To the Graduate Council:

I am submitting herewith a dissertation written by Maria Mahbub entitled "Towards Machine Reading Comprehension to Advance Biomedical Text Understanding in The Absence of Large-scale Labeled Dataset ." I have examined the final paper copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Computer Science.

Gregory D. Peterson
Major Professor

We have read this dissertation
and recommend its acceptance:

Edmon Begoli

Audris Mockus

Russell Zaretzki

Suzanne Tamang

Accepted for the Council:

Dixie Thompson
Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

TOWARDS MACHINE READING COMPREHENSION TO ADVANCE BIOMEDICAL TEXT UNDERSTANDING IN THE ABSENCE OF LARGE-SCALE LABELED DATASET

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Maria Mahbub

August 2023

Copyright © 2023 by Maria Mahbub
All rights reserved.

DEDICATION

To my parents, brother, and husband, for your unwavering support and encouragement that propelled me on this journey. And to Coco, whose love and company have been my solace during the most challenging times.

ACKNOWLEDGEMENTS

About five years ago, when I embarked on my academic journey at UTK, I had little to no knowledge about the area of machine learning and natural language processing (NLP), let alone their biomedical applications. Little did I know the immense possibilities that lay before me on this path. Reflecting upon those years now, what an incredible and priceless experience it has been for me, working alongside esteemed researchers and witnessing my growth as a research individual! Throughout this time, I have observed the remarkable progress of science and had the privilege of connecting with numerous brilliant minds.

Ph.D. is a unique odyssey, one that I would have been unable to accomplish without the invaluable assistance and support of many individuals, some of whom I would like to acknowledge here.

First and foremost, I would like to convey my deepest appreciation to my advisor Dr. Gregory Peterson, whose guidance and invaluable advice played a pivotal role in the successful completion of my dissertation. His support and mentorship were instrumental throughout the entire process. I am also grateful for his indispensable advice on scientific thinking, writing, and presentation skills.

Another person to whom I would like to express my humble gratitude is Dr. Edmon Begoli, my advisor from ORNL. He provided guidance and support every step of the way, shaping my scientific career in machine learning. His comments and advice on executing research pipelines, performing insightful analyses, writing scientific documents, presenting and effectively explaining complex concepts to non-technical audiences, and time management have been invaluable in navigating my academic expedition. Not only that, my research journey in biomedical NLP started with him at ORNL, which culminated in the completion of this dissertation. He gave me the opportunity to work on incredible projects and introduced me to some intelligent and outstanding research individuals. For that, I am immensely grateful.

Additionally, I would like to extend my heartfelt appreciation to the remaining members of the thesis committee: Dr. Russell Zaretzki, Dr. Audris Mockus, and Dr. Suzanne Tamang. My Ph.D. journey at UTK started with Dr. Zaretzki. I am thankful for his support during the initial year of my challenging transition into a new degree program in an almost entirely unfamiliar discipline and country. Furthermore, I am grateful for his continued support when I changed my academic field from Data Science and Engineering to Computer Science. I would also like to thank Dr. Mockus, whom I have known since my first year at UTK, and Dr. Tamang, who has been one of my primary collaborators for the past two and a half years, for their insightful and invaluable feedback on my research works.

Over the course of my Ph.D., I was fortunate to collaborate and interact with many excellent research professionals. I especially want to mention my colleagues and close collaborators at ORNL: Sudarshan Srinivasan, Ioana Danciu, Ian Goethert, Katie Knight, Alina Peluso, and Laura Davies. Be it pair-coding with Sudarshan or insightful research discussions with Ioana, Sudarshan, or Alina, their contributions have greatly enhanced my skills and helped my growth as a researcher. On many occasions, the consideration and emotional support I received from Ioana, Ian, Katie, Sudarshan, and Laura were irreplaceable in keeping me motivated and resilient. I feel very fortunate to have them as my colleagues.

Special thanks to my extraordinary parents: Mahbuba Begum and Mahbubar Rahman, whose unconditional love and support have shaped my very being. I am thankful for my biggest supporter, my brother: Sazan Mahbub, who always believed in me even when I doubted myself. Finally, I am grateful to my husband and best friend: Samiul Haque, for always being there by my side throughout the hardships of graduate life.

ABSTRACT

Increasing demand for retrieving accurate and timely information from unstructured text data in healthcare motivates the need for effective biomedical machine reading comprehension (MRC) systems. However, high-quality labeled data scarcity in new biomedical problem domains presents an obstacle to training state-of-the-art MRC models, thereby limiting their performance and applicability. Consequently, the application of MRC in the biomedical domain is an under-explored area of research. Employing transfer learning to transmit knowledge to the low-resource biomedical domain from a high-resource domain is a potential solution for limited data availability. Unfortunately, transferring knowledge directly between domains often hurts models' performance due to the domain shift phenomenon.

In this dissertation, we focus on advancing biomedical text understanding using MRC in the absence of large-scale annotated datasets, with particular attention to transfer learning. We study the mitigation of domain shift and propose BioADAPT-MRC, a novel framework designed to implement improved transfer learning in the biomedical domain employing a domain adaptation algorithm based on adversarial learning. BioADAPT-MRC aims to overcome data scarcity challenges and enhance the performance of biomedical MRC. We identify two new problem domains that can benefit from biomedical MRC tasks: MRC on clinical practice guidelines (CPGs) and MRC on clinical notes to extract injection drug use (IDU) information. We facilitate the evaluation and benchmarking of MRC tasks on CPGs by manually building a benchmark dataset called cpgQA. Furthermore, we create a framework named IDU-QA, enabling us to develop a reference dataset and an MRC model capable of retrieving information regarding IDU from clinical notes. We conduct a thorough evaluation of the BioADAPT-MRC framework using MRC datasets sourced from the widely-known BioASQ competition. We tackle the challenge of limited labeled data in the new problem domains by employing different transfer learning settings, along with BioADAPT-MRC. We perform a comprehensive error analysis for each problem domain, highlighting the strengths and limitations of the MRC models. We further discuss future research directions

to enhance the performance and applicability of biomedical MRC models. Overall, this dissertation contributes to the advancement of MRC systems in the biomedical domain, paving the way for improved healthcare interventions and a deeper understanding of critical biomedical information.

TABLE OF CONTENTS

Chapter 1: Introduction **1**

1.1 Contributions 3

1.2 Organization of The Dissertation 4

Chapter 2: Background & Literature Review **7**

2.1 Machine Reading Comprehension 7

2.2 Brief History of MRC Systems' Development 10

2.3 Machine Reading Comprehension Models 11

2.3.1 Question-Context Pair Representations 12

2.3.2 Deep Learning Architectures for Machine Reading Comprehension 13

2.3.3 Machine Reading Comprehension Datasets 22

2.3.4 Transfer Learning 25

Chapter 3: BioADAPT-MRC: Adversarial Learning-based Domain Adaptation Framework for Biomedical Machine Reading Comprehension **39**

3.1 Motivation 40

3.2 Material & Methods 41

3.2.1 Problem Definition 41

3.2.2 BioADAPT-MRC 43

3.3 Results & Discussion 48

3.3.1 Dataset 48

3.3.2 Metrics 50

3.3.3 Method Comparison 50

3.3.4 Experimental Results 53

3.3.5 Analysis 58

3.4 Conclusion 74

Chapter 4: cpgQA: Machine reading comprehension on clinical practice guidelines 75
4.1 Motivation ... 76
4.2 Material & Methods ... 77
4.2.1 cpgQA Dataset ... 77
4.2.2 Machine Reading Comprehension Models ... 84
4.2.3 Transfer Learning with cpgQA ... 86
4.3 Results & Discussion ... 88
4.3.1 Source-domain Datasets ... 88
4.3.2 Experimental Setup and Training Configurations ... 90
4.3.3 Evaluation Metrics ... 91
4.3.4 Experimental Results ... 92
4.3.5 Error Analysis ... 97
4.4 Limitations of cpgQA ... 103
4.5 Conclusion ... 103
Chapter 5: IDU-QA: Machine Reading Comprehension to Extract Information on Injection Drug Use from clinical notes 106
5.1 Motivation ... 107
5.2 Material & Methods ... 112
5.2.1 Problem Formulation ... 112
5.2.2 Reference Dataset Generation ... 112
5.2.3 Modeling with Machine Reading Comprehension System ... 121
5.3 Results & Discussion ... 123
5.3.1 Dataset Statistics ... 124
5.3.2 Experimental Setup ... 125
5.3.3 Metrics ... 125
5.3.4 Results on Reference Test Set ... 127
5.3.5 Temporal Out-of-Distribution Testing ... 128
5.3.6 Error Analysis ... 130
5.4 Conclusion ... 143
Chapter 6: Conclusions & Future Work 144

References 148

Vita 182

LIST OF TABLES

Table 3-1. Statistics of the datasets used in the experiments. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 54

Table 3-2. Scores from BioASQ-9b leaderboard. Here, 'Average' implies the average of mean SAcc, mean LAcc, and mean MRR scores. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 55

Table 3-3. Performance of BioADAPT-MRC compared with the best scores on BioASQ-7b, BioASQ-8b, and BioASQ-9b test sets. The best and the second-best scores are respectively highlighted in bold and italic. '–' indicates that the corresponding source did not report the scores. * denotes previously best-performing method for BioASQ-7B and BioASQ-8B. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 56

Table 3-4. Test scores for ablation experiments of BioADAPT-MRC. The best and the second-best scores are respectively highlighted in bold and italic. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) . 57

Table 3-5. Average test scores with standard deviations (across three different seeds) for the contrastive experiments of discriminator loss functions – triplet loss and distance-based loss. The best scores are highlighted in bold. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 59

Table 3-6. Average test scores with standard deviations (across three different seeds) for experiments using synthetic questions and masked questions in the target-domain training dataset. The best and the second-best scores are respectively highlighted in bold and italic. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 61
Table 3-7. Average test scores with standard deviations across experiments with three random seeds (10, 42, 2018) for initialization to measure and compare the stability of BioADAPT-MRC. The best and the second-best scores are respectively highlighted in bold and italic. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 62
Table 3-8. Test scores for experiments using varied ratios of labeled target-domain training data. The best and the second-best scores are respectively highlighted in bold and italic. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 65
Table 3-9. Mean accuracy and mean silhouette scores with standard deviations for DBSCAN clustering on the MDS components. The MDS dimensionality reduction technique was applied to reduce the dimension of the $[CLS]$ token representations for the samples in the BioASQ test sets and a set of random samples from the SQuAD training set. A $[CLS]$ token represents the accumulated representations of a sample in the dataset. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 66
Table 4-1. Comparison among cpgQA, BioASQ, and emrQA in terms of context, question, and answer lengths. (©2023 IEEE. Reprinted from [272].) 83
Table 4-2. Specifications of nine pre-trained language models used in the experiments. (©2023 IEEE. Reprinted from [272].) 89
Table 4-3. Number of target-domain training and testing samples (i.e., question-answer pairs) in the folds created for 5-fold cross-validation. (©2023 IEEE. Reprinted from [272].) 89

Table 4-4. Statistics of the source-domain datasets used in the experiments. (©2023 IEEE. Reprinted from [272].) 94
Table 4-5. Test scores on cpgQA in the unsupervised setting with nine pre-trained language models and three source domains. The highest and the second highest scores are highlighted in bold and italic, respectively. (©2023 IEEE. Reprinted from [272].) .. 94
Table 4-6. Test scores on cpgQA in the unsupervised setting with BioLinkBERT-SQuAD and BioADAPT-MRC. The highest scores are highlighted in bold. (©2023 IEEE. Reprinted from [272].) 96
Table 4-7. Test scores on cpgQA in the supervised setting. The highest scores are highlighted in bold. (©2023 IEEE. Reprinted from [272].) 99
Table 5-1. Types of information about IDU that are most likely to be inquired by clinicians from the clinical notes, categorized into nine query groups. Abbreviations: ivdu, intravenous drug use; bnz, benzodiazepines. (Reprinted from [281].) 115
Table 5-2. A list of IDU keywords/phrases provided by SMEs. Abbreviations: ssp, syringe services programs; ivda, intravenous drug abuse; ris4e, resists infection by sterile syringe safe sex and education; PWID, people who inject drugs. (Reprinted from [281].) 117
Table 5-3. A detailed list of co-occurring phrases before/after IDU keywords along with the parsing rules, targeted pattern type, and examples of answers before and after parsing. (Reprinted from [281].) 120
Table 5-4. Mappings between the query groups and the words in reference answers most likely to provide inquired information and example answers for each mapping. (Reprinted from [281].) 122
Table 5-5. Statistics of the reference dataset. 126
Table 5-6. Performance scores of baseline MRC models on the test set. The highest scores are highlighted in bold..... 129

Table 5-7. Performance scores of the best baseline (ClinicalBERT), sequential learning, and adversarial learning–based MRC models on the test set. The highest scores are highlighted in bold. 129
Table 5-8. Statistics of the additional test datasets built using Cohort-Short and Cohort-Long. The highest scores are highlighted in bold. 131
Table 5-9. Performance scores of the best baseline (ClinicalBERT), sequential learning, and adversarial learning–based MRC models on the additional test datasets built using Cohort-Short and Cohort-Long. 133
Table 5-10. Performance scores with 95% confidence intervals adversarial learning–based MRC models on the test set in the reference dataset, test datasets from Cohort-Short and Cohort-Long. 134
Table 5-11. Examples of answers predicted by the MRC model along with the question and the reference answers. Abbreviations: ble, bilateral lower extremities. 142

LIST OF FIGURES

Figure 2-1. An example context-question-answer triplet from a machine reading comprehension (MRC) dataset presented in [27]. 9
Figure 2-2. Categories of machine reading comprehension (MRC) tasks according to the classification presented in [28]. 9
Figure 2-3. Sub-domains in biomedical machine reading comprehension tasks according to the classification presented in [10]. 23
Figure 2-4. Different cases for transfer learning according to the classification presented in [24]. 27
Figure 2-5. Two-dimensional visualizations of vector representations and the distributions of the features for the samples from two domains with (a) similar and (b) different topic distributions. . 29
Figure 3-1. BioADAPT-MRC: an adversarial learning-based domain adaptation framework for biomedical machine reading comprehension task. The framework has three main components: (i) feature extractor M_F, (ii) MRC-module M_Q, and (iii) domain similarity discriminator D. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 42
Figure 3-2. Per-epoch cosine distance between source-domain training sample pairs and source- and target-domain training sample pairs across 50 epochs. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 59
Figure 3-3. Synthetic questions generated from the first context of the BioASQ-9b training set using the *question-generator* component of the AdaMRC model along with the context. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 61

Figure 3-4. Multidimensional scaling (MDS) plot for the features (extracted from the baseline (left) and the BioADAPT-MRC (right) models) of all samples from the BioASQ-9b test set (blue) and a set of random samples from the SQuAD training set (yellow). (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 65
Figure 3-5. Mean accuracy scores (left) and mean silhouette scores (right) with standard deviations for DBSCAN clustering on BioASQ test sets and SQuAD. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 66
Figure 3-6. Motivating example from the BioASQ-9b test set. The left and right plots show the word importance for answer span prediction given by the baseline and BioADAPT-MRC models. The green and red colors show positive and negative word importance, respectively. The density of color shows the amount of importance. The highlighted word in yellow represents either the predicted start or end tokens. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 68
Figure 3-7. Error analysis of BioADAPT-MRC, in comparison with the baseline model, depending on the types of questions in the BioASQ test sets. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 70
Figure 3-8. Error analysis of BioADAPT-MRC, in comparison with the baseline model, depending on the types of answers in the BioASQ test sets. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 71
Figure 3-9. Example question-answer pairs from the test sets demonstrating the strengths and weaknesses of the BioADAPT-MRC model over the baseline model. The green and red colors show correctly and incorrectly predicted answers, respectively. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].) 72

Figure 4-1. Example context and question-answer pair from the cpgQA dataset. (©2023 IEEE. Reprinted from [272].) 79
Figure 4-2. Distribution of question types in the cpgQA dataset. (©2023 IEEE. Reprinted from [272].) 81
Figure 4-3. Distribution of question types in each component of the clinical practice guideline used for cpgQA in the cpgQA dataset. (©2023 IEEE. Reprinted from [272].) 82
Figure 4-4. Distribution of context, question, and answer lengths (in number of words) in the cpgQA dataset. (©2023 IEEE. Reprinted from [272].) 82
Figure 4-5. A two-dimensional visualization of vector representations of question-context pairs using PCA. The colors represent the three sub-domains in biomedical-MRC: clinical practice guidelines (cpgQA), clinical notes (emrQA), and scholarly articles (BioASQ). (©2023 IEEE. Reprinted from [272].) 85
Figure 4-6. Drop in EM scores (top) and F1 scores (bottom) after direct knowledge transfer from the source domains to the target domain. (©2023 IEEE. Reprinted from [272].) 95
Figure 4-7. Exact match (top) and F1 (bottom) scores with confidence intervals for five folds of cpgQA in the supervised setting. (©2023 IEEE. Reprinted from [272].) 98
Figure 4-8. Error analysis for MRC modeling of cpgQA, considering the types of components in the guideline in the answers. (©2023 IEEE. Reprinted from [272].) 99
Figure 4-9. Error analysis for MRC modeling of cpgQA, considering the types of interrogative words/phrases in the questions in the answers. (©2023 IEEE. Reprinted from [272].) 100
Figure 4-10. Error analysis for MRC modeling of cpgQA, considering the length (number of words) in the answers. (©2023 IEEE. Reprinted from [272].) 101
Figure 4-11. Percentage of test samples which do not have an exact match but an overlap between words in the predicted and ground truth answers. (©2023 IEEE. Reprinted from [272].) .. 104

Figure 5-1. A sample clinical progress note featuring questions about the patient's IDU behavior with extracted IDU-related information color-coded in the note. (Reprinted from [281].) .. 111
Figure 5-2. IDU-QA: Framework for a two-part study for extracting information on IDU behavior in patients from clinical notes. The first part consists of reference dataset generation in three primary steps, and the second part consists of MRC model development from implementation to inference. (Reprinted from [281].) 113
Figure 5-3. Distribution of the query groups in the reference dataset 126
Figure 5-4. Distribution of query groups in the additional test dataset built using Cohort-Short..... 132
Figure 5-5. Distribution of query groups in the additional test dataset built using Cohort-Long. 132
Figure 5-6. Effect of context (note) length on the MRC model's performance 136
Figure 5-7. Effect of question length on the MRC model's performance.... 137
Figure 5-8. Effect of reference answer length on the MRC model's performance 138
Figure 5-9. Performance of the MRC model for each query group 139
Figure 5-10. Distributions of the ratios of the predicted answers to the contexts and the ratios of the reference answers to the contexts. The ratios here are presented in the form of percentages. 141

CHAPTER 1
INTRODUCTION

Healthcare professionals typically pose a minimum of one question for every two patients they encounter during the consultation stage of patient care [1]. These questions can be regarding the cause of a symptom, the tests to be performed, the drug of choice for a particular condition, the dose of a drug, or ways to manage a particular condition. Although they are able to find answers to around 78% of these questions, they often abandon half of their inquiries due to limited time and the uncertainty that helpful answers may be unavailable, despite the presence of abundant evidence [1, 2]. Furthermore, to enhance the quality of patient care, it is crucial for clinicians to stay informed about the latest information on diagnosis, prognosis, treatment, recommendations, and the risks and benefits of treatment within the framework of Evidence-based Medicine (EBM) [3, 4].

Because of time constraints, searching through existing resources to find high-quality, trustworthy, and pertinent information is inconvenient for clinicians – leading to the dependency on easily accessible simple-to-use general-information electronic resources like Google [5]. However, this reliance on general-purpose search engines poses a problem because their result ranking criteria do not align with the principles of EBM, lacking reliability, quality, and precision [5]. This problem is partially alleviated by traditional information retrieval (IR) systems [6–10]; however, a healthcare-information professional still requires approximately four hours to navigate and locate answers within complex biomedical resources [9].

Another information source used by healthcare professionals to provide optimized patient care is clinical practice guidelines (CPGs). These guidelines are documents with statements of recommendations developed through a rigorous review of the cutting-edge research evidence and risks and benefits evaluation of alternative care options [11]. Although these documents are concise compared to reviewing numerous literature resources, interpreting them is often deemed to be tedious and complicated by healthcare professionals [12]. Therefore, an efficient

approach is necessary to help clinicians comprehend these guidelines and save valuable time.

Furthermore, when making crucial decisions regarding comprehensive patient-centered care, it is vital to integrate information from unstructured clinical notes within electronic health records. Clinical notes encompass the comprehensive progression of patients' health conditions with different treatment measures. They provide a higher level of detail compared to structured data. While there may potentially be some intersection, these notes often capture additional information typically excluded in structured data [13–16]. Extracting beneficial information from these long clinical notes necessitates an efficient computational approach.

Machine reading comprehension (MRC), a crucial task in machine learning and natural language understanding – a subfield of natural language processing (NLP), facilitates the understanding of input text passages by machines and enables them to respond accurately to user queries with extracted relevant information (i.e., answers) from the input text, typically within seconds. MRC with input text passages from the biomedical domain is known as biomedical MRC. Unlike traditional IR systems that typically present users with a collection of documents for interpreting and finding precise answers, intelligent biomedical MRC systems offer a time-saving and effort-reducing alternative [10]. Despite the recent progress in biomedical MRC systems, their application to the clinical practice guidelines still needs to be explored. With regard to extracting valuable information from clinical notes, unlike other NLP approaches with pre-defined sets of rules [13, 17–19], MRC systems enable users to retrieve information with increased flexibility since the information extracted by MRC systems is not restricted by pre-defined rules, but the questions provided by the user [10].

With advancements in computing hardware and the availability of efficient resources, researchers have made significant progress in developing state-of-the-art MRC systems. These systems employ neural networks and demonstrate performance comparable to, or even surpassing human-level capabilities in various domains, including, but not limited to, Wikipedia articles and web search results [20–22]. However, achieving such impressive results relies heavily on training datasets that are high-quality, large-scale, and meticulously annotated by humans [21]. These MRC datasets comprise triplets containing context, question, and answer elements, where the question-answer pairs serve as labels within the dataset.

The need for large-scale, high-quality human-annotated datasets presents a significant challenge when it comes to applying MRC systems in the biomedical domain. Creating a substantial number of reliable question-answer pairs for a new biomedical domain requires expertise in the specific subject matter, making the process both costly and time-consuming.

The undeniable need for subject-matter expertise in building MRC datasets for new biomedical problem domains leads to non-existent or smaller datasets, resulting in poor performance on the MRC task [23]. Hence, despite the numerous potential application domains, the utilization of MRC systems in the biomedical field remains constrained.

Given the inherent challenges, it is of utmost importance to devise an approach that optimally leverages small-scale labeled or unlabeled biomedical MRC datasets to train the MRC models and bolster their performance. In such scenarios, transfer learning emerges as a valuable technique that facilitates the transfer of knowledge to the low-resource domain, known as the target domain, from a high-resource domain, known as the source domain [24]. High-resource domains are typically extensively explored and built using open-source text passages requiring no domain expertise. However, the direct transfer of knowledge between domains often leads to subpar performance due to discrepancies in the feature representation caused by domain shift.

This dissertation seeks to develop MRC systems for understanding and extracting information from complex biomedical texts with improved performance by incorporating unlabeled or relatively small quantities of labeled biomedical texts and harnessing the power of transfer learning.

1.1 Contributions

The contributions described in this dissertation include the following:

- A new MRC framework, BioADAPT-MRC, implemented with a domain adaptation algorithm based on adversarial learning. It comprises three key components: a feature extractor, a domain similarity discriminator equipped with an auxiliary task layer, and an MRC module. BioADAPT-MRC aims to minimize the shift between two domains: low-resource biomedical (target) and high-resource general-purpose (source). It is designed to enhance the

performance of MRC systems in the biomedical domain by reducing the domain shift and thus effectively addressing data scarcity challenges.

- Implementations of BioADAPT-MRC in unsupervised, semi-supervised, and supervised settings.
- Introduction of two new applications of biomedical MRC.
- A benchmark MRC dataset, cpgQA, for the MRC task on clinical practice guidelines, which is generated by a manual annotation process utilizing support from subject-matter experts.
- An MRC framework, IDU-QA, to extract information regarding injection drug use from patients' clinical notes. The framework integrates rule-based NLP techniques for generating a reference MRC dataset on clinical notes with injection drug use for model development.
- MRC models for comprehending and extracting information from clinical practice guidelines and clinical notes with injection drug use information.
- Extensive validation of the BioADAPT-MRC framework along with other transfer learning approaches on biomedical scholarly articles, clinical practice guidelines, and clinical notes with injection drug information.

1.2 Organization of The Dissertation

The remaining chapters of the dissertation are structured as follows.

Chapter 2 provides the necessary background material and related work on machine reading comprehension and transfer learning.

Chapter 3 describes the proposed BioADAPT-MRC framework, designed for effective domain adaptation in the biomedical field. This framework uses adversarial learning to enhance MRC models' adaptation capability to new biomedical problem domains. The domain adaptation is achieved by incorporating a domain similarity discriminator within the adversarial learning framework. The discriminator, along with the feature extractor, engages in an alternating or joint training process, resulting in the generation of feature representations with a reduced domain gap. Reduced domain gap refers to the reduced distance between the target and source domain feature representations in the embedding space. In contrast to recent approaches of domain adaptation that rely on generating

synthetic (machine-generated) labels for the purpose of augmenting training data [25, 26], our framework leverages either a limited number of labeled data or unlabeled data or both from the target biomedical domain. This unique characteristic renders BioADAPT-MRC more appropriate for scenarios where annotated datasets are scarce and synthetic label generation is resource-intensive and requires additional verification by domain experts. We extensively evaluate BioADAPT-MRC on three well-established datasets sourced from the BioASQ challenge [27]. The evaluation is performed using the recommended metrics provided by BioASQ. Furthermore, we assess the domain adaptation capability of our framework through dimensionality reduction techniques and clustering. To provide a comprehensive understanding of our framework, we end the chapter with an in-depth analysis, highlighting its strengths and limitations. This analysis demonstrates the effectiveness of BioADAPT-MRC while shedding light on areas that require further improvement. Overall, our framework represents a significant advancement in domain adaptation for biomedical MRC models, with promising potential in real-world applications.

Chapter 4 describes the pioneering benchmark dataset called cpgQA, designed explicitly for biomedical MRC on clinical practice guidelines (CPGs). This dataset, consisting of 1097 samples, is meticulously constructed and validated by subject-matter experts. We demonstrate the importance of a benchmark MRC dataset tailored to CPGs through rigorous experimentation. In particular, we conduct an extensive case study utilizing BioADAPT-MRC and various transfer learning settings with transformer-based MRC models. Furthermore, we provide a comprehensive error analysis that sheds light on the capabilities and limitations of the MRC model, specifically when applied to CPGs. This analysis offers crucial insights regarding the performance and limitations of the model, paving the way for future enhancements and research directions in the field. We conclude the chapter by discussing the limitations inherent in the cpgQA dataset. We outline the areas that require improvement and highlight the scope for future research endeavors to further enhance the quality and applicability of the dataset, opening doors for more sophisticated and effective biomedical MRC models for CPG interpretation.

Chapter 5 presents a comprehensive description of the IDU-QA framework designed to extract information about injection drug use (IDU) from clinical notes. Given the absence of evidence of IDU in structured electronic health record (EHR)

data, relying on unstructured clinical notes becomes crucial for understanding the representation of this phenomenon and improving upon NLP techniques. The primary objective of this study is to formulate the task of IDU information extraction as an MRC problem. By treating it as such, we allow for the extraction of answers that are not predetermined labels or entities, providing greater flexibility in identifying relevant information. The interdependencies and associations between the words in the answers are not constrained, enabling a more nuanced understanding of patients' IDU behavior. We consider the challenge of annotated data availability necessary to develop MRC models and, therefore, build the IDU-QA framework comprising two main parts: reference dataset generation and MRC modeling. In this chapter, we describe the techniques employed to generate the dataset. Subsequently, we present an extensive study leveraging transformer-based MRC models, including the implementation of BioADAPT-MRC and other transfer learning settings tailored to accommodate the relatively small dataset. To acquire a deeper understanding of the MRC models' performance and limitations, we end the chapter with a thorough error analysis serving as a valuable resource for guiding future research efforts in this problem domain. By developing this specialized MRC framework for IDU extraction from clinical notes, we contribute to advancing the understanding and application of NLP techniques in capturing critical information related to IDU. This research paves the way for improved healthcare interventions and a deeper understanding of the population affected by IDU.

Finally, we conclude in Chapter 6 by delving into a comprehensive analysis of our work, accompanied by a discourse on potential avenues for future research directions.

CHAPTER 2
BACKGROUND & LITERATURE REVIEW

This dissertation addresses the well-known challenge of limited, labeled data availability in biomedical machine reading comprehension (MRC) tasks. As mentioned in Chapter 1, to overcome the data scarcity issue, we focus on utilizing transfer learning and leveraging either unlabeled narrative texts or a relatively small number of labeled narrative texts from the biomedical domain. Our approach aims to enhance the performance of biomedical MRC models. Additionally, we introduce two novel applications of biomedical MRC. The first application involves MRC on clinical practice guidelines (CPGs) to extract crucial information regarding treatment recommendations, risk factors, and more. The second one involves MRC on clinical notes to extract information on patients' injection drug use (IDU) behavior.

Therefore, our work revolves around the intersection of three key research areas: MRC models, MRC datasets, and transfer learning, with a specific focus on the biomedical domain. This chapter provides relevant background information and a literature overview of recent advancements in these three research directions.

2.1 Machine Reading Comprehension

Machine reading comprehension is an essential and intricate task in machine learning and natural language understanding. Its primary objective is to comprehend textual passages and effectively retrieve pertinent information encoded within them. MRC aims to answer posed questions based on the understanding of textual narratives. Similar to the objective of reading comprehension performed by humans, the MRC task aims to evaluate the extent to which computer systems comprehend and interpret human language effectively. The input for an MRC model is a question and a piece of text excerpt known as the context, and the output is the corresponding answer. Consequently, an MRC dataset comprises numerous context-question-answer triplets (refer to Figure 2-1).

According to [28], MRC tasks can be categorized along four dimensions (Figure 2-2): (i) corpus types (textual/multi-modal), (ii) question types (natural/cloze/synthetic), (iii) answer types (natural/multi-choice), and (iv) answer sources (span/free-form).

(i) **Types of corpus:** Depending on the data types used in the task, there are two groups into which MRC tasks can be classified: textual, where the data is unstructured textual form, and multi-modal, where the data includes both image and text [28].

(ii) **Types of questions:** Depending on the types of questions, there are three main MRC task categories: natural form, cloze style, and synthetic form [28]. In *natural form*, the questions are usually interrogative sentences with "wh" question phrases. Sometimes, they can be imperative sentences such as 'Name a CFL2 mutation ...'. In *cloze style*, a question usually contains a sentence containing a placeholder. The sentence can be of any form, such as imperative, declarative, etc., and not necessarily interrogative. This type of MRC task aims to find an answer to fill the placeholder in the question correctly. For example, 'Single-agent activity for ____ reflected by response rates of 10%-30% has been reported.' In *synthetic form*, a question can be a list of words without conforming to any grammatical rules [29, 30]. For example, 'located next to body of water' [30].

(iii) **Types of answers:** Based on the types of answers, there are two types of MRC tasks: natural form and multiple-choice [28]. In the context of multiple-choice MRC tasks, the answer is usually a single or a group of phrases or an image (if multi-modal) selected from a collection of candidate answers, whereas, in natural form, there is no choice given for the answers.

(iv) **Sources of answers:** Depending on the answer sources, there are two types of MRC tasks, MRC tasks with either span answers or free-form answers [28]. In span form, the source of answers is the provided contexts where a span taken verbatim from the context is considered an answer. In contrast, the free-form answers allow for phrases or groups of phrases or images (if multi-modal) of free form as the answers, not necessarily taken from the contexts, and with no limitation in their possible sources.

Context: The aim of this study was to evaluate the presence of autoantibodies to cyclic citrullinated synthetic peptides (ACPAs) in the sputum of patients with long-standing rheumatoid arthritis.

Question: What disease is associated with Anticitrullinated peptide antibodies (ACPAs)?

Answer: rheumatoid arthritis

Figure 2-1. An example context-question-answer triplet from a machine reading comprehension (MRC) dataset presented in [27].

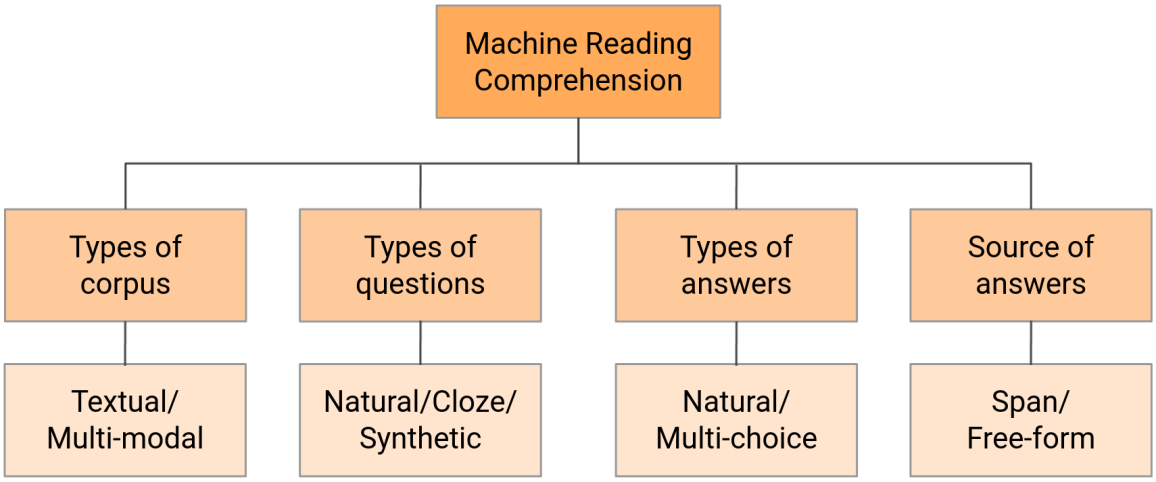


Figure 2-2. Categories of machine reading comprehension (MRC) tasks according to the classification presented in [28].

Our work focuses on MRC tasks involving a textual corpus, natural-form questions that are interrogative or imperative sentences, and natural answers in words or phrases that can be directly extracted from the context – also known as extractive MRC. In order to eliminate redundancy, throughout the remainder of this dissertation, we use the term "MRC" to refer specifically to "extractive MRC".

2.2 Brief History of MRC Systems' Development

MRC systems were first conceptualized in the 1970s, and the most noteworthy MRC system was first proposed in [31]. The authors developed an MRC system, QUALM, to understand the context in a hand-written story script and answer questions about that story [31]. QUALM's domain-specificity and small scale restricted further general application of this system. Research in the area of MRC was stagnant in the 1980s and 1990s because of the complexity of the task and unavailability of datasets [32]. In the late 1990s, Hirschman et al., in [33], revitalized the research in MRC by releasing a dataset consisting of stories from the third- to sixth-grade reading materials and of five "wh" questions.

The MRC systems, during the late 90s, were primarily based on traditional machine learning algorithms and rule-based heuristic techniques [33]. Hirschman et al., in [33], presented a bag-of-words-based approach that can answer questions from stories by identifying the best match between the questions and the stories. In another work, Riloff et al. developed Quarc [34], a rule-based MRC system that can answer questions by finding the most relevant sentence in a short story using semantic and lexical clues such as entity recognition, part-of-speech tagging, and morphological analysis, in the story and the question. The MRC systems developed in [35] using a machine-learning-based combined approach that unifies Markov logic, bootstrapping, and self-supervision.

The approaches based on rules and traditional machine learning exhibited unsatisfactory performance in contributing to real-world applications effectively. The reasons behind the low performance were the presence of inherent limitations of these approaches, such as the absence of long-range dependencies and contextual information, the non-generalizability of these approaches to other datasets, the unavailability of large-scale MRC datasets on a variety of articles, and the considerable amount of human effort required to build the datasets [21, 32, 36].

Between 2013 and 2015, the introduction of labeled MRC datasets, such as MCTest [37] and ProcessBank [38], addressed some of the issues mentioned above and thus revolutionized the research in MRC and transformed the formerly complex MRC task into a supervised learning problem [39]. Another driving force to this revolution was neural network-based (NN-based) deep learning MRC models, which exhibited their excellence in overcoming the inherent shortcomings of the previous MRC systems, such as long-range dependencies and contextual information [36]. With the combination of labeled MRC datasets, these NN-based systems outperformed the traditional approaches for solving the MRC task [20, 40]. Rajpurkar et al. made a groundbreaking contribution in 2016 by introducing the Stanford Question Answering Dataset or SQuAD, which served as the first large-scale MRC dataset [21]. The SQuAD dataset was built using Wikipedia articles, where the text passages in those articles are considered as the contexts in the dataset. The question-answer pairs, on the other hand, were annotated by crowd workers, and the answers in the dataset are extracted slices from the corresponding text passages. In recent years, the advancement of deep learning MRC models has also propagated to several specialized domains, such as the biomedical domain [10]. However, because of the need for subject-matter expertise to build high-quality datasets in the biomedical domain, there is a scarcity of large-scale datasets [10]. The lack of large-scale datasets elicits the need for transfer learning [24], a learning technique facilitating knowledge transfer from one domain to another. As mentioned earlier, our focus for this work is mainly 2-fold: applying state-of-the-art models to perform biomedical MRC tasks and leveraging transfer learning to enhance the models' performance. Therefore, in the remaining part of this chapter, we review deep learning-based MRC models and MRC datasets in the biomedical domain, along with transfer learning approaches in the MRC task focusing specifically on the biomedical domain.

2.3 Machine Reading Comprehension Models

There are three steps in solving an MRC task using deep learning architectures: creating the input vector representations of the question-context pairs using word embeddings techniques, extracting the features from the embeddings using the deep learning architectures, and predicting the answer span based on the provided context [32]. The remaining section briefly reviews research works associated with state-of-the-art techniques to achieve these steps.

2.3.1 Question-Context Pair Representations

In deep learning-based MRC models, the input question-context pairs are represented as continuous high-dimensional vector representations [20, 41–43]. In order to convert every question-context pair from its discrete textual format to a continuous vector representation, a three-step procedure is employed: tokenize the sentences(s) in the question-context pair into words or sub-words (also known as tokens), look up or train the vector embeddings of the tokenized words, and concatenate the word embeddings to represent the input question-context pair.

There are several word embeddings algorithms to transform words into vector embeddings, among which the most commonly used ones are: word2vec [41], GloVe [42], FastText [43], and Bidirectional Encoder Representations from Transformers (BERT) [20]. These aforementioned word embeddings algorithms, except BERT, represent each word in a fixed vector representation for all contexts and meanings. They leverage the words' global and local co-occurrences and morphological representations to generate word embeddings [42, 43]. Thus, they cannot encapsulate the contextual meaning of the words in a body of text. A word's contextual meaning pertains to its vectorized representation which encompasses the information surrounding that specific word.

On the other hand, BERT generates contextual word embeddings using a pre-trained bidirectional multi-layer transformer network, capturing the contextual meaning of the words and thus providing a more meaningful representation of the question-context pair [20]. Pre-training refers to the usual training process of a neural network in preparation to perform a task in the future. Consequently, MRC models trained with BERT-generated word embeddings outperform the others [20, 44, 45]. Some other examples of word embedding algorithms that generate contextual embeddings are Embeddings from Language Models (ELMo) [46] and Generative Pre-Training (GPT) [47]. ElMo generates word embeddings using a pre-trained bidirectional long short-term memory (Bi-LSTM) language model. Nevertheless, these embeddings are restricted by the limitations of LSTM. GPT generates word embeddings using a pre-trained unidirectional multi-layer transformer architecture. However, the vital shortcoming of GPT architecture is that due to being unidirectional (from left to right), it lacks the ability to capture contextual information from both directions. Thus, GPT is often unable to perform as well as other bidirectional architectures, such as BERT, in tasks where

incorporating contexts from both directions is crucial [20]. Given the resounding success of BERT over other embedding techniques in MRC tasks [48, 49], for our work, we restrict our focus only to BERT embeddings.

In addition to word embeddings, researchers have shown interest in including syntactic, semantic, and morphological information about the question-context pairs such as named entities, part-of-speech, and character-level embeddings [40, 50–54]. To solve an MRC task, Indurthi et al. obtained character-level word embeddings using convolutional neural networks (CNNs) [54]. In these embeddings, each character in each word has a fixed-dimensional embedding. In another work [53], Xu et al. used the counts of different part-of-speech tags, such as nouns, adjectives, and verbs, and named-entity-recognition tags, such as location, organization, and person, for each question-answer pair.

2.3.2 Deep Learning Architectures for Machine Reading Comprehension

There are four primary NLP deep learning architectures for the MRC task: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), networks with the attention mechanism, and transformer-based pre-trained language models (PLMs).

Convolutional Neural Networks (CNNs) Although CNNs are primarily used in computer vision applications [55], they have been shown to perform well in several NLP tasks [56]. The CNN architecture uses multiple layers of convolutional kernels or filters to extract spatial features from an image or a piece of text [56]. In an MRC model, researchers have used CNNs mainly to generate word embeddings [57] and mostly character-level embeddings [40, 54].

Chaturvedi et al., in [57], used CNN to perform MRC tasks. They employed CNN to generate word embeddings from the questions and the contexts where the CNN consists of one convolutional layer followed by one average pooling layer. In [54], Indurthi et al. proposed a CNN-based MRC model – ConZNet. Instead of considering MRC as a span prediction task focused on identifying the answer span with the highest probability score, they considered multiple spans need to be included in the correct answer. ConZNet uses a CNN architecture for generating character-level embeddings and then concatenates them with the pre-trained word embeddings for the questions and contexts. The goal here is to extract regions

in the context relevant to the question and consequently produce an answer from them.

Recurrent Neural Networks (RNNs) RNN is the class of architecture that can model sequence data such as text or speech by including relative historical information in the sequence, emulating neural activity in the human brain [58]. Prior to the success of attention-based networks, RNNs were the most popular and widely used deep learning architecture choices among researchers for solving MRC tasks [58].

RNNs have mainly two architectural varieties: (i) Long Short-Term Memory (LSTM) [59] and (ii) Gated Recurrent Unit (GRU) [60]. Both variants of RNNs overcome the vanishing gradient issue [61, 62] experienced by traditional RNNs. The vanishing gradient problem refers to a phenomenon that arises when training deep neural networks, such as RNNs. It occurs when the gradients utilized for updating the network diminish significantly or "vanish" as they propagate backward from the output layers to the earlier layers, consequently leading to minimal impact or influence on the network's learning process [63]. Originally the modeling of RNN and its varieties was restricted by the temporal order of the sequence data (from left to right) [58]. Thus, no future sequential data could be seen at the present time step. However, realizing the importance of both past and future inputs in a sequence for modeling the current one, researchers introduced the concept of bidirectional RNNs where both past and future inputs in the sequence are taken into account while modeling the current one [64]. Similar to the use cases of CNNs in MRC models, researchers have used RNNs as feature extractor component [40, 65] that captures the temporal interactions between the word embeddings and between the question-context pairs to create question-context pair representations [66, 67].

Chen et al., in [65], used a multi-layer Bidirectional-LSTM (Bi-LSTM) to encode the question-context pairs. The Bi-LSTM model takes a feature vector for each token in each context consisting of a 300-dimensional GloVe word embedding; a 3-dimensional binary feature reflecting the match between a token in the context and any tokens in the question in its lowercase, original, and lemma format; a multidimensional vector reflecting the properties related to normalized term frequency, named-entity-recognition, part-of-speech; and an attention score

depicting the similarity between a token in the context and the tokens in the questions. The Bi-LSTM also takes the 300-dimensional GloVe word embeddings of the tokens in the questions. The authors then trained two different classifiers for predicting the first and last tokens within the answer span.

In [40], Seo et al. used LSTM to extract features from question-context pairs. The LSTM network takes character-level embeddings generated by CNN and pre-trained GloVe word embeddings. The authors employed two LSTM networks in both directions for capturing the temporal interactions between the words in the sequence.

Wang et al., in [68], proposed a Multi-Perspective Context Matching (MPCM) model for the MRC task. The model uses a Bi-LSTM network to generate character-level embeddings for question-context pairs. The authors calculated matching vectors representing the comparison between the embeddings of the question and the context from multiple perspectives. They also used another Bi-LSTM network for aggregating the matching vectors and for predicting the first and last tokens within the answer span.

In [69], Kobayashi et al. introduced a neural network MRC model with a Bi-LSTM network for generating entity representations to improve the performance of the MRC task. In another work [70], Clark and Gardner applied a Bidirectional-GRU (Bi-GRU) in the last layer of their MRC model followed by a linear layer that computes the scores for the start position in the answer span. They also applied another Bi-GRU with a linear layer to calculate the probability scores for the end position. Another model named the Gated-Attention Reader (GA) was proposed in [71], which uses two Bi-GRU for extracting word embeddings and features from the contexts and the questions.

In [72], Chen et al. used shallow Bi-LSTMs to encode the contexts and the questions from the 100-dimensional pre-trained GloVe word embeddings. Ghaeini et al., on the other hand, proposed a dependent gated reading Bi-GRU network (DGR) in [73] to map the relationship between the question-context pair. Similar to [72], DGR uses Bi-GRUs for encoding the 100-dimensional GloVe embeddings of the question-context pairs.

Networks with Attention Mechanism The concept of *attention* has revolutionized many tasks (including MRC [74–76]) in NLP since its first

introduction in machine translation [77]. Due to the sequential nature of RNN architectures, they often suffer from information loss when presented with long sequences, which is reflected in their poor performance [78]. It happens because the last hidden state in an RNN encoder is tasked to contain all information related to the past hidden states, which often becomes a bottleneck. The attention mechanism alleviates this issue by facilitating access to all hidden states at each time step and improves the performance of several NLP tasks [78].

Hermann et al. proposed Attentive Readers and Impatient Readers – two architectures for solving MRC tasks using an attention mechanism [74]. Attentive Readers encode the word embeddings in the questions and the contexts using two separate single-layer Bi-LSTM networks. They then calculate the attention scores to measure the amount of attention provided by the network to a token in the context while answering the question. The Attentive Readers attend to segments of the contexts that are most likely to answer the provided query. The Impatient Readers extend the attentive readers by enabling the model to reread the context as each question token is read.

In [79], Kadlec et al. proposed an MRC model named Attention Sum Reader (AS Reader), which defines the MRC task as an answer extraction task from the given context where the answer is a single word. It uses two Bi-GRU networks as encoders and one embedding function to calculate a contextual vector representation of every individual word in the context and a fixed-length representation for each question. Then, it computes the attention score for each word in the context using these vector representations to select the most probable answer.

Seo et al., in [40], introduced BIDAF, a Bi-Directional Attention Flow network, for the MRC task. BIDAF uses a bidirectional flow of attention to assist the interactions between the questions and the contexts and builds query-aware representations of the contexts. At each time step, the attention layer in BiDAF computes the attention score, and the attended vectors are passed to the subsequent layer of the BIDAF model, reducing the information loss. The authors used a memory-less attention mechanism that captures the attention between question-context pair at the current time step, ensuring that the attention at each time step is not affected by any incorrect attention score calculated at the previous time steps.

Transformer-based Networks The transformer network was first introduced in 2017 by Vaswani et al. in [80]. The network is entirely based on the attention mechanism and caused a paradigm shift in NLP for the following reasons: Although RNN-based networks were the first choice among researchers for sequence modeling, due to the nature of their learning paradigm, their training process is computationally inefficient due to their lack of parallelization. Transformer-based networks solved this problem using self-attention mechanisms that can be efficiently parallelized. Moreover, transformer-based networks can capture the semantic correlation among the words in a sequence and generate contextualized representations of the words.

The transformer architecture comprises a stack of six encoder layers and six decoder layers. Each layer has a self-attention layer with a self-attention mechanism that captures the attention between the words in a sequence and generates contextual vector representations. Followed by that is a point-wise fully connected feed-forward neural network layer. Since the transformer network has no recurrence or convolution component to encode the information regarding the relative order of the words in the sequence, it uses positional encodings (PE) typically calculated using sine and cosine functions of varying frequencies as shown in Equation 2.1 [80]:

$$PE_{(p,2k)} = \sin \frac{p}{10000^{\frac{2k}{d_m}}}; \qquad PE_{(p,2k+1)} = \cos \frac{p}{10000^{\frac{2k}{d_m}}} \tag{2.1}$$

Here, p refers to the position of the token in the sentence, d_m is the maximum dimension used by the transformer network for word embedding, and k denotes the embedding dimension. In other words, the positional encodings utilize sinusoidal waves to represent each dimension. The wavelengths of these waves follow a geometric progression ranging from 2π to $10000 \cdot 2\pi$.

The success of the transformer network led to the development of other models. In [81], a unidirectional multi-layer transformer decoder was used for training a language model, which achieved state-of-the-art performance in the MRC task. Another notable model is BERT (Bidirectional Encoder Representations from Transformers) proposed in [20], which uses bidirectional encoders to alleviate the unidirectional constraint, masked language modeling (MLM) as the pre-training objective, and next sentence prediction (NSP) as an additional task. Pre-training corpora of the BERT model consist of Wikipedia articles and BooksCorpus [82].

There are two architectural variants of the BERT model: $BERT_{BASE}$ with 110M parameters, 12 encoder layers, 12 self-attention heads per layer, and 768 hidden size; and $BERT_{LARGE}$ with 340M parameters, 24 encoder layers, 16 self-attention heads per layer, and 1024 hidden size.

The advent of BERT has revolutionized transfer learning in NLP and, since then, has been used as a first choice among researchers to perform a myriad of NLP tasks [49]. BERT has been shown to achieve human-level performance and outperform the previous alternate methods for MRC tasks in various domains [83–86] and become the first choice for MRC tasks. When performing MRC tasks, transformer-based PLMs such as BERT are usually employed as feature extractors which extract feature representations from the input question-context pair. Later, a single/few simple linear feed-forward neural network layers are added on top of these PLMs, which works as the MRC module and predicts the first and last tokens of the answer span [20]. Note that since these PLMs are trainable, they produce contextual feature representations from the input question-context pairs.

Since 2018, researchers proposed many variants of BERT, considering aspects such as computational resources, task domains, or limitations in existing approaches. [44, 45, 83, 87, 88]. The limitation of the MLM task in BERT is neglecting the dependency between the positions that were masked during pre-training. This may cause a discrepancy between pre-training and fine-tuning, hurting performance in the downstream tasks [89]. Fine-tuning is a way of transferring learned knowledge representations to a new but similar problem. In this context, we refer to fine-tuning as a process where the parameters of a trained deep learning model (on a certain dataset for a certain task) are tuned/tweaked to solve another related task on another dataset [20].

Yang et al. proposed XLNet in [89], which uses an auto-regressive pre-training method and overcomes this limitation. In addition to the corpora used in BERT, they used Common Crawl [90], ClueWeb 2012-B [91], and Giga5 [92] for pre-training. They found that XLNet outperforms BERT on 20 NLP tasks, including the MRC task [89]. RoBERTa was proposed in [45], which replicated the study of BERT and improved upon it in hyperparameter settings and training data size. For pre-training corpora, they used STORIES [93], OPENWEBTEXT [94], and CC-NEWS [95] in addition to the ones used in BERT. RoBERTa outperformed both BERT and XLNet in the MRC task. Later, BART was introduced in [88], which

performs pre-training of an encoder-decoder architecture by corrupting text and then learning its reconstruction. They used the same training corpora as BERT. BART achieved similar performance as RoBERTa in the MRC task.

A smaller PLM DistilBERT [87] was proposed to address the problem of the computational expense required to pre-train or fine-tune large PLMs. It was shown that DistilBERT is 60% faster than BERT while being 40% smaller than it. DistilBERT also retains 97% of the capabilities for natural language understanding compared to BERT. DistilBERT achieves comparable performance to BERT in the MRC task. Although the performance of PLMs often improves with increasing model size, training larger models becomes harder due to computational constraints. Addressing this issue, Lan et al. proposed ALBERT in [44], which reduces parameters to reduce memory usage. They pre-trained ALBERT using the corpora similar to BERT and showed that it outperforms BERT in many NLP tasks, including MRC. In another work [96], Clark et al. proposed ELECTRA, which follows an efficient version of the MLM pre-training task – replaced token detection. While BERT masks some tokens in the input, ELECTRA replaces them with plausible alternatives. Then, instead of predicting the original tokens, they trained a discriminative model that predicted whether those tokens were replaced or not. The authors used similar pre-training corpora as BERT for their primary implementation of ELECTRA and later extended it to the pre-training corpora of XLNet. ELECTRA outperforms BERT and achieves comparable performance to RoBERTa and XLNet in the MRC task. Some other variants of BERT and GPT that were proposed by researchers for MRC tasks are SqueezeBERT [97], MobileBERT [98], Longformer [99], DEBERTA [100], and PEGASUS [101]

Very recently, the recognition of the impact of scaling on improving model capabilities has prompted researchers to develop large language models (LLMs) [102]. LLMs are pre-trained language models with "tens or hundreds of billions of parameters" trained on billions or trillions of tokens [102]. Training these LLMs requires massive computing resources and is sensitive to data quality and training tricks [103, 104]. Examples of such models are T5 [105], GPT-3 [106], GPT-4 [107], OPT [108], BLOOM [103], LLaMA [109], Pythia [110], PaLM [111], etc. These LLMs are trained on a vast amount of publicly available generic texts with the goal of generalizing among different domains. Therefore, no specific step is taken to address any specific domain [102]. Given the computational cost, fine-tuning these

general-purpose LLMs is preferred over pre-training of these LLMs on domain-specific texts [47, 105]. Nonetheless, fine-tuning these LLMs may still lead to more expensive inference time compared to the smaller PLMs due to the size of the LLMs while gaining minimal to no improvement in performance [112]. A rather cheaper alternative to fine-tuning an LLM is to utilize in-context learning (ICL) – a learning paradigm in natural language understanding, where LLMs observe some examples of the task as input, provided through the input prompt and decode them directly without any update to the model parameters [113].

Researchers have found that using the ICL paradigm, GPT-3 has shown competing or better performance compared to smaller PLMs on several clinical tasks, including medication extraction, co-reference resolution, and acronym disambiguation [114]. Another study has shown that ChatGPT can attain passing scores on all three steps of the United States Medical Licensing Examination (USMLE) [115]. However, OpenAI's data policies restricted the set of tasks that can evaluate GPT-3's capability [114]. GPT-3 and its variants, such as ChatGPT, which can perform as well as or even better than many fine-tuned smaller PLMs, are typically exclusively accessible via APIs offered by private enterprises like OpenAI, which may raise additional apprehensions pertaining to security and the protection of data [116, 117]. As a result, it remains uncertain whether such performance exhibited by these LLMs without any fine-tuning can be effectively translated to tasks involving clinical notes. Moreover, these LLMs have a tendency to hallucinate and generate realistic but false information [118, 119] which, if not carefully evaluated, can mislead into believing to be true, which may be especially problematic for applications in the biomedical domain [102]. A recent thorough comparative study between smaller fine-tuned PLMs and LLMs using the ICL approach demonstrated the capability of smaller models to achieve better performance with significantly fewer computational resources [102]. The findings of their extensive study indicated the importance of smaller PLMs in the biomedical domain.

This dissertation focuses on improving the performance of comparatively smaller encoder-based PLMs in the biomedical MRC task.

Domain-specific Transformer-based Pre-trained Language Models The development of transformer-based pre-trained language models or PLMs has

enabled us to achieve human-level performance in many NLP tasks, including MRC [20]. However, one of the issues with applying these models in downstream tasks, for example, MRC tasks, is the linguistic differences between the pre-training corpora and the fine-tuning corpora [83, 120]. As explained earlier, the pre-training corpora of the original BERT model consist of general-purpose English texts. Despite being the same language, many domains, such as biomedical, academic, programming, or legal, differ from the colloquial English language on various linguistic facets, such as lexicon, syntax, semantics, or noise, depending on the domain [121]. Moreover, fine-tuning PLMs, pre-trained on general-purpose domain datasets, on domain-specific datasets is often not enough for the PLMs to acquire enough domain knowledge to achieve the best performance [83, 120].

Keeping that in mind, researchers have proposed domain-specific BERT/BERT-variant models where they used domain-specific corpora to pre-train either from scratch [85, 86, 122] or using continual pre-training [123]. Some examples of such models are Finbert [84] for the finance domain; ALeaseBERT [124] and Legal-bert [122] for the legal domain; RoBERTa-News [123] for the news domain; Codebert [85], PLBART [125], Graphcodebert [126], and Cotext [127] for the programming language domain; Scibert [86] for the academic science domain; and Biobert [83], ClinicalBERT [120], PubMedBERT [128], BioElectra [129], BlueBERT[130], BioLinkBERT [131], and BioBART [132] for the biomedical domain.

This dissertation exclusively focuses on the biomedical domain. Natural language in the biomedical domain differs in many linguistic aspects, including sentence structure, semantic property, syntactic features, and lexicon uniqueness [83, 133]. As a result, the best-performing PLMs in the general-purpose domain often struggle to capture the semantics in these domain-specific corpora. Consequently, they are often unable to match the performance on various downstream biomedical NLP tasks, for example, relation extraction, named-entity-recognition (NER), and MRC [83]. PLMs that are pre-trained on corpora from the biomedical domain alleviate this issue to a great extent [134]. Some common choices of pre-training corpora among researchers are MIMIC datasets [135] and PubMed abstracts and PMC full articles (available at: https://www.ncbi.nlm.nih.gov/pmc/). The PLMs pre-trained on these corpora have reportedly outperformed the original BERT model or its variants across diverse NLP tasks in the biomedical domain [83, 120, 128, 130].

2.3.3 Machine Reading Comprehension Datasets

So far, we have seen how domain-specific corpora have been used to improve transformer-based feature extractors to generate more meaningful representations of the question-context pairs. However, to perform the downstream NLP tasks, such as MRC, by reaching the full capacity of any deep learning model, we still need high-quality, large-scale, task-specific, labeled datasets for fine-tuning, i.e., knowledge transfer [136]. In this section, we highlight the existence and availability of high-quality MRC datasets in various domains, emphasizing the biomedical domain.

In 2016, Rajpurkar et al., in [21], presented a Stanford Question Answering Dataset v1.0 (SQuAD) dataset to address the need for high-quality, large-scale MRC datasets. Before 2016, the available datasets were small-scale and not targeted to perform MRC tasks. SQuAD, created by crowdworkers, consists of 107,785 question-answer pairs with 536 unique contexts taken from Wikipedia articles (available at: https://en.wikipedia.org/wiki/). Following SQuAD, Trischler et al., in [137], presented NewsQA, a large-scale MRC dataset built using news articles from CNN, consisting of 119,633 questions posed by crowdworkers and 12,744 unique contexts. Another dataset TriviaQA was presented in [22], consisting of 662,659 samples with 95,956 question-answer pairs. Some other large-scale datasets that followed SQuAD include: QuAC [138], HotpotQA [139], and DROP [140] on Wikipedia articles; AmazonQA [141] on Amazon reviews; ROPES [142] on science textbooks and Wikipedia articles; MS MARCO [143] on Bing's web search query logs; Natural Questions (NQ) [144] on queries from Google search engine and Wikipedia articles; SpaceQA [145] on Concurrent Design Facility report; and CoQA [146] on 8000 conversations between crowdworkers about text passages from children's stories, CNN news articles, project Gutenberg literature, Wikipedia and Reddit articles, middle and high school English exams, and science articles.

While the MRC capability of NN-based models has increased to a great extent by these large-scale annotated datasets, this capability still lacks domain-specific understanding of biomedical corpora such as clinical notes or biomedical literature [83, 120, 128]. To mitigate this issue, researchers have presented several datasets in the biomedical domain.

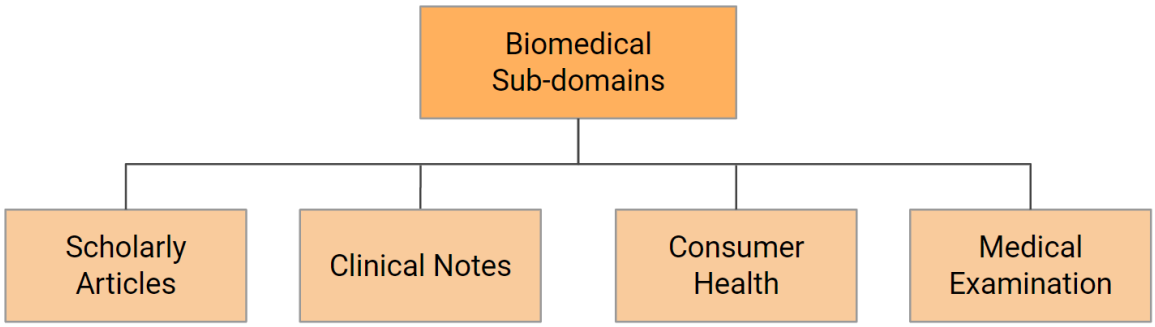


Figure 2-3. Sub-domains in biomedical machine reading comprehension tasks according to the classification presented in [10].

According to [10], based on the content, to date, there exist four types of biomedical MRC datasets[1] (Figure 2-3): clinical notes, scientific biomedical literature, medical examination, and consumer health.

The BioASQ dataset [27] focuses on addressing the challenge of extracting valuable knowledge from biomedical literature with minimal effort. However, the latest BioASQ dataset has only 4719 question-answer pairs on various PubMed abstracts [27], despite being the largest MRC dataset on biomedical scholarly articles annotated by subject-matter experts (SMEs) [10]. COVID-QA [147], CovidQA [148], and CoQUAD [149] are three additional biomedical MRC datasets on the scientific literature on COVID-19. COVID-QA and CoQUAD consist of 2019 and 150 question-answer pairs, respectively, while CovidQA comprises 124 question-article pairs. These datasets were meticulously annotated by volunteer SMEs using COVID-19-related scientific articles as contexts to provide accurate and reliable labels in the datasets. SleepQA dataset, on the other hand, consists of 5000 samples annotated from 527 sleep health-related articles [150].

In [151], Pampari et al. presented emrQA, a biomedical dataset on unstructured electronic medical records, with templated patient-specific questions that healthcare providers might ask. emrQA has 455,837 question-answer pairs, automatically generated from 2,425 unique clinical notes. However, despite being a large-scale dataset for the clinical notes sub-domain, since emrQA consists of auto-generated samples instead of domain-expert annotations, the samples often suffer from lack of diversity, unanswerable questions, and incompleteness in the answers [152] Another large-scale dataset concerning clinical case reports is CliCR [153]. However, this dataset focuses on the cloze-style MRC problem rather than span prediction. Some other large-scale cloze-style biomedical MRC datasets are BioREAD [154] and BioMRC [155]. DrugEHRQA dataset [156], on the other hand, includes medication-related queries consisting of question-answer pairs from structured and unstructured electronic health records (EHRs). Another dataset RadQA consists of 6148 samples annotated from 1009 radiology reports [157].

One other aspect of biomedical MRC tasks is consumer health, where questions are usually raised by the general public regarding various health conditions. In [158], Savery et al. presented an extractive and abstractive summarization dataset

[1]sub-domains or problem domains in the biomedical domain

MEDIQA-AnS – an extension of the task and dataset presented in MEDIQA-QA – for the consumer health MRC task [159]. MEDIQA-QA was built based on medical queries posed by the general public to the consumer health QA system CHiQA [160]. The answers in the extractive summarization MRC dataset were formed by extracting a single contiguous text span, or an accumulation of multiple detached text spans from the context. MASH-QA [161] is another consumer health MRC dataset with multiple answer spans curated from health articles.

Some other question-answering datasets are MedQuAD [162] and HealthQA [163] on consumer health, MIMICSQL [164] on SQL queries for a medical records database, and multi-choice HEAD-QA [165] on medical examination. PubMedQA dataset [166] is another large-scale biomedical dataset, consisting of 211.3k artificially generated, 1k expert-annotated, 61.2k unlabeled question-answer pairs. PubMedQA attempts to solve the problem of yes/no/maybe questions which are targeted to summarize the conclusion of PubMed abstracts.

Despite the increasing availability of MRC datasets in recent years, there are still a lot of areas in the biomedical domain where the applications of MRC can prove to be helpful. Among them is using MRC to interpret clinical practice guidelines at the point of care and extract information from clinical notes on injection drug use (IDU).

Moreover, due to expensive domain expertise, most of these datasets tend to be smaller in scale, and researchers have discovered that relying solely on these biomedical MRC datasets for fine-tuning PLMs does not yield performance comparable to that of general-purpose domains like news or Wikipedia [53, 167]. As a solution, they have adopted transfer learning, which involves transferring knowledge to the target biomedical domain from a high-resource source domain, focusing on the relevant information for MRC tasks.

2.3.4 Transfer Learning

The concept of transfer learning refers to the learning process that facilitates transferring of knowledge acquired from an extensively-explored domain or task, referred to as the source domain or task, to a new unexplored or under-explored domain or task, referred to as the target domain or task [24]. Following the definitions in [24], a domain $\mathcal{D}$ can be defined as a tuple as follows: $(\mathcal{X}, p(X))$. In this context, $\mathcal{X}$ refers to the feature space, $X \in \mathcal{X}$ denotes the data samples,

and $p(X)$ refers to the marginal probability distribution of X. A task $\mathcal{T}$, on the other hand, can be defined as the following tuple of label space $\mathcal{Y}$ for $\mathcal{D}$ and the predictor function $\mathcal{F}(.)$ [24]: $(\mathcal{Y}, \mathcal{F}(.))$. $\mathcal{F}(.)$ is learned from the training samples $(X \in \mathcal{X}, Y \in \mathcal{Y})$, and corresponds to the conditional probability distribution $p(Y \in \mathcal{Y}|X \in \mathcal{X})$. Therefore, we can express $\mathcal{T}$ as $(\mathcal{Y}, p(Y|X))$.

Usually, in NLP, there can be several cases [24] where we can apply transfer learning (Figure 2-4):

(i) Case I: The target and source domains have distinct feature spaces, i.e., $\mathcal{X}_t \neq \mathcal{X}_s$. For example, multi-lingual feature spaces.

(ii) Case II: The target and source domains have distinct marginal probability distributions, i.e., $p(X_t \in \mathcal{X}_t) \neq p(X_s \in \mathcal{X}_s)$. For example, despite sharing the same language, the topics in both domains are dissimilar. The differences in topics between two domains can occur when some words are used in one domain more than others because some words have a stronger relationship with the domain topic. Figure 2-5a shows that two domains have the same distribution of features for the samples if they have similar topic distributions. On the other hand, Figure 2-5b shows two different domains with different distributions of features for the samples resulting from different topic distributions.

(iii) Case III: The target and source domains have distinct label spaces, i.e., $\mathcal{Y}_t \neq \mathcal{Y}_s$ – for example, binary and multi-class classification tasks.

(iv) Case IV: The target and source domains have distinct label distributions, i.e., $p(Y_t \in \mathcal{Y}_t|X_t) \neq p(Y_s \in \mathcal{Y}_s|X_s)$. For example, a source domain could have 100 positive samples, whereas the target domain has 10 positive samples.

The unavailability of high-quality, large-scale labeled data in the biomedical domain is often an obstruction for models to fully achieve their full potential performance. As a result, researchers have resorted to Case II of transfer learning to enhance models' generalizability to the target tasks in the biomedical domain, using obtained knowledge from both low-resource target biomedical domain $\mathcal{D}_t$ and high-resource source domain $\mathcal{D}_s$ – which is also the focus of this dissertation.

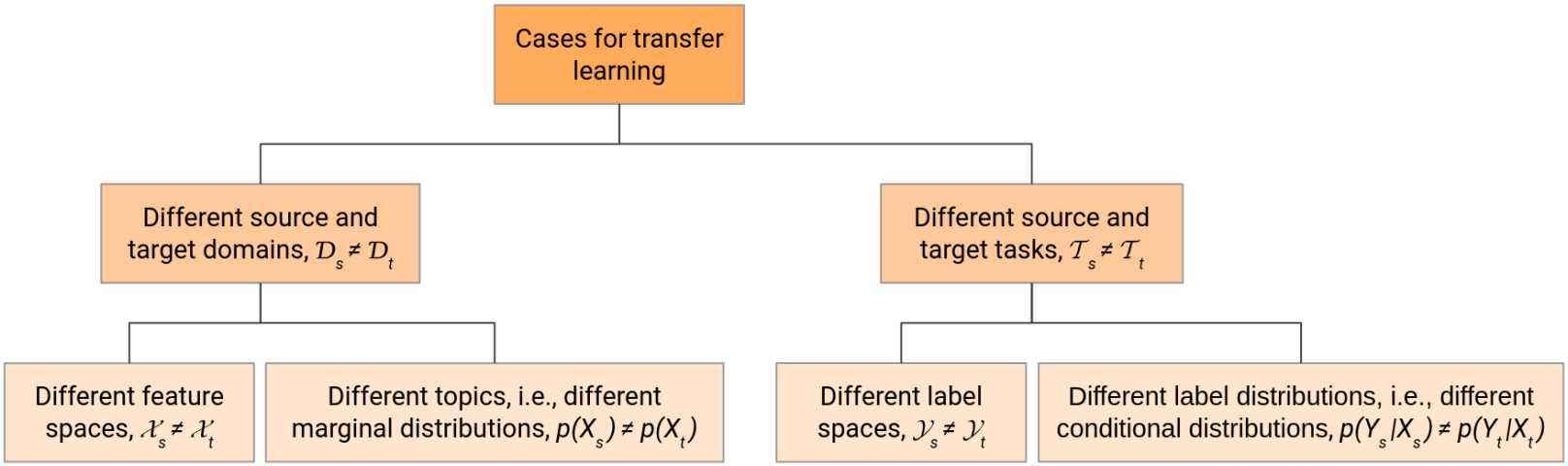


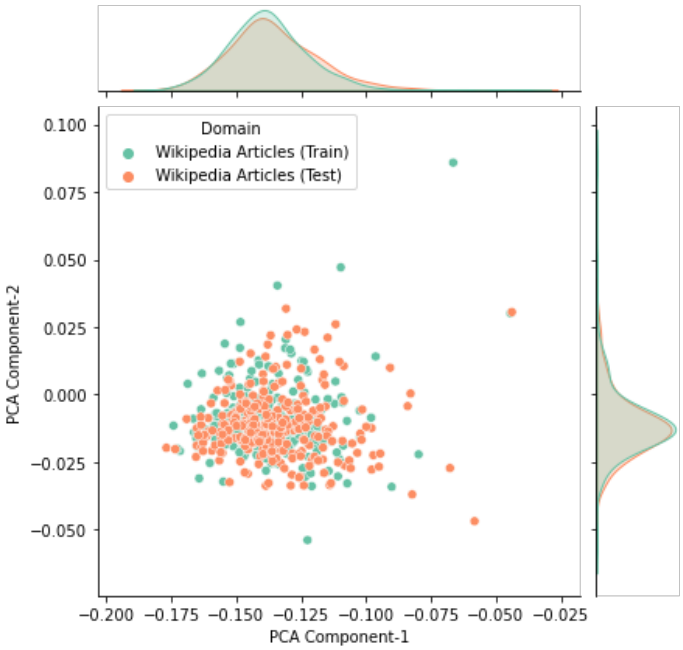
Figure 2-4. Different cases for transfer learning according to the classification presented in [24].

The learning paradigms of transfer learning can be classified into three primary categories [24]: (i) Inductive, (ii) Unsupervised, and (iii) Transductive.

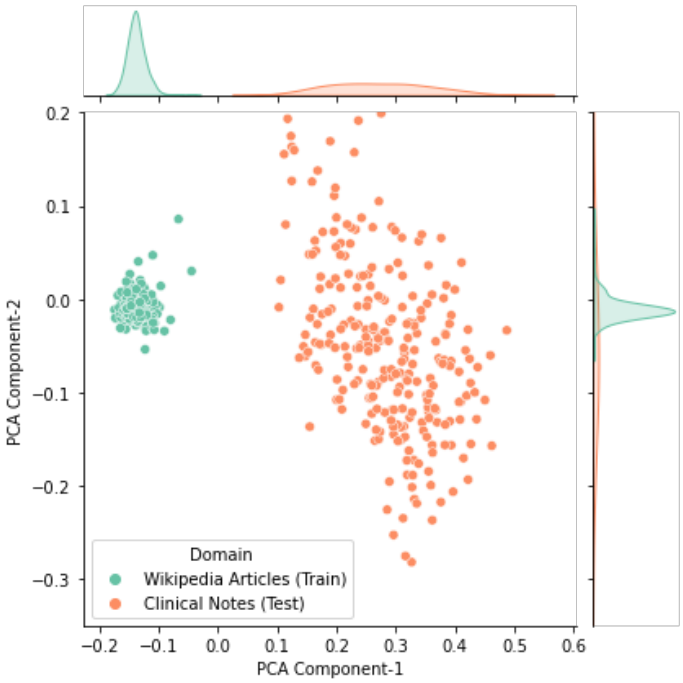
(i) **Inductive transfer learning:** In this learning paradigm, the target and source tasks differ, whereas their domains can be dissimilar or similar. It requires the target domain to have labeled data for inducing an objective predictor function $\mathcal{F}_t(.) = p(Y_t|X_t)$ to be used in the target task. According to [24], inductive transfer learning can be defined as follows: Given $\mathcal{D}_t$, $\mathcal{D}_s$ and $\mathcal{T}_t$, $\mathcal{T}_s$, the aim here is to improve the learning of $\mathcal{F}_t(.)$ for $\mathcal{D}_t$, by leveraging knowledge acquired from $\mathcal{D}_s$ and $\mathcal{T}_s$, where $\mathcal{T}_t \neq \mathcal{T}_s$. Here, $\mathcal{D}_t$ is the target domain, and $\mathcal{T}_t$ is the associated learning task. $\mathcal{D}_s$ is the source domain, and $\mathcal{T}_s$ is the associated learning task. Based on different scenarios regarding the availability of labeled data in the source domain, inductive transfer learning can be sub-divided into two cases:

Case I: The source domain has a labeled dataset that is large-scale and of high quality. This case leads to the multi-task learning setting, which seeks to facilitate knowledge transfer from the source domain by trying to learn the target and source domain tasks simultaneously to achieve high performance in the target task. One alternative approach to multi-task learning is sequential transfer learning. In this setting, a model is trained individually and sequentially on the source and target domain tasks, aiming to achieve the same objective as the multi-task learning setting. Sequential learning is chosen over multi-task learning in the following circumstances: (i) labeled data for both target and source domains are not simultaneously available, (ii) the source domain has too large a dataset in comparison with the target domain inducing data imbalance issue and hurting the performance in the target task. Nevertheless, the 'catastrophic forgetting' phenomenon is one of the major concerns when opting for sequential transfer learning instead of multi-task learning. This phenomenon refers to the fact that deep NN-based models trained on a source-domain dataset tend to 'forget' learned knowledge representations when further trained on the target domain dataset. Consequently, this phenomenon results in a significant drop in the source domain's performance.

Case II: Labeled data in the source domain is unavailable. This case leads to the self-taught learning setting, where a representation is learned using only



(a)Domains with similar topic distributions



(b)Domains with different topic distributions

Figure 2-5. Two-dimensional visualizations of vector representations and the distributions of the features for the samples from two domains with (a) similar and (b) different topic distributions.

unlabeled source domain data and later applied to the labeled target domain data [168].

(ii) **Unsupervised transfer learning:** In this learning paradigm, both target domain and source domain data used for training are unlabeled, while their corresponding tasks are different but related. This setting focuses on solving unsupervised target domain learning tasks such as clustering or dimensionality reduction [169, 170]. According to [24], unsupervised transfer learning can be defined in the following manner: Given $\mathcal{D}_t$, $\mathcal{D}_s$ and $\mathcal{T}_t$, $\mathcal{T}_s$, the aim here is to improve the learning of $\mathcal{F}_t(.)$ for $\mathcal{D}_t$, by leveraging knowledge acquired from $\mathcal{D}_s$ and $\mathcal{T}_s$, where $\mathcal{T}_t \neq \mathcal{T}_s$. Here, $\mathcal{D}_t$ is the target domain, and $\mathcal{T}_t$ is the associated learning task. $\mathcal{D}_s$ is the source domain, and $\mathcal{T}_s$ is the associated learning task. Given target domain label space $\mathcal{Y}_t$ and source domain label space $\mathcal{Y}_s$, unsupervised transfer learning assumes that $\mathcal{Y}_t$ and $\mathcal{Y}_s$ are not observable.

(iii) **Transductive transfer learning:** Within the paradigm of transductive learning, the target and source domains differ, whereas their corresponding tasks are identical. Furthermore, in the given scenario, the source domain constitutes a large amount of labeled data, while the target domain lacks labeled data. According to the definition in [24], we can describe transductive transfer learning as follows: Given $\mathcal{D}_t$, $\mathcal{D}_s$ and $\mathcal{T}_t$, $\mathcal{T}_s$, the aim here is to improve the learning of $\mathcal{F}_t(.)$ for $\mathcal{D}_t$, by leveraging knowledge acquired from $\mathcal{T}_s$ and $\mathcal{D}_s$, where $\mathcal{T}_t = \mathcal{T}_s$ but $\mathcal{D}_t \neq \mathcal{D}_s$. Here, $\mathcal{D}_t$ is the target domain, and $\mathcal{T}_t$ is the associated learning task. $\mathcal{D}_s$ is the source domain, and $\mathcal{T}_s$ is the associated learning task. Depending on different scenarios, the transductive transfer learning paradigm can be sub-divided into two cases:

Case I: The target and source domains have distinct feature spaces, which leads to cross-lingual learning or adaptation.

Case II: The target domain and source domain marginal probability distributions are dissimilar, despite their identical feature spaces. This case arises when the two domains have distinct topics, such as ours. It is otherwise known as domain adaptation [24].

Machine learning models require data for training and inference – training data and test data – where it is typically assumed that the data used for

training and testing are "independent and identically distributed" (i.i.d.) [121, 171]. Nonetheless, many practical applications in the real world necessitate the application of machine learning models to datasets exhibiting distributions different from those encountered during the training stage. Typically, this scenario arises when our aim is to achieve satisfactory model performance in a domain that lacks or has a limited amount of labeled data, and we can leverage a somewhat related domain that has an adequate number of training samples. However, when this happens, often due to distributional discrepancies between domains, the domain introduced at the test time suffers from performance degradation in machine learning models. This issue is tackled by domain adaptation. Domain adaptation is a learning technique with the objective of increasing the generalization capability of machine learning models so that they can attain comparable performance in the domains that have either unlabeled or very limited labeled data [172, 173]. For increasing generalization capability, domain adaptation mainly utilizes target domain data with few or no labels and a substantial amount of source domain data with labels. In NLP, various efforts have been put forth in developing various domain adaptation methods such as data selection and representation learning [24].

Data selection Researchers have found that all source domain instances are not equally beneficial for training a system that would yield the best performance in the target domain [174, 175]. One solution is to apply automated data selection techniques to choose the samples from the source domain that closely resemble the samples from the target domain. Thus, properly implementing data selection techniques relies primarily on two components: an instance representation and a similarity metric. Plank and Noord in [176] used character tetragrams, relative frequencies of words, and topic models as instance representations. For the similarity metrics, they experimented with both 'probabilistically-motivated' and 'geometrically-motivated' distance functions – Jensen-Shannon divergence, Kullback-Leibler (KL) divergence, cosine distance, euclidean distance, and Manhattan distance functions. For instance representations, researchers also used pre-trained word embeddings [177, 178], sentence embeddings [179–181], and autoencoder representations [172, 182].

Representation learning Another way to implement domain adaptation is representation learning. Its objective is to learn a function facilitating adaptation from the distribution of training data to the distribution of test data by projecting the samples from two domains into a common latent space. In [172], Glorot et al. introduced a method that learns a common representation by training a shared stacked denoising autoencoder that can reconstruct noisy source domain and target domain texts. The stacked denoising autoencoder was used for domain adaptation in other studies as well [183, 184]. These approaches *implicitly* reduce the distance between the source and target instance representations by using the same model for text reconstruction. Further improvements in the approaches for representation learning – representation distance minimization [182, 185] and adversarial learning [186] – *explicitly* enforce the overlap between two domains in the latent spaces and are more efficient.

Representation distance minimization focuses on explicitly minimizing the distance between the target and source instance representations in the latent space. To diminish this distance, the researchers in [185, 187, 188] used maximum mean discrepancy, whereas the authors in [182] used KL divergence to achieve the same objective.

Another domain adaptation approach was introduced in [173]. This approach employs adversarial learning for domain adaptation. It uses an encoder followed by a domain classifier that can predict the domain from the latent representations of the target and source samples. This approach reverses the gradients of the domain classifier when back-propagated to the encoder – a technique that encourages the encoder to learn representations that can confuse the domain classifier into misidentifying the domains. The adversarial learning approach has been shown to effectively reduce the discrepancy between the distributions of the target and source-domain samples [173, 189]. However, similar to the learning of GAN, adversarial learning makes the training procedure unstable [190, 191]. Researchers have used various approaches, such as minimizing the Kantorovich–Rubinstein metric (the Wasserstein/Earth mover's distance) between target and source domain distributions, to solve this issue and make the learning process more stable [192]. While the above representation learning techniques for domain

adaptation focus on minimizing the distance between two domains by learning the mapping of domain-invariant features, another school of research focuses on individual properties of each domain as well as the domain-invariant ones [193, 194]. They identified that, in performing the task, some individual domain-specific properties may play a vital role. An example of this can be provided using clinical notes. The word 'negative' often means a positive outcome of a test result for a patient, whereas generally, in movie reviews, the same word usually bears its literal meaning. Here, domain-invariant refers to the fact that the instance representations from the target and source domains exhibit higher proximity in the latent embedding space. In [193], Bousmalis et al. partitioned the latent space for feature representations into two components – shared domain-invariant space that can be obtained by the representation learning techniques mentioned above and private domain-specific space that is orthogonal to the shared space, preserving the information unique to the respective domains.

Transfer learning in biomedical MRC Researchers have applied transfer learning techniques to many NLP tasks, for example, neural machine translation [195], speech recognition [196], sentiment classification [172], and named entity recognition (NER) [197]. Compared to these tasks, applying transfer learning to the MRC task requires handling an additional challenge. Other NLP tasks usually deal with datasets consisting of a single label per sample, e.g., a sentiment per review in a sentiment classification dataset and a translation per text in a machine translation dataset. However, for the MRC task, in an MRC dataset, we consider the questions as a part of the labels in addition to the answers as the labels. Therefore, during transfer learning, along with the missing answers in the target domain, we must also deal with the missing questions. To tackle this challenge, over the past few years, a variety of methods have been suggested by researchers to generate synthetic labels using unlabeled contexts, such as using conditional probability, inside–outside–beginning (IOB) tagger, and Bi-LSTM [25]; NER and Bi-LSTM [26]; seq2seq model with an attention mechanism [198–213]; pre-trained language models [214]; and variational autoencoder (VAE) [215–217]. Researchers have also used the roundtrip consistency algorithm (RTC) [218], reinforcement learning [219], and self-training [220] for synthetic question-answer generation.

Synthetic or pseudo labels refer to machine-generated labels. Only a few of these works focused on the biomedical domain [198, 213, 214, 216, 217].

However, while it is crucial to generate synthetic labels to save domain experts' valuable effort and time, the computational expense required to generate these synthetic data should also be counted. Moreover, these question-answer generation techniques focus on one sub-domain of the biomedical domain at a time, lacking generalization to other sub-domains and often over-fitting on specific datasets [198, 221]. Hence, our research endeavors to avoid developing yet another synthetic question-answer generator limited to a specific biomedical sub-domain. Instead, we aim to extend the applicability of biomedical MRC tasks through the process of transfer learning. Primarily, we concentrate on leveraging adversarial learning-based domain adaptation technique [173] to tackle the challenge of scarce labeled data within the target biomedical domain. In a similar vein, a previous study [26] employed adversarial learning for domain adaptation while incorporating synthetic question-answer pairs. Although the domain adaptation approach was demonstrated to improve performance in more general domains like news, movie reviews, web search, and Wikipedia [26], its effectiveness within the biomedical domain remains to be determined.

In the rest of the section, we comprehensively review existing literature on transfer learning in biomedical MRC. Based on the publicly available resources, we examine the related works along two key dimensions: (i) the utilization of labeled target and labeled source domains and (ii) the utilization of unlabeled target and labeled source domains.

With the emergence of the BioASQ dataset, researchers in the biomedical domain have shown significant interest in transfer learning [10]. Typically, researchers tend to focus on transfer learning by sequential fine-tuning [53, 167, 222–227]. In this setting, MRC models are sequentially trained on a large-scale labeled source domain (typically general-purpose domain) dataset and a small-scale labeled target biomedical domain dataset. Wiese et al., in [222], trained the FastQA model sequentially on SQuAD and BioASQ and achieved state-of-the-art performance in the BioASQ test set. They also addressed the catastrophic forgetting phenomenon using an additional forgetting cost regularizer and an l2 weight regularizer. The effect of catastrophic forgetting was explored in more detail in [228]. Another study used SQuAD and BioASQ as source domains and

COVID-QA as the target domain for sequential fine-tuning of the RoBERTa model [229]. A similar study with COVID-QA used only SQuAD as the source-domain dataset to sequentially fine-tune the BERT model [230]. In [231], Hu et al. proposed a new MRC model that employs dynamic routing and answer voting along with sequential transfer learning with SQuAD and BioASQ.

A Hierarchical Question-Aware Context Learning (HQACL) model was introduced in [232], which incorporates a multi-level attention mechanism to improve the ability of language representation and a data augmentation technique to inflate the SQuAD training set. HQACL uses sequential transfer learning with two source domains and one target domain dataset. It uses SQuAD and CoQA as the source, and for the target, it uses BioASQ. In another study [233], Du et al. used a hierarchical multi-layer transfer learning model for knowledge transfer to BioASQ from SQuAD using sequential training and a forgetting cost regularizer to handle catastrophic forgetting. In a recent study, the BioBERT model with a novel self-attention enriching mechanism was sequentially fine-tuned on SQuAD and BioASQ datasets to achieve better performance on BioASQ [234]. In this model, the self-attention enriching mechanism integrated named and biomedical entity information, derived from the context and question, into the BioBERT model to enhance its self-attention layer. In [235], Soni and Roberts explored the impact of different sequences of fine-tuning datasets during sequential learning on the performance of biomedical MRC involving clinical notes. They separately fine-tuned four PLMs using multiple combinations of SQuAD, CliCR, and emrQA datasets. To assess the effectiveness of the fine-tuned models, the authors employed held-out test sets from the emrQA and CliCR datasets.

Researchers not only applied the transfer learning approach using pre-trained language models to accomplish knowledge transfer from the MRC task leveraging large-scale MRC datasets but also extended this approach to transfer the knowledge of the natural language inference task to the biomedical MRC task [167, 236, 237]. They fine-tuned the transformer models on the natural language inference task before fine-tuning them on the MRC task.

Although researchers have made efforts to tackle the task of utilizing small labeled datasets in the target domain, the focus on scenarios with very few or absent labeled data in MRC has been limited [25, 26]. Moreover, there has been even less emphasis on addressing this issue within the biomedical domain [238, 239].

Considering the gaps mentioned above, although this dissertation primarily focuses on biomedical MRC, we also review some highly pertinent literature from the non-biomedical domain.

Golub et al., in [25], proposed a novel two-stage synthesis network (SynNet) to transfer knowledge from a high-resource labeled domain to a low-resource domain where no labeled data is available. They used an answer synthesis module and a question synthesis module to generate synthetic question-answer pairs. They used a Bi-LSTM model as an IOB tagger for the answer synthesis module, predicting whether each word in the context is an answer. The question synthesis module used an encoder-decoder network with attention, a Bi-LSTM network, and an LSTM decoder. As the MRC module, they used a BiDAF model. They evaluated their framework on SQuAD and NewsQA datasets by alternatively using them as source and target domains.

Wang et al., in [26], extended SynNet to achieve effective domain adaptation by implementing adversarial learning to learn transferable domain-invariant features. Their framework, Adversarial Domain Adaptation for MRC (AdaMRC), uses source domain data with manually-generated labels and target domain data with synthetically-generated labels to train an MRC model and a domain classifier in the adversarial setting. As the answer synthesis module, the authors used an NER system to identify possible answer spans. As the question generation module, they used a sequence-to-sequence model equipped with a BiLSTM encoder, an attention mechanism, and an LSTM decoder. For the MRC model, they experimented with SAN, BiDAF, BERT, and ELMo models. The authors evaluated their framework by alternatively using SQuAD, NewsQA, and MS MARCO as source and target domains.

Cao et al., in [240], introduced Conditional Adversarial Self-training (CASe), another domain adaptation framework focusing on non-biomedical domains. This framework tackles the challenge of utilizing partially labeled target domain data consisting of questions only. It incorporates the fine-tuning of a BERT model on the labeled source-domain dataset, followed by self-training with pseudo answers for the target domain and conditional adversarial learning. To assess the effectiveness of their approach, the authors evaluated CASe on multiple datasets, including SQuAD, CNN/Daily Mail [74], NewsQA, CoQA, and DROP.

Some relatively recent studies in late 2022 have also used adversarial learning-based domain adaptation [241, 242]. In [241], Bai et al. proposed a model named Indirect Domain Adaptive Multi-Task Transformer (Indirect DAMT), which uses adversarial learning along with MLM and NSP with source and target domain contexts. They used SQuAD, NewsQA, and ReviewRC [243] datasets for transfer learning. In another study [242], Wu and Xu proposed a multi-task learning approach TrustMRC, where pseudo answers for the partially unlabeled (consisting of only questions) target domain data are obtained by training an MRC model on the source-domain dataset. A conditional calibration strategy was then introduced to estimate the predictive uncertainty of these answers and increase the reliability of the confidence in the predicted answers by adding calibration loss constraint. A conditional adversarial learning strategy was also incorporated. This strategy considers the uncertainty mismatches for source and target domain feature representations by integrating Maximum Mean Discrepancy (MMD) as the uncertainty-aware constraint in the domain classifier. The authors evaluated their approach by alternatively using SQuAD, NewsQA, and NQ as the target and source domains.

In [244], a simple zero-shot method was proposed to facilitate transfer learning from the source domain to the unlabeled target domain. This method employed several base models with similar structures but separate training datasets with an accuracy-based weight ensemble technique. The training of these models required no target domain data. This method was evaluated on four target domain datasets, one of which is BioASQ. Dhingra et al., in [245], used transfer learning with limited labeled data in multiple target domains, one of which is the biomedical domain. As the dataset for the source domain, they used two large-scale cloze-style MRC datasets built from Wikipedia articles and PubMed articles, respectively. They experimented with two MRC models: GA Reader [71] and BiDAF + Self-Attention (SA) [246]. They used a small set of labeled samples from the SQuAD, TriviaQA, and BioASQ datasets as the target domain datasets.

Nishida et al. proposed a multi-task transfer learning approach in [238]. This approach performs the language modeling and MRC tasks simultaneously. Their methodology entails adapting the BERT model to the target domain through language modeling on the unlabeled target domain contexts, followed by fine-tuning the model using the labeled source-domain dataset. The effectiveness of

their approach was assessed by conducting experiments on various datasets. The source-domain datasets employed for validation include SQuAD and NewsQA, whereas the target domain datasets consist of BioASQ, DuoRC [247], and NQ. Xu et al., in [248], also used a multi-task learning framework and validated their framework on four datasets – SQuAD, NewsQA, MS MARCO, and Who-Did-What. Another work used unlabeled target domain (BioASQ) and labeled source domain (SQuAD) data for transfer learning with BioBERT and SciBERT [239]. They used a named entity extractor to identify possible answers from the contexts. For synthetic questions, they generated cloze-style questions using a denoising approach that focused on learning about entity mentions in the contexts by finding incorrect entity mentions in a corrupted context.

It is evident that although some research endeavors in MRC have focused on transferring knowledge representations from domains with abundant resources to those with limited resources, the exploration of learning domain-invariant feature representations has been quite limited. Furthermore, the biomedical domain remains largely unexplored in this particular context.

CHAPTER 3

BIOADAPT-MRC: ADVERSARIAL LEARNING-BASED DOMAIN ADAPTATION FRAMEWORK FOR BIOMEDICAL MACHINE READING COMPREHENSION

Much of the content in this chapter is taken without alteration from the published article ***M. Mahbub, S. Srinivasan, E. Begoli, and G. D. Peterson, "BioADAPT-MRC: Adversarial Learning-based Domain Adaptation Improves Biomedical Machine Reading Comprehension Task," Bioinformatics, Jul. 2022, btac508, issn: 1367-4803.*** *[249] by Maria Mahbub, Sudarshan Srinivasan, Edmon Begoli, and Gregory D Peterson. For permission to reuse this material, please visit http://global.oup.com/academic/rights. The authors' contributions to the published article are as follows: MM conceptualized and designed the study, pre-processed data, implemented methodology, performed experiments, analyzed results, and wrote the article with comments and reviews from SS, EB, and GDP. EB and GDP supervised the primary author.*

In this chapter, we present BioADAPT-MRC, a transfer learning framework that focuses on learning domain-invariant feature representations to achieve high performance in the biomedical machine reading comprehension (MRC) task. BioADAPT-MRC stands for **A**dversarial learning-based **D**omain ad**APT**ation framework for **Bio**-medical **M**achine **R**eading **C**omprehension. Domain-invariant features imply that the feature representations extracted from the source- and the target-domain samples are closer in the feature embedding space. We start the chapter with a brief overview of the motivation behind the framework, followed by a formal problem definition. We then provide a detailed description of BioADAPT-MRC and report the outcomes of extensive validations of the framework for performing biomedical MRC tasks on three commonly used and

publicly available benchmark datasets—BioASQ-7b, BioASQ-8b, and BioASQ-9b. Finally, we end the chapter with a thorough error analysis of the approach to identify its strengths and weaknesses. To facilitate reproducibility, we make the BioADAPT-MRC framework freely available as an open-source project at https://github.com/mmahbub/BioADAPT-MRC.

3.1 Motivation

As explained in Chapter 2, the availability of large-scale, high-quality annotated datasets plays a crucial role in unleashing the full potential of state-of-the-art MRC models that employ deep neural networks, enabling them to excel in the MRC task and achieve high performance. Building such datasets in the biomedical domain is challenging as it requires domain knowledge from (subject-matter experts) SMEs, which is expensive to manage. Therefore, the issue posed by labeled-data insufficiency in a new or less-explored biomedical problem domain still needs to be tackled. To resolve this issue, we use transfer learning that transmits knowledge from a widely-explored source domain with labeled samples to the target biomedical domain with unlabeled or few labeled samples. Nonetheless, as elaborated in Chapter 2, the samples in the source and target domains are often not i.i.d. due to the variances in the topics between the domains. According to the assumptions of machine/deep learning, this phenomenon results in the discrepancy in marginal distributions between the domains. As a result, when we try to apply transfer learning with direct transferring, i.e., transmitting the feature representations learned by a model (from being trained on a source domain) directly to the biomedical domain, it often degrades the performance of the model [173].

Our framework, BioADAPT-MRC, attempts to solve this issue by computationally reducing the distributional dissimilarity. It uses adversarial learning to generate domain-invariant feature representations for better domain adaptation in biomedical MRC models. In the BioADAPT-MRC framework, we train two adversaries – feature extractor and discriminator – jointly against one another with labeled data from the source domain and unlabeled data from the target domain. In that way, BioADAPT-MRC attempts to generate feature representations that are discriminative from the viewpoint of the MRC task in the source domain and indiscriminative from the perspective of the shift in the marginal distributions between the source and target domains [173]. Thus, it

relaxes the need for generating labels for training a well-performing MRC model for a new biomedical/clinical problem domain and the need for expensive domain expertise for data annotations.

It is worth noting that some recent domain adaptation techniques for the MRC task concentrate on generating artificial question-answer pairs to augment the training dataset, as observed in studies like [25, 26]. However, our approach exclusively relies on the unlabeled contexts from the target domain. This aspect makes our framework more appropriate in scenarios where obtaining annotated data is challenging and generating synthetic question-answer pairs is computationally demanding and necessitates domain experts' further assessment.

We validate BioADAPT-MRC on three benchmark datasets from the BioASQ [27], built from PubMed articles and commonly used for biomedical question-answering. We employed the recommended evaluation metrics by BioASQ to validate our approach. Our experimental results show that BioADAPT-MRC can achieve state-of-the-art performance on these datasets, even without any labeled data from the biomedical domain, whether synthetic or human-annotated. To further assess the domain adaptation capability of our approach, we utilize clustering and dimensionality reduction techniques. Moreover, we extend our framework to a semi-supervised scenario and assess its performance with varying ratios of labeled target-domain data. Lastly, we conduct a comprehensive error analysis of our proposed framework to demonstrate its strengths and limitations.

3.2 Material & Methods

This section discusses the problem formulation and the BioADAPT-MRC framework (Figure 3-1), which consists of three primary components: feature extractor, domain similarity discriminator, and MRC-module.

3.2.1 Problem Definition

We assume that there exists an unlabeled biomedical problem domain (i.e., target domain), denoted as $\mathcal{D}_t$, and a labeled source domain, denoted as $\mathcal{D}_s$, each associated with their own learning tasks, $\mathcal{T}_t$ and $\mathcal{T}_s$, respectively. We assume that the tasks are equivalent, i.e., $\mathcal{T}_t = \mathcal{T}_s$, but the domains are distinct because the marginal probability distribution of the target domain, $p(X_t \in \mathcal{X})$, is not the same as that of the source domain, $p(X_s \in \mathcal{X})$.

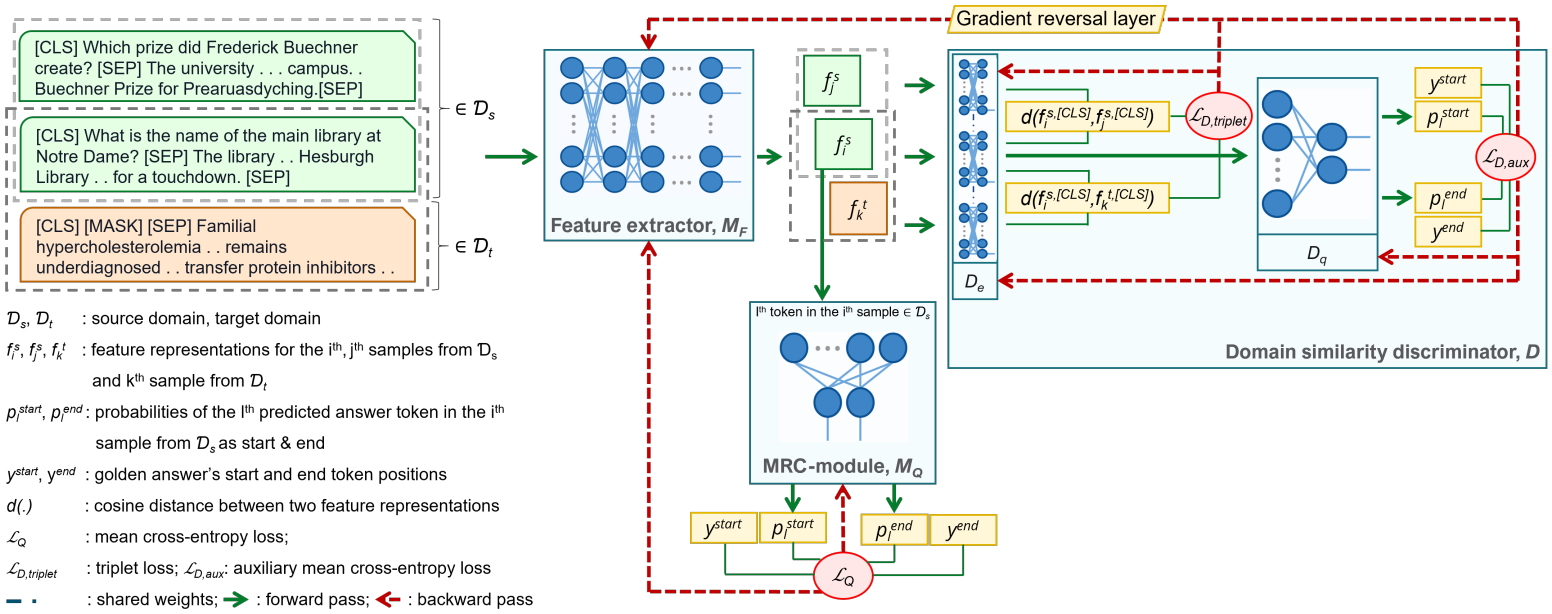


Figure 3-1. BioADAPT-MRC: an adversarial learning-based domain adaptation framework for biomedical machine reading comprehension task. The framework has three main components: (i) feature extractor M_F, (ii) MRC-module M_Q, and (iii) domain similarity discriminator D. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

Here, $p(.)$ denotes the marginal probability distribution, and X_t and X_s are learning samples from the target and source domains, correspondingly. The differences in $p(X_s)$ and $p(X_t)$ are attributed to dissimilarities in the topics and hence, the word distributions of the learning samples between the two domains, despite both domains using the same language.

In this work, we designate a general-purpose domain (such as Wikipedia articles) as the source domain, which contains distinct topics from the biomedical problem domains (such as biomedical scholarly articles). As a result, we postulate that $p(X_t) \neq p(X_s)$. The task for both domains is MRC (i.e., $\mathcal{T}_t = \mathcal{T}_s$).

Given a question $Q = \{q_1, q_2, ..., q_n\}$ and a context $C = \{c_1, c_2, ..., c_m\}$, MRC predicts the start and end positions a^{start} and a^{end}, respectively, of the answer span $A = \{c_i\}_{a^{start}}^{a^{end}}$ in C such that there exists one and only one answer span consisting of continuous tokens in the context. Here, q_i denotes the i^{th} token in the question, c_i denotes the i^{th} token in the context, and n, m respectively denote the number of tokens in Q and C.

3.2.2 BioADAPT-MRC

The proposed BioADAPT-MRC framework accepts the labeled and unlabeled instances, respectively, from the source and target domains and endeavors to accomplish two primary objectives, namely (i) the prediction of answer spans from the provided contexts, and (ii) the mitigation of the differences in marginal distributions between the inputs from the source and target domains by generating domain-invariant features. BioADAPT-MRC consists of three primary components, illustrated in Figure 3-1.

- *Feature extractor* takes a sequence of text as input. It then transforms the text sequence into a continuous vector representation with high dimensionality.
- *MRC-module* receives the output of the feature extractor, which is the vector representation of the input sample from either the source domain (during training) or the target domain (during testing) as input. It then predicts the start and end positions of the answer span A within C using the input vector representation.

- *Domain similarity discriminator* receives the vector representations (the outputs of the feature extractor) of the samples from the source and target domains. It then learns to differentiate between source and target-domain representations. The goal of the discriminator is to encourage the feature extractor to generate domain-invariant vector representations.

3.2.2.1 Feature Extractor

The feature extractor $M_F(.)$ maps the extracted feature representation f_i for the i^{th} input sample X_i from either $\mathcal{D}_s$ or $\mathcal{D}_t$ into a shared feature space F:

$$f_i = M_F(X_i), \;\; X_i \in \;\; \mathcal{D}_s \cup \mathcal{D}_t \tag{3.1}$$

As explained in Chapter 2, the encoders in the state-of-the-art PLMs are the first choice among researchers to use as trainable feature extractors in many downstream NLP tasks due to their unmatched performance. The feature extractor used in our work is the encoder of the PLM BioELECTRA [129]. We opted for BioELECTRA for the following reason: at the time of this work, it achieved the highest performance scores in various biomedical NLP tasks compared to other existing biomedical domain-specific BERT models such as BioBERT [83], SciBERT [86], and PubMedBERT [128] – according to the *Biomedical Language Understanding and Reasoning Benchmark (BLURB)* [128].

The vector representations produced by the BioELECTRA encoder are high-dimensional trainable word embeddings that are dependent on the surrounding contexts and extracted from the input question-context pairs. The BioELECTRA encoder generates these word embeddings by utilizing the WordPiece tokenization algorithm [250] and the architecture from the ELECTRA encoder—a transformer-based deep neural network which is also a variant of the BERT model [96]. The ELECTRA encoder model has 12 layers, with a hidden size of 768, 3072 for the feed-forward network (FFN) inner hidden size and 12 attention heads per layer. Although BioELECTRA uses the architecture of the ELECTRA model, it is pre-trained on a biomedical corpus with 3.2 million full-text articles from PubMed Central and 22 million abstracts from PubMed, making it more suitable for biomedical NLP tasks. The pre-training task involves predicting replaced tokens. Its vocabulary consists of 30522 words, and the maximum number of tokens that

can be used per input is 512, with each token having an embedding dimension of 768.

The final tokenized input to the BioELECTRA encoder model for each question-context pair consists of $\{[CLS], Q, [SEP], C, [SEP]\}$, where Q and C are tokens from the question and context respectively. The special token $[CLS]$ represents an accumulated representation of the input sequence [20] and is utilized for classification tasks, while $[SEP]$ is used to separate consecutive sequences. It should be noted that in the absence of labels in the target domain samples, the special token $[MASK]$ is used in place of the question tokens Q to maintain consistency in the structure of the tokenized samples. The $[MASK]$ token ensures that the model understands the absence of the question tokens.

3.2.2.2 MRC-module

To implement the MRC-module $M_Q(.)$, we incorporate a simple fully-connected neural network layer with a hidden size H of 768 above the feature extractor $M_F(.)$. We utilize the softmax activation function, as outlined in Equation 3.2, to generate probability distributions for the positions of start and end tokens in the answer spans.

$$p_l^{start} = \frac{exp^(W^{start} \cdot h_l)}{\sum_{k=1}^{n_{seq}} exp^{W^{start} \cdot h_k}}; \quad p_l^{end} = \frac{exp^(W^{end} \cdot h_l)}{\sum_{k=1}^{n_{seq}} exp^{W^{end} \cdot h_k}} \tag{3.2}$$

Here, p_l^{start} and p_l^{end} are the probabilities of the l^{th} token to be predicted as *start* and *end* respectively, $h_l \in \mathbb{R}^H$ is the hidden representation vector of the l^{th} token, $W^{start} \in \mathbb{R}^H$ and $W^{end} \in \mathbb{R}^H$ are two trainable weight matrices, n_{seq} is the input sequence length. We use the cross-entropy loss $\mathcal{L}_Q$ on the predicted answer positions as the objective function for the M_Q. Since for each answer span prediction, we get two predicted outputs for the start and end positions; we average the total cross-entropy loss as shown in Equation 3.3.

$$\mathcal{L}_Q = -\frac{1}{2}(logp_{y^{start}}^{start} + logp_{y^{end}}^{end}) \tag{3.3}$$

Here the golden answer's start and end token positions are represented by y^{start} and y^{end}, respectively. During the test phase, the predicted answer span is selected based on the positions of the highest probabilities from $p_{k\in[1,n_{seq}]}^{start}$ and $p_{k\in[1,n_{seq}]}^{end}$.

3.2.2.3 Domain Similarity Discriminator

The domain similarity discriminator $D(.)$ addresses the domain variance between two domains (caused by the discrepancies in the marginal probability distributions) as follows: In the adversarial setting, $D(.)$ learns to distinguish between the feature representations of the source-domain and target-domain samples generated by the feature extractor. $D(.)$ then penalizes the feature extractor for producing domain-variant feature representations and thus promotes the generation of domain-invariant features. $D(.)$ uses cosine distance between the feature representations of the input samples to distinguish between the domains. We consider that two samples are closer in the embedding space and thus have a greater chance to be in the same domain if their feature representations have a smaller cosine distance between them and vice-versa. The input of $D(.)$ is a triplet (f_k^t, f_i^s, f_j^s) where f_k^t, f_i^s, and f_j^s are respectively the feature representations of the k^{th} sample X_k^t from the target domain and i^{th} sample X_i^s and j^{th} sample X_j^s from the source domain extracted by $M_F(.)$. The triplet is then split into two distinct pairs, consisting of (f_i^s, f_k^t) and (f_i^s, f_j^s). As indicated in Figure 3-1, upon receiving each triplet, D accomplishes two tasks: i) measures the similarity between (f_i^s, f_j^s) and dissimilarity between (f_i^s, f_k^t), ii) performs MRC task similar to M_Q for the source sample X_i^s.

We introduce a Siamese network [251] D_e with a single transformer encoder layer for the first task. D_e acts as a function that helps estimate the similarity and dissimilarity between the received pairs. Considering the success of the $BERT$ models in many NLP tasks, for the Siamese network, we adopt the same architecture as any encoder layer in the $BERT_{Base}$ model, which has 12 attention heads, 768 embedding dimensions, 3072 FFN inner hidden size with 10% dropout rate and 'GeLU' activation function. We encode the input pairs (f_i^s, f_k^t) and (f_i^s, f_j^s), using the same encoder network D_e. Considering the role of the special token $[CLS]$ as explained in 3.2.2.1, to let $D(.)$ differentiate whether the pairs are from the same domain or not, we extract the $[CLS]$ token representations from D_e for f_k^t, f_i^s, and f_j^s. We then use these $[CLS]$ token representations to calculate the domain-similarity and dissimilarity via triplet loss function [252] and use it as the learning objective of the discriminator $D(.)$ as shown in Equation 3.4.

$$\begin{aligned}\mathcal{L}_{D,triplet} = max\{d(f_i^{s,[CLS]}, f_j^{s,[CLS]}) \\ -d(f_i^{s,[CLS]}, f_k^{t,[CLS]}) + \alpha, 0\}\end{aligned} \tag{3.4}$$

Here, $f_i^{s,[CLS]}, f_j^{s,[CLS]}$, $f_k^{t,[CLS]}$ are respectively the $[CLS]$ token representations of i^{th} and j^{th} samples from the source domain and the k^{th} sample from the target domain where $i \neq j$, $d(.) = 1.0 - CosSim(.)$ is the cosine distance where $CosSim(.)$ is the cosine similarity, α is the non-negative margin representing the minimum difference between $d(f_i^{s,[CLS]}, f_j^{s,[CLS]})$ and $d(f_i^{s,[CLS]}, f_k^{t,[CLS]})$ that is required for the triplet loss to be 0. To optimize $D(.)$, $\mathcal{L}_{D,triplet}$ aims at minimizing $d(.)$ between the samples from the source domain and maximizing $d(.)$ between the samples from the source and target domains. Using triplet loss, our discriminator efficiently employs both similar and dissimilar information extracted by the feature extractor component of the model.

While using triplet loss in the adversarial setting, we also consider that the triplet loss might make representations from the source domain dissimilar. Therefore, as an additional experiment, we optimize the discriminator by minimizing a distance-based loss, which is equivalent to just minimizing the distance $d(.)$ between the source- and the target-domain sample representations in the adversarial setting. We demonstrate the comparison between these two approaches in Section 3.3.5.1.

Adversarial learning for domain adaptation is known to be unstable [190, 191]. To stabilize the training process, we use the concept of AC-GAN [253]. AC-GAN uses an auxiliary task layer on top of the discriminator and appears to stabilize the adversarial learning procedure and improve the performance [253].

Following this concept, for the second task of $D(.)$, we introduce an MRC-module D_q similar to M_Q on top of D_e as an auxiliary task layer. D_q enforces that the discriminator does not lose task-specific information while learning to encode domain-variant features. Later in Sections 3.3.4 and 3.3.5.2, we demonstrate the effectiveness of the auxiliary layer by performing an ablation study and a stability analysis. The input of D_q is the output of D_e for f_i^s, and the output is the probability distributions for the start and end token positions of the answer span, similar to (p_i^{start}, p_j^{end}). Thus, the loss function $\mathcal{L}_{D,aux}$ for D_q is the same as $\mathcal{L}_Q$. The final loss function $\mathcal{L}_D$ for $D(.)$ is shown in Equation 3.5.

$$\mathcal{L}_D = \mathcal{L}_{D,triplet} + \mathcal{L}_{D,aux} \tag{3.5}$$

3.2.2.4 Cost Function

To eliminate domain shift by learning domain-invariant features, we integrate $M_F(.)$, $M_Q(.)$, and $D(.)$ into an adversarial learning framework where we update M_F and M_Q to maximize $\mathcal{L}_D$ and minimize $\mathcal{L}_Q$ while updating D to minimize $\mathcal{L}_D$. Thus, the cost function $\mathcal{L}_{total}$ of the BioADAPT-MRC framework consists of $\mathcal{L}_Q$ and $\mathcal{L}_D$ as shown in Equation 3.6 and is optimized end-to-end:

$$\mathcal{L}_{total} = \mathcal{L}_Q - \lambda \mathcal{L}_D \tag{3.6}$$

Here, λ is a regularization parameter to balance $\mathcal{L}_Q$ and $\mathcal{L}_D$. Unlike the original adversarial learning framework proposed in GAN, where the adversaries are updated alternately [254], we perform joint optimization for all three components of our model using the gradient-reversal layer [173], as suggested by [255].

3.3 Results & Discussion

We perform an extensive study to evaluate the BioADAPT-MRC framework and compare it with the state-of-the-art biomedical MRC methods on a collection of publicly available and widely used benchmark biomedical MRC datasets.

3.3.1 Dataset

To demonstrate the effectiveness of our framework, we evaluate BioADAPT-MRC and compare it with the state-of-the-art methods on three biomedical MRC datasets from the BioASQ annual challenge [27]. The BioASQ competition has been organized since 2013 and consists of two large-scale biomedical NLP tasks, one of which is question-answering (task B), which involves four types of questions – Yes/No, Factoid, List, and Summary. Among them, the factoid questions resemble the biomedical MRC task. As such, we utilize only the factoid MRC data from the BioASQ challenges held in 2019 (BioASQ-7b), 2020 (BioASQ-8b), and 2021 (BioASQ-9b) as the target-domain datasets to verify our model. These datasets were created from the search engine for biomedical literature, PubMed, with the help of domain experts. Note that, for training, our framework requires only unlabeled contexts in the target domain. As such, we only consider the contexts in the BioASQ-7b,8b, and 9b training sets and disregard the question-answer pairs. The details on the availability of the training data and the pre-processing steps are provided below:

We reinforce the reliability of the comparison of experimental results with other methods and validate the efficacy of the BioADAPT-MRC framework by using the pre-processed training set for BioASQ-7b and 8b, provided by [256] and [167], respectively. For BioASQ-9b, we pre-process the training data by retrieving full abstracts from PubMed using their provided PMIDs. We use these retrieved abstracts as the contexts in the BioASQ-9b training set. Since our framework requires no label (question-answer pair) in the training set of the target domain, for all BioASQ training sets, we disregard each question by assigning a $[MASK]$ token and each answer by replacing it with an empty string. At test time, we use the golden enriched test sets – BioASQ-7b, 8b, and 9b – from the BioASQ challenges.

Similar to the previous studies [167, 256], as the source-domain dataset, we use SQuAD-1.1 [21], which was developed from Wikipedia articles by crowd-workers. Table 3-1 shows the basic statistical information of all datasets used in the experiments. As shown, the number of training data samples in the source domain is noticeably higher than that of the target domain. The details on experimental setup and training configurations are provided below:

The BioADAPT-MRC framework consists of three main components: feature extractor, MRC-module, and discriminator. In the implementation of the framework with PyTorch [257], we initialize the feature extractor with the parameters from the pre-trained BioELECTRA model using the huggingface API [258]. We perform random initialization for the parameters in the MRC-module and the discriminator. For tokenization, we set the maximum query length to 64, the maximum answer length to 30, the maximum sequence length to 384, and the document stride to 128, as suggested in [20, 83, 129]. For optimizing the hyperparameters – learning rate, batch size, and the regularization parameter λ, we use a set of randomly selected 5000 samples from the SQuAD development dataset. We then empirically determined that learning rate 5e-5 and batch size 35 are the best choices for our experiments and for the hardware resources we are using. We also empirically set the regularization parameter λ to 0 and then increase it by 0.01 every 10 epochs up to 0.04.

For each step in the training epoch, we randomly select two samples from the source domain dataset SQuAD and one from the target domain dataset BioASQ-7b/8b/9b. We run all our experiments on a Linux server with Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz and a single Tesla V100-SXM2-16GB GPU.

3.3.2 Metrics

For evaluation, we use three metrics used in the MRC task in the official BioASQ challenge: strict accuracy (SAcc), lenient accuracy (LAcc), and mean reciprocal rank (MRR). The BioASQ challenge requires the participant systems to predict the five best-matched answer spans extracted from the context(s) in a decreasing order based on the confidence score. In the test set, for each question, the biomedical experts in the BioASQ team provided one golden answer extracted from the context. Both golden answers and predicted answer spans are used to calculate the SAcc, LAcc, and MRR scores, as shown in Equations 3.7, 3.8, and 3.9. SAcc shows the models' capability to find the exact answer location, LAcc determines the models' understanding of the predicted answers' range, and MRR reflects the quality of the predicted answer spans [27]:

$$SAcc = \frac{c_1}{n_{test}} \tag{3.7}$$

$$LAcc = \frac{c_5}{n_{test}} \tag{3.8}$$

$$MRR = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \frac{1}{r(i)} \tag{3.9}$$

Here, c_1 is the number of questions correctly answered by the predicted answer span with the highest confidence score, c_5 is the number of questions answered correctly by any of the five predicted answer spans, n_{test} is the number of questions in the test set, and $r(i)$ is the rank of the golden answer among all five predicted answer spans for the i^{th} question. If the golden answer does not belong to the five predicted answer spans, we consider $\frac{1}{r(i)} = 0$, i.e., $r(i) \to +\infty$ We implement these metrics by leveraging the publicly available tools provided by the official BioASQ challenge at https://github.com/BioASQ/Evaluation-Measures.

3.3.3 Method Comparison

We compare the test-time performance of BioADAPT-MRC on BioASQ-7b and 8b with six best-performing models selected based on related published articles: Google [259], BioBERT [256], UNCC [260], Umass [239], KU-DMIS-2020 [167], and BioQAExternalFeatures [53]. For BioASQ-9b, we pick the best-performing

system Ir_sys2 from the BioASQ-9b leaderboard (available at: http://participants-area.bioasq.org/results/9b/phaseB/).

We also consider a hypothetical system that we would get for BioASQ-9b if that system would achieve the highest SAcc, LAcc, and MRR scores on the leaderboard in all five batches of this test set. Note that, in reality, no individual system in the competition achieved the highest scores for all three metrics in all the batches.

In addition to these models, we also compare the performance of BioADAPT-MRC with AdaMRC [26], a state-of-the-art domain adaptation method for the MRC task. We provide brief descriptions of these models mentioned above below:

- *Google:* Authors in [259] have used sequential transfer learning for biomedical MRC tasks by fine-tuning a BERT PLM on two labeled general-purpose source-domain datasets – NQ [144] and CoQA [146], and a labeled target-domain dataset – BioASQ [27].
- *BioBERT:* Authors in [256] have taken a similar approach as [259] but fine-tuned a BioBERT PLM and used SQuAD as the labeled general-purpose source-domain dataset. Their system secured first place in the BioASQ-7b challenge.
- *UNCC:* In addition to the approach taken in [256], authors in [260] have used lexical answer types (LAT) as additional input features.
- *Umass:* Authors in [239] have used labeled SQuAD data and unlabeled BioASQ data for transfer learning with BioBERT and SciBERT, along with a focus on a self-denoising task that helps learn mentions of biomedical named entities in the contexts.
- *KU-DMIS-2020:* Authors in [167] have introduced the natural language inference (NLI) task in the pipeline of sequential transfer learning, where it fine-tunes BioBERT on labeled SQuAD, MNLI, and BioASQ datasets.
- *BioQAExternalFeatures:* The current state-of-the-art method on BioASQ-7b and 8b was proposed by authors in [53]. It uses externally extracted syntactic and lexical features in addition to the approach taken by authors in [256], possibly exposing to different adversarial attacks that may leverage syntactic and lexical knowledge-base from the dataset [261].

- *BioASQ-9b Challenge - Best system (Ir_sys2):* For BioASQ-9b, we pick the best performing model *Ir_sys2* from the BioASQ-9b leaderboard (available at http://participants-area.bioasq.org/results/9b/phaseB/). Each year the BioASQ question-answering challenge publishes five batches of the test set (e.g., Task 9b - Test batch 1, Test batch 2, etc.), and the participants submit their system evaluations for each of these batches separately. To find the best-performing system from the leaderboard for BioASQ-9b, we take the following steps: First, we discard the systems that did not report scores for all five batches. Then we calculate the weighted average of SAcc, LAcc, and MRR scores (across five batches) of the remaining systems, which is equivalent to reporting the results on the whole BioASQ-9b dataset (see Table 3-2). We do this because the published articles on the compared state-of-the-art models (for BioASQ-7b and 8b) reported SAcc, LAcc, and MRR scores on the whole dataset. Finally, to select the best system for BioASQ-9b, we average these mean SAcc, mean LAcc, and mean MRR scores (shown in the last column of Table 3-2) for each system and find that Ir_sys2 is the best-performing system. Ir_sys2 achieved the highest SAcc and MRR scores. For the LAcc score, Ir_sys2 achieved the second-best score with a slight difference of 0.006.

- *BioASQ-9b Challenge - Hypothetical system:* As another compared baseline for BioASQ-9b, we also considered a hypothetical system with the best SAcc, LAcc, and MRR scores in all five batches of the BioASQ-9b test set. Note that no system on the leaderboard achieved the highest scores in all three metrics for all five batches in the test set. Later in Section 3.3.4, we show that BioADAPT-MRC outperforms even this hypothetical system.

- *AdaMRC:* Authors in [26] have used an adversarial domain adaptation framework that uses a binary domain classifier, a synthetic question generator, and an MRC-module. As the MRC module, AdaMRC uses SAN, BiDAF, BERT, and ELMo. AdaMRC has been shown to improve performance for target domains such as Wikipedia (SQuAD dataset), news (NewsQA dataset [137]), and web search log (MS MARCO dataset [262]).

3.3.4 Experimental Results

Table 3-3 shows the comparison of BioADAPT-MRC with the state-of-the-art biomedical MRC methods on BioASQ-7b, 8b, and 9b. As shown, BioADAPT-MRC improves on both LAcc and MRR when tested on all three BioASQ test sets and achieves the best performance. We also notice that while our model achieves the highest SAcc score for BioASQ-9b, it achieves the second-best SAcc scores for BioASQ-7b and 8b. The higher SAcc and LAcc scores imply that our model is able to correctly extract complete answers from the given contexts more frequently than the previous methods. The higher MRR scores, on the other hand, reflect our model's ability to extract complete answers with higher probability than the previous methods. In contrast to the previous works, our method uses no label information (question-answer pairs) during the training process and has still been able to achieve good performance, implying the effectiveness of our proposed framework.

As explained in Section 3.2, in the framework, we propose a domain similarity discriminator with an auxiliary task layer that aims to promote the generation of domain-invariant features in the feature extractor and thus improve the model's performance. To show the effectiveness of the discriminator and the auxiliary task layer, we perform an ablation study and report the experimental results in Table 3-4. For a fair comparison, we perform all experiments under the same hyperparameter settings. The baseline model shown in Table 3-4 consists of only the feature extractor and the MRC-module and was trained on the labeled source domain dataset, SQuAD.

For the remaining two models, we use the labeled SQuAD and the unlabeled BioASQ training datasets simultaneously. The addition of the discriminator enables the feature extractor in the baseline model to use the unlabeled BioASQ training datasets for generating domain-invariant feature representations. This is achieved by using the dissimilarity measurements between the feature representations of the SQuAD and BioASQ training data. As shown, after adding only the discriminator without the auxiliary task layer, the performance of the model improves from the baseline, suggesting the influence of the discriminator.

Table 3-1. Statistics of the datasets used in the experiments. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

Dataset name	Train (raw)	Train (pre-processed)	Target to source train set	Test set
SQuAD-1.1	87,599	87,599	–	–
BioASQ-7b	779	5,537	∼1:16	162
BioASQ-8b	941	10,147	∼1:9	151
BioASQ-9b	1,092	13,178	∼1:7	163

Table 3-2. Scores from BioASQ-9b leaderboard. Here, 'Average' implies the average of mean SAcc, mean LAcc, and mean MRR scores. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

System	Mean SAcc	Mean LAcc	Mean MRR	Average
Ir_sys2	0.503065	0.662573	0.566660	0.577433
lalala	0.466259	0.668706	0.544162	0.559709
Ir_sys1	0.478536	0.638055	0.544579	0.553723
Ir_sys4	0.447864	0.644181	0.518606	0.536884
LASIGE_ULISBOA	0.411042	0.650313	0.505514	0.522290
Ir_sys3	0.423315	0.631908	0.497349	0.517524
LASIGE_ULISBOA_3	0.404896	0.631904	0.497517	0.511439
LASIGE_ULISBOA_2	0.404911	0.631901	0.482426	0.506413
KU-DMIS-5	0.404922	0.607362	0.480883	0.497722
KU-DMIS-2	0.404922	0.601233	0.480654	0.495603
bio-answerfinder-2	0.453973	0.539867	0.492339	0.495393
bio-answerfinder	0.453973	0.539867	0.492339	0.495393
KU-DMIS-3	0.404902	0.595119	0.481271	0.493764
KU-DMIS-1	0.398782	0.582811	0.469935	0.483843
KU-DMIS-4	0.386517	0.588975	0.463591	0.479695
Best factoid	0.368108	0.527612	0.430671	0.442130
AUEB-System4	0.368087	0.515336	0.426766	0.436729
Best yesno	0.355813	0.503080	0.414328	0.424407
AUEB-System3	0.349711	0.484645	0.408483	0.414280
AUEB-System2	0.368114	0.453992	0.407663	0.409923
UoT_baseline	0.337436	0.484642	0.397855	0.406644
UoT_multitask_learn	0.337421	0.478522	0.392840	0.402928
AUEB-System1	0.318998	0.429440	0.365547	0.371328
UoT_allquestions	0.214723	0.380357	0.276910	0.290663
BioASQ_Baseline	0.079759	0.325160	0.167685	0.190868

Table 3-3. Performance of BioADAPT-MRC compared with the best scores on BioASQ-7b, BioASQ-8b, and BioASQ-9b test sets. The best and the second-best scores are respectively highlighted in bold and italic. '–' indicates that the corresponding source did not report the scores. * denotes previously best-performing method for BioASQ-7B and BioASQ-8B. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

Model	BioASQ-7b			BioASQ-8b			BioASQ-9b		
	SAcc	LAcc	MRR	SAcc	LAcc	MRR	SAcc	LAcc	MRR
Google [259]	0.4201	0.5822	0.4798	–	–	–	–	–	–
BioBERT [256]	0.4367	0.6274	0.5115	–	–	–	–	–	–
UNCC [260]	0.3554	0.4922	0.4063	–	–	–	–	–	–
Umass [239]	–	–	–	0.3133	0.4798	0.3780	–	–	–
KU-DMIS-2020 [167]	**0.4510**	0.6245	0.5163	0.3819	0.5719	0.4593	–	–	–
BioQAExternalFeatures [53]*	0.4444	*0.6419*	*0.5165*	**0.3937**	*0.6098*	*0.4688*	–	–	–
BioASQ-9b Challenge - Best system (Ir_sys2)	–	–	–	–	–	–	0.5031	0.6626	0.5667
BioASQ-9b Challenge - Hypothetical system	–	–	–	–	–	–	**0.5399**	*0.7300*	*0.6017*
AdaMRC [26]	0.4321	0.6235	0.5136	0.3510	0.5828	0.4455	*0.5337*	0.7117	0.6001
BioADAPT-MRC	*0.4506*	**0.6420**	**0.5289**	*0.3841*	**0.6225**	**0.4749**	**0.5399**	**0.7423**	**0.6187**

Table 3-4. Test scores for ablation experiments of BioADAPT-MRC. The best and the second-best scores are respectively highlighted in bold and italic. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

Model	BioASQ-7b			BioASQ-8b			BioASQ-9b		
	SAcc	LAcc	MRR	SAcc	LAcc	MRR	SAcc	LAcc	MRR
Baseline	0.4136	*0.6296*	0.5056	0.3642	0.5960	0.4602	0.5092	0.7362	0.6010
BioADAPT-MRC (no auxiliary layer)	*0.4259*	*0.6296*	*0.5146*	*0.3775*	*0.6026*	*0.4679*	*0.5276*	**0.7485**	*0.6142*
BioADAPT-MRC	**0.4506**	**0.6420**	**0.5289**	**0.3841**	**0.6225**	**0.4749**	**0.5399**	*0.7423*	**0.6187**

We explain this influence on the feature extractor more elaborately later in Section 3.3.5.5. For the final experiment in the ablation study (Table 3-4), we use our whole model consisting of the domain similarity discriminator with the auxiliary task layer and notice an even further performance improvement. The auxiliary task layer, in this study, constrains the changes in the task-relevant features in the domain similarity discriminator during training.

Thus, the improvement in model performance after incorporating the auxiliary task layer suggests that with the task layer, the domain similarity discriminator can better promote the generation of domain-invariant features that are simultaneously discriminative from the viewpoint of the MRC task in the source domain. Moreover, as explained in Section 3.2.2.3, we further demonstrate the stabilizing capability of the auxiliary task layer in Section 3.3.5.2.

3.3.5 Analysis

In this section, we analyze different components of the proposed framework. We also study the domain adaptation capability and the strengths and weaknesses of the BioADAPT-MRC model.

3.3.5.1 Triplet vs. Distance-based Loss

BioADPT-MRC uses triplet loss to optimize the discriminator. As explained in Section 3.2.2.3, we also consider using distance-based loss in place of triplet loss. Table 3-5 shows the results of the contrastive experiments of loss functions – triplet loss and distance-based loss. As shown in Table 3-5, the model with triplet loss outperforms the one with the distance-based loss with higher mean SAcc, LAcc, MRR, and lower standard deviation.

To further analyze this performance gap, we examine the trend of the distance between source-domain training sample pairs and between source- and target-domain training sample pairs per epoch across 50 training epochs (Figure 3-2).

Note that both experiments in Figure 3-2 are performed under the same seed. As shown in Figure 3-2, when we use distance-based loss, the cosine distance between either source-domain training sample pairs or source- and target-domain training sample pairs tends to be higher than when we use the triplet loss. It may happen because in the adversarial framework, while distance-based loss focuses only on minimizing the distance between the source- and target-domain training samples

Table 3-5. Average test scores with standard deviations (across three different seeds) for the contrastive experiments of discriminator loss functions – triplet loss and distance-based loss. The best scores are highlighted in bold. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

Discriminator Loss	BioASQ-9b		
	SAcc	LAcc	MRR
Distance-based	0.5256±0.0202	0.7219±0.0104	0.6038±0.0105
Triplet	**0.5358±0.0029**	**0.7321±0.0077**	**0.6140±0.0035**

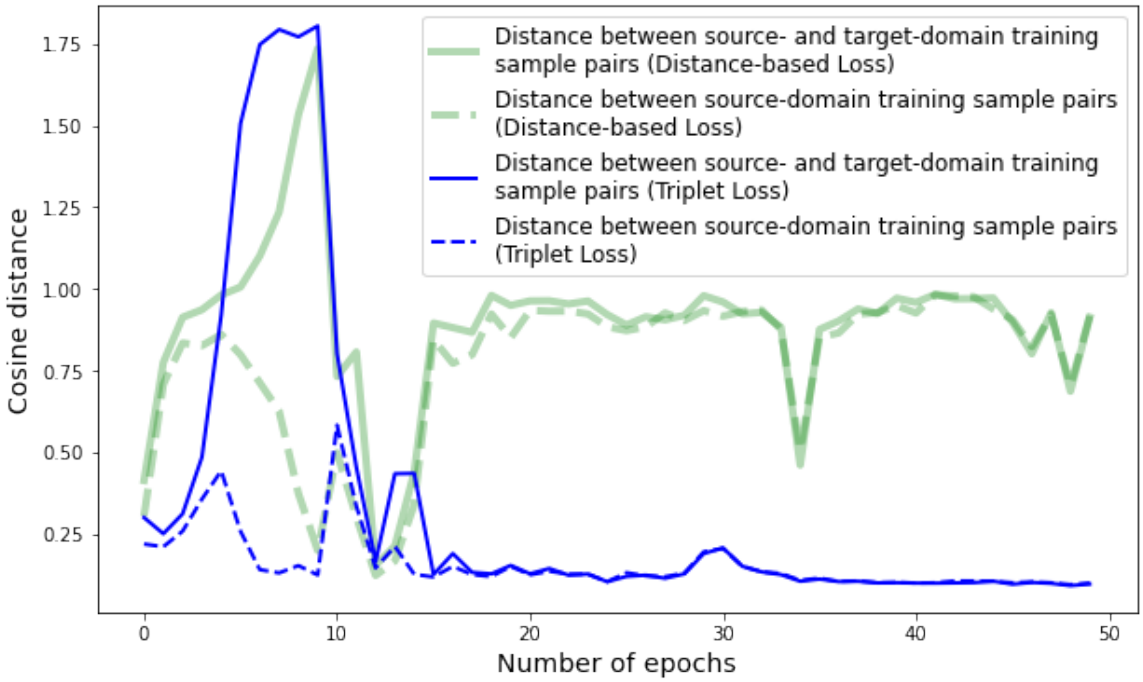


Figure 3-2. Per-epoch cosine distance between source-domain training sample pairs and source- and target-domain training sample pairs across 50 epochs. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

without considering the distance between the source samples, triplet loss focuses on balancing both [263, 264]. As a result, triplet loss can minimize the domain shift to a greater extent than the distance-based loss and thus enable our framework to achieve higher performance.

3.3.5.2 Stability Analysis

We examine the stability of BioADAPT-MRC and perform an error analysis of its performance by repeating the experiments for three different random seeds (10, 42, 2018). Table 3-7 shows the test scores averaged across three seeds with standard deviations for BioASQ-7b, 8b, and 9b.

We also report the averaged scores with standard deviations for our baseline, the BioADAPT-MRC model with no auxiliary layer, and a state-of-the-art method AdaMRC to compare model stability. As shown in Table 3-7, BioADAPT-MRC outperforms the other models with lower standard deviations, indicating higher stability of our framework. Moreover, the scores from the BioADAPT-MRC models with and without the auxiliary layer indicate that the auxiliary task layer helps increase both performance and overall model stability.

3.3.5.3 Masked vs. Synthetic Questions

Recall that, BioADAPT-MRC uses a special token $[MASK]$ in place of the question tokens Q for the unlabeled target-domain training samples. The $[MASK]$ tokens are used to inform the encoder model that the question tokens are missing and maintain consistency in the structure of the tokenized samples. Another approach to address the issue of missing questions in the target-domain training samples is to use synthetic questions [26]. In Table 3-6, we present the results of the contrastive experiments of these two approaches – masked and synthetic questions. Inspired by the success of the AdaMRC question-generator in various target domains such as news, Wikipedia, and web search log, we use it to generate synthetic questions.

Table 3-6 shows the average test scores for BioASQ-9b with standard deviations across three seeds. We find that although the synthetic questions can noticeably improve performance over the baseline (see results for 'Baseline' in Table 3-7), BioADAPT-MRC with synthetic questions is unable to achieve better performance than with masked questions. It may happen because the biomedical domain differs

Table 3-6. Average test scores with standard deviations (across three different seeds) for experiments using synthetic questions and masked questions in the target-domain training dataset. The best and the second-best scores are respectively highlighted in bold and italic. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

Questions	BioASQ-9b		
	SAcc	LAcc	MRR
Synthetic	*0.5337±0.0087*	*0.7198±0.0126*	*0.6069±0.0059*
Masked	**0.5358±0.0029**	**0.7321±0.0077**	**0.6140±0.0035**

Context:
Psoriasiform dermatitis in a case of newly diagnosed locally advanced pyriform sinus tumour: Bazex syndrome revisited. Acrokeratosis paraneoplastica of Bazex is a rare but important paraneoplastic dermatosis, usually manifesting as psoriasiform rashes over the acral sites. It often precedes diagnosis of the associated malignancy, usually that of upper aerodigestive tract squamous cell carcinoma. We present the case of a patient with a newly diagnosed pyriform sinus tumour and associated acrokeratosis paraneoplastica. To the best of our knowledge, this is the first reported case in the local literature.

Synthetic questions:
- ❑ What is the first rare?
- ❑ Why was the first patient of a patient in a patient diagnosed with the patient in the first patient to be a patient in the?
- ❑ What did advanced dermatitis in order dermatitis?
- ❑ What case did advanced cell of locally in a case?
- ❑ To what name is the first diagnosed important in advanced cell of locally?
- ❑ How upper case did the first sites in the local literature of the local literature?
- ❑ What case was reported in the first case in local literature?
- ❑ What was a <UNK> of the upper case of <UNK> first?
- ❑ What kind of advanced sites tumour?
- ❑ What is one of the first advanced cell sites?

Figure 3-3. Synthetic questions generated from the first context of the BioASQ-9b training set using the *question-generator* component of the AdaMRC model along with the context. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

Table 3-7. Average test scores with standard deviations across experiments with three random seeds (10, 42, 2018) for initialization to measure and compare the stability of BioADAPT-MRC. The best and the second-best scores are respectively highlighted in bold and italic. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

Model	BioASQ-7b		
	SAcc	LAcc	MRR
AdaMRC	*0.4300±0.0029*	0.6152±0.0116	*0.5083±0.0076*
Baseline	0.4156±0.0127	0.6173±0.0101	0.5038±0.0033
BioADAPT-MRC (no auxiliary layer)	0.4177±0.0058	*0.6193±0.0105*	0.5038±0.0076
BioADAPT-MRC	**0.4465±0.0029**	**0.6379±0.0029**	**0.5237±0.0037**
Model	BioASQ-8b		
	SAcc	LAcc	MRR
AdaMRC	0.3422±0.0083	0.5960±0.0143	0.4425±0.0031
Baseline	0.3554±0.0125	0.5960±0.0162	0.4547±0.0152
BioADAPT-MRC (no auxiliary layer)	*0.3664±0.0113*	*0.5982±0.0031*	*0.4618±0.0080*
BioADAPT-MRC	**0.3797±0.0031**	**0.6137±0.0083**	**0.4750±0.0024**
Model	BioASQ-9b		
	SAcc	LAcc	MRR
AdaMRC	0.5276±0.0050	0.7239±0.0100	0.6068±0.0062
Baseline	0.5174±0.0058	0.7198±0.0126	0.6018±0.0011
BioADAPT-MRC (no auxiliary layer)	*0.5337±0.0050*	*0.7280±0.0153*	*0.6127±0.0012*
BioADAPT-MRC	**0.5358±0.0029**	**0.7321±0.0077**	**0.6140±0.0035**

from other domains, such as news or Wikipedia, in many linguistic dimensions such as syntax, lexicon, and semantics [83, 133]. As a result, while the question-generator can generate meaningful questions for domains such as web, news, and movie reviews [26], it mostly generates incoherent questions for the biomedical domain (as shown in Figure 3-3), which can eventually hurt the performance of the model. Moreover, synthetic-question generation also requires additional computational time – generating questions from around 10,000 contexts using a trained question generator took approximately 3 hours with our computational resources. Given the findings, we think that generating synthetic questions for the biomedical domain requires more attention and consider it as a future study.

3.3.5.4 Semi-supervised Setting

As an additional experiment, we evaluate the BioADAPT-MRC framework under a semi-supervised setting where we combine labeled and unlabeled target-domain training data. We perform four experiments where the ratios of labeled target-domain training samples are 0%, 10%, 50%, and 80% of the total target-domain training data. Note that we choose the labeled target-domain data by random sampling. Table 3-8 shows the test scores on BioASQ-9b when trained with varying ratios of labeled target-domain data. As shown, with an increased ratio of labeled samples in the target-domain training data, the performance scores also increase, which is expected. These results suggest that our proposed framework is also effective in a semi-supervised setting.

Note that although multiple labeled datasets in various sub-domains of biomedical MRC (such as scientific literature and clinical notes) have been made available in the past few years [27, 151], there is still a severe scarcity of labeled data in some other sub-domains that are linguistically different (e.g., consumer health biomedical MRC) [10, 265].

3.3.5.5 Domain Adaptation

We show the influence of the domain similarity discriminator by plotting (Figure 3-4) all samples from the BioASQ-9b test set and a set of random samples from the SQuAD training set. We pick random samples from the SQuAD training set to match the number of samples in the BioASQ-9b test set. As explained in Section 3.2.2.1, we use the feature representation of the $[CLS]$ token as an accumulated representation of the whole input sequence. Each feature representation of the

[CLS] token has a dimension of 768. To reduce these dimensions into two for visualization, we use multidimensional scaling (MDS) [266]. We use MDS because it reduces the dimensions by preserving the dissimilarities between two data points in the original high-dimensional space. Since we use cosine distance in the discriminator to measure the dissimilarity between two domains, as the dissimilarity measure in the MDS, we use the pairwise cosine distance. The feature representations of the [CLS] token on the left plot and the right plot in Figure 3-4 are generated by the feature extractors from the baseline model and the BioADAPT-MRC model, respectively.

For a fair comparison, the selection of random SQuAD training samples is the same for the baseline and BioADAPT-MRC models. As shown, the features generated by the baseline model create two separate clusters for SQuAD and BioASQ-9b. The features generated by the BioADAPT-MRC model, on the other hand, form two overlapping clusters implying reduced dissimilarities between the source and target domains. Interestingly, we notice that the data points from the BioASQ are closer to its cluster than those from the SQuAD. It may be because, unlike SQuAD, the data in the BioASQ originate from one single domain, and thus the feature representations are more similar to one another.

To further analyze the quality of the clusters before and after introducing the domain similarity discriminator to the framework and thus quantify the effect of domain adaptation, we perform DBSCAN clustering [267]. We perform clustering on the MDS components of the features for the [CLS] tokens for the samples in the BioASQ test sets and the random samples from the SQuAD training set. Considering the bias of random sampling, for each BioASQ test set, we select five sets of random samples from the SQuAD training set and report the mean accuracy and silhouette scores with standard deviation in Table 3-9 and Figure 3-5. We use the DBSCAN clustering because it views clusters as high-density regions where the distance between the samples is measured by a distance metric, providing flexibility in shapes and numbers of clusters. We describe the selected hyperparameters for the DBSCAN algorithm implementation below:

As the distance measure for DBSCAN, we choose pairwise cosine distance. We implement the DBSCAN algorithm using the scikit-learn tool [268] and tune the two essential hyperparameters: eps and $min_samples$. The details on these two hyperparameters can be found in the scikit-learn documentation. After

Table 3-8. Test scores for experiments using varied ratios of labeled target-domain training data. The best and the second-best scores are respectively highlighted in bold and italic. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

Ratio	BioASQ-9b		
	SAcc	LAcc	MRR
0%	0.5399	**0.7423**	0.6187
10%	0.5460	0.7178	0.6118
50%	*0.5644*	*0.7301*	*0.6264*
80%	**0.5828**	*0.7301*	**0.6369**

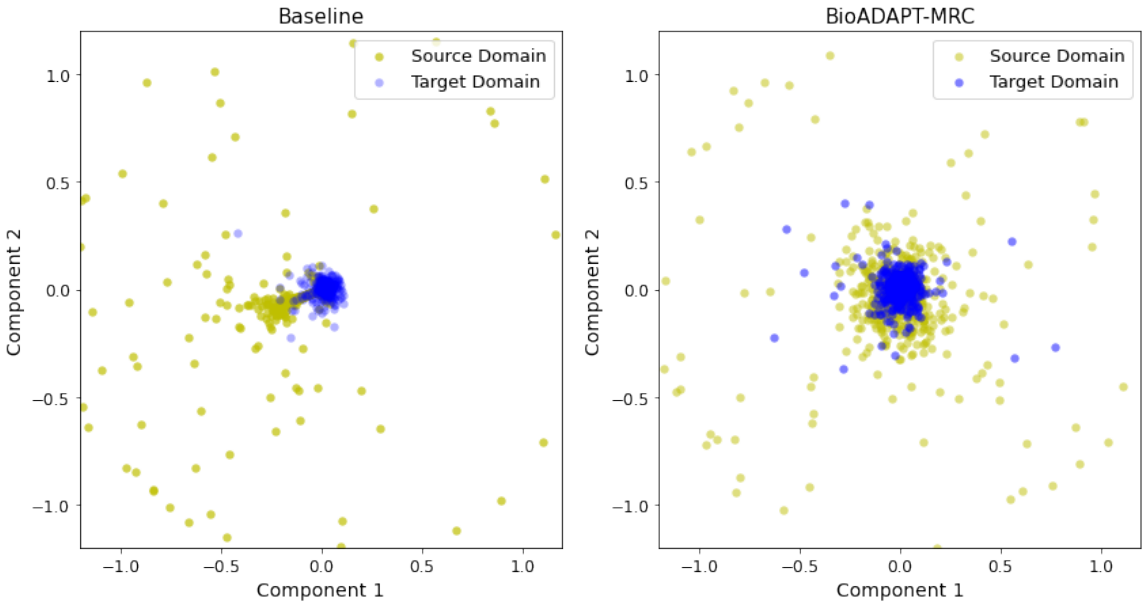


Figure 3-4. Multidimensional scaling (MDS) plot for the features (extracted from the baseline (left) and the BioADAPT-MRC (right) models) of all samples from the BioASQ-9b test set (blue) and a set of random samples from the SQuAD training set (yellow). (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

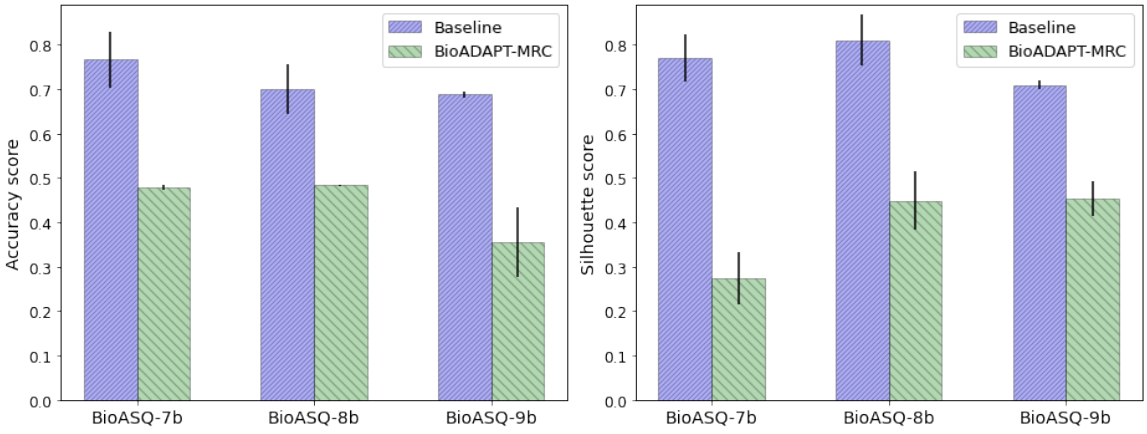


Figure 3-5. Mean accuracy scores (left) and mean silhouette scores (right) with standard deviations for DBSCAN clustering on BioASQ test sets and SQuAD. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

Table 3-9. Mean accuracy and mean silhouette scores with standard deviations for DBSCAN clustering on the MDS components. The MDS dimensionality reduction technique was applied to reduce the dimension of the $[CLS]$ token representations for the samples in the BioASQ test sets and a set of random samples from the SQuAD training set. A $[CLS]$ token represents the accumulated representations of a sample in the dataset. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

	Mean accuracy (±standard deviation)		Mean silhouette score (±standard deviation)	
	Baseline	BioADAPT-MRC	Baseline	BioADAPT-MRC
BioASQ-7b	0.7664(±0.0625)	0.4786(±0.0051)	0.7714(±0.0535)	0.2745(±0.0585)
BioASQ-8b	0.7011(±0.0563)	0.4838(±0.0015)	0.8111(±0.0577)	0.4484(±0.0660)
BioASQ-9b	0.6880(±0.0075)	0.3561(±0.0782)	0.7100(±0.0089)	0.4540(±0.0398)

optimization, we select 0.005, 0.01, and 0.001 as the *eps* for BioASQ-7b, 8b, and 9b, respectively, and 20 as the *min_samples* for all three test sets. Note that, for a fair comparison, for each test set, the choice of hyperparameters is the same for the baseline and BioADAPT-MRC models.

Table 3-9 and Figure 3-5 show that DBSCAN can identify two clusters with high accuracy when the features of the samples are extracted from the baseline model. The accuracy goes down when the features of the same samples are extracted from the BioADAPT-MRC model as they form a single cluster. Moreover, we analyze the silhouette scores to understand the separation distance between clusters.

The range of silhouette score is $[-1, 1]$. A score of 1 indicates that the clusters are highly dense and clearly distinguishable from each other, whereas -1 refers to incorrect clustering. A score of zero or near zero indicates indistinguishable or overlapping clusters. As shown in Table 3-9 and Figure 3-5, in this case, the high silhouette scores (closer to 1) for the baseline model reflect that the feature representations of the samples from the same domain are very similar to its own cluster compared to the other one. On the contrary, the low silhouette scores (closer to zero) for the BioADAPT-MRC model indicate that the feature representations of the samples from both domains are very similar to one another. These results show the effectiveness of the domain similarity discriminator in the BioADAPT-MRC framework.

3.3.5.6 Motivating Example

Considering the variability of the predicted answers in an MRC task, we present a motivating example to demonstrate how the word importance may impact the answer predictions, thus, the performance of the biomedical MRC task. The example in Figure 3-6 shows the effectiveness of BioADAPT-MRC over the baseline model for the given sample.

This example is selected randomly from the BioASQ-9b test set from the samples incorrectly predicted by the baseline model and correctly predicted by the BioADAPT-MRC model. The question is separated from the context by the $[SEP]$ token, and the correct answer, in this case, is *cortisol*. The predicted start and end tokens are highlighted in yellow. The positive and negative importance of the words is highlighted in green and red, respectively. The intensity of the color shows the amount of importance. Higher intensity reflects higher importance.

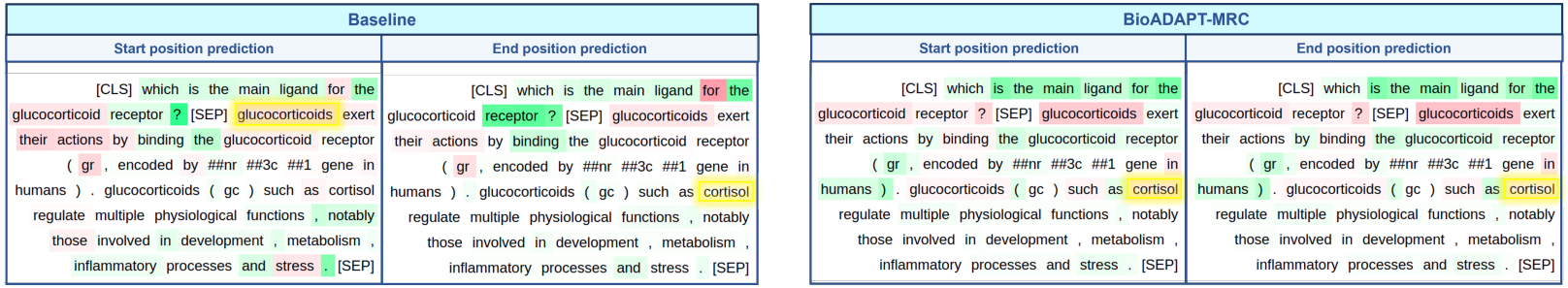


Figure 3-6. Motivating example from the BioASQ-9b test set. The left and right plots show the word importance for answer span prediction given by the baseline and BioADAPT-MRC models. The green and red colors show positive and negative word importance, respectively. The density of color shows the amount of importance. The highlighted word in yellow represents either the predicted start or end tokens. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

We calculate the word importance using the *Integrated Gradients* algorithm [269] implemented in the Captum tool [270]. Integrated gradients measure the importance of the word (in predicting the answer span) by computing the gradient of the predicted output with respect to the specified input word [269].

In this particular example, the start and end token being the same, we expect that the correct prediction of the MRC model should provide similar importance to the words for the start and end position predictions. However, while this is true for the BioADAPT-MRC model (right plot), for the baseline (left plot), we notice a different pattern in word importance while predicting for start and end positions and thus result in wrong answer span prediction. It shows the effectiveness of our model over the baseline for this example.

3.3.5.7 Error Analysis

We analyze the strengths and weaknesses of our approach by performing a two-fold error analysis. We focus on two aspects of the MRC dataset – types of questions and answers. Through this error analysis, we aim to answer the following questions: (i) What types of questions can be answered after domain adaptation that could not be answered before? (ii) What types of answers can be identified by the proposed approach after domain adaptation? (iii) What types of items does the model struggle with, even with the domain adaptation component?

We use the SAcc score to perform the error analysis since the ultimate goal for any MRC system is to predict correct answer spans with the highest probability, reflected in the SAcc score. We categorize the test sets based on the following types of questions: *which, what, how, where, when*, and *name* (an example question for *name* – "Name a CFL2 mutation which is associated with nemaline myopathy?"). We find that the most prevalent question types in the test sets are *what* (43%) and *which* (44%) (embedded in Figure 3-7). We also find that after adding the domain adaptation module, the SAcc score increases by 52%, 20%, 17%, and 4% for question types – *when, name, what,* and *which*, respectively (Figure 3-7). It indicates that the domain adaptation module can help increase the model's capability to answer these types of questions with higher probability. For question types *how* and *where*, we do not notice further improvement, even with the domain adaptation module.

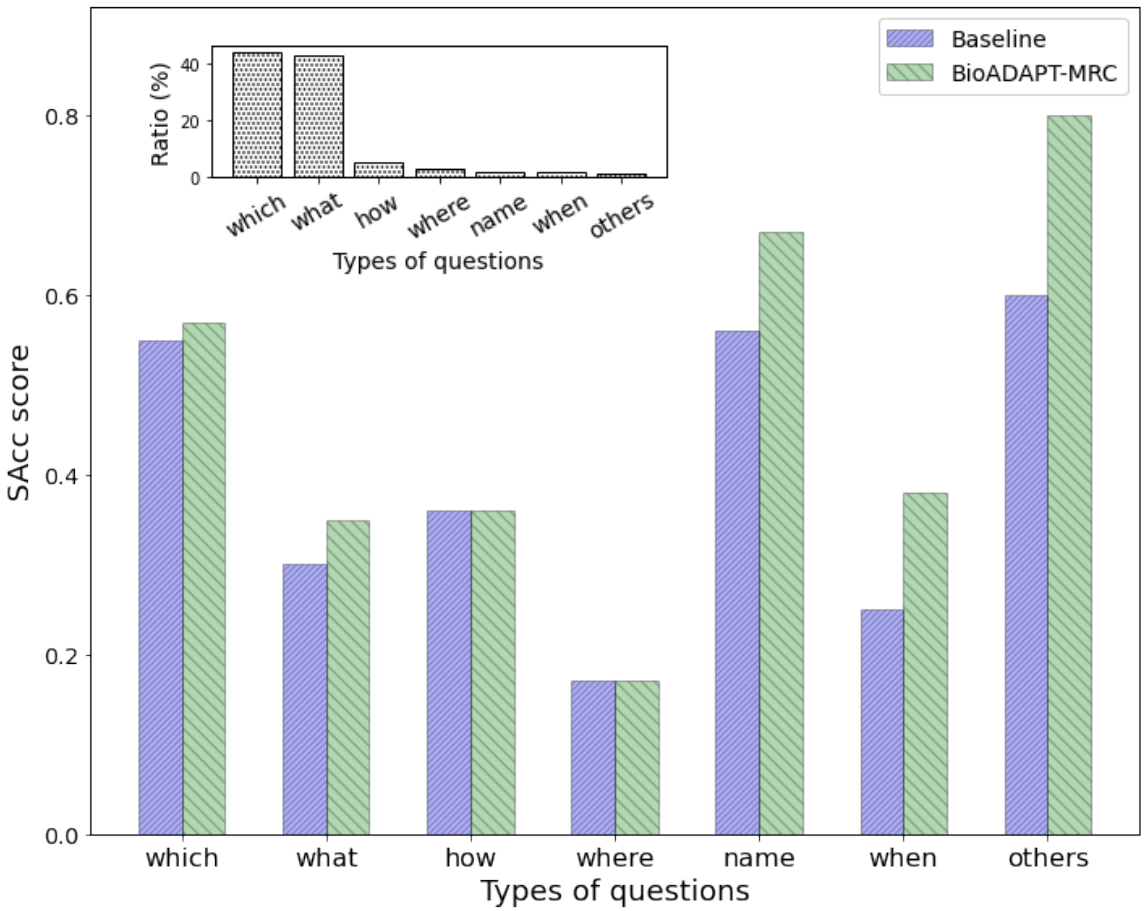


Figure 3-7. Error analysis of BioADAPT-MRC, in comparison with the baseline model, depending on the types of questions in the BioASQ test sets. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

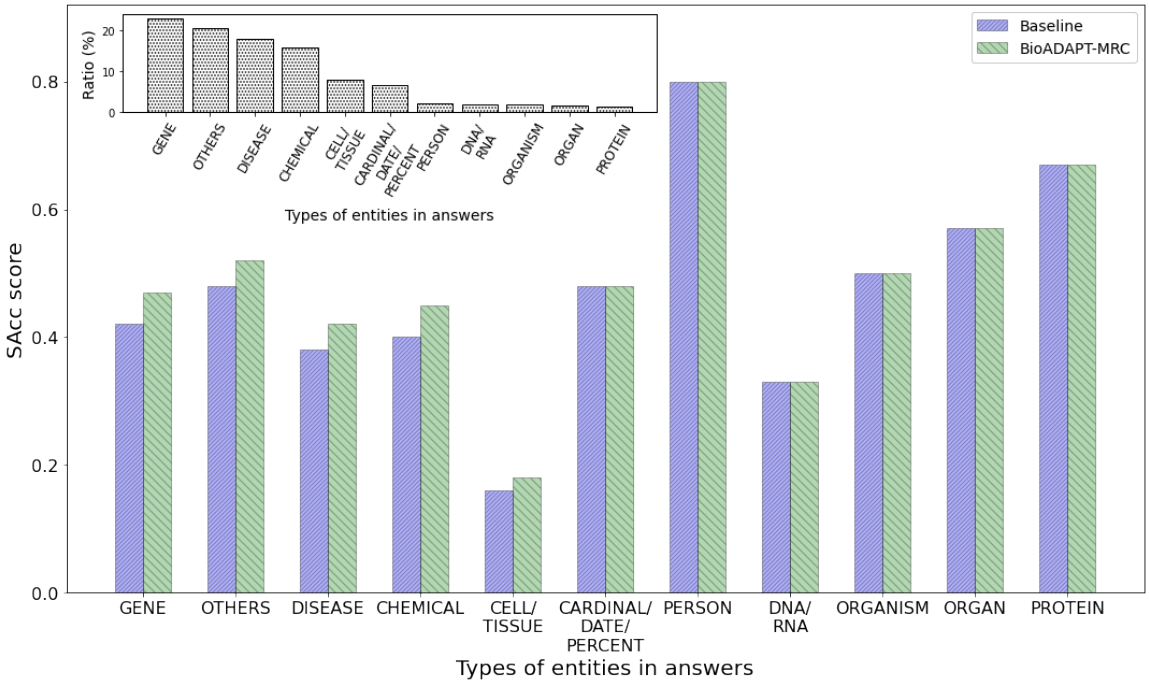


Figure 3-8. Error analysis of BioADAPT-MRC, in comparison with the baseline model, depending on the types of answers in the BioASQ test sets. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

Example 1:
Question: What is the most common monogenic cause of common variable immunodeficiency (CVID) in Europeans?
Golden Answer: Heterozygous loss-of-function variants in NFKB1
Baseline Model: NFKB1 (probability: 0.4902)
BioADAPT-MRC Model: Heterozygous loss-of-function variants in NFKB1 (probability: 0.5527)

Example 2:
Question: What does "28" stand for in the Disease Activity Score DAS28?
Golden Answer: 28 joints
Baseline Model: subjective components (probability: 0.4456)
BioADAPT-MRC Model: 28 joints (probability: 0.4951)

Example 3:
Question: Which disease is treated with Anti–Siglec-8 Antibody?
Golden Answer: Eosinophilic Gastritis and Duodenitis
Baseline Model: Eosinophilic Gastritis (probability: 0.3602)
BioADAPT-MRC Model: Eosinophilic Gastritis and Duodenitis (probability: 0.8733)

Example 4:
Question: What impacts stability of genomic imprinting in mouse pluripotent stem cells?
Golden Answer: A Susceptibility Locus on Chromosome 13
Baseline Model: A Susceptibility Locus (probability: 0.4693)
BioADAPT-MRC Model: A Susceptibility Locus on Chromosome 13 (probability: 0.5629)

Example 5:
Question: What antibiotic is currently used as the standard of care for Clostridium Difficile infection as of 2018?
Golden Answer: Fidaxomicin
Baseline Model: Fidaxomicin (probability: 0.8742)
BioADAPT-MRC Model: Fidaxomicin (probability: 0.9727)

Example 6:
Question: Approximately how many genes are contained in the X chromosome's non-pseudoautosomal region (non-PAR)?
Golden Answer: 783
Baseline Model: 783 (probability: 0.8582)
BioADAPT-MRC Model: 783 (probability: 0.9869)

Example 7:
Question: What is anophthalmia?
Golden Answer: absence of one or both eyes
Baseline Model: autosomal dominant syndrome caused by mutations in the SOX2 gene and clinically characterized by severe eye malformations (probability: 0.3139)
BioADAPT-MRC Model: severe eye malformations (probability: 0.4540)

Example 8:
Question: What is the indication for KYMRIAH?
Golden Answer: children and young adults with refractory or relapse B cell precursor acute lymphoblastic leukemia
Baseline Model: refractory or relapse (R/R) B cell precursor acute lymphoblastic leukemia (probability: 0.5414)
BioADAPT-MRC Model: treatment of pediatric patients and young adults with refractory or relapse (R/R) B cell precursor acute lymphoblastic leukemia (probability: 0.4813)

Figure 3-9. Example question-answer pairs from the test sets demonstrating the strengths and weaknesses of the BioADAPT-MRC model over the baseline model. The green and red colors show correctly and incorrectly predicted answers, respectively. (©The Author(s) 2022. Published by Oxford University Press. Reprinted from [249].)

To analyze the types of answers that can be identified after introducing the domain adaptation module, we categorize the test sets based on the named entities in the answers. We use named entity recognition (NER) algorithms from the widely used spaCy library [271]. Figure 3-8 shows that the answers in the test sets mainly consist of entities such as *GENE (23%), DISEASE (18%), CHEMICAL (16%), CELL/TISSUE (8%),* and *CARDINAL/DATE/PERCENT (7%).*

As shown in Figure 3-8, after adding the domain adaptation module, the model has been able to identify the *CHEMICAL, CELL/TISSUE* entities with the highest SAcc scores (improvement by 13% from the baseline). We also notice an improvement in the SAcc scores for *GENE* and *DISEASE* entities. However, we notice no improvement in the SAcc scores for *CARDINAL/DATE/PERCENT, PERSON, DNA/RNA, ORGANISM, ORGAN, and PROTEIN* entities even after adding the domain adaptation module. Given the results, we would like to emphasize that in both of the aforementioned analyses, for some categories, we do not have a large enough sample set to draw a definite conclusion about the effectiveness of the domain adaptation module and hence requires further future investigation.

To provide further evidence about what the model learned, we present eight example question-answer pairs demonstrating the strengths and weaknesses of the BioADAPT-MRC model over the baseline model (Figure 3-9). We randomly select two examples from each of the following four categories: (i) mispredicted by baseline, correctly predicted by BioADAPT-MRC, (ii) incompletely predicted by baseline, correctly predicted by BioADAPT-MRC, (iii) correctly predicted by both baseline and BioADAPT-MRC, and (iv) mispredicted by both baseline and BioADAPT-MRC. In examples 1-8, the answer spans respectively contain GENE, CARDINAL, DISEASE, CELL, CHEMICAL, CARDINAL, ORGAN, and DISEASE entities. As shown in examples 1-6, the BioADAPT-MRC model can identify the answer span better than the baseline model with a higher probability score. However, the probability scores (i.e., the prediction capability) can be further improved. Moreover, examples 7 and 8 provide additional motivation for future investigation of the reason behind the misprediction of the model.

The results from the overall error analysis indicate that while the BioADAPT-MRC model does well under various scenarios, there is still significant room for potential improvement.

3.4 Conclusion

We have proposed a framework for biomedical machine reading comprehension, BioADAPT-MRC, addressing the issue of domain shift by using a domain adaptation technique in an adversarial learning setting. We have used a labeled MRC dataset from a general-purpose domain (source domain) along with unlabeled contexts from the biomedical domain (target domain) as our training data. We have introduced a domain similarity discriminator, aiming to reduce the domain shift between the general-purpose domain and the biomedical domain to help boost the performance of the biomedical MRC model. We have validated our proposed framework on three widely used benchmark datasets from the biomedical question answering and semantic indexing challenge, BioASQ. We have comprehensively demonstrated that without any label information in the target domain during training, the BioADAPT-MRC framework can achieve state-of-the-art performance on these datasets. We have performed an extensive quantitative study on the domain adaptation capability using dimensionality reduction and clustering techniques and have shown that our framework can learn domain-invariant feature representations. Additionally, we have extended our framework to a semi-supervised setting and demonstrate that our framework can be efficiently applied even with varying ratios of labeled data. We have performed a two-fold error analysis to investigate the shortcomings of our framework and provide motivation for further future investigation and improvement.

CHAPTER 4

CPGQA: MACHINE READING COMPREHENSION ON CLINICAL PRACTICE GUIDELINES

Much of the content in this chapter is taken without alteration from the published article ***M. Mahbub, E. Begoli, S. Martins, A. Peluso, S. Tamang, and G. Peterson, "Cpgqa: A benchmark dataset for machine reading comprehension tasks on clinical practice guidelines and a case study using transfer learning," IEEE Access, 2023.*** *[272] by Maria Mahbub, Edmon Begoli, Susana Martins, Alina Peluso, Suzanne Tamang, and Gregory D Peterson (©2023 IEEE. Reprinted, with permission, from [272]). The authors' contributions to the published article are as follows: MM and EB conceptualized the task. MM designed the study, curated data, investigated, developed methodology, performed experiments, and wrote the article with comments and reviews from EB, SM, AP, ST, and GDP. MM and AP performed the formal analysis. SM validated the data. EB and GDP supervised the primary author.*

This chapter focuses on expanding the application of biomedical machine reading comprehension (MRC) to a previously unexplored area – MRC on clinical practice guidelines (CPGs). The ultimate goal of this study is to assist clinicians in efficiently interpreting the guidelines by leveraging MRC systems. We develop a benchmark dataset called "cpgQA" to accomplish this objective. The cpgQA dataset is manually annotated and validated by subject-matter experts (SMEs)

"In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of The University of Tennessee, Knoxville's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink. If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation."

specifically for the biomedical MRC task on CPGs. Its primary purpose is to evaluate the performance of intelligent systems undertaking MRC tasks on CPGs.

We start the chapter with a brief overview of the motivation, followed by a dataset description. We then provide a case study employing state-of-the-art MRC models to evaluate the proposed dataset extensively. We address the lack of training data in this newly defined domain by utilizing transfer learning, including but not limited to BioADAPT-MRC introduced in Chapter 3. Finally, we end the chapter with a thorough error analysis of the best-performing MRC model to identify its strengths and weaknesses in performing the MRC task on CPGs. To facilitate further research efforts in this problem domain, we release the cpgQA dataset at https://github.com/mmahbub/cpgQA.

4.1 Motivation

CPGs are a pivotal component of clinical decision-making at the point of care. These guidelines are formulated by meticulously reviewing the most up-to-date research evidence to provide recommendations for patient-specific care based on the best available evidence, value judgments on risks and benefits, alternative care options, patient management, and practice experience. Their ultimate objective is to support healthcare professionals in delivering optimal care and implementing best practices for their patients [11]. Despite the conciseness compared to reviewing multiple resources, clinicians often find the CPGs lengthy, complex, and time-consuming to use [12]. The goal of the biomedical MRC task on CPGs is to save clinicians valuable time and effort by assisting them in comprehending the complex narratives in the CPGs while providing targeted selected information to support their clinical practice. While recent advances in biomedical MRC consider text data from resources such as clinical notes and scholarly articles, the CPGs still need to be explored.

In this work, we attempt to bridge the gap in MRC for CPGs by introducing it as a new biomedical MRC problem domain. In this context, utilizing a benchmark dataset is a critical factor for evaluating the abilities of intelligent systems in reading and comprehending the narratives in CPGs and then answering queries from them. To the best of our knowledge, no existing benchmark MRC dataset focuses on or includes CPGs. This chapter presents a manually built, SME-validated benchmark MRC dataset – cpgQA, with 1097 samples, using a clinical

practice guideline. cpgQA can be viewed as the pioneer dataset for biomedical MRC on CPGs that can extend existing MRC models to enable efficient and accurate interpretation of the CPGs.

We also present an extensive case study using the state-of-the-art technique for the low-resource biomedical MRC task – *transfer learning with transformer-based pre-trained language models*. As explained in Chapter 2, in real-world applications, MRC tasks on new problem domains, such as this, suffer from a lack of high-quality, large-scale labeled training datasets, which is a bottleneck for the high-performance of these models. Moreover, acquiring such datasets in the biomedical field requires subject-matter expertise, causing the process to be expensive and time-consuming. Transfer learning comes into play in scenarios where large-scale datasets are unavailable for training.

This chapter further experimentally demonstrates the necessity for a benchmark MRC dataset on CPGs. We also perform a thorough error analysis that depicts the strengths and weaknesses of the state-of-the-art approach in light of the MRC task on CPGs. Lastly, we explain the limitations of this dataset and delineate the scope of improvements.

4.2 Material & Methods

The primary goal of this research is three-fold: (i) introduce a new problem domain in biomedical MRC – MRC on CPGs, (ii) publish a reliable benchmark dataset called cpgQA for evaluating the effectiveness of MRC systems in comprehending CPGs and answering questions related to them, (iii) present a comprehensive case study using transfer learning with state-of-the-art machine reading comprehension approaches and cpgQA. In this section, we describe the aforementioned parts of this study in detail.

4.2.1 cpgQA Dataset

CPGs are "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" [273]. The foundation of the CPGs is a systematic review of the available research evidence targeted to answer specific clinical questions on a certain condition with an emphasis on the strength of the evidence used for confident clinical decision-making [11]. CPG developers consider a variety of questions when they review

multiple sources to compile a guideline from them – "the identification of risk factors for conditions; diagnostic criteria for conditions; prognostic factors with and without treatment; the benefits and harms of different treatment options; the resources associated with different diagnostic or treatment options; and patients' experiences of healthcare interventions" [274]. These characteristics make CPGs different from other unstructured biomedical text data such as scientific literature or clinical notes.

4.2.1.1 Dataset Construction:

The proposed cpgQA dataset consists of 1097 questions with answers and contexts taken verbatim from the *VA/DoD CPG on Opioid Therapy for Chronic Pain* [275] (available at https://www.healthquality.va.gov/guidelines/Pain/cot/VADoDOTCPGPocketCard022817.pdf). Similar to other CPGs, this guideline contains the most prevailing information, collated from multiple relevant resources into one concise document, and follows the same structure [275].

cpgQA is manually annotated and released in collaboration with SMEs as the first standard benchmark MRC dataset in the biomedical sub-domain, CPGs. Each context in the dataset is a paragraph from the CPG. The questions and answers are created from the contexts with the help of SMEs by manually reading through the document paragraph-by-paragraph to ensure high quality, focusing on the following five primary components of the CPG. The numbers in the parentheses indicate the count of data instances in cpgQA for that specific component:

- *Introductory information* on the guideline (44 instances)
- *Background information* on the subject-matter (233 instances)
- *Features and overview* of the guideline (280 instances)
- *Algorithm* that accommodates the "understanding of the clinical pathway and decision making process" (129 instances)
- *Recommendations* under the consideration of "confidence in the quality of the evidence, balance of desirable and undesirable outcomes (i.e., benefits and harms), patient or provider values and preferences, and other implications, as appropriate (e.g., resource use, equity, acceptability)" (411 instances)

Figure 4-1 shows an example sample from the cpgQA dataset.

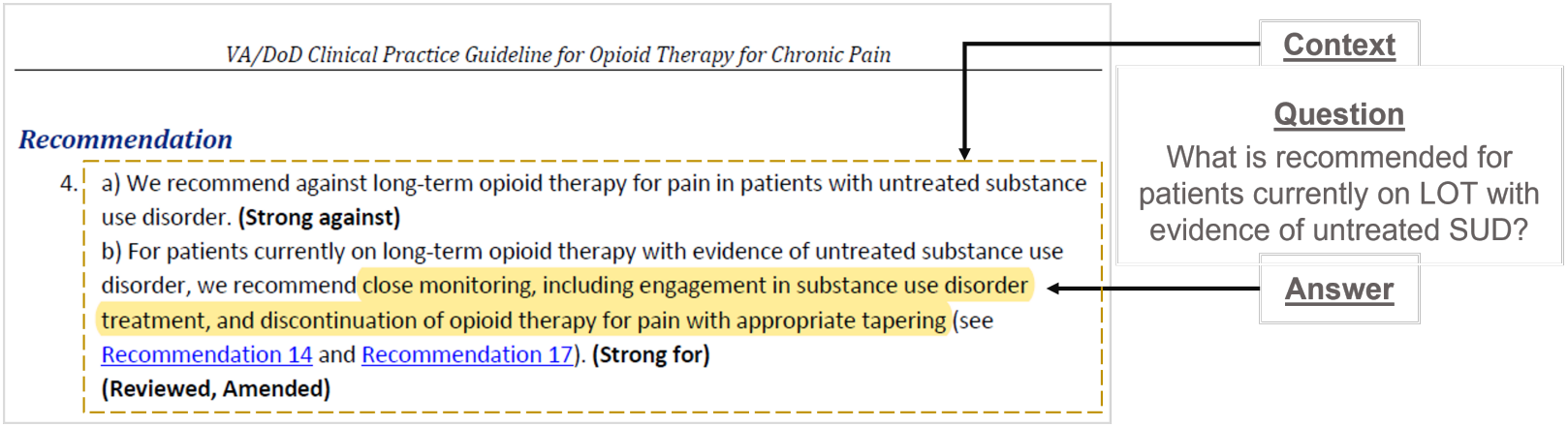


Figure 4-1. Example context and question-answer pair from the cpgQA dataset. (©2023 IEEE. Reprinted from [272].)

4.2.1.2 Dataset Statistics

The cpgQA dataset has 190 unique contexts and 1,097 question-answer pairs. We further analyze the characteristics of the cpgQA dataset based on two linguistic aspects: (i) distribution of types of questions (based on interrogative words/phrases) and (ii) distribution of the number of words in the questions, answers, and contexts. Figure 4-2 shows that there are eight types of questions in the dataset, among which approximately 61% are "What" type questions, dominating the dataset, while one-third of the dataset is equally shared by the "When", "Which", "How", "Who" type questions.

The dominance of the "What" question is also shown in Figure 4-3, which details the distribution of the question types for each of the five components in the CPG: introductory information, background information, features and overview, algorithm and recommendations. 75% of the questions and answers in the cpgQA dataset consist of less than 18 and 17 words, respectively (Figure 4-4). On the other hand, 75% of the contexts consist of 247 words or less.

In Table 4-1, we also show the comparison between cpgQA and two other biomedical MRC datasets – BioASQ [276] and emrQA [151] test sets, in terms of context, question, and answer lengths.

4.2.1.3 Comparison with Other Biomedical Sub-domains

To explain why we need a dataset for CPGs, we compare the CPG sub-domain with two other biomedical sub-domains that have MRC datasets – scholarly articles and clinical notes/EMRs. For scholarly articles and EMRs, we choose the question-context pairs from the test sets of BioASQ-9b (Factoid) [276], and emrQA (Relation subset) [151], respectively. BioASQ-9b (Factoid) is an SME-annotated MRC dataset published in the BioASQ challenge 2021, where the contexts are snippets extracted from PubMed/MedLine articles [1]. emrQA is an automatically-annotated MRC dataset where the contexts are from "longitudinal EMRs" of patients [151]. Among four subsets of emrQA, we choose the relation subset based on the experiments performed by [152].

We demonstrate the domain difference by plotting the vector representations of the question-context pairs from these domains in 2D (Figure 4-5), adapting the approach described in [277]. In this approach, given a single question-context pair,

[1] https://pubmed.ncbi.nlm.nih.gov/

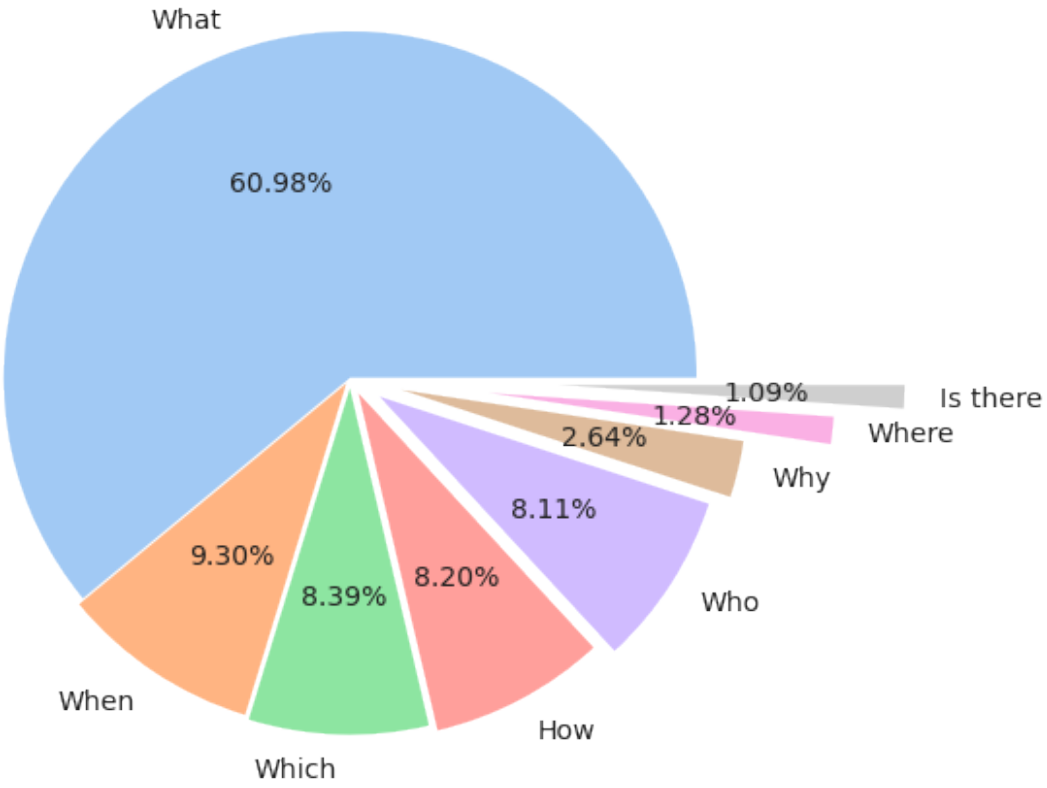


Figure 4-2. Distribution of question types in the cpgQA dataset. (©2023 IEEE. Reprinted from [272].)



Figure 4-3. Distribution of question types in each component of the clinical practice guideline used for cpgQA in the cpgQA dataset. (©2023 IEEE. Reprinted from [272].)

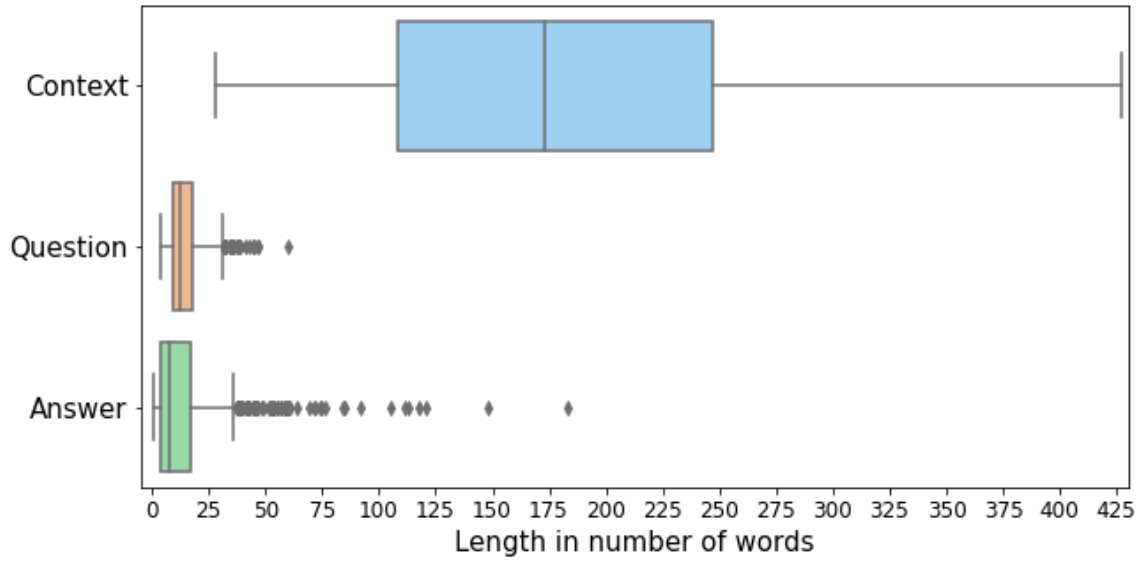


Figure 4-4. Distribution of context, question, and answer lengths (in number of words) in the cpgQA dataset. (©2023 IEEE. Reprinted from [272].)

Table 4-1. Comparison among cpgQA, BioASQ, and emrQA in terms of context, question, and answer lengths. (©2023 IEEE. Reprinted from [272].)

Dataset	Length of 75% of the contexts	Length of 75% of the questions	Length of 75% of the answers
cpgQA	≤247	≤18	≤17
BioASQ	≤42	≤12	≤2
emrQA	≤1669	≤10	≤16

we use the last hidden state of the BERT (Bidirectional Encoder Representations from Transformers) encoder [20] to create 768-dimensional vector representations for each of the tokens in that pair. BERT is one of the state-of-the-art bidirectional multi-layer transformer network for modeling language representations, which is pre-trained on Wikipedia articles and BookCorpus [20]. Then, we calculate the average over these token representations, which results in a 768-dimensional vector representation for the question-context pair. To visualize, we then perform dimensionality reduction with 2-component PCA over the vector representations of all question-context pairs. For a fair comparison, we use the original BERT model [20] and do not fine-tune on any of the biomedical sub-domains.

Figure 4-5 clearly shows that CPGs are linguistically very different from scholarly articles and clinical notes, necessitating the introduction of CPGs as a different sub-domain in biomedical MRC.

4.2.2 Machine Reading Comprehension Models

As explained in Chapter 2, MRC models aim to understand biomedical narratives to correctly answer questions posed by users (such as clinicians, patients, or general people, depending on the application) from these narratives. An input to the MRC model is a question-context pair, and the output is the positions of the start and end tokens in the answer span taken from the context.

More formally, Given a question Q with n_q tokens (words) and a context C with n_c tokens, an MRC model respectively predicts the start token position a^s and end token position a^e of the answer span $A_{a^s}^{a^e} \in C$ such that there exists a unique answer span composed of n_a (with $n_a \leq n_c$) consecutive tokens in the context.

For evaluating the state-of-the-art transformer-based models on cpgQA, we pick the top nine best-performing models from the *Biomedical Language Understanding and Reasoning Benchmark (BLURB)*: BioLinkBERT [131], PubMedBERT [128], ClinicalBERT [120], BioELECTRA [129], SciBERT [86], BioBERT [83], RoBERTa [45], BERT [20], and BlueBERT [130].

We provide the specifications of these models in Table 4-2. All nine encoder models have 12 layers, 12 attention heads per layer, 768 hidden nodes, and 3,072 FFNN inner hidden nodes. For these models, the embedding dimension of each token is 768. A maximum of 512 tokens can be provided as one input sequence in these models, except RoBERTa, which can accept 514 tokens per input.

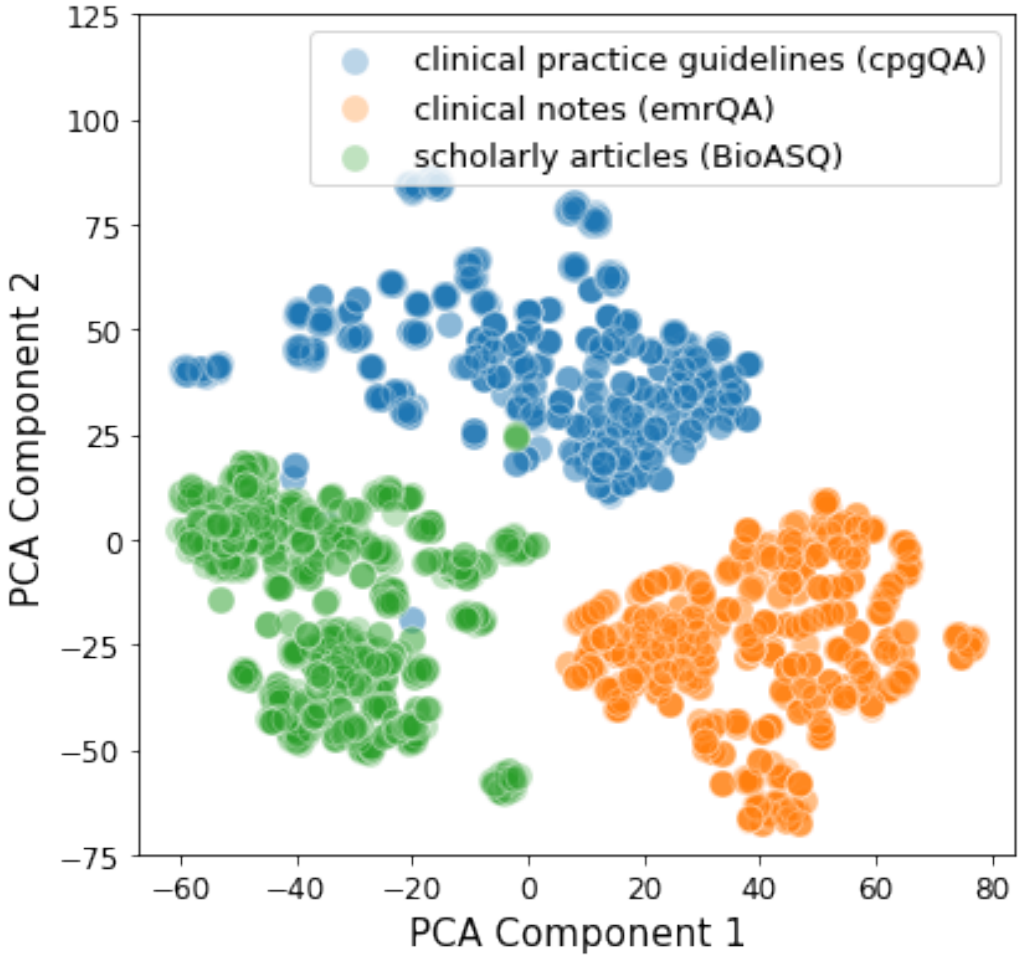


Figure 4-5. A two-dimensional visualization of vector representations of question-context pairs using PCA. The colors represent the three sub-domains in biomedical-MRC: clinical practice guidelines (cpgQA), clinical notes (emrQA), and scholarly articles (BioASQ). (©2023 IEEE. Reprinted from [272].)

We use a single-layer FFNN with 768 hidden nodes as the task-specific layer. The layer calculates the probability distributions for start and end token positions using the softmax activation function, as shown in Equation (4.1).

$$p_i^s = \frac{\exp(W^s \cdot h_i)}{\sum_{k=1}^{n_l} \exp(W^s \cdot h_k)}; \quad p_i^e = \frac{\exp(W^e \cdot h_i)}{\sum_{k=1}^{n_l} \exp(W^e \cdot h_k)} \tag{4.1}$$

Here, n_l is the sequence length of the input, $h_i \in \mathbb{R}^H$ is the hidden representation vector of the i^{th} token, $W^s, W^e \in \mathbb{R}^H$ and $\in \mathbb{R}^H$ are two trainable weight matrices, and p_i^s and p_i^e refer the probabilities of the i^{th} token being predicted as *start* and *end* respectively,

To optimize the MRC model, we employ the cross-entropy (CE) loss $\mathcal{L}$ on the predicted answer positions. For each sample, we average the total cross-entropy loss of two predicted outputs for the start and end positions, following Equation 4.2.

$$\mathcal{L} = -\frac{1}{2}(\log p_{y^s}^s + \log p_{y^e}^e) \tag{4.2}$$

Here y^s and y^e respectively denote the ground truth answer's start and end token positions. The answer in the test phase is predicted by selecting the sequence of tokens in the interval defined by the two highest probabilities from the distributions $p_{k\in[1,n_l]}^s$ and $p_{k\in[1,n_l]}^e$.

4.2.3 Transfer Learning with cpgQA

As explained in Chapter 2, transfer learning is the first choice among researchers in the absence of no labeled training data. As a new biomedical sub-domain, no existing labeled training dataset focuses on/including CPGs, hence the need for transfer learning. In this work, we present a case study on MRC modeling of cpgQA with transfer learning, where we consider the CPGs as the target domain. For the source domain, we explore three different ones: (i) biomedical scholarly articles, (ii) clinical notes, and (ii) Wikipedia. Our case study is two-fold, considering two transfer settings:

(1) **Unsupervised setting:** In this setting, we consider that no labeled training data is available in the target domain. Hence, we only use the source domain to train the MRC models in Section 4.2.2. We then perform inference on the

cpgQA dataset using the trained models. We repeat this experiment for three different source domains, as mentioned above.

(2) **Supervised setting:** While no labeled training data is available for CPGs, to demonstrate the potential of using small-scale labeled training data from the target domain in the transfer learning setting, we also perform MRC modeling in the supervised setting. Here, we consider that some labeled target-domain training data are available and use subsets from the cpgQA dataset. As the source domain and the MRC model, we pick the best-performing ones based on the experiments performed in the unsupervised setting.

To apply transfer learning in this setting, we consider two sub-settings:

(i) *Sequential learning:* In this setup, the MRC model is sequentially trained on the source and target domains. As mentioned in Chapter 2, deep learning models require large-scale datasets to learn better representations of the data instances during training and thus perform better in the task. Sequential learning ensures the model's exposure to a large-scale dataset during training by sequentially training it on a source domain with a large-scale dataset and then on the target domain with a small-scale dataset. Thus, it increases the efficiency of the MRC model in performing the MRC task in the target domain [278, 279].

(ii) *Simultaneous learning:* In this setup, the MRC model is simultaneously trained on the source and target domains. Similar to sequential learning, simultaneous learning ensures that the model is exposed to a large-scale dataset during training.

To evaluate the model on the cpgQA dataset in the supervised setting, we perform a 5-fold cross-validation as follows: We divide the cpgQA dataset into five disjoint subsets (folds) based on 190 unique contexts to avoid data leakage in the test sets. For sequential learning, we subsequently train the MRC model on the source-domain dataset and all but one of the folds of the target-domain dataset. At the end of the training, the remaining target-domain subset is then used as the test set. We repeat this five times, each with a different subset excluded from training and reserved for testing. Table

4-3 shows the number of target-domain training and testing samples (i.e., question-answer pairs) in each fold.

We follow the same process for simultaneous learning, except that we train the MRC model simultaneously on the source and target domain samples. As the baseline and to show the effect of transfer learning, we also perform a 5-fold cross-validation using only the target-domain dataset.

Additionally, we experiment with BioADAPT-MRC for both of these settings [249]. In Chapter 2, we explained how using a trained model (on the source domain) for inference in the target domain in the aforementioned unsupervised setting often hurts the model's performance. We demonstrate this phenomenon in Section 4.3.4.1.

BioADAPT-MRC addresses this issue and improves model performance using a deep learning framework incorporating adversarial learning. This approach leverages both the source domain with a large-scale labeled dataset and the target domain with an unlabeled or limited labeled dataset for learning. The details of this methodology are described in Chapter 3. In this study, we experimentally explore the potential of BioADAPT-MRC in performing the MRC task on CPGs.

4.3 Results & Discussion

In this section, we describe the specifications of the source-domain datasets used in the experiments, the metrics used to measure the performance of the MRC models, and the experimental setup. We further report the experimental results of the case study on MRC modeling of cpgQA using transfer learning. We then report a thorough error analysis to demonstrate the strengths and weaknesses of the current state-of-the-art approach in performing the MRC task on cpgQA.

4.3.1 Source-domain Datasets

As mentioned in Section 4.2.3, we consider three source domains for this study: (i) biomedical scholarly articles, (ii) clinical notes, and (iii) Wikipedia. As datasets from these domains, we use BioASQ-9B (Factoid) [276], emrQA (Relation subset) [151], and SQuAD-1.1 [21], respectively. While BioASQ and emrQA are biomedical domain-specific datasets, SQuAD is a general-purpose dataset with the contexts from Wikipedia articles and crowd-workers developed the question-answer pairs

Table 4-2. Specifications of nine pre-trained language models used in the experiments. (©2023 IEEE. Reprinted from [272].)

Model name	Pre-training corpora	Pre-trained from	Vocabulary size
BioLinkBERT	PubMed + Citation Links Between PMC Articles	BERT	28895
PubMedBERT	PubMed + PMC	Scratch	30522
BioELECTRA	PubMed	Scratch	30522
BioBERT	PubMed + PMC	BERT	28996
SciBERT	Biomedical + Computer Science Literature	Scratch	31090
ClinicalBERT	MIMIC III	BERT	28996
RoBERTa	Web crawl	Scratch	50265
BlueBERT	PubMed + MIMIC III	BERT	30522
BERT	Wikipedia + BookCorpus	Scratch	30522

Table 4-3. Number of target-domain training and testing samples (i.e., question-answer pairs) in the folds created for 5-fold cross-validation. (©2023 IEEE. Reprinted from [272].)

Fold	Target-domain training set	Target-domain test set
1	1012	85
2	939	158
3	856	241
4	869	228
5	712	385

[21]. Table 4-4 provides the number of contexts and question-answer pairs in the training and test sets of these datasets.

The SQuAD-1.1 dataset can be found in the Wolfram data repository[2]. For BioASQ-9b, we use the training and test sets from the BioASQ challenge website[3]. BioASQ dataset includes four types of questions – yes/no, factoid, list, and summary. Among them, the factoid question-answering task closely relates to extractive MRC. Hence, we pre-process the training and test sets to keep only the factoid questions. In place of text passages, the contexts in the BioASQ training set consist of PMIDs. Therefore, we further pre-process the BioASQ training data by retrieving full abstracts from PubMed utilizing the PMIDs. Then, we use these abstracts as the contexts in the training set. We also remove the entries (from the training and test sets) that do not have an answer in the context. For emrQA (relation subset)[4], we pre-process the training and development sets following [152]. For both SQuAD and emrQA, we use the development datasets as the test sets.

Target-domain training dataset for BioADAPT-MRC: As explained in Section 4.2.3, BioADAPT-MRC can utilize unlabeled training datasets from the target domain. An unlabeled MRC dataset implies that the dataset does not contain the labels, i.e., question-answer pairs, but the contexts. Hence, as the unlabeled target-domain training data, we use 10,987 paragraphs, automatically extracted using *Regular Expressions*, from 21 other VA/DoD CPGs[5]. Along with the benchmark dataset cpgQA, we release this unlabeled training set in our GitHub repository[6] for reproducibility.

4.3.2 Experimental Setup and Training Configurations

We implement the MRC models using PyTorch [257] using the huggingface API [258]. According to the hyperparameter choices for the MRC tasks provided in [20, 45, 83, 86, 120, 128–131], we select the following ones for all the experiments except the ones with BioADAPT-MRC: for tokenization – 384 as the maximum sequence length, 64 as the maximum query length, 128 as the document stride; for training – $3e^{-5}$ as the learning rate, 24 as the batch size, and 3 as the number of training

[2]https://datarepository.wolframcloud.com/resources/SQuAD-v1.1
[3]http://participants-area.bioasq.org/datasets/
[4]https://github.com/panushri25/emrQA
[5]https://www.healthquality.va.gov/
[6]https://github.com/mmahbub/cpgQA

epochs. For the maximum answer length, we choose 200, given the distribution of answer length in cpgQA (Figure 4-4). To implement the BioADAPT-MRC model, we follow the default implementation process and hyperparameter setting provided in Chapter 3. We performed all experiments on a Linux virtual machine with a single Tesla V100-SXM2-16GB GPU and Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz.

4.3.3 Evaluation Metrics

For measuring the performance of the MRC models on cpgQA, we use two metrics widely used in the MRC tasks: Exact Match (EM) and F1-Score (F1).

Exact match: Each question in the cpgQA dataset corresponds to exactly one correct answer that can be a word, a phrase, or single/multiple sentences. For each question-answer pair, if the predicted answer matches strictly with the ground truth answer to the character, then for that data instance, EM is calculated as 1, otherwise 0. Thus, being off by a single character in the predicted answers results in an EM score of zero. In this way, for the whole dataset, the EM score is calculated following Equation 4.3.

$$\text{EM} = \frac{\sum_{i=1}^{N} \text{em}_i}{N} \tag{4.3}$$

Here, N is the length of the dataset, i.e., the total number of QA pairs in the dataset, and em_i is the exact match score for the i^{th} pair.

F1-score: F1-score is a well-known classification metric used in cases where precision and recall should be provided equal importance. The basis of the F1-score in MRC is the number of shared words between the ground truth and the predicted answers. For each question-answer pair, the F1-score can be calculated using Equations 4.4 and 4.5.

$$\text{f1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}};\ \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.5}$$

Here, TP is the number of tokens shared between the predicted answer and ground truth, FP is the number of tokens in the predicted answer but not in the ground truth, and FN denotes the number of tokens in the ground truth answer but not in the predicted one.

F1-score for the whole dataset is calculated following Equation 4.6.

$$\text{F1} = \frac{\sum_{i=1}^{N} \text{f1}_i}{N} \tag{4.6}$$

Here, f1_i is the F1-score calculated using Equation 4.4 for the i^{th} QA pair.

The range of EM and F1 scores are {0,1} and [0,1], respectively.

Cross-validation score: We perform 5-fold cross-validation (CV) for experiments in the supervised setting. Since the folds are not evenly distributed (Table 4-3), we calculate the weighted CV mean (for both EM and F1 calculated by Equations 4.3 and 4.6) across all five folds following Equation 4.7. We also calculate the error rates of EM and F1 scores using weighted standard deviation following Equation 4.8.

$$\text{CV Mean, } \mu = \frac{\sum_{i=1}^{5} s_i f_i}{\sum_{i=1}^{5} f_i} \tag{4.7}$$

$$\text{CV Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{5} f_i (s_i - \mu)^2}{\frac{(M-1)}{M} \sum_{i=1}^{5} f_i}} \tag{4.8}$$

Here, M is the number of folds. For a 5-fold CV, $M = 5$. s_i is either EM or F1 score calculated over the samples in fold i and f_i is the number of samples in fold i.

4.3.4 Experimental Results

In this section, we discuss the experimental results of the MRC models on cpgQA.

4.3.4.1 Unsupervised Setting

Table 4-5 and Figure 4-6 show the experimental results of MRC models on cpgQA in the unsupervised setting. As depicted by the low EM and F1 scores on

cpgQA in Table 4-5 and the drop in performance from source to target domains in Figure 4-6, the models poorly perform when we transfer knowledge directly from the biomedical scholarly articles (BioASQ) and clinical notes (emrQA). It indicates that the CPGs are linguistically different from articles or clinical notes. These results reaffirm the need for the new biomedical sub-domain – CPGs. We also notice that when we use a more general-purpose domain such as Wikipedia (SQuAD) as the source domain, the performance of the MRC models on cpgQA is higher. It may happen because compared to BioASQ and emrQA, SQuAD consists of more diverse documents and does not focus on a narrow domain. Moreover, as shown in Table 4-4, while emrQA is the largest of three datasets with more than 600k samples in the training set, the diversity in the contexts is approximately 64 times less than that of the contexts in the SQuAD dataset. Consequently, MRC models that are trained on SQuAD may generate comparatively better generalizable feature representations.

Table 4-6 show the results from the BioADAPT-MRC model and the best-performing model in Table 4-5 – BioLinkBERT trained on SQuAD. In the rest of the chapter, we denote the model name BioLinkBERT by its original name hyphenated with the name of the training data to avoid redundancy. As shown, BioADAPT-MRC can achieve a 62% EM score with 95% confidence interval (CI): 59%-65% and 82% F1-score with 95% CI: 81%-84% in the target domain while retaining high-performance scores in the source domain. This is because the unlabeled target-domain dataset and adversarial learning enable BioADAPT-MRC to generate features that reduce domain shift and thus reduce the gap between the performance of MRC models in the source and target domains.

4.3.4.2 Supervised Setting

Table 4-7 shows the cross-validation scores with 95% CI of MRC models on cpgQA in the supervised setting. We also show the trends of EM and F1 scores across five folds of cpgQA used in the experiments (Figure 4-7). As shown, when trained with only the target domain dataset, the BioLinkBERT-cpgQA model achieves the worst performance. It may happen because the training dataset is too small to use the full potential of the model. Nonetheless, when we train the MRC Model on SQuAD and cpgQA (BioLinkBERT-SQuAD-cpgQA), it achieves higher EM and F1 scores.

Table 4-4. Statistics of the source-domain datasets used in the experiments. (©2023 IEEE. Reprinted from [272].)

Source-domain Datasets	Training set		Test set	
	#QA Pairs	#Contexts	#QA Pairs	#Contexts
BioASQ-9b (Factoid)	10,584	3,703	937	930
emrQA (Relation)	621,428	296	101,700	42
SQuAD-1.1	87,599	18,891	10,570	2,067

Table 4-5. Test scores on cpgQA in the unsupervised setting with nine pre-trained language models and three source domains. The highest and the second highest scores are highlighted in bold and italic, respectively. (©2023 IEEE. Reprinted from [272].)

Model	Trained on BioASQ-9b		Trained on emrQA		Trained on SQuAD	
	EM	F1	EM	F1	EM	F1
BioLinkBERT	**0.1668**	**0.3816**	*0.0055*	0.3311	*0.5497*	**0.7815**
PubMedBERT	*0.1076*	*0.2844*	**0.0082**	0.3022	**0.5515**	*0.7742*
BioELECTRA	0.0319	0.1457	0.0027	0.1988	0.4941	0.7411
BioBERT	0.0501	0.1855	0.0027	0.2881	0.5369	0.7680
SciBERT	0.0811	0.2387	0.0018	0.3229	0.5232	0.7532
ClinicalBERT	0.0301	0.1497	0.0027	0.2742	0.4950	0.7321
RoBERTa	0.0866	0.2761	0.0027	**0.3373**	0.4950	0.7335
BlueBERT	0.0137	0.1187	0.0036	*0.3327*	0.4038	0.6524
BERT	0.0228	0.1533	*0.0055*	0.318	0.4859	0.7226

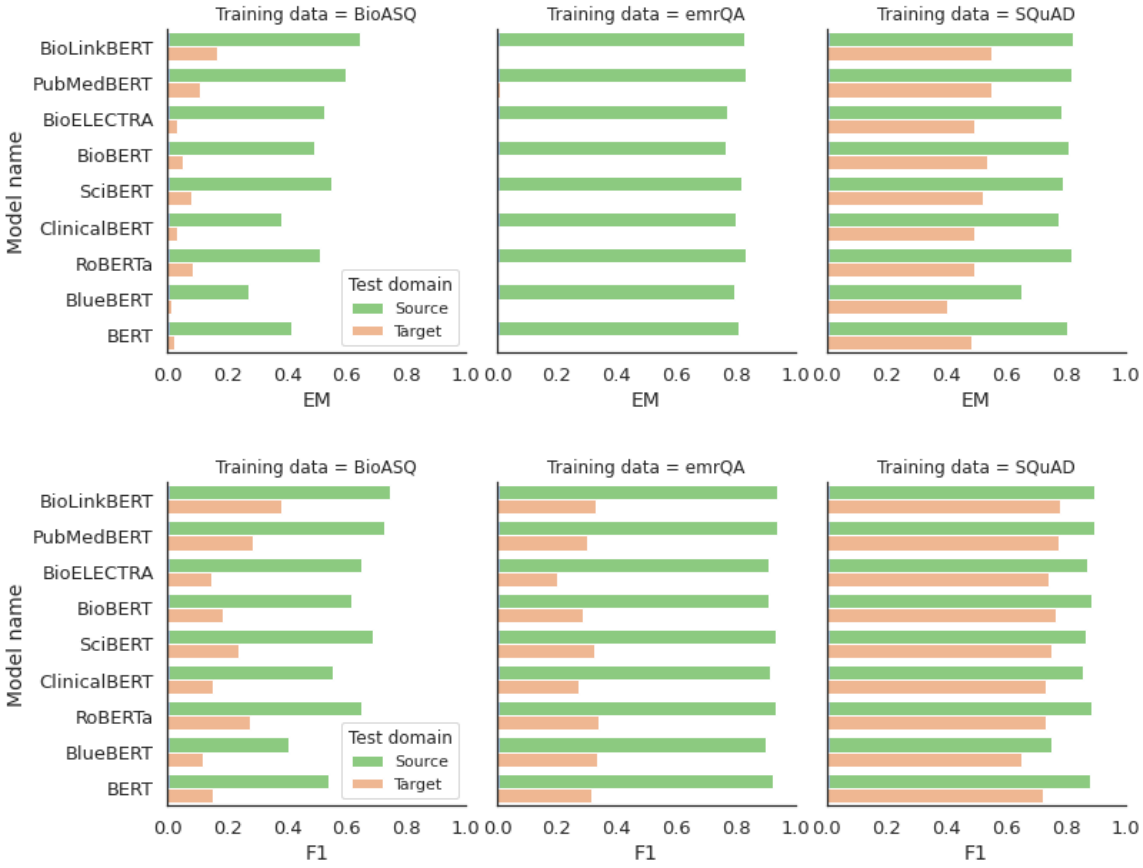


Figure 4-6. Drop in EM scores (top) and F1 scores (bottom) after direct knowledge transfer from the source domains to the target domain. (©2023 IEEE. Reprinted from [272].)

Table 4-6. Test scores on cpgQA in the unsupervised setting with BioLinkBERT-SQuAD and BioADAPT-MRC. The highest scores are highlighted in bold. (©2023 IEEE. Reprinted from [272].)

Model	Source Domain				Target Domain			
	EM	EM (95% CI)	F1	F1 (95% CI)	EM	EM (95% CI)	F1	F1 (95% CI)
BioLinkBERT-SQuAD	0.8239	0.8167-0.8312	0.8978	0.8928-0.9028	0.5497	0.5202-0.5792	0.7815	0.7634-0.7995
BioADAPT-MRC	0.7812	0.7733-0.7891	0.8712	0.8657-0.8767	**0.6180**	**0.5893-0.6468**	**0.8229**	**0.8062-0.8397**

Comparing the scores reported in Table 4-7, we can say that the addition of the source domain dataset, SQuAD, in the training process helps improve the learning of the MRC model. We also notice that sequential training is better than simultaneous training – indicating a learning pattern in the transformer-based MRC model. We also show the experimental results using BioADAPT-MRC (Table 4-7). Even in the default setting without any hyperparameter optimization, BioADAPT-MRC outperforms BioLinkBERT-cpgQA and BioLinkBERT-SQuAD-cpgQA with higher CV mean and lower CV standard deviation, achieving higher stability in performance across five folds (Figure 4-7). It shows the potential of this state-of-the-art approach in performing MRC tasks on CPGs.

4.3.5 Error Analysis

In this section, we present a comprehensive three-fold error analysis of the best-performing MRC model from our experiments by considering the following aspects of the cpgQA dataset: (i) components of the clinical practice guideline, (ii) types of questions, (iii) length of answers.

According to our experimental results presented in Section 4.3.4, we choose the BioADAPT-MRC model (in the supervised setting) for performance analysis.

Components: As mentioned in Section 4.2.1, there are five components in the guideline which was used to build the dataset: (1) Introductory information, (2) Background information, (3) Features and overview, (4) Algorithm, and (5) Recommendations.

To analyze which parts of the guideline are correctly comprehended and interpreted by the model, we divide the dataset into five subsets according to these components and then calculate the EM and F1 scores for each of these subsets. An ideal MRC model that can understand CPGs should be able to answer questions correctly from each of these parts of the CPG and achieve an exact match score of 1 for each of these parts. Figure 4-8 shows that the model has been able to answer approximately 80% of the questions from each subset with an exact match to the ground truth answers. Nonetheless, to build the MRC system more reliable for future deployment in the healthcare setting, further performance improvement is required for each of the components of the guideline.

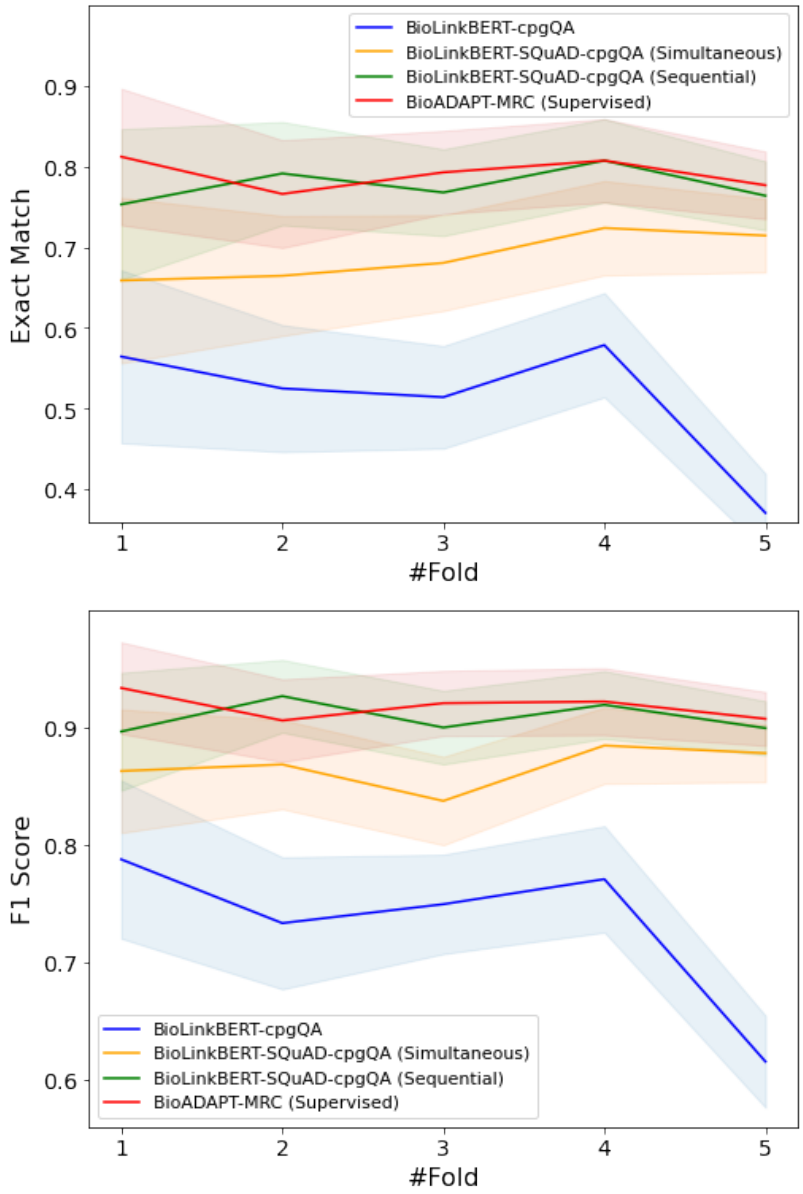


Figure 4-7. Exact match (top) and F1 (bottom) scores with confidence intervals for five folds of cpgQA in the supervised setting. (©2023 IEEE. Reprinted from [272].)

Table 4-7. Test scores on cpgQA in the supervised setting. The highest scores are highlighted in bold. (©2023 IEEE. Reprinted from [272].)

Model	CV Mean ± CV Standard Deviation and Confidence Interval			
	EM	EM (95% CI)	F1	F1 (95% CI)
BioLinkBERT-cpgQA	0.4831±0.0853	0.4535-0.5127	0.7076±0.0691	0.6860-0.7292
BioLinkBERT-SQuAD-cpgQA (Simultaneous)	0.6974±0.0240	0.6701-0.7246	0.8678±0.0173	0.8525-0.8830
BioLinkBERT-SQuAD-cpgQA (Sequential)	0.7767±0.0185	0.7520-0.8013	0.9071±0.0112	0.8937-0.9206
BioADAPT-MRC (Supervised)	**0.7876±0.0155**	**0.7634-0.8118**	**0.9149±0.0087**	**0.9019-0.9279**

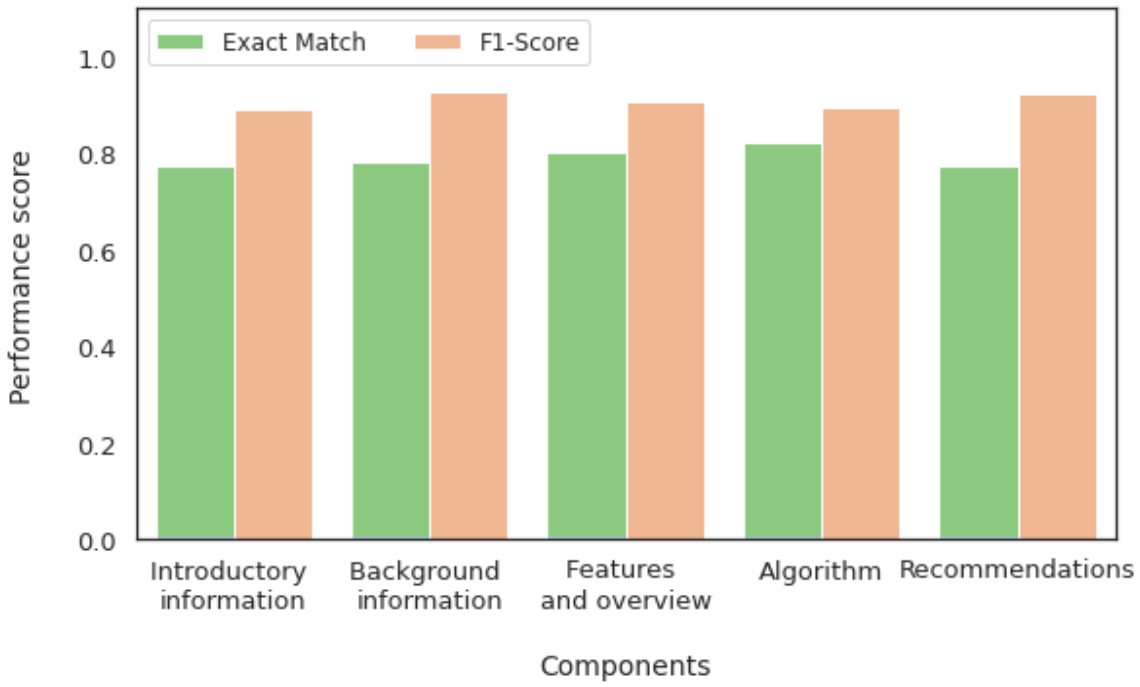


Figure 4-8. Error analysis for MRC modeling of cpgQA, considering the types of components in the guideline in the answers. (©2023 IEEE. Reprinted from [272].)

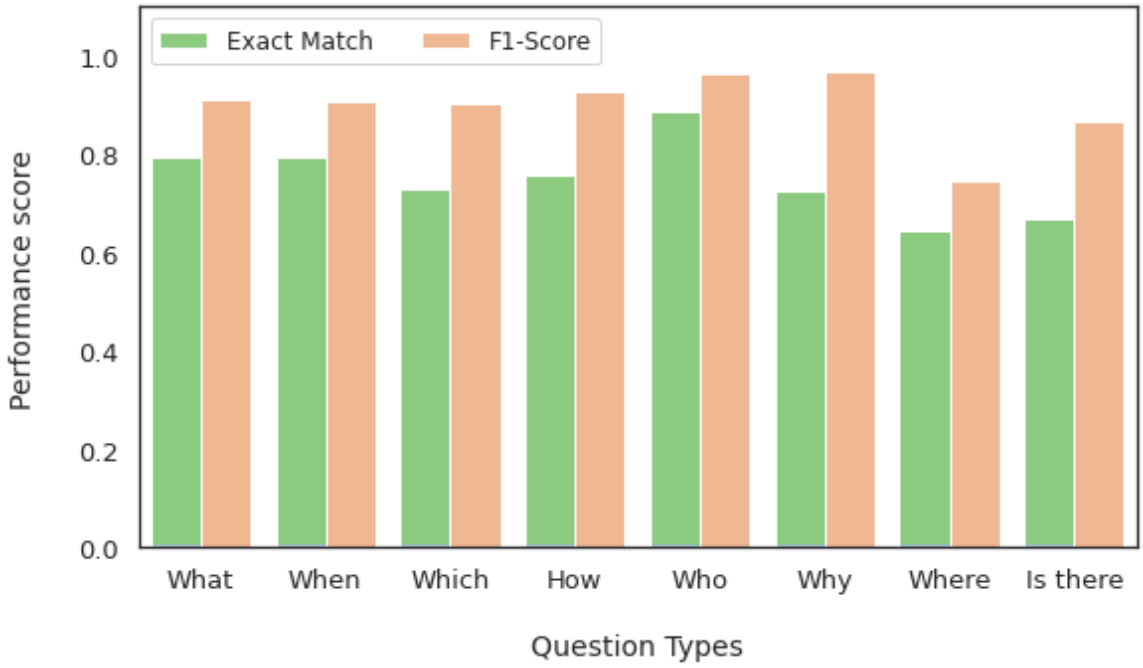


Figure 4-9. Error analysis for MRC modeling of cpgQA, considering the types of interrogative words/phrases in the questions in the answers. (©2023 IEEE. Reprinted from [272].)

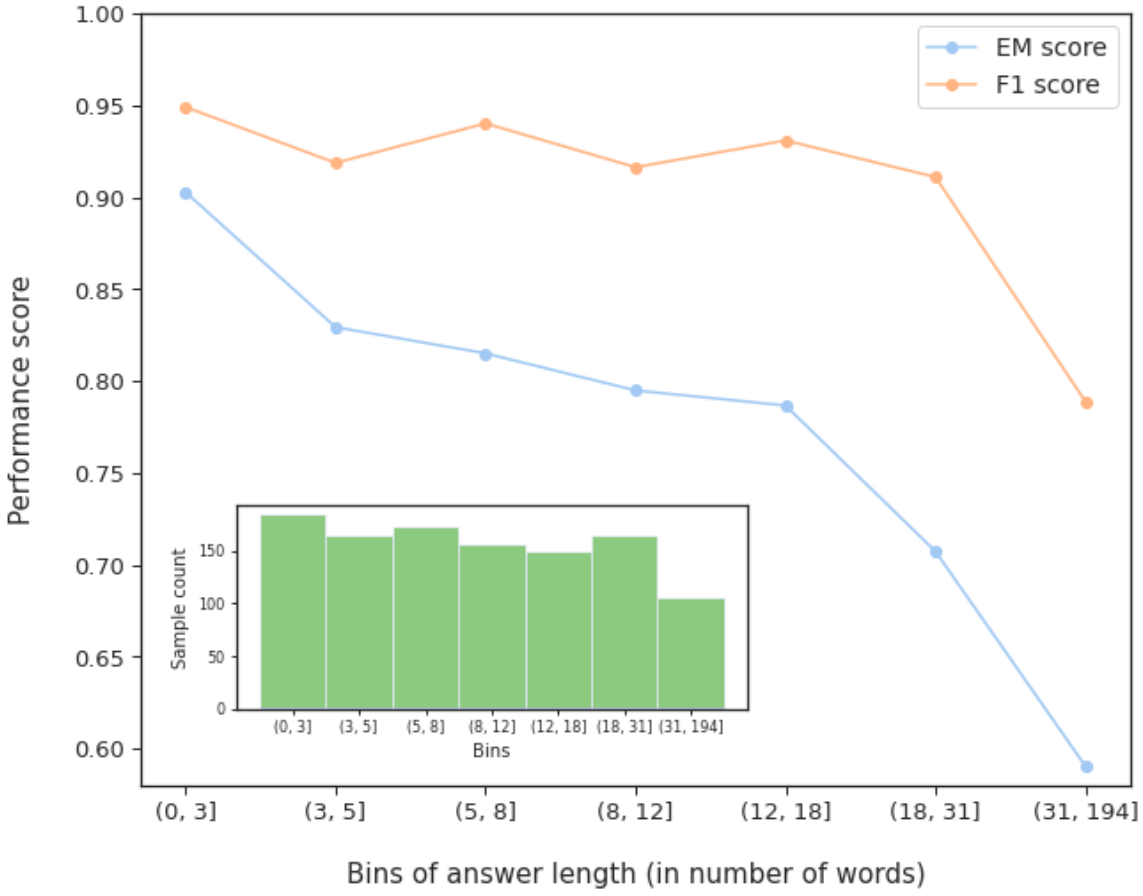


Figure 4-10. Error analysis for MRC modeling of cpgQA, considering the length (number of words) in the answers. (©2023 IEEE. Reprinted from [272].)

Question types: The cpgQA dataset consists of eight types of questions based on eight different interrogative words/phrases: "What, When, Which, How, Who, Why, Where, Is there". Figure 4-9 shows the capability of the MRC model to answer different types of questions present in the dataset. According to the EM scores, while the model can correctly answer most of the questions with the interrogative word "Who", it struggles the most with questions with the interrogative word "Where". There is also plenty of room for improvement for other question types such as "Is there", "Why", "Which", etc.

Answer length: cpgQA dataset consists of answer spans ranging from 1 word to 194 consecutive words. An ideal MRC model should be able to capture all ranges of answer spans. To show the influence of answer length on the performance of the model, in this experiment, we divide the dataset based on "binned" answer length. As the dividing criteria, we use 15 percentiles for granularity, and we group the dataset into seven disjoint bins as follows:

bin-1: $(0^{th} - 15^{th}\ percentile] \implies (0 - 3\ words]$

bin-2: $(10^{th} - 30^{th}\ percentile] \implies (3 - 5\ words]$

bin-3: $(30^{th} - 45^{th}\ percentile] \implies (5 - 8\ words]$

bin-4: $(45^{th} - 60^{th}\ percentile] \implies (8 - 12\ words]$

bin-5: $(60^{th} - 75^{th}\ percentile] \implies (12 - 18\ words]$

bin-6: $(75^{th} - 90^{th}\ percentile] \implies (18 - 31\ words]$

bin-7: $(90^{th} - 100^{th}\ percentile] \implies (31 - 194\ words]$

Figure 4-10 shows that the model does remarkably well with smaller answer spans, i.e., short answers. As the answer length increases, the model struggles to predict the start and end token positions (of the answer span) that exactly match the ground truth.

Model's capability to identify ground truth answer locations: Last but not least, we demonstrate whether the model can identify the location of the ground truth answer in cases where it is unable to find the exact match. We do this by calculating the percentage of test samples which do not have an exact match but an overlap between words in the predicted and ground truth answers. Figure

4-11 shows that approximately 6.7% mispredicted answer spans have 100% overlap with the ground truth, whereas only 1.3% mispredicted samples have no overlap. It indicates that while the model is struggling to find the exact answer spans for all the questions in the test set, it is able to identify their location most of the time.

The overall error analysis indicates that while, in this study, the best-performing model performs well in various scenarios, there is still much room for potential improvement.

4.4 Limitations of cpgQA

While the cpgQA dataset contains most parts of the guideline, it omits the tables embedded in the appendix of the guideline. These tables contain additional information on diagnosis, treatment, recommendations, etc. Thus, a future research direction that can stem from this work is incorporating the tabular data into the text-based cpgQA dataset to generate a multi-modal dataset, similar to the Finance dataset presented in [280]. Furthermore, we used only one guideline to build the benchmark cpgQA dataset due to resource constraints, which resulted in a smaller dataset. Additionally, the dataset, in its current state, does not contain cases where no answer can be found in the provided context. While cpgQA provides a well-informed baseline for the MRC task on CPGs, including more guidelines and sample cases with no answers will enlarge and diversify the dataset and help us increase the reliability of the MRC models.

Disclaimer: The sole purpose of cpgQA is to evaluate the state-of-the-art machine reading comprehension models and pioneer the research in the sub-domain – *clinical practice guidelines* – in biomedical machine reading comprehension. The dataset is not intended as a resource for patient care and should not be used as such.

4.5 Conclusion

Clinical practice guidelines are intricate biomedical narratives that are crucial resources at the point of care, as they provide the most up-to-date and authoritative recommendations necessitated by a consistent and well-defined clinical decision-making process. In recent years, several research studies have dedicated their efforts to the biomedical MRC task, primarily focusing on resources such as

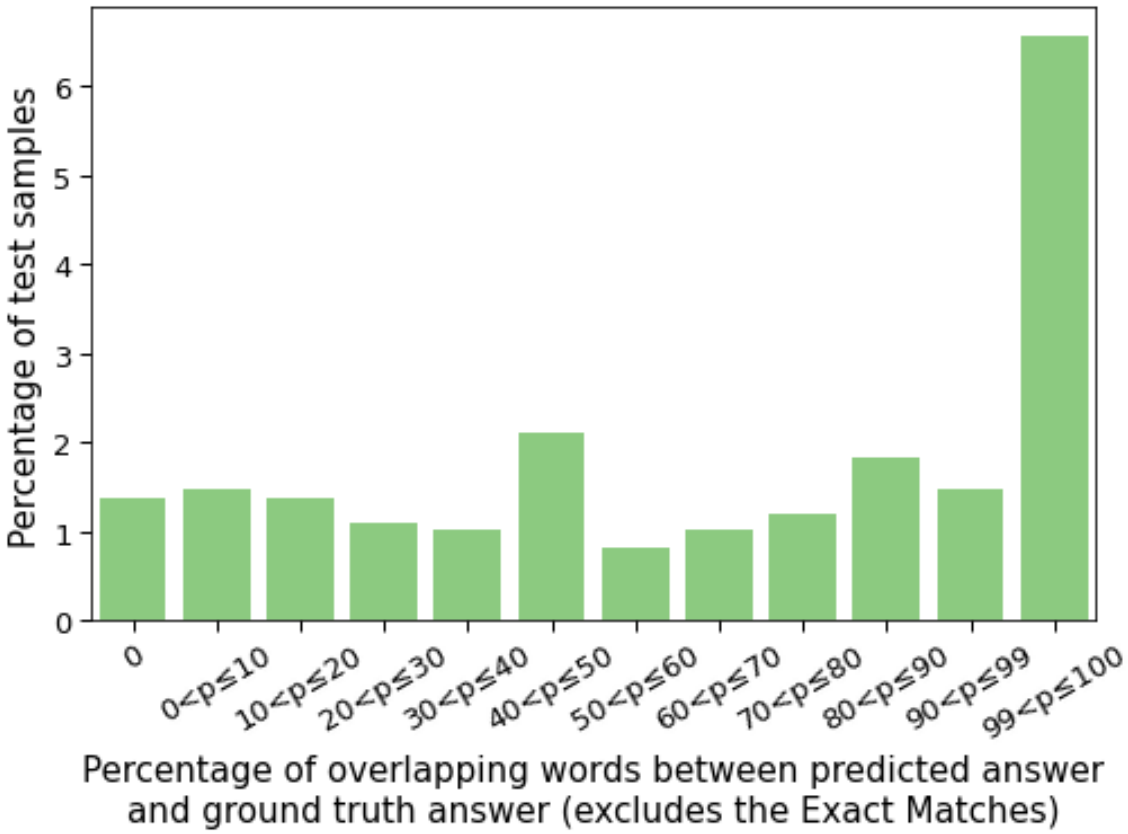


Figure 4-11. Percentage of test samples which do not have an exact match but an overlap between words in the predicted and ground truth answers. (©2023 IEEE. Reprinted from [272].)

scholarly articles and clinical notes. However, an unexplored area that still awaits investigation is using clinical practice guidelines. In this work, we have explored the clinical practice guidelines for the biomedical MRC task and identified them as a new problem domain. We have presented a benchmark dataset – cpgQA – manually annotated using a guideline with assistance from SMEs. We have then evaluated the dataset by presenting a thorough case study on transfer learning with state-of-the-art transformer-based language models. Finally, we have investigated the shortcomings of the state-of-the-art approach in performing the MRC task on cpgQA and identified possible future research directions by performing a three-fold error analysis.

CHAPTER 5

IDU-QA: MACHINE READING COMPREHENSION TO EXTRACT INFORMATION ON INJECTION DRUG USE FROM CLINICAL NOTES

The results discussed in this chapter are derived using clinical notes obtained from MIMIC-III [135] and have yet to be officially published. Nonetheless, the methods employed in this chapter have been implemented in a real-world scenario [281], the results from which have been made available as a pre-print publication [281] along with the methodologies. Therefore, much of the content in this chapter is taken without alteration from the pre-print article [281] by Maria Mahbub, Ian Goethert, Ioana Danciu, Kathryn Knight, Sudarshan Srinivasan, Suzanne Tamang, Karine Rozenberg-Ben-Dror, Hugo Solares, Susana Martins, Edmon Begoli, and Gregory D. Peterson. The authors' contributions to the pre-print article are as follows: MM, ID, and EB conceptualized the task. MM designed the study, generated data, investigated, developed methodology, performed experiments, and wrote the article with comments and reviews from ID, KK, SS, ST, KR, SM, EB, and GDP. MM and ID performed the formal analysis. IG curated data. ST and KR assisted in dataset generation. HS validated data. EB and GDP supervised the primary author.

In this chapter, we expand the application of biomedical machine reading comprehension (MRC) to another unexplored area – MRC on clinical notes containing information related to injection drug use (IDU). This study aims to extract patients' IDU information from clinical notes using an MRC system. Extracting IDU behaviors from patients' electronic health records (EHR) is difficult because there is no International Classification of Disease (ICD) code, and the only place IDU information can be indicated is unstructured free-text clinical notes. Although NLP can efficiently extract this information from unstructured data, no validated tools exist. To address this gap in clinical information, we design and

demonstrate an MRC framework, IDU-QA, to extract information on IDU from clinical notes. Unlike other methods discussed in the literature, the MRC model can extract various types of information without being constrained by predefined entities, relations, or concepts. Due to the unavailability of an MRC dataset for this specific problem, we have designed our study as a two-part investigation: (1) generating a reference MRC dataset and (2) developing and testing the MRC model. To build the dataset, we use a combination of manual exploration, consultation with subject-matter experts, and rule-based NLP techniques. However, given the sensitivity of the subject matter and use cases of the MRC dataset, the final dataset still needs to be validated by subject-matter experts (SMEs) to ensure high quality. Therefore, we call this MRC dataset a "reference dataset". In our experiments, this reference dataset serves as the reference against which we measure the MRC models' performance using the evaluation metrics.

We start the chapter with a brief overview of the motivation and a description of the pipeline used for generating the reference dataset. Subsequently, we delve into the process of building and evaluating MRC systems using this dataset. To enhance the performance of the MRC model within this new problem domain with a comparatively smaller dataset, we employ transfer learning, including but not limited to the BioADAPT-MRC framework introduced in Chapter 3. Additionally, we demonstrate the MRC model's ability to extract IDU-related information from temporally out-of-distribution data. Finally, we end the chapter with a thorough error analysis of the MRC model to identify its strengths and weaknesses in effectively extracting IDU-related information from clinical notes.

5.1 Motivation

IDU is a critical health concern in the United States and internationally [282]. Most people begin using illicit drugs through other modes of administration, such as smoking, intranasal absorption, or oral ingestion. As dependence grows, individuals tend to prefer the intravenous (IV) route of drug administration, injecting drugs directly into the veins, as it offers stronger and more immediate effects [283]. The number of people who use IV drugs increased almost fivefold from 2011 to 2018 according to estimates in [284], whereas the number of IDU-related overdoses increased eightfold from 2000 to 2018 [285].

IDU is a highly dangerous practice, which can lead to complicated medical conditions such as abscesses and cutaneous infections, scarring and needle tracks, endocarditis, HIV/AIDS, Hepatitis C, overdose and deaths [282, 286–289]. An increase in IDU is also associated with an increase in morbidity and mortality [290–292].

Accurately identifying IDU behaviors in patients is crucial for risk assessment and detection of patients who can benefit from harm reduction interventions to potentially prevent IDU-related morbidity and mortality potentially [13, 293]. In the literature, IDU-related information extraction has been studied along with other socio-behavioral determinants of health (SBDH). Considering and including SBDH such as prior incarceration, substance use (regardless of administration mode), treatment attitude, psychological distress, and interpersonal violence improve patient mortality and enhances the prediction of medication adherence, hospital readmission, and suicide attempts [294, 295].

Despite the growing interest, SBDH such as IDU is not identifiable in patients' electronic health records (EHRs) through ICD codes; although not systematically assessed, it can be documented in clinical notes [15, 16]. While structured data fields derived from EHRs may provide some information about risky drug use behaviors and morbidities related to IDU, the clinical progress note is the only place it can be explicitly documented [13]. Despite being clinically meaningful and having the potential to identify patients that can benefit from harm reduction interventions, care providers often struggle to retrieve these data points from EHRs, and evidently, the exclusion of this data may result in overall reduced quality of care [296, 297].

Natural language processing (NLP) can help extract SBDH-related information from clinical notes and expand the utility of such information in patient care [298–300]. NLP is a branch of computer science that involves automated learning, understanding, and generation of natural languages, enabling the interactions between machines and human languages. Although NLP deals with a variety of tasks involving unstructured text data (e.g., event prediction [14], entity recognition [301], information extraction [272], and relation extraction [302]); in this chapter, we use MRC task to automatically extract information related to IDU from clinical notes in EHRs.

Although not specific to IDU, several studies have focused on identifying clinical concepts or information on substance use disorders (SUD) using NLP [19, 300]. Various NLP techniques have been used in these studies to extract SUD-related information. The stemming algorithm has been used to identify words and phrases associated with mental illness and substance use in clinical notes [303, 304]. Dependency structure has been utilized to capture relationships between phrases and tokens in the substance use statement [19]. Word-embedding models have been employed to identify alcohol and substance abuse status [299]. MRC has been applied to extract predefined sets of clinical concept categories and relation categories, such as relations of medications with adverse drug events and SBDH [300]. Multi-label text classification and sequence labeling have been used to identify sentences containing labeled arguments about drug use [305]. Topic modeling and keyword matching techniques have been leveraged to extract drug use–related information [18]. Techniques such as active learning [306], multi-label classification [307–310], concept extraction, and joint extraction of entities and relations have been employed to extract information about drug use [17]. Researchers have also focused on identifying drug use information by using NLP-specific techniques to detect opioid use disorder and predict overdose [311–322]. In the literature, we found one research study focused exclusively on IDU. The study has utilized rule-based algorithms, such as regular expressions (RegEx), NegEx [323], and N-grams to search for IDU-related terms, with the objective of identifying people who inject drugs (PWIDs) [13].

Although these techniques have been shown to perform relatively well in extracting drug use–related information from clinical notes, they are constrained by a set of predefined entities, labels, or relations and, as such, may not generalize well when presented with unseen types of entities/labels/relations [17, 302, 307, 324]. Moreover, only a few of these works focus solely on IDU-related information extraction [13], which is the theme of our study. Since evidence of IDU cannot be found in structured EHR data and therefore must be inferred from clinical notes, this study's sole focus on IDU aims to help understand how this phenomenon is represented in unstructured notes data and augment techniques that have used NLP techniques less generalizable to this population.

We focus on extracting IDU-related information from clinical notes by formulating the problem as an MRC task. To the best of our knowledge, there has

been no published attempt at developing an MRC algorithm to extract IDU-related information from clinical notes.

In this MRC task, given a query and a clinical progress note, an MRC model would return the relevant answer verbatim from the note as the extracted information. Thus, an MRC system is tasked with learning to read and comprehend the clinical progress note provided a query and then to extract information consisting of consecutive words from the notes relevant to the query from that note (Figure 5-1).

Because extracted answers are not derived using a predefined set of labels or entities and the dependencies/relationships between the words in the answers are not constricted in any way, this method provides more flexibility in identifying relevant information.

The most streamlined way to solve the MRC task is by using state-of-the-art Bidirectional Encoder Representations from Transformers (BERT)-based models [20, 83]. BERT models are efficient deep neural networks that can evidently achieve comparable performance in extracting targeted information from different types of biomedical documents, such as scholarly articles [83, 249], clinical practice guidelines [272], electronic medical records [151], etc. Nonetheless, evidence suggests that these supervised deep learning models require high-quality and large-scale annotated datasets to achieve comparable performance [20–22] and the absence of such a dataset for our targeted MRC task poses a critical challenge. An annotated MRC dataset consists of data samples where each sample is a triplet with a context (e.g., clinical progress note), a question, and an answer taken verbatim from the context (i.e., the extracted information).

To overcome the challenge of annotated MRC data availability, we build a high-quality reference MRC dataset that includes clinical notes as contexts and question-answer pairs specific to IDU in collaboration with subject matter experts (SMEs). We then build and evaluate an effective MRC system using this reference dataset. We also perform an error analysis to identify the strengths and weaknesses of our MRC system to inform future research.

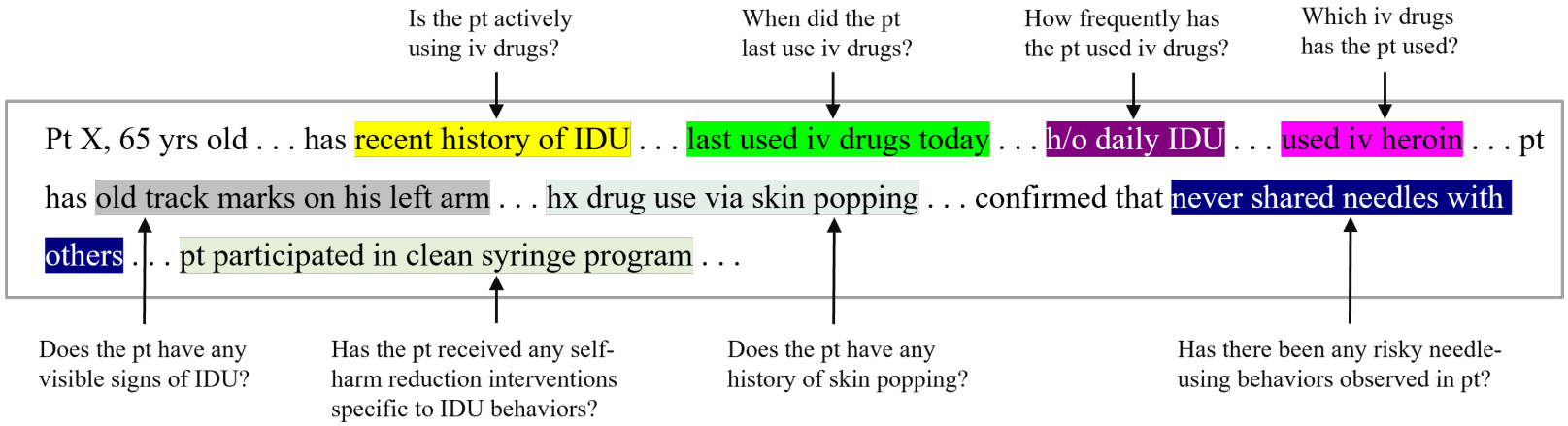


Figure 5-1. A sample clinical progress note featuring questions about the patient's IDU behavior with extracted IDU-related information color-coded in the note. (Reprinted from [281].)

5.2 Material & Methods

In this section, we elaborate on the formulation of this study, the IDU-QA framework, along with its two components: (i) Reference dataset generation and (ii) modeling (Figure 5-2).

5.2.1 Problem Formulation

We formulate the information extraction task as an MRC problem in NLP in the following manner: Given a question on patients' behavior about IDU and a clinical progress note with IDU-related information (i.e., the context), an MRC system retrieves the relevant information (i.e., the answer) from the provided note.

For example, given the following question *Does the patient have a history of IDU?* and clinical progress note *pt X, 200 yrs old . . . he has a history of smoking with 50 pack years, quit 10 years ago . . . social ethanol user . . . no history of idu . . . remote history of marijuana use . . . family hx: . . . physical exam: . . . provider: name.* – the MRC system is expected to return the answer *no history of idu* – verbatim from the note.

5.2.2 Reference Dataset Generation

MRC is a supervised NLP learning task and, as such, requires an annotated reference dataset for model development and inference. In an MRC dataset, each sample comprises a triplet consisting of the context, a question, and an answer, with the question-answer pairs serving as annotations. To generate a reference dataset from clinical notes, which serve as the context, we employ a three-stage process outlined in Figure 5-2: (1) question collection, (2) context processing, and (3) reference-answer extraction.

5.2.2.1 Question Collection

We initialize the process of question collection for the dataset by asking SMEs about the kind of information on IDU they are interested in from the clinical notes. We then generate a set of questions based on their interest. Table 5-1 shows the nine categories of interest. In the rest of the chapter, we use the term "Query Group" to imply categories of interest.

Table 5-1 also provides sample questions and answers for each query group.

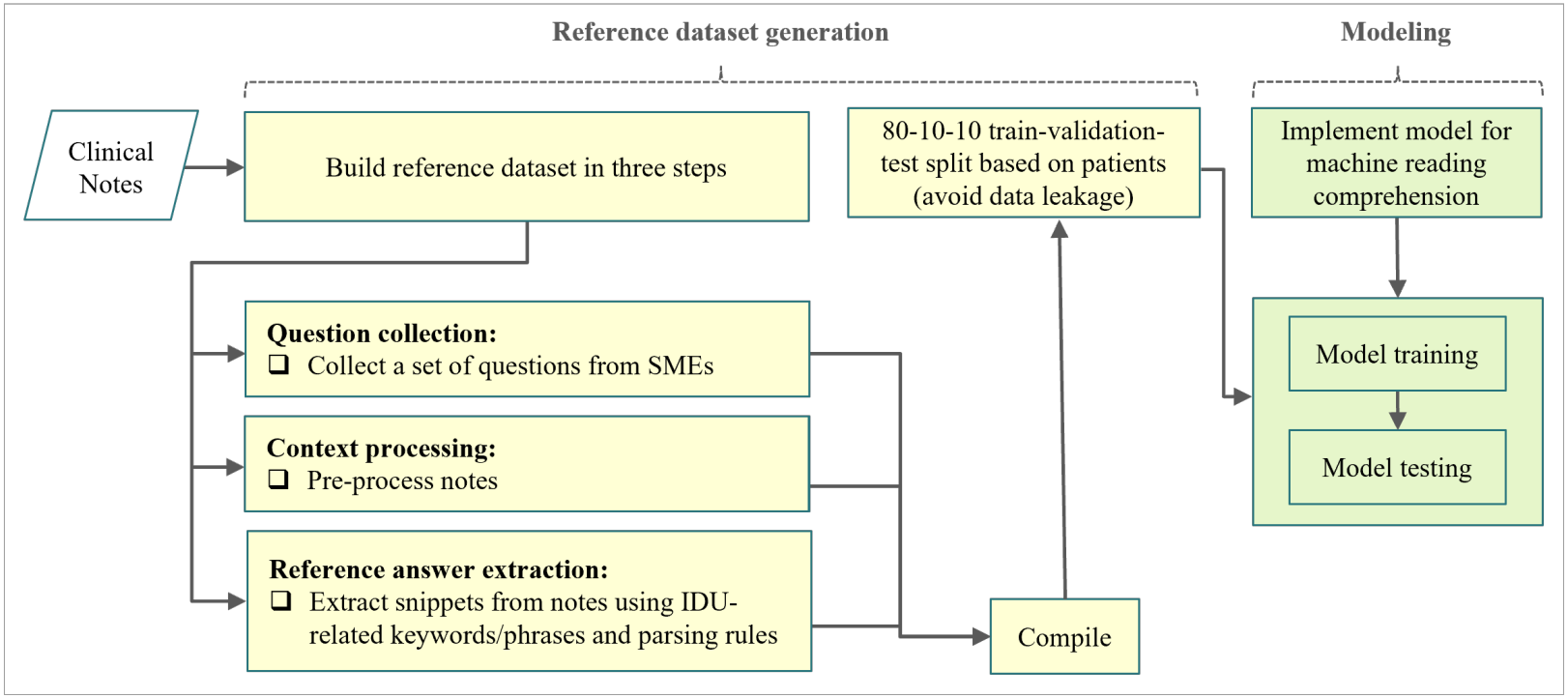


Figure 5-2. IDU-QA: Framework for a two-part study for extracting information on IDU behavior in patients from clinical notes. The first part consists of reference dataset generation in three primary steps, and the second part consists of MRC model development from implementation to inference. (Reprinted from [281].)

One query group targets to extract one type of information from the notes pertaining to that group. Thus, one question per query group should suffice to extract IDU-related information for that group. For example, for the query group "drug names", we can use the question "Which IV drugs has the patient used?" to extract any information about IV drug names from the notes. However, we anticipate our system to be used as a standalone application – a more user-friendly MRC tool to collect IDU evidence – and to be capable of handling different styles of questions posed by clinicians. Moreover, we also want our model to understand the questions from each query group rather than memorize them, and incorporating only one question from each query group would encourage memorizing behavior.

Therefore, to ensure that the MRC model can handle a wide range of user queries and extract information from clinical notes accurately, we include multiple versions of the same question for each query group. These different versions are designed to extract the same information from the notes, which helps to increase the models' flexibility, comprehensiveness, and robustness. We achieve this by including appropriate abbreviations, synonyms, and syntactical variations:

- Abbreviations: "Is the patient actively using intravenous drugs?" → "Is the pt actively using intravenous drugs?", "Is the patient actively using intravenous drugs?" → "Is the patient actively using iv drugs?", etc.
- Synonyms: "Does the pt have a history of using intravenous drugs?" → "Does the pt have a history of using injection drugs?", "Does the pt have a history of IDU?" → "Does the pt have a history of IVDU?",
- Syntactical variations: "Which iv drugs has the patient used?" → "To which iv drugs has the patient been exposed?", "Does the pt have a history of IVDU?" → "Has the pt ever used IV drugs?"

It should be noted that when identifying abbreviations and synonyms to be used in questions, we only choose terms and variants that clinicians commonly use. Examples of these terms and variants include "patient" and "pt", "intravenous" and "iv", "history" and "hx", and "IVDU" and "IDU". And, to ensure that we were able to accurately capture the nuances of possible language usage in the questions with regard to syntactical variations, we sought the guidance of SMEs.

Table 5-1. Types of information about IDU that are most likely to be inquired by clinicians from the clinical notes, categorized into nine query groups. Abbreviations: ivdu, intravenous drug use; bnz, benzodiazepines. (Reprinted from [281].)

Query Groups	**Sample Question** **→ Sample Answer**
Drug names	Which iv drugs has the patient used? → hx of iv heroin abuse, cocaine, and bnz
Visible signs of IDU	Does the patient have any needle track marks? → track marks noted over bilateral upper extremities
Risky needle-using behavior	Has the patient ever shared needles? → hx of ivdu, has shared needles in the past few weeks
Active/historical use	Is the patient actively using iv drugs? → h/o active iv drug use
Frequency of use	How frequently has the patient used iv drug? → history of ivdu (reports daily use of heroin)
Last use	When did the patient last use iv drugs? → daily use of iv heroin with last use 4 days prior to admission
Skin popping	Does the patient have any history of skin popping? → diffuse scarring from skin popping on lower extremities
Self-harm reduction interventions	Has the patient been counseled on safe injection technique? → discussed the importance of using clean needles with patient should he continue to inject drugs
Existence of IDU	Does the patient have any history of IDU? → no ivdu or h/o sharing needles

5.2.2.2 Context Processing

The contexts in the reference dataset are clinical notes that contain some IDU-related information. To guarantee that the clinical notes include information relevant to IDU and narrow down the notes accordingly, we use a list of keywords/phrases indicative of IDU (refer to Table 5-2), which has been developed by SMEs. SMEs followed an iterative approach to create this list. They began by compiling a list of common terms related to IDU, which they then refined by reviewing the associated snippet. They removed terms that caused excessive noise, such as "slamming" and "drug paraphernalia," and added terms like "skin popping" to enhance granularity. The experts received extensive training to sort and/or define the snippet categories, and they validated the terms to ensure their accuracy.

For our study, we assumed that the presence of any of these IDU-related keywords indicates the presence of relevant information pertaining to IDU in the note. Hence, we discard the notes that do not contain any of the words/phrases provided in Table 5-2, suggesting the possible non-existence of any IDU-related information in that note. As shown in Table 5-2, this list can be categorized into the following groups: IV drug names, visible signs of IDU, risky needle-using behavior, skin popping, self-harm reduction interventions, and generic IDU terms.

Our final set of clinical notes undergoes minimal cleaning to preserve crucial information. The preprocessing involves removing multiple consecutive new lines, white spaces, periods, dashes, and underscores. Additionally, new lines trailing internal punctuation marks such as commas, colons, semicolons, apostrophes, and quotation marks are also removed.

5.2.2.3 Reference Answer Extraction

The next step in our dataset generation process is to extract reference answers (i.e., information related to IDU) from the clinical notes. clinical notes are inherently lengthy, and manually extracting the reference answers from them requires a substantial amount of time, rendering the process unfeasible. Therefore, we devise an automated step-by-step extraction process that incorporates rule-based NLP techniques, considerably reducing manual effort while retaining high quality. We base the process on three assumptions:

1. Our MRC task only tackles information extraction (i.e., answering questions) from one single place (a sentence) in the note at a time.

Table 5-2. A list of IDU keywords/phrases provided by SMEs. Abbreviations: ssp, syringe services programs; ivda, intravenous drug abuse; ris4e, resists infection by sterile syringe safe sex and education; PWID, people who inject drugs. (Reprinted from [281].)

Keyword Groups	**Keywords/phrases**
IV drug names	iv/intravenous/inject(s/ed) heroin/meth/cocaine/crack, speedball
Visible signs of IDU	track marks, skin popping
Risky needle-using behavior	sharing/shared/dirty needle
Skin popping	skin popping
Self-harm reduction interventions	community/clean/safe syringe service/program, ssp, ris4e, counseled on safe(r) injection, safe injection technique
Generic IDU terms	ivdu, idu, ivda, iv/intravenous/injection drug use/abuse, inject/injected drug, drug(s) by injection, iv/intravenous drug injector/injection, illicit iv/intravenous drug, iv/intravenous drug paraphernalia, suspect injecting, pwid

2. The inquired information can be found in a single sentence in the note.

3. If the note contains IDU-related information in multiple locations, each is considered a separate answer string. Furthermore, multiple answer strings from the same note are expected to contain different kinds of information that should be answered by different questions. For example, in the note snippet – *pt has a history of smoking with 50 pack years, quit 10 years ago . . . social ethanol user . . . has h/o ivdu . . . remote history of marijuana use . . . last used iv meth 2 years ago . . .*, there are two locations where IDU-related information can be found – *has h/o ivdu* and *last used iv drugs 2 years ago*. In such cases, we consider them as separate answers that are retrieved when asked the following questions: *Does the pt have a history of IDU?* and *When did the pt last use iv drugs?*.

Given clinical notes, we extract the automated reference answers using rule-based NLP techniques as follows:

Step 1: Tokenize the sentences in the notes.

Step 2: Identify sentences that contain any of the IDU keywords from Table 5-2 using regular expression string matching and discard the rest.

Step 3: At this point, the sentences containing the IDU keywords can be ideally considered the reference answers (i.e., extracted information relevant to IDU). Nonetheless, our primary aim is to extract IDU-related information from the notes, but we also want the extracted information to be as precise as possible, containing lesser nonessential information. A full-sentence answer is most likely to include nonessential information, which can be further reduced by using parsing rules.[1] An example is transforming the sentence *social history: pt lives with family in [location], quit smoking 10 y ago, occ etoh, denies ivdu* into the phrase *denies ivdu*.

To create the parsing rules in this study, we randomly sample a set of sentences and focus on identifying specific phrases that occur together before or after the IDU keywords and modify or provide information that is crucial to the IDU-related history of the patient (refer to Table 5-1). These phrases can be adjacent to or distant from the keywords. For example, *pt. lives with*

[1]Parsing rules refer to NLP techniques that can identify specific patterns of text within a string that represent the concepts of interest while ignoring the remaining text.

family, denies ivdu. vs *pt lives with family, denies any tobacco, etoh or ivdu.* In this example, the phrase "denies" provides crucial information on the IDU behavior of the patient.

In Table 5-3, we provide a detailed list of these phrases along with the targeted pattern type, parsing rules, and examples of how they help reduce the nonessential information from the answers. The parsing rules mainly focus on identifying patterns stating negative IDU mentions, temporal information, opioid use disorder specific to IDU, and status of track marks.

As we apply each parsing rule, we check a set of random samples for quality control. Following the application of these parsing rules, we randomly sample 50 answers and manually analyze their correctness. Although our answer extraction process can usually identify the correct, concise reference answers from the clinical notes, we have observed that in certain cases, irrelevant words are also present in the identified answers. We address this issue in Section 5.3.6.

5.2.2.4 Question-to-Answer Mapping

Finally, to generate the labels (question-answer pairs) of our reference dataset, we create mappings between the questions from Section 5.2.2.1 and the reference answers from Section 5.2.2.3. We achieve this by considering the query groups in Table 5-1. For each query group, we identify a group of words in the reference answers that are most likely to provide the information inquired by that query group. Depending on the kind of information we are interested in (reflected by the query groups), these words can be either the keywords in Table 5-2 or the words (Table 5-3) that co-occur with the keywords and can help convey the information inquired by the user. For example, co-occurring words "daily" and "last" describes the "frequency of use" and the "last use" of IDU, respectively.

Thus, for each query group, we decide on a group of words that are most likely to help convey the inquired information and map the answers that contain these words to the questions in that query group (Table 5-1).

In Table 5-4, we present the mappings between the query groups and the words in reference answers that are most likely to provide inquired information. We also demonstrate sample answers for each mapping. Note that the answers in one query group and the answers in a different query may not be mutually exclusive.

Table 5-3. A detailed list of co-occurring phrases before/after IDU keywords along with the parsing rules, targeted pattern type, and examples of answers before and after parsing. (Reprinted from [281].)

Targeted Patterns	**Co-occurring Phrases Before/After IDU Keywords**	**Parsing Rules**	**Example Answers Before and After Parsing**
Negations (nullifying the existence of IDU/skin popping/track marks/needle sharing)	Negation Phrases (NP): den(ying/ies/ied), no, never	**Sentence:** <text before NP> NP <text after NP containing keywords> **Rule:** NP <text after NP containing keywords>	**Before:** 65y/o m w cardiac procedures, or recent surgical procedures, admits to drinking alcohol daily for the past 10 years, denies any history of ivdu **After:** denies any history of ivdu
Temporal information (active/historical/last/ frequency of use)	Temporal Phrases (TemP): past/remote/distant/prior/previous/former/ active/current/recent/last/long history/hx (or h/o), daily, occasional, past, remote, distant, prior, previous, former, active, current, recent, last, long, intermittent, hpi, history, hx, h/o, pmh	**Sentence:** <text before TemP> TemP <text after TemP containing keywords> **Rule:** TemP <text after TemP containing keywords>	**Before:** pt smokes cannabis, has a h/o ivdu but none now, went to rehab 2070 **After:** h/o ivdu but none now, went to rehab 2070
Additional temporal information	Additional Temporal Phrases (ATP): year(s)/yr(s)/month(s)/mnth(s)/day(s)/d/wk/ wks/mos ago	**Sentence:** <TemP> <text after TemP and before ATP containing keywords> ATP <text after ATP> **Rule:** <TemP> <text after TemP and before ATP containing keywords> ATP	**Before:** last ivdu was 10 days ago, snorts cocaine occasionally **After:** last ivdu was 10 days ago
Opioid/substance use disorder specific to IDU	Phrases related to Substance use disorder (SP): substance/polysubstance use/abuse disorder, sud, psud, oud, polysubstance, opioid use disorder, opioid, opiate	**Sentence:** <text before SP> SP <text after SP containing keywords> **Rule:** SP <text after SP containing keywords>	**Before:** 200m w niddm, htn, bipolar disorder and oud (iv heroin) on methadone maintenance, recent heroin relapse **After:** oud (iv heroin) on methadone maintenance, recent heroin relapse
Status of track marks	Phrases related to Track Mark status (TMP): arm(s)/abnormal/multiple/many/several/ healing/healed/old/diffuse/localized/visible/ red/iv/fresh/dark/needle/notable	**Sentence:** <text before TMP> TMP <text after TMP containing "track marks"> **Rule:** TMP <text after TMP containing "track marks">	**Before:** comments: extremities: mid line in upper right arm, scars and old track marks noted on mid arm **After:** old track marks noted on mid arm

This is because if we find words in an answer that belong to multiple query groups, then that answer is mapped to all questions from these query groups. For example, the first sample answer from Table 5-4 – "recent iv meth use" – contains the words "recent" and "iv meth" from query groups "active/historical use" and "drug names", respectively. Hence, this answer will be mapped to all questions in these two query groups.

5.2.3 Modeling with Machine Reading Comprehension System

In the next step of our study, we develop the MRC model for extracting IDU-related information using the reference MRC dataset from Section 5.2.2. We use BERT-based deep learning MRC models where the feature extractor is a trainable pre-trained BERT-based language model, and the MRC task layer is a single-layer feed-forward neural network.

Baseline modeling: We experiment with four state-of-the-art pre-trained language models – BERT [20], BioBERT [83], BlueBERT [130], and ClinicalBERT [120] – as trainable feature extractors and develop four MRC models.

Provided a sequence of tokens (words or pieces of words) in a question and a clinical progress note, the MRC model returns the start and end token of the answer span. Any text between the start and end tokens included is then considered as the answer (i.e., the extracted information). Together with the question and the note, the maximum allowable number of input tokens in these BERT-based MRC models is 512. To handle samples with longer clinical notes, we follow a widely known technique in MRC modeling – sliding window with a document stride [20].

Below we provide a brief description of this technique: Given an input question consisting of 20 tokens, the remaining allowable number of input tokens for the note is limited to 492 (which is 512 minus the 20 tokens in the question). If the note exceeds this limit, we employ a sliding window technique to split it into multiple chunks using a document stride of 128 tokens. The document stride determines the starting token of each subsequent chunk. After this preprocessing step, each chunk prepended with the original question tokens is considered a separate data sample.

Modeling with transfer learning: As described in Chapter 2, previous studies have demonstrated the effectiveness of using an additional high-quality, large-scale

Table 5-4. Mappings between the query groups and the words in reference answers most likely to provide inquired information and example answers for each mapping. (Reprinted from [281].)

Query Groups	**Words in Reference Answers Most Likely to Provide Inquired Information**	**Sample Answers**
Drug names	IV drug names from Table 5-2, opioid, opiate, oud	recent iv meth use
Visible signs of IDU	Phrases for visible signs of IDU from Table 5-2	multiple track marks over extremities
Risky needle-using behavior	Phrases for risky needle-using behavior from Table 5-2	h/o sharing needles with gf
Active/historical use	Temporal phrases from Table 5-3, remission	active iv drug user up to day of admission
Frequency of use	daily, occasional, regularly, often, sometimes, frequently, intermittent	- iv cocaine daily, ∼$5–40/day
Last use	Additional temporal phrases from Table 5-3, last, quit, since, clean	last ivdu >30 years ago
Skin popping	Phrases for skin popping from Table 5-2	h/o drug injections - skin popping
Self-harm reduction interventions	Phrases for self-harm reduction interventions from Table 5-2	patient participates in clean syringe program
Existence of IDU	Negation phrases from Table 5-3	denies any ivdu for many years
Existence of IDU	Remaining answers	iv drug user

reference dataset from the source domain in the training pipeline of the biomedical MRC models [167, 249]. The reason for this choice is the smaller size of biomedical MRC datasets often poses an obstacle for the MRC models to achieve comparable performance. These studies have used various transfer learning settings where the model's knowledge is transferred from the source domain to a new less-explored target biomedical domain with a smaller reference dataset [249, 272]. Thus transfer learning enables new problem domains to have a smaller dataset and still be able to achieve good performance in the MRC task.

For our study, we also use two state-of-the-art transfer learning settings – sequential learning [167, 272] and adversarial learning [249]. In sequential transfer learning, the MRC model is sequentially trained on the source- and target-domain instances with the expectation of transferring the model's knowledge that is acquired by learning the MRC task on the source domain to the MRC task in the target domain.

In adversarial learning, the MRC model is simultaneously trained on the source- and target-domain instances in a game-theoretic learning environment involving the feature extractor that generates feature representations for both source- and target-domain samples and a domain discriminator that discriminates these domains, competing them against each other. The domain discriminator encourages the feature extractor in the model to reduce the variance in source- and target-domain feature representations while maintaining good performance in their common task, MRC. The idea is to use a source domain in which the MRC models have already achieved human-level performance [21] and retain that performance for the target domain by bringing the feature representations from the source and target domains closer to each other. Overall, the purpose of adversarial learning is to ensure that the MRC model concentrates on performing the MRC task rather than the domain. A detailed description of the adversarial learning process is provided in Chapter 3.

In both transfer learning settings, as the source domain, we use the widely explored domain of Wikipedia articles. And as the target, we use our problem domain – clinical notes with IDU-related information.

5.3 Results & Discussion

In this section, we outline the specifications of our reference dataset, along with the metrics used to assess the performance of the MRC models. We also describe

the experimental setup, followed by a report of the findings from the experiments with MRC models. Furthermore, we conduct a comprehensive error analysis to demonstrate the capabilities and limitations of the MRC models in extracting information related to IDU from clinical notes.

5.3.1 Dataset Statistics

We use clinical notes sourced from the MIMIC-III dataset [135] to construct the reference dataset. MIMIC-III is a large database that contains health-related data from over forty thousand patients who were treated in critical care units at the Beth Israel Deaconess Medical Center between 2001 and 2012. The data in MIMIC-III is freely available and has been deidentified to comply with Health Insurance Portability and Accountability Act (HIPAA) standards. The deidentification process involved removing identifying information such as patient names, contact details, and dates. Dates in the database were shifted into the future using a random offset for each patient, preserving intervals and creating stays between 2100 and 2200. Additionally, the dates of birth for patients older than 89 were shifted to protect their true age, resulting in ages over 300 years in the database [135].

The NOTEEVENTS table in the database contains a total of 2,083,180 notes pertaining to 46,146 patients. These notes span between 7^{th} December 2097 and 1^{st} October 2210. For our study, we specifically focus on notes that include any of the IDU-related keywords listed in Table 5-2. Additionally, we narrow down our selection to notes documented within the initial fifty-year period. As a result, the timeline considered for our experiment ranges from 22^{nd} June 2100 to 9^{th} June 2150.

To reduce computational overhead during training and because unusually large notes (determined by the outliers in the distribution of note lengths) may contain templated nonessential information that is not relevant to any specific patient, we remove some outlier notes based on the interquartile range of the note lengths. We later show in Section 5.3.6 that note length does not affect the performance of the model at the time of inference.

Table 5-5 shows the statistics of our reference dataset. Our cohort consists of 1209 patients with a total of 3265 notes that have an average length of 879.19 words.

In addition, we examine the distribution of the query groups outlined in Tables 5-1 and 5-4 within the reference dataset. This analysis is illustrated by the pie chart depicted in Figure 5-3. The dataset is dominated by MRC pairs related to the "existence of IDU", as demonstrated. Following closely behind are MRC pairs about "active/historical use", whereas the least frequent MRC pairs in the dataset are those pertaining to "frequency of use" and "risky needle-using behavior". It is important to mention that within our set of notes, we did not come across any instances specifically addressing "skin popping" or "self-harm reduction interventions."

5.3.2 Experimental Setup

For experimentation, we divide our reference dataset into train, test, and validation sets using an 80-10-10 split based on patients to avoid any data leakage. To implement the MRC models, we use PyTorch [257]. We use the pre-trained language models from the huggingface API [258].

Based on the statistics of our reference dataset, we choose 512 as the maximum sequence length, 20 as the query length, and 100 as the answer length. After reviewing the hyperparameters utilized in various MRC tasks as outlined in [20, 45, 83, 86, 120, 128–131, 272], we set the document stride to 128 and opted for a batch size of 32, a learning rate of $3e^{-5}$ and a training epoch count of 3 for the training configurations. As the adversarial learning-based MRC model, we use the BioADAPT-MRC model from Chapter 3 and follow the default implementation process [249]. As the feature extractor in BioADAPT-MRC, we use the best-performing model, ClinicalBERT, from the baseline modeling experiments. As the source-domain dataset in adversarial and sequential learning–based MRC models, we use the SQuAD dataset [21], which has 87599 samples, built from Wikipedia articles and manually annotated by crowd workers. We performed all experiments using a single GPU on a Linux virtual machine with two GRID V100-32C GPUs.

5.3.3 Metrics

To assess how well the MRC models can extract information related to IDU from the reference dataset, we employ three widely used metrics in MRC tasks: Exact Match (EM) Score, Perfect Recall Score, and F1 Score.

Table 5-5. Statistics of the reference dataset.

Property	**Statistics**
#Patients	1209
#Notes	3265
#Notes per patient (average)	2.70
#Samples in dataset	33028
#MRC per note (average)	10.12
Note length (in words)	879.19 (average), 924 (median), 1604 (max)
Question length (in words)	6.33 (average), 7 (median), 14 (max)
Answer length (in words)	7.41 (average), 5 (median), 37 (max)

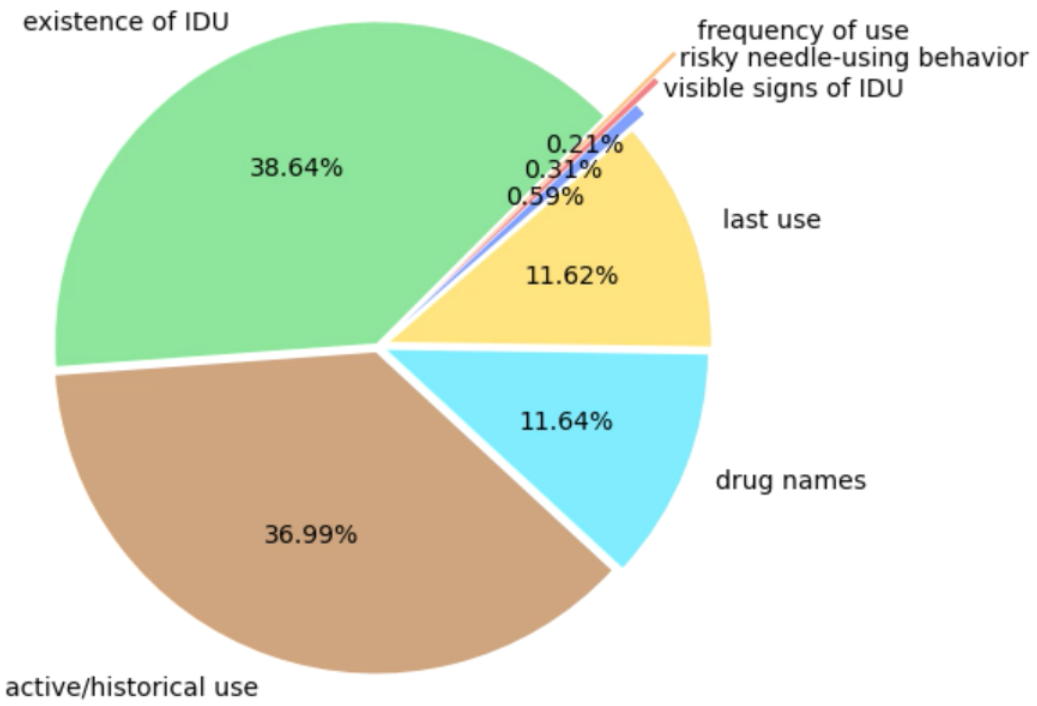


Figure 5-3. Distribution of the query groups in the reference dataset

The EM score for a data instance is determined by whether the predicted answer exactly matches the reference answer character-for-character. If there is a strict match, the EM score is 1; otherwise, it is 0. This means that even being off by a single character will result in an EM score of zero.

Perfect recall examines if the predicted answer contains the reference answer. If that is the case, then the perfect recall score is 1; otherwise, it is 0. Although this metric is less strict than the EM score, there is a potential issue with the predicted answer being the entire context. We address this concern in Section 5.3.6, where we compare the percentage overlap of the predicted answer with the context to that of the reference answer with the context.

The F1 score is a widely used classification metric that assigns equal importance to precision and recall, making it more lenient than EM or Perfect Recall score. An MRC model can make two types of errors: False Positive (FP) when the model extracts words that are not in the reference answer and False Negative (FN) when the model fails to extract words that are in the reference answer. The F1-score takes into account both of these errors by taking the harmonic mean of precision and recall in its measurement, providing a balanced measure of how the MRC model is performing. Precision is calculated as the ratio of correctly extracted words in the predicted answer to the total number of words in the predicted answer, whereas recall is calculated as the ratio of correctly extracted words in the predicted answer to the total number of words in the reference answer. Equations 4.4 and 4.5 can be used to calculate the F1 score for each question-answer pair.

The range of EM, Perfect Recall and F1 scores are {0,1}, {0,1}, and [0,1], respectively.

5.3.4 Results on Reference Test Set

This section focuses on examining the experimental outcomes of the MRC models and demonstrates their performance on the test set of our reference dataset. As shown in Table 5-6, in our baseline modeling experiments, ClinicalBERT outperforms other BERT-based baseline models. An exact match of ∼65% and a perfect recall of ∼78% for ClinicalBERT implies that ClinicalBERT can extract IDU-related information 78% of the time and 65% of the time the extracted information is an exact match with the reference answer. A high F1 score of ∼82%

indicates that the ClinicalBERT model can extract a high percentage of correct information while achieving high precision in those extracted answers.

We use ClinicalBERT as the baseline model for our next experiments.

Table 5-7 shows the experimental results for the transfer learning settings – sequential and adversarial learning, using the best-performing ClinicalBERT model and SQuAD dataset. The results indicate that transfer learning, especially adversarial learning, improves the overall performance of the MRC model from the baseline.

5.3.5 Temporal Out-of-Distribution Testing

The writing style of clinical notes may change over time because of changes in clinicians, health care facilities, patients, etc. [325]. Given the purpose of our MRC model, it is imperative to examine whether the performance of our MRC models is retained over time. Therefore, we perform additional testing of the models from Table 5-7 on unseen data. We examine the MRC models' short-term and long-term information extraction capabilities by testing on clinical notes from two additional cohorts.

To select these cohorts, we consider the clinical notes in the NOTEEVENTS table, which were recorded after 9^{th} June 2150 (the cutoff date for our reference dataset as described in Section 5.3.1) and contain at least one of the IDU-related keywords listed in Table 5-2. The chart dates of these notes span a 58-year time window, ranging from 24^{th} June 2150 to 30^{th} April 2208.

For testing the short-term capability, we use notes from the first 29-year period spanning from 24^{th} June 2150 to 14^{th} June 2179. And, to assess the long-term capability, we use notes from the last 29-year period ranging from 24^{th} June 2179 to 30^{th} April 2208.

To avoid data leakage, we use patients and their notes that did not appear in the training set of our reference dataset. We use the method described in Section 5.2.2 for building the test datasets using these notes. For the rest of the chapter, we use the terms "Cohort-Short" and "Cohort-Long" to represent temporally out-of-distribution notes in the first 29-year period and the last 29-year period, respectively. Table 5-8 shows the statistics of the test datasets built using Cohort-Short and Cohort-Long. We also show the distribution of query groups in these test

Table 5-6. Performance scores of baseline MRC models on the test set. The highest scores are highlighted in bold.

Model	EM	Perfect Recall	F1
BERT	61.66	77.11	80.99
BioBERT	63.26	**78.70**	81.13
BlueBERT	55.96	74.11	77.99
ClinicalBERT	**64.93**	78.20	**81.73**

Table 5-7. Performance scores of the best baseline (ClinicalBERT), sequential learning, and adversarial learning–based MRC models on the test set. The highest scores are highlighted in bold.

Model	EM	Perfect Recall	F1
Baseline	64.93	78.20	81.73
Sequential Learning	66.42	**79.86**	82.59
Adversarial Learning	**66.72**	78.30	**83.20**

datasets in Figures 5-4 and 5-5. As shown, the distribution of the query groups is similar for the additional test sets and our original reference dataset (refer to Figure 5-3).

Table 5-9 shows the performance of the baseline, sequential learning, and adversarial learning–based models. As shown, for both test datasets, our MRC models perform with high scores, reflecting their possible competence in extracting information over time. Nonetheless, more experiments are necessary to draw any conclusion about this capability. The comparative analysis of the three models in Table 5-9 shows that overall, the adversarial learning–based model outperforms the others, indicating the effectiveness of adversarial learning–based transfer learning in our MRC task. However, more experiments are needed to ascertain the effectiveness of adversarial learning in this scenario.

5.3.6 Error Analysis

In this section, we provide a comprehensive analysis of the strengths and weaknesses of our best-performing model, which is the adversarial learning–based MRC model, in extracting IDU-related information. We perform a fivefold analysis as follows: Examine the (i) confidence intervals of the performance scores, the effect of (ii) note length, (iii) question length, and (iv) reference answer length on the performance of the MRC model, and (v) the performance of the MRC model for each query group. Furthermore, we showcase the proficiency of the MRC model in retrieving information related to IDU. For our error analysis, we consider all three of our test sets – the test set in our reference dataset and the test datasets from Cohort-Short and Cohort-Long.

Confidence Intervals of Performance Scores: We calculate the confidence intervals (CI) for EM, Perfect Recall, and F1 scores achieved by the best-performing MRC model to represent how "good" these estimates are and thus quantify their uncertainty. Smaller confidence intervals demonstrated in Table 5-10 indicate that our estimates are precise with a high level (95%) of confidence.

Effect of Note Length: Clinical notes have varying lengths – they can be as short as eight words up to as long as approximately 5787 words, based on the statistics of our test datasets. Therefore, we want the MRC model to perform consistently well for all lengths of clinical notes.

Table 5-8. Statistics of the additional test datasets built using Cohort-Short and Cohort-Long. The highest scores are highlighted in bold.

Property	**Statistics**	
	Test Dataset (Cohort-Short)	**Test Dataset (Cohort-Long)**
#Patients	1031	799
#Notes	2186	1852
#Notes per patient (average)	2.12	2.32
#Samples in dataset	22507	15693
#MRC per note (average)	10.30	8.47
Note length (in words)	1300.80 (average), 1170 (median), 5379 (max)	1373.61 (average), 1213 (median), 5787 (max)
Question length (in words)	6.32 (average), 7 (median), 14 (max)	6.27 (average), 7 (median), 14 (max)
Answer length (in words)	7.19 (average), 5 (median), 36 (max)	7.36 (average), 5 (median), 37 (max)

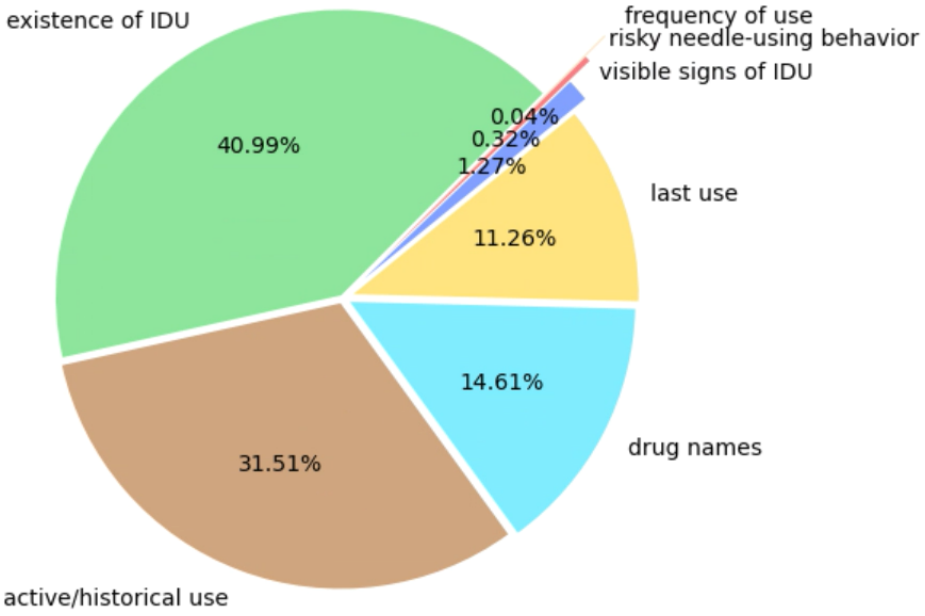


Figure 5-4. Distribution of query groups in the additional test dataset built using Cohort-Short.

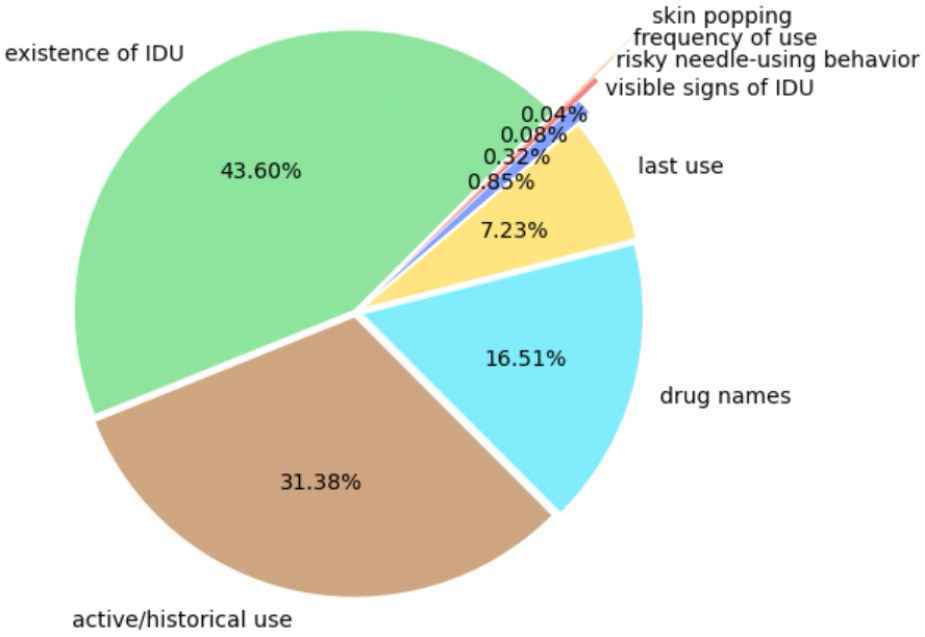


Figure 5-5. Distribution of query groups in the additional test dataset built using Cohort-Long.

Table 5-9. Performance scores of the best baseline (ClinicalBERT), sequential learning, and adversarial learning–based MRC models on the additional test datasets built using Cohort-Short and Cohort-Long.

Model	**Test dataset (Cohort-Short)**		
	EM	**Perfect Recall**	**F1**
Baseline	63.78	**80.77**	81.76
Sequential Learning	63.20	80.13	81.57
Adversarial Learning	**64.43**	80.12	**82.15**
Model	**Test dataset (Cohort-Long)**		
	EM	**Perfect Recall**	**F1**
Baseline	65.14	77.81	**80.74**
Sequential Learning	65.28	76.72	80.57
Adversarial Learning	**65.48**	**77.87**	80.34

Table 5-10. Performance scores with 95% confidence intervals adversarial learning–based MRC models on the test set in the reference dataset, test datasets from Cohort-Short and Cohort-Long.

Model	**EM (95% CI)**	**Perfect Recall (95% CI)**	**F1 (95% CI)**
Test Set (Reference)	66.72 (65.36 - 68.05)	78.30 (77.09 - 79.44)	83.20 (82.38 - 84.03)
Test Dataset (Cohort-Short)	64.43 (63.80 - 65.06)	80.12 (79.59 - 80.64)	82.15 (81.75 - 82.56)
Test Dataset (Cohort-Long)	65.48 (64.73 - 66.22)	77.87 (77.21 - 78.51)	80.34 (79.84 - 80.84)

To identify the effect of note length on the MRC model's performance, we calculate the length of the contexts (i.e., notes) in the three test sets and bin them into four quartiles based on their ascending lengths. The x-axis in Figure 5-6 shows the length range of these bins, whereas the green bars with the right y-axis show the sample count for each bin. We find that note length does not affect the model's performance scores, depicted on the left y-axis of Figure 5-6.

Effect of Question Length: We also examine the effect of question length on the performance of the MRC model. For this analysis, we adopt the same binning approach as the one on note length. Figure 5-7 shows that similar to note length, question length also has no notable effect on the model's performance scores.

Effect of Reference Answer Length: In our test sets, we have varying lengths for the reference answers (i.e., extracted information). For successful implementation, it is essential for the MRC model to be able to extract different lengths of information from the clinical notes. Using the binning approach described earlier for the analysis on note length, we find that the MRC model struggles more with longer answers than shorter answers – demonstrated in Figure 5-8. There is a higher chance that longer answers may include information that is not relevant to the specific inquiry (see Section 5.2.2.3). As a result, we believe that the presence of irrelevant words in these longer answers contributes to the decrease in the performance of the MRC model for these particular samples. Therefore, the subsequent phase of this research involves a manual examination of the lengthy answers in order to eliminate any irrelevant information.

Performance for Query Groups: Based on the information we are interested in extracting from the clinical notes, we create nine query groups as shown in Table 5-1. The green bars along with the right y-axis in Figure 5-8 show the sample count for the query groups in the test sets. The query group "existence of IDU" dominates the test sets, followed by the query group "active/historical use". Interestingly we find that the model performs the best on the query group "existence of IDU" and struggles the most with the group "skin popping" with the lowest exact match score. However, for a more fair comparison among the query groups, we need more test samples for the least appearing groups in our test sets.

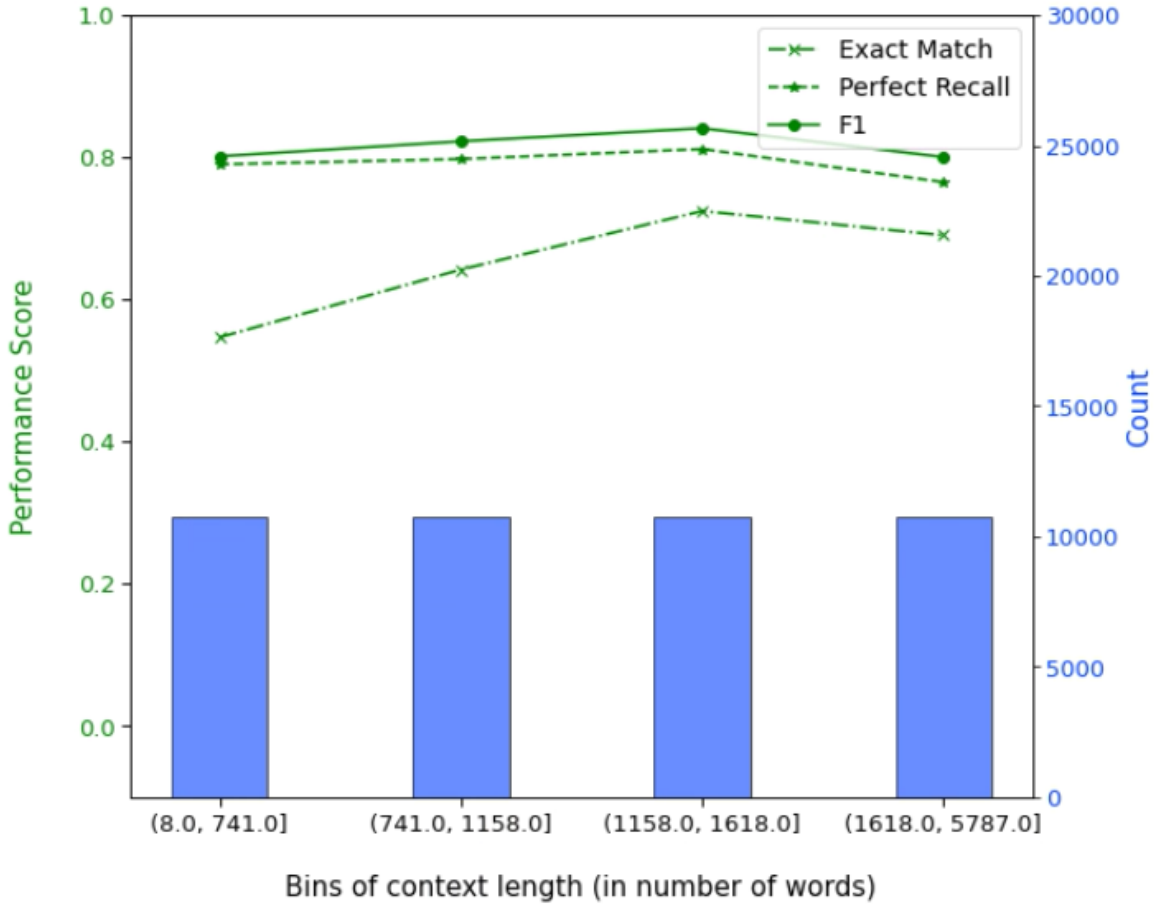


Figure 5-6. Effect of context (note) length on the MRC model's performance

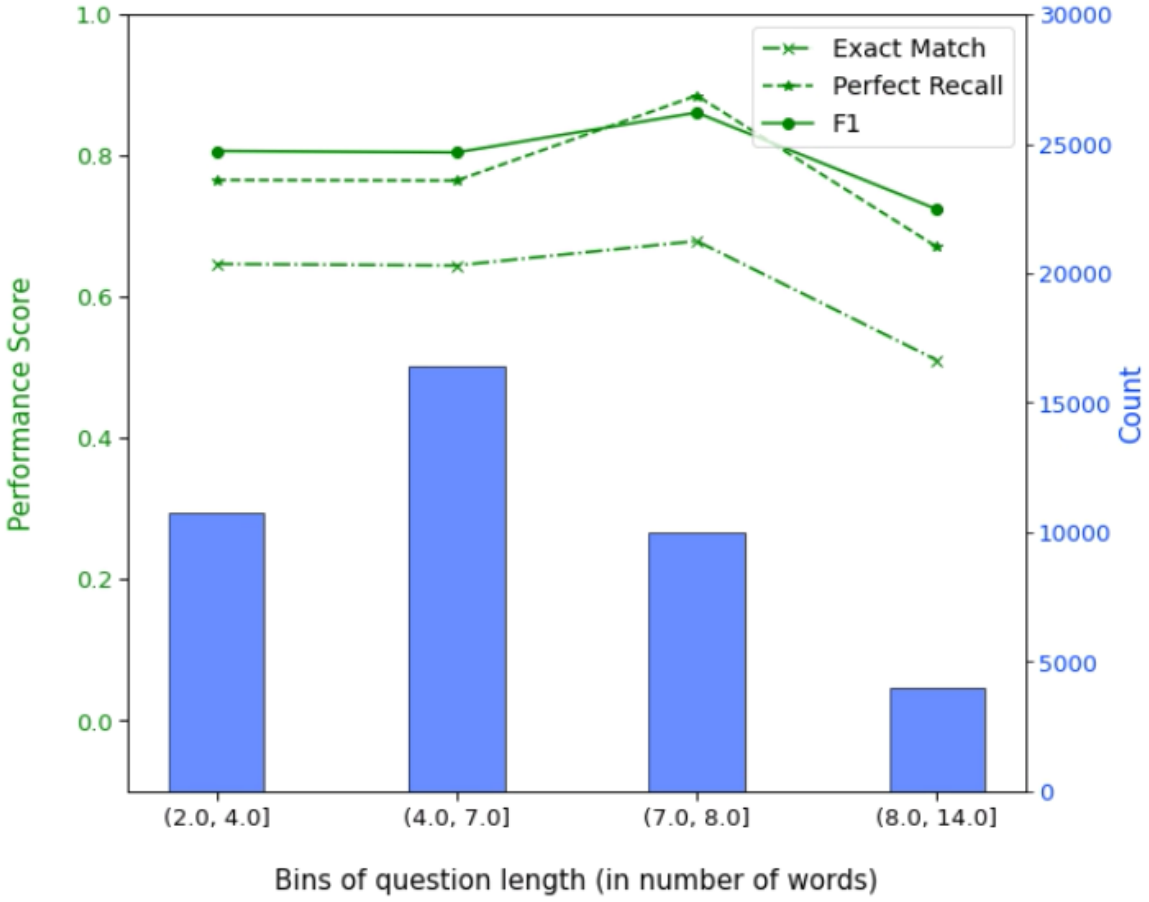


Figure 5-7. Effect of question length on the MRC model's performance

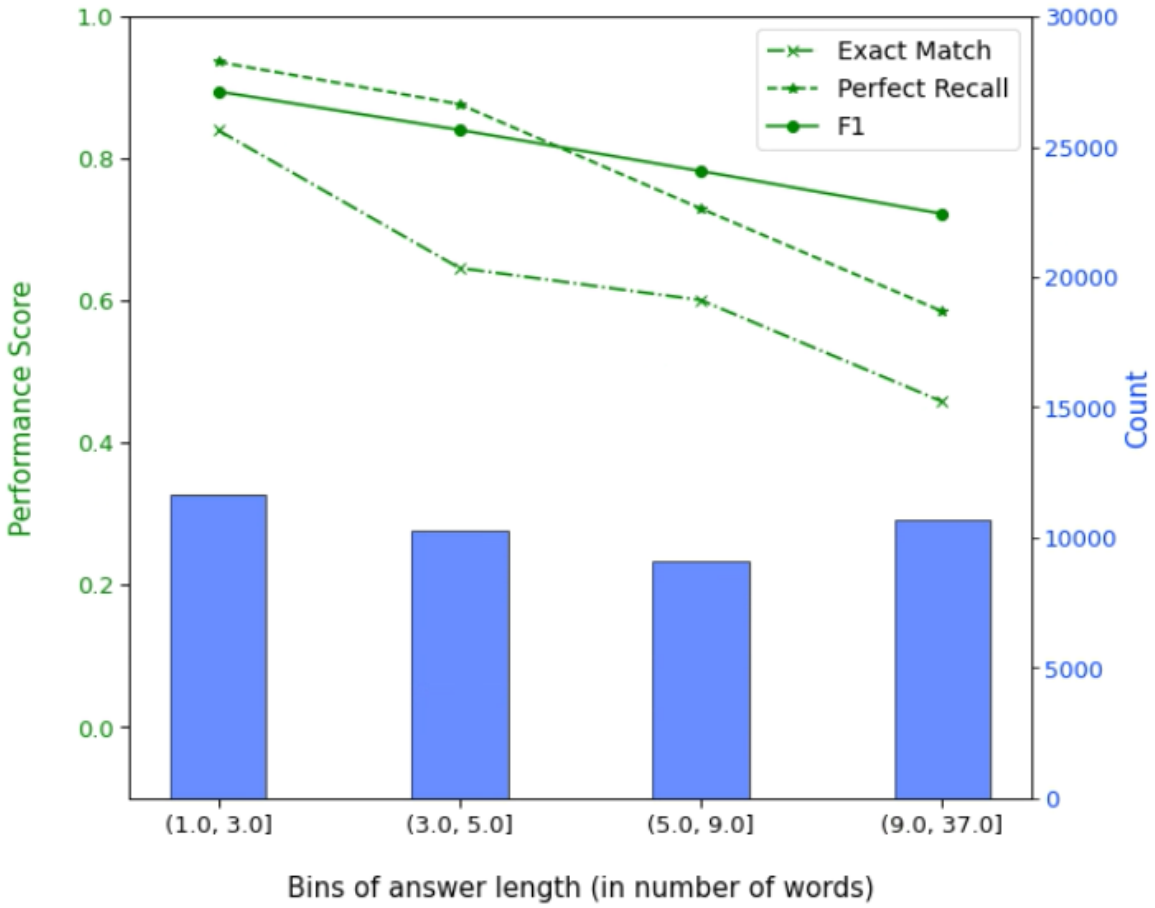


Figure 5-8. Effect of reference answer length on the MRC model's performance

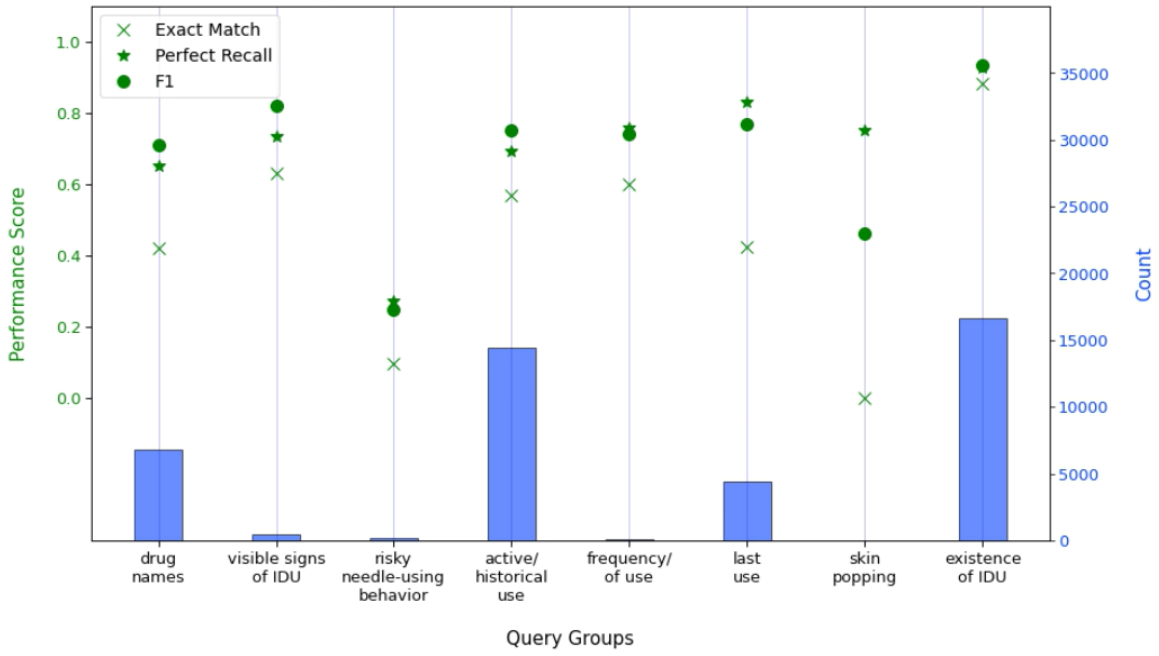


Figure 5-9. Performance of the MRC model for each query group

Analysis of Perfect Recall: As discussed in Section 5.3.3, the performance metric Perfect Recall has a potential issue with the predicted answers being the entire context in cases where they do not exactly match with the reference answers. To address this concern, we compare the ratio of the predicted answers (that do not have an exact match of 1) to the contexts with the ratio of the reference answer to the contexts. Figure 5-10 shows that the distribution of the percentage ratios of the predicted answers to the contexts is similar to that of the reference answers to the contexts.

Examples of Predicted Answers: We demonstrate the capability of the MRC model by showing some randomly selected examples of the predicted answers along with the questions and reference answers in Table 5-11.

Analysis of Model's Capability to Identify Whether a Note Contains IDU-related Information or Not: Our study focuses on extracting IDU-related information from clinical notes, but ideally, we also want our MRC model to identify whether the note contains IDU-related information or not. As such, as an additional analysis, we examine the MRC model's ability to identify clinical notes that do not contain any mention of IDU keywords (Table 5-2) and, as such, are assumed to have no information about IDU. We hypothesize that given a clinical progress note with no mentions of IDU, the MRC model should return an empty string because it could not find the information it was asked to retrieve.

To test this, we use the test set in the reference dataset from Section 5.2.2. Recall that in our context processing step in Section 5.2.2.2, we remove notes that do not contain any IDU keywords. For this analysis, we incorporate 362 notes from 181 patients with no mentions of IDU keywords. We ensure that these 181 patients do not have any notes in the data used for training the MRC model but only in the test set from Section 5.3.2.

To annotate these notes, we use the query group "existence of IDU" as questions and empty strings as answers. For example, given a note with no mentions of IDU and the question "Does the patient have a history of IDU?", the MRC model should return an empty string. To measure the performance, we consider only the exact match. Thus, if the predicted answer matches with the empty string, we consider that a success (exact match score = 1) and otherwise a failure (exact match score = 0).

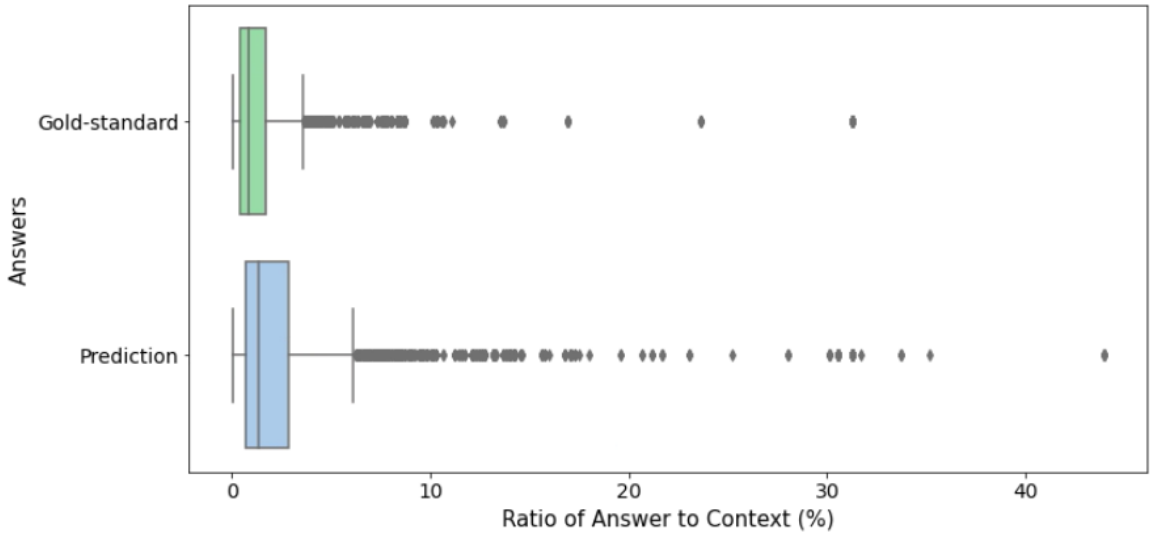


Figure 5-10. Distributions of the ratios of the predicted answers to the contexts and the ratios of the reference answers to the contexts. The ratios here are presented in the form of percentages.

Table 5-11. Examples of answers predicted by the MRC model along with the question and the reference answers. Abbreviations: ble, bilateral lower extremities.

Question	**Reference Answer**	**Predicted Answer**
Which IV drugs has the pt used?	ivda (cocaine, heroin)	ivda (cocaine, heroin)
Does the pt have any visible signs of IVDU?	noted needle track marks on arms bil	noted needle track marks on arms bil
Does the patient have a hx of sharing needles?	no active iv drug use (thinks he contracted hep c in college with a shared needle)	no active iv drug use (thinks he contracted hep c in college with a shared needle)
Active vs historical IDU?	remote history of iv drug use	remote history of iv drug use
When did the patient last use IV drugs?	past ivdu abuse in 90s	past ivdu abuse in 90s - poor peripheral veins
Does the patient have any history of skin popping?	he also has one on his rt arm that he says is from skin popping	one on his rt arm that he says is from skin popping
Does the patient have a history of IDU?	no etoh or tobacco, no illicit or ivdu	no illicit or ivdu

We find that our MRC model can identify 94.20% of the clinical notes that do not contain any IDU-related information. Therefore, we can say that while our MRC model can extract IDU-related information from clinical notes, it also has the potential to identify the notes that do not contain any.

Reference Answer Validation by SMEs: In Section 5.2.2.3, we explained that although the automated rule-based reference answer extraction process has succeeded, some samples still require further nonessential information reduction. Additionally, it is possible that some answers may contain information related to family members who use IV drugs instead of the patients themselves. For instance, "patient reported sadness as his daughter recently died from iv heroin overdose".

To ensure that our reference answers are of high quality and accuracy and to maintain the study's credibility, it is crucial to validate the correctness and precision of the data using domain knowledge. Therefore, in the future, we plan to collaborate with SMEs to validate the answers in our reference dataset and the additional test sets.

5.4 Conclusion

Detection of injection drug use or IDU behavior among patients is crucial for informed patient care. In this chapter, we have tackled the challenging task of extracting IDU-related information from clinical notes. We have built a framework IDU-QA to generate a reference MRC dataset using clinical notes and develop an MRC model that takes in a clinical progress note and end-user query on IDU to return the information on IDU extracted from the note. The dataset generation process involved a combination of manual exploration and rule-based NLP techniques. The outcomes of our MRC model are only influenced by the questions asked by the user and not by any predefined entities, relations, or concepts. We have conducted a comprehensive evaluation of the MRC model using our generated dataset. Furthermore, we have performed an additional evaluation to assess the capability of our trained MRC model to extract information from notes that are temporally out-of-distribution. Finally, we have thoroughly examined the strengths and limitations of the MRC model by conducting an error analysis to identify potential avenues for future research.

CHAPTER 6
CONCLUSIONS & FUTURE WORK

Unstructured text data in the biomedical field encompasses various types of documents, including scientific literature, clinical notes, psychotherapy notes, medical examinations, and practice guidelines. These unstructured data hold significant value for healthcare professionals as they can aid in clinical decision-making, enhance patient care and outcomes, and provide deeper insights. However, unlike structured data that is organized and easily searchable, text data is often irregular and disorganized. Consequently, extracting meaningful information from unstructured text data poses unique challenges and requires additional effort.

Recent advancements in machine learning and natural language processing techniques have facilitated the extraction of valuable information from unstructured biomedical text data in an efficient manner. One notable technique is machine reading comprehension, which emphasizes understanding complex biomedical documents and extracting information from them.

While extensive research has been conducted on extracting information using MRC models from non-biomedical text data, there remains a lack of exploration in various problem domains involving unstructured biomedical text data. A major impediment to applying these models to new biomedical problem domains is the unavailability of large-scale, high-quality annotated datasets required for training high-performing MRC models. Acquiring subject-matter expertise, which is necessary for creating such datasets, contributes to this scarcity. One approach to address this issue is through transfer learning.

This dissertation makes significant contributions toward implementing machine reading comprehension systems for biomedical problem domains in the absence of large-scale annotated MRC datasets with transfer learning.

We developed BioADAPT-MRC to enhance performance in the biomedical MRC task through domain adaptation with adversarial learning. The framework uses a dataset with a substantial amount of labeled samples from a widely explored

source domain and another dataset with few or no labeled samples from the target biomedical domain. We evaluated BioADAPT-MRC on three benchmark datasets from the BioASQ challenge. We demonstrated that our framework achieved overall superior performance on these datasets, surpassing previous benchmarks, without using labels in the target-domain training dataset. The performance improvement was approximately 7% compared to other state-of-the-art methods. Additionally, we conducted a comprehensive quantitative study using clustering and dimensionality reduction techniques to assess the domain adaptation capability, confirming that our framework has the ability to learn feature representations that are invariant across two domains, albeit to a certain extent. Furthermore, we extended our framework to a semi-supervised setting and showcased its efficiency when different proportions of labeled target-domain data were utilized. We identified the limitations of our framework by conducting a thorough error analysis to provide insights for future improvement and investigation. We concluded that BioADAPT-MRC has the potential to benefit healthcare systems by efficiently retrieving information from biomedical narratives, thereby preserving the crucial time and energy of healthcare professionals.

We identified two novel areas of biomedical research where the MRC task can be advantageous. Firstly, we explored the application of the MRC task on clinical practice guidelines (CPGs) to extract crucial information such as risk factors, diagnostic criteria, treatment recommendations, treatment benefits, and harms. Secondly, we focused on the MRC task applied to clinical notes containing information about injection drug use (IDU) to identify instances of IDU behavior in patients.

Given the importance of CPGs in healthcare and the difficulty of efficiently comprehending lengthy CPGs by healthcare professionals, we identified biomedical MRC tasks on CPGs as a new problem domain that can assist clinicians in interpreting them. We curated a benchmark validation dataset called cpgQA, which consists of 1097 context-question-answer triplets. The dataset was meticulously annotated with the help of subject-matter experts and a guideline. To thoroughly evaluate the dataset, we employed state-of-the-art pre-trained language models. Considering the scarcity of labeled data, we utilized transfer learning and the BioADAPT-MRC framework. The results showed that 78% of the answers predicted by our best-performing MRC model match the answers in the cpgQA

dataset. Additionally, we comprehensively conducted an error analysis to examine the shortcomings of the existing method for the MRC task on the cpgQA dataset and proposed potential directions for future research. Our objective in providing this dataset is to foster the development of MRC systems that enable efficient and intelligent interpretation of CPGs in the healthcare domain.

Furthermore, we tackled the complex task of retrieving information from clinical notes on injection drug use (IDU) by formulating it as an MRC task and building a framework IDU-QA. Our developed MRC framework incorporates rule-based NLP techniques and manual exploration to generate a reference dataset and an MRC model that accepts a user query regarding IDU and a clinical progress note as input to retrieve the relevant information on IDU from the note. We utilized a dataset constructed from clinical notes from MIMIC-III to assess the performance of the model. We applied sequential transfer learning and the BioADAPT-MRC framework to enhance model performance. The results showed that 67% of the answers predicted by our best-performing MRC model match the reference answers. It also showed that 78% of the predicted answers contain reference answers surrounded by additional words. Additionally, we conducted a further evaluation to assess the model's ability to comprehend temporally out-of-distribution notes and extract information from them. We found that the performance of the model remains consistent over time. Through a rigorous error analysis, we identified the strengths and limitations of our MRC model and outlined promising avenues for further study. Our aim with this study is to contribute to efficiently identifying IDU behaviors in patients from clinical notes, thereby supporting clinicians in their work.

We have identified several avenues for future work and outlined them here:

- Building MRC models and expanding the application of BioADAPT-MRC: The application of MRC models along with BioADAPT-MRC could be extended to extract various types of information from clinical notes, including but not limited to the severity and frequency of drug use (e.g., methamphetamine use). This extension could further enhance the capabilities of the BioADAPT-MRC framework and other MRC models.
- Integration of MRC models into real-world applications: The MRC models could be integrated into real-world applications focusing on identifying

patients who inject drugs and analyzing their IDU behavior patterns. This practical integration may help clinicians effectively address these issues.

- Incorporating tabular data in the cpgQA dataset: As tabular data is integral to CPGs, we recognize the importance of integrating it into the cpgQA dataset. One future direction can be incorporating tabular data along with text and transforming the cpgQA dataset into a multi-modal dataset. This expansion might enable the development of an MRC system capable of handling multi-modal inputs.
- Expanding the cpgQA dataset: To increase the reliability and diversity of the cpgQA dataset, more guidelines and data samples that may not have answers could be included to expand the dataset. This expansion may contribute to improving the performance of MRC models for CPGs.
- Developing a synthetic question-answer generator: To further enhance the performance of the BioADAPT-MRC framework, one could develop an efficient synthetic question-answer generator specifically designed for the biomedical domain to aid in generating additional training data for MRC models.
- Applying the BioADAPT-MRC framework to other NLP applications: The application of the BioADAPT-MRC framework could be extended to other biomedical NLP tasks that may encounter a shortage of labeled data, such as clinical negation detection and biomedical named-entity-recognition, to improve the accuracy by addressing labeled data scarcity issues in these areas.
- Robustness analysis of MRC models: The robustness of the MRC models developed in this research could be analyzed against carefully tailored adversarial attack scenarios by leveraging lexical and syntactic information present within the dataset. Such analysis might provide insights into the models' reliability and help identify areas for improvement.

The overarching objective of all these future endeavors is to advance the utilization of MRC techniques within the biomedical domain, ultimately leading to improved efficiency in healthcare practices.

REFERENCES

[1] G. Del Fiol, T. E. Workman, and P. N. Gorman, "Clinical questions raised by clinicians at the point of care: A systematic review," *JAMA internal medicine*, vol. 174, no. 5, pp. 710–718, 2014.

[2] H. Bastian, P. Glasziou, and I. Chalmers, "Seventy-five trials and eleven systematic reviews a day: How will we ever keep up?" *PLoS medicine*, vol. 7, no. 9, e1000326, 2010.

[3] I. Masic, M. Miokovic, and B. Muhamedagic, "Evidence based medicine–new approaches and challenges," *Acta Informatica Medica*, vol. 16, no. 4, p. 219, 2008.

[4] D. L. Sackett, "Evidence-based medicine," in *Seminars in perinatology*, Elsevier, vol. 21, 1997, pp. 3–5.

[5] P. N. Hider, G. Griffin, M. Walker, and E. Coughlan, "The information-seeking behavior of clinical staff in a large health care organization," *Journal of the Medical Library Association: JMLA*, vol. 97, no. 1, p. 47, 2009.

[6] W. R. Hersh, M. K. Crabtree, D. H. Hickam, L. Sacherek, C. P. Friedman, P. Tidmarsh, C. Mosbaek, and D. Kraemer, "Factors associated with success in searching medline and applying evidence to answer clinical questions," *Journal of the American Medical Informatics Association*, vol. 9, no. 3, pp. 283–293, 2002.

[7] P. Jacquemart and P. Zweigenbaum, "Towards a medical question-answering system: A feasibility study," in *The New Navigators: From Professionals to Patients*, IOS Press, 2003, pp. 463–468.

[8] M. Lee, J. Cimino, H. R. Zhu, C. Sable, V. Shanker, J. Ely, and H. Yu, "Beyond information retrieval—medical question answering," in *AMIA annual symposium proceedings*, American Medical Informatics Association, vol. 2006, 2006, p. 469.

[9] T. Russell-Rose and J. Chamberlain, "Expert search strategies: The information retrieval practices of healthcare information professionals," *JMIR medical informatics*, vol. 5, no. 4, e7680, 2017.

[10] Q. Jin, Z. Yuan, G. Xiong, Q. Yu, H. Ying, C. Tan, M. Chen, S. Huang, X. Liu, and S. Yu, "Biomedical question answering: A survey of approaches and challenges," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–36, 2022.

[11] J. H. Barth, S. Misra, K. M. Aakre, M. R. Langlois, J. Watine, P. J. Twomey, and W. P. Oosterhuis, "Why are clinical practice guidelines not followed?" *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 54, no. 7, pp. 1133–1139, 2016.

[12] S. Gupta, N. Rai, O. Bhattacharrya, A. Y. Cheng, K. A. Connelly, L.-P. Boulet, A. Kaplan, M. C. Brouwers, and M. Kastner, "Optimizing the language and format of guidelines to improve guideline uptake," *CMAJ*, vol. 188, no. 14, E362–E368, 2016.

[13] D. Goodman-Meza, A. Tang, B. Aryanfar, S. Vazquez, A. J. Gordon, M. Goto, M. B. Goetz, S. Shoptaw, and A. A. Bui, "Natural language processing and machine learning to identify people who inject drugs in electronic health records," in *Open Forum Infectious Diseases*, Oxford University Press US, vol. 9, 2022, ofac471.

[14] M. Mahbub, S. Srinivasan, I. Danciu, A. Peluso, E. Begoli, S. Tamang, and G. D. Peterson, "Unstructured clinical notes within the 24 hours since admission predict short, mid & long-term mortality in adult icu patients," *Plos one*, vol. 17, no. 1, e0262182, 2022.

[15] B. G. Patra, M. M. Sharma, V. Vekaria, P. Adekkanattu, O. V. Patterson, B. Glicksberg, L. A. Lepow, E. Ryu, J. M. Biernacka, A. Furmanchuk, *et al.*, "Extracting social determinants of health from electronic health records using natural language processing: A systematic review," *Journal of the American Medical Informatics Association*, vol. 28, no. 12, pp. 2716–2727, 2021.

[16] D. J. Feller, O. J. B. Don't Walk, J. Zucker, M. T. Yin, P. Gordon, N. Elhadad, *et al.*, "Detecting social and behavioral determinants of health with structured and free-text clinical data," *Applied clinical informatics*, vol. 11, no. 01, pp. 172–181, 2020.

[17] K. Lybarger, N. J. Dobbins, R. Long, A. Singh, P. Wedgeworth, O. Ozuner, and M. Yetisgen, "Leveraging natural language processing to augment structured social determinants of health data in the electronic health record," *arXiv preprint arXiv:2212.07538*, 2022.

[18] D. J. Feller, J. Zucker, M. T. Yin, P. Gordon, and N. Elhadad, "Using clinical notes and natural language processing for automated hiv risk assessment," *Journal of acquired immune deficiency syndromes (1999)*, vol. 77, no. 2, p. 160, 2018.

[19] Y. Wang, E. S. Chen, S. Pakhomov, E. Arsoniadis, E. W. Carter, E. Lindemann, I. N. Sarkar, and G. B. Melton, "Automated extraction of substance use information from clinical texts," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2015, 2015, p. 2121.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, 2019, pp. 4171–4186.

[21] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. DOI: 10.18653/v1/D16-1264. [Online]. Available: https://aclanthology.org/D16-1264.

[22] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1601–1611. DOI: 10.18653/v1/P17-1147. [Online]. Available: https://aclanthology.org/P17-1147.

[23] G. Pergola, E. Kochkina, L. Gui, M. Liakata, and Y. He, "Boosting low-resource biomedical QA via entity-aware masking strategies," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds., Association for Computational Linguistics, 2021, pp. 1977–1985. DOI: 10.18653/v1/2021.

eacl-main.169. [Online]. Available: https://doi.org/10.18653/v1/2021.eacl-main.169.

[24] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[25] D. Golub, P.-S. Huang, X. He, and L. Deng, "Two-stage synthesis networks for transfer learning in machine comprehension," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 835–844. DOI: 10.18653/v1/D17-1087. [Online]. Available: https://aclanthology.org/D17-1087.

[26] H. Wang, Z. Gan, X. Liu, J. Liu, J. Gao, and H. Wang, "Adversarial domain adaptation for machine reading comprehension," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2510–2520. DOI: 10.18653/v1/D19-1254. [Online]. Available: https://aclanthology.org/D19-1254.

[27] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, *et al.*, "An overview of the bioasq large-scale biomedical semantic indexing and question answering competition," *BMC bioinformatics*, vol. 16, no. 1, pp. 1–28, 2015.

[28] C. Zeng, S. Li, Q. Li, J. Hu, and J. Hu, "A survey on machine reading comprehension-tasks, evaluation metrics and benchmark datasets," *Applied Sciences*, vol. 10, no. 21, p. 7640, 2020.

[29] J. Welbl, P. Stenetorp, and S. Riedel, "Constructing datasets for multi-hop reading comprehension across documents," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 287–302, 2018.

[30] D. Hewlett, A. Lacoste, L. Jones, I. Polosukhin, A. Fandrianto, J. Han, M. Kelcey, and D. Berthelot, "Wikireading: A novel large-scale language understanding task over wikipedia," *arXiv preprint arXiv:1608.03542*, 2016.

[31] W. Lehnert, "The process of question answering. research report no. 88.," 1977.

[32] R. Baradaran, R. Ghiasi, and H. Amirkhani, "A survey on machine reading comprehension systems," *Natural Language Engineering*, pp. 1–50, 2020.

[33] L. Hirschman, M. Light, E. Breck, and J. D. Burger, "Deep read: A reading comprehension system," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, 1999, pp. 325–332.

[34] E. Riloff and M. Thelen, "A rule-based question answering system for reading comprehension tests," in *ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, 2000.

[35] H. Poon, J. Christensen, P. Domingos, O. Etzioni, R. Hoffmann, C. Kiddon, T. Lin, X. Ling, A. Ritter, S. Schoenmackers, *et al.*, "Machine reading at the university of washington," in *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, 2010, pp. 87–95.

[36] S. Liu, X. Zhang, S. Zhang, H. Wang, and W. Zhang, "Neural machine reading comprehension: Methods and trends," *Applied Sciences*, vol. 9, no. 18, p. 3698, 2019.

[37] M. Richardson, C. J. Burges, and E. Renshaw, "Mctest: A challenge dataset for the open-domain machine comprehension of text," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 193–203.

[38] J. Berant, V. Srikumar, P.-C. Chen, A. Vander Linden, B. Harding, B. Huang, P. Clark, and C. D. Manning, "Modeling biological processes for reading comprehension," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1499–1510.

[39] T. Hastie, R. Tibshirani, and J. Friedman, "Overview of supervised learning," in *The elements of statistical learning*, Springer, 2009, pp. 9–41.

[40] M. J. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=HJ0UKP9ge.

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2013. [Online]. Available: http://arxiv.org/abs/1301.3781.

[42] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[43] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

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

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

[46] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "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, Jun. 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. [Online]. Available: https://aclanthology.org/N18-1202.

[47] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[48] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "Ammu: A survey of transformer-based biomedical pretrained language models," *Journal of biomedical informatics*, p. 103 982, 2021.

[49] ——, "Ammus: A survey of transformer-based pretrained models in natural language processing," *arXiv preprint arXiv:2108.05542*, 2021.

[50] Z. Zhang, Y. Wu, J. Zhou, S. Duan, H. Zhao, and R. Wang, "Sg-net: Syntax-guided machine reading comprehension," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 9636–9643.

[51] S. Verberne, L. Boves, N. Oostdijk, and P. Coppen, "Using syntactic information for improving why-question answering," 2008.

[52] X. Li, "Syntactic features in question answering," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 383–384.

[53] G. Xu, W. Rong, Y. Wang, Y. Ouyang, and Z. Xiong, "External features enriched model for biomedical question answering," *BMC bioinformatics*, vol. 22, no. 1, pp. 1–19, 2021.

[54] S. Indurthi, S. Yu, S. Back, H. Cuayáhuitl, *et al.*, "Cut to the chase: A context zoom-in network for reading comprehension," Association for Computational Linguistics, 2018.

[55] X. Feng, Y. Jiang, X. Yang, M. Du, and X. Li, "Computer vision algorithms and hardware implementations: A survey," *Integration*, vol. 69, pp. 309–320, 2019.

[56] Y. Chen, "Convolutional neural network for sentence classification," M.S. thesis, University of Waterloo, 2015.

[57] A. Chaturvedi, O. Pandit, and U. Garain, "CNN for text-based multiple choice question answering," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 272–277. DOI: 10.18653/v1/P18-2044. [Online]. Available: https://aclanthology.org/P18-2044.

[58] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[59] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[60] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[61] S. Hochreiter, "Studies on dynamic neural networks," *Diploma, Technical University of Munich*, vol. 91, no. 1, 1991.

[62] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber, *et al.*, *Gradient flow in recurrent nets: The difficulty of learning long-term dependencies*, 2001.

[63] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[64] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[65] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to answer open-domain questions," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1870–1879. DOI: 10.18653/v1/P17-1171. [Online]. Available: https://aclanthology.org/P17-1171.

[66] E. Choi, D. Hewlett, J. Uszkoreit, I. Polosukhin, A. Lacoste, and J. Berant, "Coarse-to-fine question answering for long documents," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 209–220. DOI: 10.18653/v1/P17-1020. [Online]. Available: https://aclanthology.org/P17-1020.

[67] J. Li, B. Li, and X. Lv, "Machine reading comprehension based on the combination of bidaf model and word vectors," in *Proceedings of the 2nd International Conference on Computer Science and Application Engineering*, 2018, pp. 1–4.

[68] Z. Wang, H. Mi, W. Hamza, and R. Florian, "Multi-perspective context matching for machine comprehension," *ArXiv*, vol. abs/1612.04211, 2016.

[69] S. Kobayashi, R. Tian, N. Okazaki, and K. Inui, "Dynamic entity representation with max-pooling improves machine reading," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 850–855. DOI: 10.18653/v1/N16-1099. [Online]. Available: https://aclanthology.org/N16-1099.

[70] C. Clark and M. Gardner, "Simple and effective multi-paragraph reading comprehension," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 845–855. DOI: 10.18653/v1/P18-1078. [Online]. Available: https://aclanthology.org/P18-1078.

[71] B. Dhingra, H. Liu, Z. Yang, W. Cohen, and R. Salakhutdinov, "Gated-attention readers for text comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1832–1846. DOI: 10.18653/v1/P17-1168. [Online]. Available: https://aclanthology.org/P17-1168.

[72] D. Chen, J. Bolton, and C. D. Manning, "A thorough examination of the CNN/Daily Mail reading comprehension task," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 2358–2367. DOI: 10.18653/v1/P16-1223. [Online]. Available: https://aclanthology.org/P16-1223.

[73] R. Ghaeini, X. Z. Fern, H. Shahbazi, and P. Tadepalli, "Dependent gated reading for cloze-style question answering," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 3330–3345.

[74] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," *Advances in neural information processing systems*, vol. 28, 2015.

[75] A. Sordoni, P. Bachman, A. Trischler, and Y. Bengio, "Iterative alternating neural attention for machine reading," *arXiv preprint arXiv:1606.02245*, 2016.

[76] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu, "Attention-over-attention neural networks for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 593–602. DOI: 10.18653/v1/P17-1055. [Online]. Available: https://aclanthology.org/P17-1055.

[77] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.0473.

[78] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 10, pp. 4291–4308, 2020.

[79] R. Kadlec, M. Schmid, O. Bajgar, and J. Kleindienst, "Text understanding with the attention sum reader network," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 908–918. DOI: 10.18653/v1/P16-1086. [Online]. Available: https://aclanthology.org/P16-1086.

[80] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[81] R. Alec, N. Karthik, S. Tim, and S. Ilya, "Improving language understanding with unsupervised learning," *Technical report, OpenAI*, 2018.

[82] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[83] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[84] Y. Yang, M. C. S. Uy, and A. Huang, "Finbert: A pretrained language model for financial communications," *arXiv preprint arXiv:2006.08097*, 2020.

[85] Z. Feng, D. Guo, D. Tang, N. Duan, X. Feng, M. Gong, L. Shou, B. Qin, T. Liu, D. Jiang, *et al.*, "Codebert: A pre-trained model for programming and natural languages," *arXiv preprint arXiv:2002.08155*, 2020.

[86] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. DOI: 10.18653/v1/D19-1371. [Online]. Available: https://aclanthology.org/D19-1371.

[87] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[88] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[89] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[90] "Common crawl. common crawl.," 2019. [Online]. Available: http://http://commoncrawl.org.

[91] J. Callan, M. Hoy, C. Yoo, and L. Zhao, "Clueweb09 data set," 2009.

[92] R. Parker, D. Graff, J. Kong, K. Chen, and K. Maeda, "English gigaword fifth edition, linguistic data consortium," *Technical report, Technical Report. Linguistic Data Consortium, Philadelphia, Tech. Rep.*, 2011.

[93] T. H. Trinh and Q. V. Le, "A simple method for commonsense reasoning," *arXiv preprint arXiv:1806.02847*, 2018.

[94] A. Gokaslan and V. Cohen, "Openwebtext corpus," 2019. [Online]. Available: http://web.archive.org/save/http://Skylion007.github.io/OpenWebTextCorpus.

[95] S. Nagel, "Cc-news," 2016. [Online]. Available: http://web.archive.org/save/http://commoncrawl.org/2016/10/newsdataset-available.

[96] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *ICLR*, 2020. [Online]. Available: https://openreview.net/pdf?id=r1xMH1BtvB.

[97] F. N. Iandola, A. E. Shaw, R. Krishna, and K. W. Keutzer, "Squeezebert: What can computer vision teach nlp about efficient neural networks?" *arXiv preprint arXiv:2006.11316*, 2020.

[98] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "Mobilebert: A compact task-agnostic bert for resource-limited devices," *arXiv preprint arXiv:2004.02984*, 2020.

[99] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[100] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," *arXiv preprint arXiv:2006.03654*, 2020.

[101] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*, PMLR, 2020, pp. 11 328–11 339.

[102] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[103] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.

[104] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, *et al.*, "Glm-130b: An open bilingual pre-trained model," *arXiv preprint arXiv:2210.02414*, 2022.

[105] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.

[106] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[107] OpenAI, *Gpt-4 technical report*, 2023. arXiv: 2303.08774 [cs.CL].

[108] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[109] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[110] S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, *et al.*, "Pythia: A suite for analyzing large language models across training and scaling," *arXiv preprint arXiv:2304.01373*, 2023.

[111] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[112] E. Lehman, E. Hernandez, D. Mahajan, J. Wulff, M. J. Smith, Z. Ziegler, D. Nadler, P. Szolovits, A. Johnson, and E. Alsentzer, "Do we still need clinical language models?" *arXiv preprint arXiv:2302.08091*, 2023.

[113] O. Rubin, J. Herzig, and J. Berant, "Learning to retrieve prompts for in-context learning," *arXiv preprint arXiv:2112.08633*, 2021.

[114] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag, "Large language models are zero-shot clinical information extractors," *arXiv preprint arXiv:2205.12689*, 2022.

[115] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, *et al.*, "Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models," *PLoS digital health*, vol. 2, no. 2, e0000198, 2023.

[116] H. Li, D. Guo, W. Fan, M. Xu, and Y. Song, "Multi-step jailbreaking privacy attacks on chatgpt," *arXiv preprint arXiv:2304.05197*, 2023.

[117] M. Sallam, "Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns," in *Healthcare*, MDPI, vol. 11, 2023, p. 887.

[118] L. De Angelis, F. Baglivo, G. Arzilli, G. P. Privitera, P. Ferragina, A. E. Tozzi, and C. Rizzo, "Chatgpt and the rise of large language models: The new ai-driven infodemic threat in public health," *Frontiers in Public Health*, vol. 11, p. 1 166 120, 2023.

[119] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja, *et al.*, "'"so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy," *International Journal of Information Management*, vol. 71, p. 102 642, 2023.

[120] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Association for Computational Linguistics, 2019, pp. 72–78.

[121] S. Bashath, N. Perera, S. Tripathi, K. Manjang, M. Dehmer, and F. E. Streib, "A data-centric review of deep transfer learning with applications to text data," *Information Sciences*, vol. 585, pp. 498–528, 2022.

[122] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Legal-bert: The muppets straight out of law school," *arXiv preprint arXiv:2010.02559*, 2020.

[123] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020.

[124] S. Leivaditi, J. Rossi, and E. Kanoulas, "A benchmark for lease contract review," *arXiv preprint arXiv:2010.10386*, 2020.

[125] W. U. Ahmad, S. Chakraborty, B. Ray, and K.-W. Chang, "Unified pre-training for program understanding and generation," *arXiv preprint arXiv:2103.06333*, 2021.

[126] D. Guo, S. Ren, S. Lu, Z. Feng, D. Tang, S. Liu, L. Zhou, N. Duan, A. Svyatkovskiy, S. Fu, *et al.*, "Graphcodebert: Pre-training code representations with data flow," *arXiv preprint arXiv:2009.08366*, 2020.

[127] L. Phan, H. Tran, D. Le, H. Nguyen, J. Anibal, A. Peltekian, and Y. Ye, "Cotext: Multi-task learning with code-text transformer," *arXiv preprint arXiv:2105.08645*, 2021.

[128] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[129] K. Raj Kanakarajan, B. Kundumani, and M. Sankarasubbu, "Bioelectra: Pretrained biomedical text encoder using discriminators," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 143–154.

[130] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets," *arXiv preprint arXiv:1906.05474*, 2019.

[131] M. Yasunaga, J. Leskovec, and P. Liang, "Linkbert: Pretraining language models with document links," *arXiv preprint arXiv:2203.15827*, 2022.

[132] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, and S. Yu, "BioBART: Pretraining and evaluation of a biomedical generative language model," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 97–109. [Online]. Available: https://aclanthology.org/2022.bionlp-1.9.

[133] K. Verspoor, K. B. Cohen, and L. Hunter, "The textual characteristics of traditional and open access scientific journals are similar," *BMC bioinformatics*, vol. 10, no. 1, pp. 1–16, 2009.

[134] Q. Jin, B. Dhingra, W. W. Cohen, and X. Lu, "Probing biomedical embeddings from language models," *arXiv preprint arXiv:1904.02181*, 2019.

[135] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[136] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020.

[137] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman, "Newsqa: A machine comprehension dataset," *arXiv preprint arXiv:1611.09830*, 2016.

[138] E. Choi, H. He, M. Iyyer, M. Yatskar, W.-t. Yih, Y. Choi, P. Liang, and L. Zettlemoyer, "Quac: Question answering in context," *arXiv preprint arXiv:1808.07036*, 2018.

[139] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," *arXiv preprint arXiv:1809.09600*, 2018.

[140] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner, "Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs," *arXiv preprint arXiv:1903.00161*, 2019.

[141] M. Gupta, N. Kulkarni, R. Chanda, A. Rayasam, and Z. C. Lipton, "Amazonqa: A review-based question answering task," *arXiv preprint arXiv:1908.04364*, 2019.

[142] K. Lin, O. Tafjord, P. Clark, and M. Gardner, "Reasoning over paragraph effects in situations," *arXiv preprint arXiv:1908.05852*, 2019.

[143] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, *et al.*, "Ms marco: A human generated machine reading comprehension dataset," *arXiv preprint arXiv:1611.09268*, 2016.

[144] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, *et al.*, "Natural questions: A

benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, 2019.

[145] A. Garcia-Silva, C. Berrio, J. M. Gomez-Perez, J. A. Martıénez-Heras, A. Donati, and I. Roma, "Spaceqa: Answering questions about the design of space missions and space craft concepts," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 3306–3311.

[146] S. Reddy, D. Chen, and C. D. Manning, "Coqa: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019.

[147] T. Möller, A. Reina, R. Jayakumar, and M. Pietsch, "COVID-QA: A question answering dataset for COVID-19," in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online: Association for Computational Linguistics, Jul. 2020. [Online]. Available: https://aclanthology.org/2020.nlpcovid19-acl.18.

[148] R. Tang, R. Nogueira, E. Zhang, N. Gupta, P. Cam, K. Cho, and J. Lin, "Rapidly bootstrapping a question answering dataset for covid-19," *arXiv preprint arXiv:2004.11339*, 2020.

[149] S. Raza, B. Schwartz, and N. Ondrusek, "A question-answering system on covid-19 scientific literature," in *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, IEEE, 2022, pp. 1331–1336.

[150] I. Bojic, Q. C. Ong, M. Thakkar, E. Kamran, I. Y. Le Shua, J. R. E. Pang, J. Chen, V. Nayak, S. Joty, and J. Car, "Sleepqa: A health coaching dataset on sleep for extractive question answering," in *Machine Learning for Health*, ser. Proceedings of Machine Learning Research, vol. 193, PMLR, Nov. 2022, pp. 199–217. [Online]. Available: https://proceedings.mlr.press/v193/bojic22a.html.

[151] A. Pampari, P. Raghavan, J. Liang, and J. Peng, "EmrQA: A large corpus for question answering on electronic medical records," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 2357–2368. DOI: 10.18653/v1/D18-1258. [Online]. Available: https://aclanthology.org/D18-1258.

[152] X. Yue, B. J. Gutierrez, and H. Sun, "Clinical reading comprehension: A thorough analysis of the emrqa dataset," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds., Association for Computational Linguistics, 2020, pp. 4474–4486. DOI: 10.18653/v1/2020.acl-main.410. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.410.

[153] S. Šuster and W. Daelemans, "Clicr: A dataset of clinical case reports for machine reading comprehension," *arXiv preprint arXiv:1803.09720*, 2018.

[154] D. Pappas, I. Androutsopoulos, and H. Papageorgiou, "Bioread: A new dataset for biomedical reading comprehension," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[155] D. Pappas, P. Stavropoulos, I. Androutsopoulos, and R. McDonald, "Biomrc: A dataset for biomedical machine reading comprehension," in *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, 2020, pp. 140–149.

[156] J. Bardhan, A. Colas, K. Roberts, and D. Z. Wang, "Drugehrqa: A question answering dataset on structured and unstructured electronic health records for medicine related queries," *arXiv preprint arXiv:2205.01290*, 2022.

[157] S. Soni, M. Gudala, A. Pajouhi, and K. Roberts, "RadQA: A question answering dataset to improve comprehension of radiology reports," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, Jun. 2022, pp. 6250–6259. [Online]. Available: https://aclanthology.org/2022.lrec-1.672.

[158] M. Savery, A. B. Abacha, S. Gayen, and D. Demner-Fushman, "Question-driven summarization of answers to consumer health questions," *Scientific Data*, vol. 7, no. 1, pp. 1–9, 2020.

[159] A. B. Abacha, C. Shivade, and D. Demner-Fushman, "Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019, pp. 370–379.

[160] D. Demner-Fushman, Y. Mrabet, and A. Ben Abacha, "Consumer health information and question answering: Helping consumers find answers to

their health-related information needs," *Journal of the American Medical Informatics Association*, vol. 27, no. 2, pp. 194–201, 2020.

[161] M. Zhu, A. Ahuja, D.-C. Juan, W. Wei, and C. K. Reddy, "Question answering with long multiple-span answers," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 3840–3849.

[162] A. Ben Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–23, 2019.

[163] M. Zhu, A. Ahuja, W. Wei, and C. K. Reddy, "A hierarchical attention retrieval model for healthcare question answering," in *The World Wide Web Conference*, 2019, pp. 2472–2482.

[164] P. Wang, T. Shi, and C. K. Reddy, "Text-to-sql generation for question answering on electronic medical records," in *Proceedings of The Web Conference 2020*, 2020, pp. 350–361.

[165] D. Vilares and C. Gòmez-Rodrìguez, "Head-qa: A healthcare dataset for complex reasoning," *arXiv preprint arXiv:1906.04701*, 2019.

[166] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering," *arXiv preprint arXiv:1909.06146*, 2019.

[167] M. Jeong, M. Sung, G. Kim, D. Kim, W. Yoon, J. Yoo, and J. Kang, "Transferability of natural language inference to biomedical question answering," *Working notes of CLEF 2020 conference and labs of the evaluation forum*, 2020.

[168] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 759–766.

[169] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Self-taught clustering," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 200–207.

[170] Z. Wang, Y. Song, and C. Zhang, "Transferred dimensionality reduction," in *Joint European conference on machine learning and knowledge discovery in databases*, Springer, 2008, pp. 550–565.

[171] A. L'heureux, K. Grolinger, H. F. Elyamany, and M. A. Capretz, "Machine learning with big data: Challenges and approaches," *Ieee Access*, vol. 5, pp. 7776–7797, 2017.

[172] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *ICML*, 2011.

[173] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, PMLR, 2015, pp. 1180–1189.

[174] B. Tan, Y. Zhang, S. Pan, and Q. Yang, "Distant domain transfer learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[175] M. Van Der Wees, A. Bisazza, and C. Monz, "Dynamic data selection for neural machine translation," *arXiv preprint arXiv:1708.00712*, 2017.

[176] B. Plank and G. Van Noord, "Effective measures of domain similarity for parsing," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 1566–1576.

[177] C. S. Perone, R. Silveira, and T. S. Paula, "Evaluation of sentence embeddings in downstream and linguistic probing tasks," *arXiv preprint arXiv:1806.06259*, 2018.

[178] S. Ruder, P. Ghaffari, and J. G. Breslin, "Data selection strategies for multi-domain sentiment analysis," *arXiv preprint arXiv:1702.02426*, 2017.

[179] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[180] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[181] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," *Advances in neural information processing systems*, vol. 28, 2015.

[182] F. Zhuang, X. Cheng, P. Luo, S. J. Pan, and Q. He, "Supervised representation learning: Transfer learning with deep autoencoders," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[183] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," *arXiv preprint arXiv:1206.4683*, 2012.

[184] G. Zhou, Z. Xie, X. Huang, and T. He, "Bi-transferring deep neural networks for domain adaptation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 322–332.

[185] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[186] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[187] Z. Wang, Y. Qu, L. Chen, J. Shen, W. Zhang, S. Zhang, Y. Gao, G. Gu, K. Chen, and Y. Yu, "Label-aware double transfer learning for cross-specialty medical named entity recognition," *arXiv preprint arXiv:1804.09021*, 2018.

[188] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "Adaptive semi-supervised learning for cross-domain sentiment classification," *arXiv preprint arXiv:1809.00530*, 2018.

[189] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[190] M. Wulfmeier, A. Bewley, and I. Posner, "Addressing appearance change in outdoor robotics with adversarial domain adaptation," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2017, pp. 1551–1558.

[191] A. Rios, R. Kavuluru, and Z. Lu, "Generalizing biomedical relation classification with neural adversarial domain adaptation," *Bioinformatics*, vol. 34, no. 17, pp. 2973–2981, 2018.

[192] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, PMLR, 2017, pp. 214–223.

[193] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," *Advances in neural information processing systems*, vol. 29, 2016.

[194] Y.-B. Kim, K. Stratos, and D. Kim, "Adversarial adaptation of synthetic or stale data," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1297–1307.

[195] B. Thompson, J. Gwinnup, H. Khayrallah, K. Duh, and P. Koehn, "Overcoming catastrophic forgetting during domain adaptation of neural machine translation," 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)*, 2019, pp. 2062–2068.

[196] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[197] T.-T. Vu, D. Phung, and G. Haffari, "Effective unsupervised domain adaptation with adversarially trained language models," Jan. 2020, pp. 6163–6173. DOI: 10.18653/v1/2020.emnlp-main.497.

[198] X. Yue, X. F. Zhang, Z. Yao, S. M. Lin, and H. Sun, "Cliniqg4qa: Generating diverse questions for domain adaptation of clinical question answering," *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 580–587, 2021.

[199] X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," *arXiv preprint arXiv:1705.00106*, 2017.

[200] Q. Zhou, N. Yang, F. Wei, C. Tan, H. Bao, and M. Zhou, "Neural question generation from text: A preliminary study," in *National CCF Conference on Natural Language Processing and Chinese Computing*, Springer, 2017, pp. 662–671.

[201] X. Sun, J. Liu, Y. Lyu, W. He, Y. Ma, and S. Wang, "Answer-focused and position-aware neural question generation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3930–3939.

[202] Y. Zhao, X. Ni, Y. Ding, and Q. Ke, "Paragraph-level neural question generation with maxout pointer and gated self-attention networks," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3901–3910.

[203] P. Nema, A. K. Mohankumar, M. M. Khapra, B. V. Srinivasan, and B. Ravindran, "Let's ask again: Refine network for automatic question generation," *arXiv preprint arXiv:1909.05355*, 2019.

[204] L. A. Tuan, D. Shah, and R. Barzilay, "Capturing greater context for question generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 9065–9072.

[205] Z. Yang, J. Hu, R. Salakhutdinov, and W. W. Cohen, "Semi-supervised qa with generative domain-adaptive nets," *arXiv preprint arXiv:1702.02206*, 2017.

[206] X. Du and C. Cardie, "Harvesting paragraph-level question-answer pairs from wikipedia," *arXiv preprint arXiv:1805.05942*, 2018.

[207] C. Alberti, D. Andor, E. Pitler, J. Devlin, and M. Collins, "Synthetic qa corpora generation with roundtrip consistency," *arXiv preprint arXiv:1906.05416*, 2019.

[208] S. Zhang and M. Bansal, "Addressing semantic drift in question generation for semi-supervised question answering," *arXiv preprint arXiv:1909.06356*, 2019.

[209] J. Kang, H. P. S. Roman, and S.-H. Myaeng, "Let me know what to ask: Interrogative-word-aware question generation," *arXiv preprint arXiv:1910.13794*, 2019.

[210] J. Cho, M. Seo, and H. Hajishirzi, "Mixture content selection for diverse sequence generation," *arXiv preprint arXiv:1909.01953*, 2019.

[211] B. Liu, H. Wei, D. Niu, H. Chen, and Y. He, "Asking questions the human way: Scalable question-answer generation from text corpus," in *Proceedings of The Web Conference 2020*, 2020, pp. 2032–2043.

[212] B. Ghanem, L. L. Coleman, J. R. Dexter, S. M. von der Ohe, and A. Fyshe, "Question generation for reading comprehension assessment by modeling how and what to ask," *arXiv preprint arXiv:2204.02908*, 2022.

[213] E. Lehman, V. Lialin, K. Y. Legaspi, A. J. R. Sy, P. T. S. Pile, N. R. I. Alberto, R. R. R. Ragasa, C. V. M. Puyat, I. R. I. Alberto, P. G. I. Alfonso, *et al.*, "Learning to ask like a physician," *arXiv preprint arXiv:2206.02696*, 2022.

[214] H. Assem, R. Sarkar, and S. Dutta, "Qasar: Self-supervised learning framework for extractive question answering," in *2021 IEEE International Conference on Big Data (Big Data)*, IEEE, 2021, pp. 1797–1808.

[215] S. Shen, Y. Li, N. Du, X. Wu, Y. Xie, S. Ge, T. Yang, K. Wang, X. Liang, and W. Fan, "On the generation of medical question-answer pairs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 8822–8829.

[216] S. Soni and K. Roberts, "A paraphrase generation system for ehr question answering," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019, pp. 20–29.

[217] ——, "Paraphrasing to improve the performance of electronic health records question answering," *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 626, 2020.

[218] S. Rennie, E. Marcheret, N. Mallinar, D. Nahamoo, and V. Goel, "Unsupervised adaptation of question answering systems via generative self-training," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1148–1157.

[219] A. Maharana and M. Bansal, "Adversarial augmentation policy search for domain and cross-lingual generalization in reading comprehension," *arXiv preprint arXiv:2004.06076*, 2020.

[220] H.-G. Lee, Y. Jang, and H. Kim, "Machine reading comprehension framework based on self-training for domain adaptation," *IEEE Access*, vol. 9, pp. 21 279–21 285, 2021.

[221] Z. Zhang, H. Zhao, and R. Wang, "Machine reading comprehension: The role of contextualized language models and beyond," *arXiv preprint arXiv:2005.06249*, 2020.

[222] G. Wiese, D. Weissenborn, and M. Neves, "Neural domain adaptation for biomedical question answering," *arXiv preprint arXiv:1706.03610*, 2017.

[223] D. Pappas, P. Malakasiotis, and I. Androutsopoulos, "Data augmentation for biomedical factoid question answering," *arXiv preprint arXiv:2204.04711*, 2022.

[224] T. Mairittha, N. Mairittha, and S. Inoue, "Improving fine-tuned question answering models for electronic health records," in *Adjunct Proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM International Symposium on Wearable Computers*, 2020, pp. 688–691.

[225] W. Yoon, R. Jackson, A. Lagerberg, and J. Kang, "Sequence tagging for biomedical extractive question answering," *Bioinformatics*, vol. 38, no. 15, pp. 3794–3801, 2022.

[226] J. Bai, C. Yin, J. Zhang, Y. Wang, Y. Dong, W. Rong, and Z. Xiong, "Adversarial knowledge distillation based biomedical factoid question answering," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.

[227] S. Kamath, B. Grau, and Y. Ma, "An adaption of bioasq question answering dataset for machine reading systems by manual annotations of answer spans.," in *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, 2018, pp. 72–78.

[228] Y. Xu, X. Zhong, A. J. J. Yepes, and J. H. Lau, "Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–8.

[229] A. Kumar, R. Bhargava, and M. Jayabalan, "Covid qa network: A specific case of biomedical question answering," in *2023 15th International Conference on Developments in eSystems Engineering (DeSE)*, IEEE, 2023, pp. 333–338.

[230] N. Rachmawati and E. Yulianti, "Transfer learning for closed domain question answering in covid-19," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 12, 2022.

[231] Z. Hu, P. Yang, B. Li, Y. Sun, and B. Yang, "Biomedical extractive question answering based on dynamic routing and answer voting," *Information Processing & Management*, vol. 60, no. 4, p. 103 367, 2023.

[232] Y. Du, W. Guo, and Y. Zhao, "Hierarchical question-aware context learning with augmented data for biomedical question answering," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2019, pp. 370–375.

[233] Y. Du, B. Pei, X. Zhao, and J. Ji, "Hierarchical multi-layer transfer learning model for biomedical question answering," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2018, pp. 362–367.

[234] Z. Kaddari and T. Bouchentouf, "A novel self-attention enriching mechanism for biomedical question answering," *Expert Systems with Applications*, vol. 225, p. 120 210, 2023.

[235] S. Soni and K. Roberts, "Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 5532–5538.

[236] M. M. Campos and F. M. Couto, "Post-processing biobert and using voting methods for biomedical question answering.," in *CLEF (Working Notes)*, 2021, pp. 258–273.

[237] S. Alrowili and V. Shanker, "Large biomedical question answering models with albert and electra.," in *CLEF (Working Notes)*, 2021, pp. 213–220.

[238] K. Nishida, K. Nishida, I. Saito, H. Asano, and J. Tomita, "Unsupervised domain adaptation of language models for reading comprehension," *arXiv preprint arXiv:1911.10768*, 2019.

[239] V. Kommaraju, K. Gunasekaran, K. Li, T. Bansal, A. McCallum, I. Williams, and A.-M. Istrate, "Unsupervised pre-training for biomedical question answering," in *CLEF (Working Notes)*, 2020. [Online]. Available: http://ceur-ws.org/Vol-2696/paper_144.pdf.

[240] Y. Cao, M. Fang, B. Yu, and J. T. Zhou, "Unsupervised domain adaptation on reading comprehension," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 7480–7487.

[241] Z. Bai, B. Wang, Z. Wang, C. Yuan, and X. Wang, "Domain adaptive multi-task transformer for low-resource machine reading comprehension," *Neurocomputing*, vol. 509, pp. 46–55, 2022.

[242] Z. Wu and H. Xu, "Trustworthy machine reading comprehension with conditional adversarial calibration," *Applied Intelligence*, pp. 1–18, 2022.

[243] H. Xu, B. Liu, L. Shu, and P. Yu, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," 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)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2324–2335. DOI: 10.18653/v1/N19-1242. [Online]. Available: https://aclanthology.org/N19-1242.

[244] R. Baradaran and H. Amirkhani, "Zero-shot estimation of base models' weights in ensemble of machine reading comprehension systems for robust generalization," in *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, IEEE, 2021, pp. 1–5.

[245] B. Dhingra, D. Pruthi, and D. Rajagopal, "Simple and effective semi-supervised question answering," *arXiv preprint arXiv:1804.00720*, 2018.

[246] C. Clark and M. Gardner, "Simple and effective multi-paragraph reading comprehension," *arXiv preprint arXiv:1710.10723*, 2017.

[247] A. Saha, R. Aralikatte, M. M. Khapra, and K. Sankaranarayanan, "Duorc: Towards complex language understanding with paraphrased reading comprehension," *arXiv preprint arXiv:1804.07927*, 2018.

[248] Y. Xu, X. Liu, Y. Shen, J. Liu, and J. Gao, "Multi-task learning with sample re-weighting for machine reading comprehension," *arXiv preprint arXiv:1809.06963*, 2018.

[249] M. Mahbub, S. Srinivasan, E. Begoli, and G. D. Peterson, "BioADAPT-MRC: Adversarial Learning-based Domain Adaptation Improves Biomedical Machine Reading Comprehension Task," *Bioinformatics*, Jul. 2022, btac508, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac508. eprint: https://academic.oup.com/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btac508/45060117/btac508.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btac508.

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

[251] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," *Advances in neural information processing systems*, vol. 6, 1993.

[252] K. Q. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *Advances in neural information processing systems*, vol. 18, 2005.

[253] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *International conference on machine learning*, PMLR, 2017, pp. 2642–2651.

[254] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[255] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger, "Adversarial deep averaging networks for cross-lingual sentiment classification," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 557–570, 2018.

[256] W. Yoon, J. Lee, D. Kim, M. Jeong, and J. Kang, "Pre-trained language model for biomedical question answering," in *Machine Learning and Knowledge Discovery in Databases - International Workshops of ECML PKDD 2019, Proceedings*, P. Cellier and K. Driessens, Eds., Springer, 2020, pp. 727–740. DOI: 10.1007/978-3-030-43887-6_64.

[257] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[258] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[259] S. Hosein, D. Andor, and R. McDonald, "Measuring domain portability and error propagation in biomedical qa," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2019, pp. 686–694.

[260] S. K. Telukuntla, A. Kapri, and W. Zadrozny, "Uncc biomedical semantic question answering systems. bioasq: Task-7b, phase-b," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2019, pp. 695–710.

[261] F. Qi, M. Li, Y. Chen, Z. Zhang, Z. Liu, Y. Wang, and M. Sun, "Hidden killer: Invisible textual backdoor attacks with syntactic trigger," in *ACL/IJCNLP (1)*, 2021, pp. 443–453. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.37.

[262] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "Ms marco: A human generated machine reading comprehension dataset," in *CoCo@ NIPS*, 2016.

[263] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 403–412.

[264] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep metric learning with angular loss," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2593–2601.

[265] V. Nguyen, "Question answering in the biomedical domain," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 2019, pp. 54–63.

[266] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[267] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *kdd*, vol. 96, 1996, pp. 226–231.

[268] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, pp. 2825–2830, 2011.

[269] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17, Sydney, NSW, Australia: JMLR.org, 2017, pp. 3319–3328.

[270] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, *et al.*, "Captum: A unified and generic model interpretability library for pytorch," *arXiv preprint arXiv:2009.07896*, 2020.

[271] M. Honnibal and I. Montani, "Spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, vol. 7, no. 1, pp. 411–420, 2017.

[272] M. Mahbub, E. Begoli, S. Martins, A. Peluso, S. Tamang, and G. Peterson, "Cpgqa: A benchmark dataset for machine reading comprehension tasks on clinical practice guidelines and a case study using transfer learning," *IEEE Access*, 2023.

[273] M. J. Field, K. N. Lohr, *et al.*, "Clinical practice guidelines," *Directions for a new program*, p. 1990, 1990.

[274] S. Woolf, H. J. Schünemann, M. P. Eccles, J. M. Grimshaw, and P. Shekelle, "Developing clinical practice guidelines: Types of evidence and outcomes; values and economics, synthesis, grading, and presentation and deriving recommendations," *Implementation Science*, vol. 7, no. 1, pp. 1–12, 2012.

[275] J. M. Rosenberg, B. M. Bilka, S. M. Wilson, and C. Spevak, "Opioid therapy for chronic pain: Overview of the 2017 us department of veterans affairs and us department of defense clinical practice guideline," *Pain Medicine*, vol. 19, no. 5, pp. 928–941, 2018.

[276] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, L. Gasco, M. Krallinger, and G. Paliouras, "Overview of bioasq 2021: The ninth bioasq challenge on large-scale biomedical semantic indexing and question answering," in *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2021, pp. 239–263.

[277] R. Aharoni and Y. Goldberg, "Unsupervised domain clusters in pretrained language models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 7747–7763. DOI: 10 . 18653 / v1 / 2020 . acl - main . 692. [Online]. Available: https://aclanthology.org/2020.acl-main.692.

[278] J. Phang, T. Févry, and S. R. Bowman, "Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks," *arXiv preprint arXiv:1811.01088*, 2018.

[279] D. Yogatama, C. d. M. d'Autume, J. Connor, T. Kocisky, M. Chrzanowski, L. Kong, A. Lazaridou, W. Ling, L. Yu, C. Dyer, *et al.*, "Learning and evaluating general linguistic intelligence," *arXiv preprint arXiv:1901.11373*, 2019.

[280] F. Zhu, W. Lei, Y. Huang, C. Wang, S. Zhang, J. Lv, F. Feng, and T.-S. Chua, "Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance," *arXiv preprint arXiv:2105.07624*, 2021.

[281] M. Mahbub, I. Goethert, I. Danciu, K. Knight, S. Srinivasan, S. Tamang, K. Rozenberg-Ben-Dror, H. Solares, S. Martins, E. Begoli, and G. D. Peterson, *Question-answering system extracts information on injection drug use from clinical progress notes*, 2023. arXiv: 2305.08777 [cs.AI].

[282] N. Goel, L. B. Munshi, and B. Thyagarajan, "Intravenous drug abuse by patients inside the hospital: A cause for sustained bacteremia," *Case Reports in Infectious Diseases*, vol. 2016, 2016.

[283] C. P. O'Brien, "Drug addiction and drug abuse," *Goodman and Gilman's the pharmacological basis of therapeutics*, vol. 11, pp. 607–627, 2006.

[284] H. Bradley, E. W. Hall, A. Asher, N. W. Furukawa, C. M. Jones, J. Shealey, K. Buchacz, S. Handanagic, N. Crepaz, and E. S. Rosenberg, "Estimated number of people who inject drugs in the united states," *Clinical Infectious Diseases*, vol. 76, no. 1, pp. 96–102, 2023.

[285] E. W. Hall, E. S. Rosenberg, C. M. Jones, A. Asher, E. Valverde, and H. Bradley, "Estimated number of injection-involved drug overdose deaths, united states, 2000–2018," *Drug and alcohol dependence*, vol. 234, p. 109 428, 2022.

[286] C. Cornford and H. Close, *The physical health of people who inject drugs: Complexities, challenges, and continuity*, 2016.

[287] L. R. Marks, N. S. Nolan, S. Y. Liang, M. J. Durkin, and M. B. Weimer, "Infectious complications of injection drug use," *Medical Clinics*, vol. 106, no. 1, pp. 187–200, 2022.

[288] D. Powell, A. Alpert, and R. L. Pacula, "A transitioning epidemic: How the opioid crisis is driving the rise in hepatitis c," *Health Affairs*, vol. 38, no. 2, pp. 287–294, 2019.

[289] S. A. Strathdee, I. Kuo, N. El-Bassel, S. Hodder, L. R. Smith, and S. A. Springer, "Preventing hiv outbreaks among people who inject drugs in the united states: Plus ça change, plus ça même chose," *AIDS (London, England)*, vol. 34, no. 14, p. 1997, 2020.

[290] A. G. Wurcel, J. E. Anderson, K. K. Chui, S. Skinner, T. A. Knox, D. R. Snydman, and T. J. Stopka, "Increasing infectious endocarditis admissions among young people who inject drugs," in *Open forum infectious diseases*, Oxford University Press, vol. 3, 2016.

[291] M. Sredl, A. T. Fleischauer, Z. Moore, D. L. Rosen, and A. J. Schranz, "Not just endocarditis: Hospitalizations for selected invasive infections among persons with opioid and stimulant use diagnoses-north carolina, 2010–2018," *The Journal of Infectious Diseases*, vol. 222, no. Supplement_5, S458–S464, 2020.

[292] I. See, R. H. Gokhale, A. Geller, M. Lovegrove, A. Schranz, A. Fleischauer, N. McCarthy, J. Baggs, and A. Fiore, "National public health burden estimates of endocarditis and skin and soft-tissue infections related to injection drug use: A review," *The Journal of Infectious Diseases*, vol. 222, no. Supplement_5, S429–S436, 2020.

[293] A. E. Edwards and C. B. Collins Jr, "Exploring the influence of social determinants on hiv risk behaviors and the potential application of structural interventions to prevent hiv in women," *Journal of health disparities research and practice*, vol. 7, no. SI2, p. 141, 2014.

[294] A. E. Nijhawan, L. R. Metsch, S. Zhang, D. J. Feaster, L. Gooden, M. K. Jain, R. Walker, S. Huffaker, M. J. Mugavero, P. Jacobs, *et al.*, "Clinical and sociobehavioral prediction model of 30-day hospital readmissions among people with hiv and substance use disorder: Beyond electronic health record data," *Journal of acquired immune deficiency syndromes (1999)*, vol. 80, no. 3, p. 330, 2019.

[295] M. Chen, X. Tan, and R. Padman, "Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review," *Journal of the American Medical Informatics Association*, vol. 27, no. 11, pp. 1764–1773, 2020.

[296] L. M. Gottlieb, K. J. Tirozzi, R. Manchanda, A. R. Burns, and M. T. Sandel, "Moving electronic medical records upstream: Incorporating social determinants of health," *American Journal of Preventive Medicine*, vol. 48, no. 2, pp. 215–218, 2015.

[297] C. R. Weir, N. Staggers, B. Gibson, K. Doing-Harris, R. Barrus, and R. Dunlea, "A qualitative evaluation of the crucial attributes of contextual information necessary in ehr design to support patient-centered medical home care," *BMC medical informatics and decision making*, vol. 15, no. 1, pp. 1–8, 2015.

[298] C. J. Hayes, M. A. Cucciare, B. C. Martin, T. J. Hudson, K. Bush, W. Lo-Ciganic, H. Yu, E. Charron, and A. J. Gordon, *Using data science to improve outcomes for persons with opioid use disorder*, 2022.

[299] M. Topaz, L. Murga, O. Bar-Bachar, K. Cato, and S. Collins, "Extracting alcohol and substance abuse status from clinical notes: The added value of nursing data," in *MEDINFO 2019: Health and Wellbeing e-Networks for All*, IOS Press, 2019, pp. 1056–1060.

[300] C. Peng, X. Yang, Z. Yu, J. Bian, W. R. Hogan, and Y. Wu, "Clinical concept and relation extraction using prompt-based machine reading comprehension," *arXiv preprint arXiv:2303.08262*, 2023.

[301] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2020.

[302] M. Eberts and A. Ulges, "Span-based joint entity and relation extraction with transformer pre-training," *arXiv preprint arXiv:1909.07755*, 2019.

[303] J. P. Ridgway, A. Uvin, J. Schmitt, T. Oliwa, E. Almirol, S. Devlin, and J. Schneider, "Natural language processing of clinical notes to identify mental illness and substance use among people living with hiv: Retrospective cohort study," *JMIR Medical Informatics*, vol. 9, no. 3, e23456, 2021.

[304] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, 2011.

[305] M. Torii, I. M. Finn, S. Doan, P. Wang, E. W. Yang, and D. S. Zisook, "Task formulation for extracting social determinants of health from clinical narratives," *arXiv preprint arXiv:2301.11386*, 2023.

[306] K. Lybarger, M. Ostendorf, and M. Yetisgen, "Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction," *Journal of Biomedical Informatics*, vol. 113, p. 103 631, 2021.

[307] S. Han, R. F. Zhang, L. Shi, R. Richie, H. Liu, A. Tseng, W. Quan, N. Ryan, D. Brent, and F. R. Tsui, "Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing," *Journal of Biomedical Informatics*, vol. 127, p. 103 984, 2022.

[308] Z. Yu, X. Yang, Y. Guo, J. Bian, and Y. Wu, "Assessing the documentation of social determinants of health for lung cancer patients in clinical narratives," *Frontiers in Public Health*, vol. 10, 2022.

[309] D. J. Feller, J. Zucker, *et al.*, "Towards the inference of social and behavioral determinants of sexual health: Development of a gold-standard corpus with semi-supervised learning," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2018, 2018, p. 422.

[310] H. Ahsan, E. Ohnuki, A. Mitra, and H. You, "Mimic-sbdh: A dataset for social and behavioral determinants of health," in *Machine Learning for Healthcare Conference*, PMLR, 2021, pp. 391–413.

[311] D. S. Carrell, D. Cronkite, R. E. Palmer, K. Saunders, D. E. Gross, E. T. Masters, T. R. Hylan, and M. Von Korff, "Using natural language processing to identify problem usage of prescription opioids," *International journal of medical informatics*, vol. 84, no. 12, pp. 1057–1064, 2015.

[312] M. Afshar, B. Sharma, S. Bhalla, H. M. Thompson, D. Dligach, R. A. Boley, E. Kishen, A. Simmons, K. Perticone, and N. S. Karnik, "External validation of an opioid misuse machine learning classifier in hospitalized adult patients," *Addiction Science & Clinical Practice*, vol. 16, no. 1, pp. 1–11, 2021.

[313] M. Afshar, B. Sharma, D. Dligach, M. Oguss, R. Brown, N. Chhabra, H. M. Thompson, T. Markossian, C. Joyce, M. M. Churpek, *et al.*, "Development and multimodal validation of a substance misuse algorithm for referral to treatment using artificial intelligence (smart-ai): A retrospective deep learning study," *The Lancet Digital Health*, vol. 4, no. 6, e426–e435, 2022.

[314] J. M. Lingeman, P. Wang, W. Becker, and H. Yu, "Detecting opioid-related aberrant behavior using natural language processing," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2017, 2017, p. 1179.

[315] S. V. Blackley, E. MacPhaul, B. Martin, W. Song, J. Suzuki, and L. Zhou, "Using natural language processing and machine learning to identify hospitalized patients with opioid use disorder," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2020, 2020, p. 233.

[316] V. J. Zhu, L. A. Lenert, K. S. Barth, K. N. Simpson, H. Li, M. Kopscik, and K. T. Brady, "Automatically identifying opioid use disorder in non-cancer patients on chronic opioid therapy," *Health Informatics Journal*, vol. 28, no. 2, p. 14 604 582 221 107 808, 2022.

[317] M. N. Poulsen, P. J. Freda, V. Troiani, A. Davoudi, and D. L. Mowery, "Classifying characteristics of opioid use disorder from hospital discharge summaries using natural language processing," *Frontiers in Public Health*, p. 1278, 2022.

[318] P. J. Ward, P. J. Rock, S. Slavova, A. M. Young, T. L. Bunn, and R. Kavuluru, "Enhancing timeliness of drug overdose mortality surveillance: A machine learning approach," *PloS one*, vol. 14, no. 10, e0223318, 2019.

[319] J. Badger, E. LaRose, J. Mayer, F. Bashiri, D. Page, and P. Peissig, "Machine learning for phenotyping opioid overdose events," *Journal of biomedical informatics*, vol. 94, p. 103 185, 2019.

[320] B. Hazlehurst, C. A. Green, N. A. Perrin, J. Brandes, D. S. Carrell, A. Baer, A. DeVeaugh-Geiss, and P. M. Coplan, "Using natural language processing of clinical text to enhance identification of opioid-related overdoses in electronic health records data," *Pharmacoepidemiology and drug safety*, vol. 28, no. 8, pp. 1143–1151, 2019.

[321] D. R. Harris, C. Eisinger, Y. Wang, and C. Delcher, "Challenges and barriers in applying natural language processing to medical examiner notes from fatal opioid poisoning cases," in *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, 2020, pp. 3727–3736.

[322] D. Goodman-Meza, C. L. Shover, J. A. Medina, A. B. Tang, S. Shoptaw, and A. A. Bui, "Development and validation of machine models using natural language processing to classify substances involved in overdose deaths," *JAMA Network Open*, vol. 5, no. 8, e2225593–e2225593, 2022.

[323] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries," *Journal of biomedical informatics*, vol. 34, no. 5, pp. 301–310, 2001.

[324] N. Perera, M. Dehmer, and F. Emmert-Streib, "Named entity recognition and relation detection for biomedical information extraction," *Frontiers in cell and developmental biology*, p. 673, 2020.

[325] J. J. Gong, H. Soleimani, S. G. Murray, and J. Adler-Milstein, "Characterizing styles of clinical note production and relationship to clinical work hours among first-year residents," *Journal of the American Medical Informatics Association*, vol. 29, no. 1, pp. 120–127, 2022.

VITA

Maria Mahbub was born and brought up in Dhaka, Bangladesh. She completed her education at the University of Dhaka, earning a Bachelor of Science in Mathematics in 2016 and a Master of Science in Applied Mathematics in 2017. In the pursuit of higher education, she joined the University of Tennessee, Knoxville, in 2018, where she obtained a Master of Science in Computer Science in 2022 and continued her Ph.D. She has been working at ORNL since 2019, where she started her research endeavors in biomedical natural language processing and robust machine learning in 2021. Since then, she has developed efficient machine learning models for predictive modeling and biomedical information extraction. Furthermore, she has dedicated her efforts to advancing the field of robust machine learning, collaborating on the development of adversarial attack algorithms designed to pinpoint vulnerabilities in machine learning models. Her primary research interests include machine reading comprehension, information extraction, predictive modeling, and robust machine learning. Her dedication to these areas has contributed to her accomplishments and ongoing contributions in the field.