and the parallel between results from the initial model and the results from the re-ranking process. It illustrates the desired behavior from the re-ranking: documents annotated with various relevant concepts are prioritized in the ranking.

T07 Outlier detection based on low-rank and sparse representations.

Narrative: looking for reports that classify signals based on features or elements extracted from low-rank and sparse representations of them.

In this case, we observe a relative improvement in the precision of 30%, which is not only due to the addition of new relevant documents to the top 10 but also the re-ranking of documents that were originally part of it. Specifically, we are bringing up five results that were outside the initial top-10 ranked documents. On average, the number of documents shared by the two ranks is 5.65.

Although these results are limited, they are indicative of the potential performance improvement that can be achieved by enhancing queries and documents. The examples described are representative of the effect of this enhancement in characteristic cases: scenarios like the one illustrated by topic *T 05*, where the approach does not perform better than the baseline, and others where there is a clear difference between the results.

B.5 Supplementary results on SDR

To offer a description of our approach in Chapter 6, we will use two queries: "*Digital assistance*" corresponding to press articles "*Digital assistants like Siri and Alexa entrench gender biases says UN*", and "*privacy*" corresponding to the topic titled "*Apple contractors 'regularly hear confidential details' on Siri recordings*."

The task proposed by SanJuan et al. [95] aims to gather references to concepts that are mentioned in press articles. Figure B.5 shows two examples of queries corresponding to press articles and candidate documents with the relevant content highlighted.

With our approach in Chapter 6, we aim to analyze the distribution of themes the collection can provide for a given query and study the effect of tailoring the search results according to specific patterns. Being able to assign granular topics to the documents, we exploit this tool in further steps.

Figure B.5 is also an illustration of how documents are relevant according to concepts that are used for annotation. In general, we hypothesize that by establishing relevant themes to a given query, the task of content selection could benefit from trimming down the retrieved documents to a more focused or concentrated set of documents.

Figure B.5 shows examples of content selection based on the diversity of themes. For the example queries, we infer the topics highlighted are relevant from its themes distribution feedback. Ideally, the selected content should focus on those topics since dispersed content does not show a clear importance to the task.

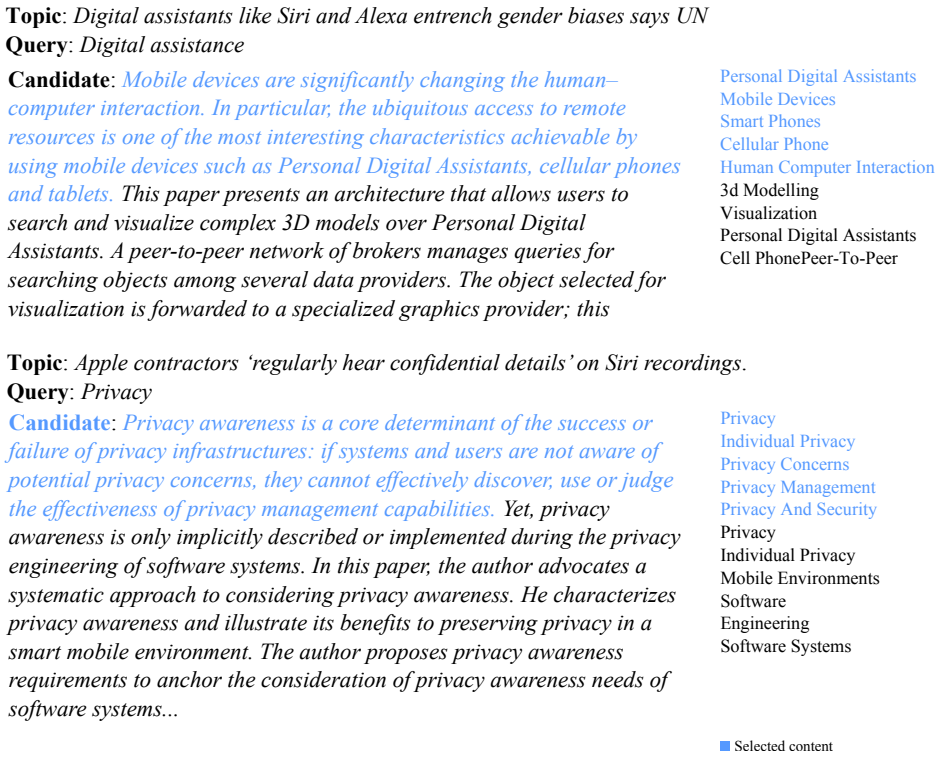


Fig. B.5 Example of content selection for the SDR task. We highlight the relevant content for two examples given query examples. Relevant document expansions are also highlighted.

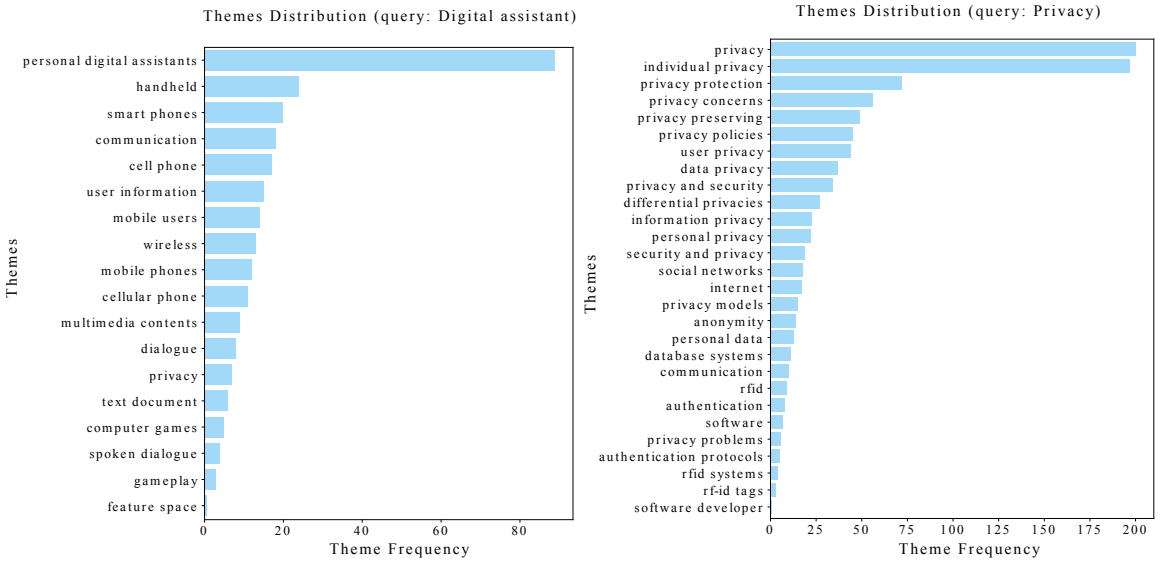
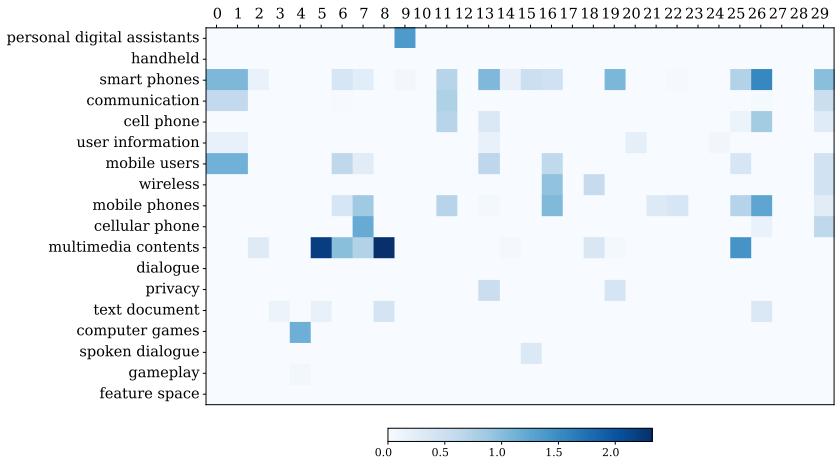


Fig. B.6 Frequency of document expansions on retrieved documents by query examples. BM25 retrieves a set of documents for a given query; since we previously expanded documents with concepts from an ontology, we can measure how frequent expansion terms are in a retrieved set of documents.

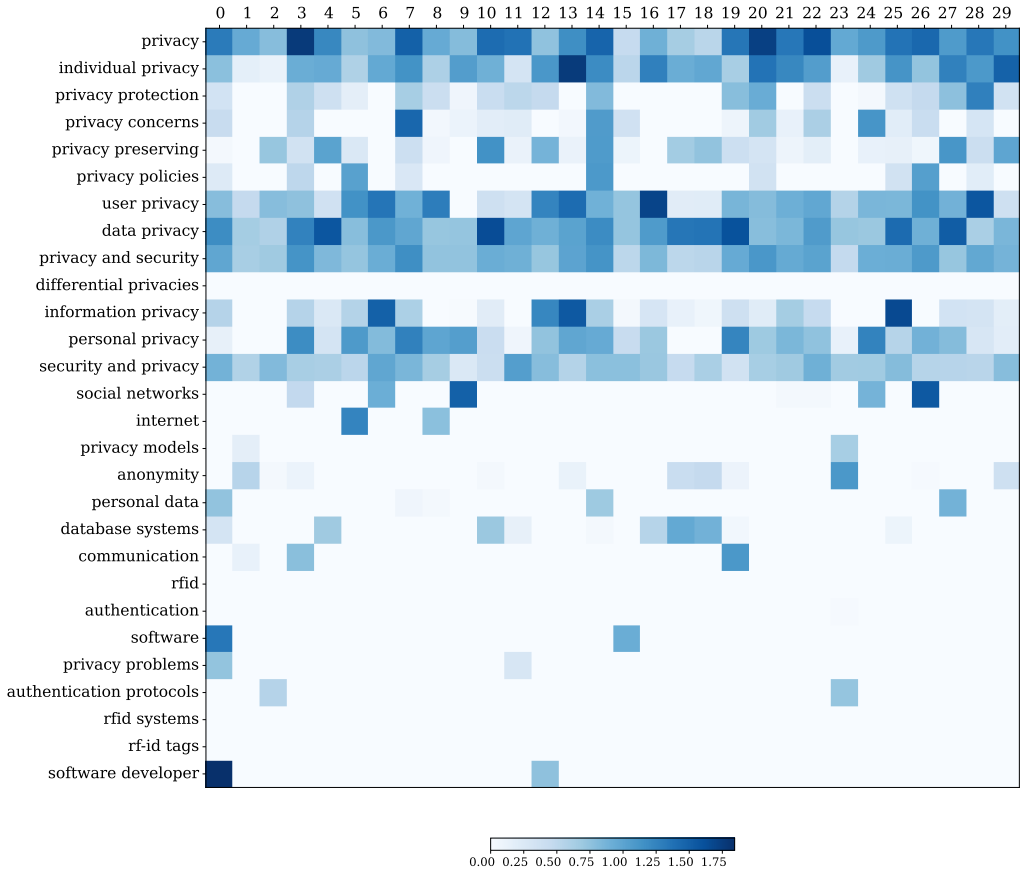
Figure B.6 shows the frequency of themes given a set of retrieved documents with the example queries mentioned previously. The query "*privacy*" matches documents mostly about "*personal digital assistants*," which is a potential keyword-based query for searching within the collection.

Figure B.7 shows heat maps for the topic distribution of the first 30 ranked documents retrieved using the examples queries (each row then shows how likely each document is to be annotated with a theme). These two examples show the contrast of possible results the collection can offer to specific queries.

We can see how documents are relatively diverse by looking at the distributions, such as those described for the example queries. They are ranked around a specific keyword but still exhibit diversity that does not necessarily help to achieve the task of gathering supporting content. Considering this, we then look at themes at a different level of granularity of the retrieved documents to decide whether specific snippets are even more concentrated in the relevant themes.



(a) "Digital assistance"



(b) "Privacy"

Fig. B.7 Expansions scores for documents retrieved with the query examples. Heat maps for the topic distribution of the first 30 ranked retrieved documents