UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA ANIMAL

Development of a Corpus for User-based Scientific Question Answering

Miguel Ângelo Conde Vieira

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:

Professor Doutor Francisco José Moreira Couto

Doutor André Francisco Martins Lamúrias

2021

Acknowledgements

I would like to thank Professor Doutor Francisco Couto and Doutor André Lamúrias not only for their guidance and expertise but, more importantly, because of their never-ending patience, support and permanent availability to tackle any issues or pitfalls that inevitably occur in this type of work. Also a big thank you to all the wonderful colleagues at FCUL with whom I had the opportunity to exchange ideas, share knowledge and the privilege to create lasting professional and personal bonds. To my family, in particular my parents and my godmother which were a constant supporting shoulder and a source of motivation for finishing this challenge.

Abstract

In recent years Question & Answering (QA) tasks became particularly relevant in the research field of natural language understanding. However, the lack of good quality datasets has been an important limiting factor in the quest for better models. Particularly in the biomedical domain, the scarcity of gold standard labelled datasets has been a recognized obstacle given its idiosyncrasies and complexities often require the participation of skilled domain-specific experts in producing such datasets.

To address this issue, a method for automatically gather Question-Answer pairs from online QA biomedical forums has been suggested yielding a *corpus* named BiQA. The authors describe several strategies to validate this new dataset but a human manual verification has not been conducted.

With this in mind, this dissertation was set out with the objectives of performing a manual verification of a sample of 1200 questions of BiQA and also to expanding these questions, by adding features, into a new *corpus* of text - BiQA2 - with the goal of contributing with a new *corpus* for biomedical QA research.

Regarding the manual verification of BiQA, a methodology for its characterization was laid out and allowed the identification of an array of potential problems related to the nature of its questions and answers aptness for which possible improvement solutions were presented.

Concomitantly, the proposed new BiQA2 *corpus* - created upon the validated questions and answers from the perused samples from BiQA - builds new features similar to those observed in other biomedical *corpus* such as the BioASQ dataset.

Both BiQA and BiQA2 were applied to deep learning strategies previously submitted to the BioASQ competition to assess their performance as a source of training data. Although the results achieved with the models created using BiQA2 exhibit limited capability pertaining to the BioASQ challenge, they also show some potential to contribute positively to model training in tasks such as Document re-ranking and answering to ‘yes/no’ questions.

Keywords: Biomedical Literature, Question-Answering, *corpus*, Deep Learning.

Resumo

Nos últimos anos, a investigação no domínio de tarefas de “Question & Answering” (QA) tem ocupado um lugar de particular destaque no campo da ciência da computação e compreensão de linguagem natural.

Estes sistemas são uma forma especializada ou avançada de tarefas de extração de informação que tentam elaborar respostas a questões colocadas por humanos na sua língua natural. Um sistema de QA é, em poucas palavras, um sistema computacional automatizado capaz de responder a questões humanas. Os progressos galopantes que se têm verificado nesta área devem-se não só à crescente disponibilidade de poder de computação, e concomitante decréscimo de custos, mas igualmente à emergência de modelos de aprendizagem profunda que têm vindo a superar as técnicas previamente aplicadas. No entanto, outro fator determinante na pesquisa e desenvolvimento em sistemas de QA é a disponibilidade de dados de treino de alta qualidade. A escassez de dados tem sido um dos maiores fatores de limitação no desenvolvimento de melhores modelos.

No caso particular das ciências biomédicas – com complexidades e desafios próprios - a falta de dados de alta qualidade para aprendizagem supervisionada, é reconhecidamente, um grande obstáculo dado que a compilação de corpos de texto desta natureza requer, muito frequentemente, a cooperação de peritos altamente qualificados no universo biomédico. A compilação manual de perguntas e respostas para treino de sistemas QA no domínio da literatura biomédica, além de morosa, apresenta desafios próprios. Uma frequente fonte de dados a que os investigadores recorrem são os fóruns *online* onde surgem múltiplas dificuldades na angariação quer de perguntas, quer das respostas. Por exemplo, é bastante comum utilizadores diferentes perguntarem a mesma questão. Ou seja, para uma mesma resposta podem surgir perguntas com formulações muito díspares, com léxico distinto. Outra dificuldade frequente é o desvio ou divagação de raciocínio aquando da elaboração de uma determinada resposta em que os utilizadores frequentemente se envolvem em discussões paralelas não relacionadas com a pergunta original. A elaboração de corpos de texto para sistemas biomédicos em QA são relativamente recentes. Por exemplo, só em 2007 os responsáveis pela competição TREC decidiram incluir um corpo de texto para extração de documentos relacionados com genómica. Por volta da mesma altura, a competição QA4MRE também resolveu incluir uma tarefa de QA com cerca de 40 amostras. Mais recentemente, alguns autores propuseram metodologias de angariação automática de questões e respostas de modo a compilar dados em grande escala. Exemplo disso são o corpo de texto emrQA relativo a bases de dados médicos e o corpo de texto BioRead. Em 2019 foi publicado o PubMedQA que é já um corpo de larga escala com a particularidade de conter uma quantidade substancial de amostras validadas manualmente.

Tendo em mente o número limitado de dados para treino, foi recentemente apresentado um método de compilação automática de pares questão-resposta que resultou num novo corpo de texto biomédico que os autores intitularam de BiQA. A sua abordagem consistiu em angariar perguntas e respetivas respostas a partir fóruns biomédicos *online*. O objetivo seria extrair perguntas reais vindas directamente dos utilizadores enquanto, ao mesmo tempo, se percorrem as respostas correspondentes, por outros colegas utilizadores do fórum, em busca de possíveis referências a artigos científicos que constem da base de dados PubMed. Estas referências são então traduzidas para o correspondente código identificador de modo a adquirir o seu resumo, se disponível ao público. Assume-se que o texto nestes resumos contenham respostas à respetiva pergunta. A premissa da metodologia de BiQA é que os sistemas de QA podem beneficiar de publicações colaborativas, com perguntas formuladas por utilizadores do mundo real, como uma abordagem automática de geração de corpos de texto biomédico. O *corpus* gerado é composto por pares sob a forma de pergunta-documento afixado no 'StackExchange' de Biologia e Ciências Médicas e Fórum de Nutrição do Reddit. Os autores de BiQA exploram diversas estratégias na tentativa de validar o seu corpo de texto, porém nenhuma inclui a verificação manual das questões, e da pertinência das respostas, por um anotador humano.

Como tal, um dos objetivos desta tese é a avaliação manual de uma amostra de pares questão-resposta de BiQA de modo a compreender melhor a natureza e as características particulares deste *corpus* biomédico e propor melhorias, se possível. Para isso 1200 questões foram analisadas e aferiu-se se contêm qualidade semântica suficiente para integrar num corpo de texto desta natureza. Os artigos identificados por BiQA como tendo relevância para estas perguntas são igualmente avaliados na sua capacidade de, efetivamente, providenciar respostas objetivas para as questões em causa. Adicionalmente – e tendo como ponto de partida as questões validadas na amostra de BiQA – esta tese propõe um novo *corpus* - BiQA2. Este corpo de texto tem como base somente as questões validadas e as suas respetivas respostas em BiQA, a partir das quais se constrói atributos semelhantes aos atributos encontrados no corpo de texto desenvolvido no âmbito da competição biomédica de QA intitulado BioASQ. Esta competição é uma iniciativa financiada pela União Europeia no domínio da pesquisa semântica biomédica com o objetivo de fornecer uma plataforma publuca de avaliação de indexação de textos e sistemas QA biomédicos.

Adicionalmente, e no sentido de validar BiQA2 como fonte de dados de treino para modelos capazes de competir em tarefas da competição BioASQ, esta dissertação implementa várias arquiteturas de aprendizagem profunda que já concorreram nesta competição em anos transatos, nomeadamente na execução das sub-tarefas de "Document re-ranking" e "Snippets retrieval". As arquiteturas utilizadas seguem uma pipeline proposta concorrente ao BioASQ6 baseada em redes convolucionais. Esta pipeline inclui duas arquiteturas distintas. A primeira destinada à sub-tarefa de Document re-ranking a partir da qual se re-ordenam, por ordem de importancia, uma série de N documentos cuja rede identifica como relevantes para responder a determinada pergunta. A segunda rede destina-se a extrair pedaços de texto relevantes a partir desses N documentos. Ambas as redes foram, originalmente, concebidas como redes convolucionais. No entanto, nesta tese, foi explorada a tarefa de extração de texto dos artigos através da implementação de modelos baseados numa arquitetura que tem tido ótimos resultados no campo de processamento de linguagem natural - BERT. A análise resultante do trabalho desenvolvido no que respeita o objetivo de validação manual de uma amostra de BiQA, mostra que uma quantidade significativa

de questões se revelaram de compreensão difícil ou cujo sentido ou intenção não são claros devido ao facto de serem formuladas de maneira ambígua, subjectiva ou demasiado complexa em que, por exemplo, várias questões são incluídas numa só. Relativamente aos artigos dados como potencial resposta às perguntas, verificou-se que menos de 50% realmente contêm informação pertinente, capaz de responder objetivamente às respetivas questões. As evidências apresentadas relativamente às características de BiQA permitiram a identificação de potenciais problemas ou obstáculos, tanto na natureza das perguntas como na relevância das respostas, que podem interferir na prestação de BiQA como corpo de texto para treino de modelos na tarefa de ”Document re-ranking”.

Relativamente ao objetivo de criar um novo corpo de texto a partir de BiQA, BiQA2 resultou num corpo de texto de dimensões ainda reduzidas visto que foi elaborado apenas numa amostra de BiQA. A geração de modelos de aprendizagem profunda utilizando arquiteturas concorrentes ao BioASQ mostrou capacidade limitada de ser usado como treino em tarefas da competição BioASQ nomeadamente na sub-tarefa de extração de pedaços de texto a partir dos resumos dos artigos. Por outro lado, os resultados obtidos nas tarefas de “Document re-ranking” e resposta a questões do tipo “yes/no”, permitem concluir que, apesar dessas limitações, BiQA2 contribui positivamente como fonte de dados complementar ao dataset do BioASQ nestas sub-tarefas.

Palavras Chave: Literatura Biomédica, Question & Answering, Corpo de texto, Aprendizagem Profunda.

Contents

Acronyms **1**

1 Introduction **3**

1.1 Motivation . . . 3

1.1.1 Datasets for Question-Answering . . . 5

1.1.2 Biomedical QA systems and Corpora . . . 6

1.2 Objectives . . . 8

1.3 Methodology . . . 9

1.4 Contributions . . . 10

2 Related Work **11**

2.1 Corpora/Datasets . . . 11

2.2 BioASQ and BiQA . . . 14

2.2.1 BioASQ challenge and corpus . . . 14

2.2.2 BiQA . . . 17

2.3 Question-Answering Systems . . . 19

2.4 Deep Learning architectures in QA systems . . . 24

2.4.1 Convolutional Neural Networks . . . 24

2.4.2 AUEB's CNN-based QA system for BioASQ . . . 27

2.4.3 Transformers/BERT . . . 29

3 BiQA Verification **33**

3.1 Verification methodology . . . 33

3.1.1 Queries verification . . . 33

3.1.2 Abstracts verification . . . 34

3.2 BiQA characterization . . . 35

4 BiQA2 **39**

4.1 Methodology . . . 39

4.2 Characterization and features . . . 43

4.3 Validation with Deep Learning models in BioASQ tasks . . . 45

4.3.1 Document Re-ranking with AUEB's system . . . 45

4.3.1.1 TERM-PACRR implementation and modelling details 45
4.3.1.2 TERM-PACRR results 50
4.3.2 Snippets retrieval with AUEB's system 52
4.3.2.1 TERM-PACRR + BCNN implementation and modelling details 52
4.3.2.2 TERM-PACRR + BCNN results 54
4.3.3 Snippets retrieval with TERM-PACRR + BERT 55
4.3.3.1 TERM-PACRR + BERT implementation and modelling details 55
4.3.3.2 TERM-PACRR + BERT results 57
4.3.4 'yes/no' BioASQ phase B 57

5 Discussion **61**
5.1 Queries-Abstracts verification in BiQA and challenges in BiQA2 61
5.2 BiQA2 Question types 63
5.3 Document re-ranking task 64
5.3.1 'Ideal' inputs 65
5.3.2 BiQA and BiQA2 combinations 65
5.4 Snippets 66
5.5 'yes/no' BioASQ Phase B 67

6 Conclusions and Future Work **69**
6.1 Conclusions 69
6.2 Future Work 70

References **73**

List of Figures

2.1 Datasets for QA 12
2.2 Task B dataset Format. 16
2.3 QA systems workflow general overview 19
2.4 Techniques, algorithms, frameworks and tools in QA systems. 21
2.5 Basic CNN structure. 25
2.6 CNN kernel. 25
2.7 Max pooling schema 26
2.8 AUEB’s complete system for Task B Phase B 28
2.9 TERM-PACRR 28
2.10 BCNN scoring snippets relative to a query 29
2.11 BERT architecture overview. 31
2.12 BERTs embeddings inputs. 31
2.13 BERT applications 32

3.1 Distribution of Abstracts relevance 36
3.2 Distribution of Abstracts relevance to respective queries per forum 37
3.3 Queries grammatical features 38
3.4 Queries grammatical features per forum. 38

4.1 Snippets statistics 44
4.2 BiQA2 question type per forum 44
4.3 % difference in mean MAP Scores 51

List of Tables

2.1 BiQA samples 18
2.2 BiQA doc re-rank results 18

3.1 Validation of BiQA QA sampled pairs. 35

4.1 Medical forum examples of uncertain type questions 42
4.2 Examples of BiQA2 types of questions. 42
4.3 BiQA2 Questions and Abstracts per forum. 43
4.4 yes/no/uncertain question type samples 45
4.5 Baseline experiments for TERM-PACRR training. ’pos samples’: number of positive samples yielded by each *corpus* according to the methodology in TERM-PACRR; ’queries in *corpus*’: number of total available queries in each *corpus*; ’queries in training’: number of queries from which the positive samples are computed. 48
4.6 Document re-rank model training datasets versions with ’ideal’ positive samples 49
4.7 Additional Document re-rank models 49
4.8 TERM-PACRR model for Document re-ranking MAP results per test batch 50
4.9 Number of samples in BCNN training 53
4.10 MAP scores for BCNN model 55
4.11 Snippets retrieval MAP scores for training with Bert_base_uncased 58
4.12 Snippets retrieval MAP scores for training with pubmedBert_base_uncased 58
4.13 F1 scores for ’yes/no’ exact answer models 59

5.1 Comparison between pairs of MAP results. 66

Acronyms

ANN Artificial Neural Network.
AUEB Athens University Economics and Business School.

BCNN Basic Convolutional Neural Network.
BERT Bidirectional Encoder Representations from Transformers.

CNN Convolutional Neural Network.
CQA Community Question and Answering.

DL Deep Learning.

FC Fully Connected.

IDF Inverse Document Frequency.
IR Information Retrieval.

LSTM Long Short-Term Memory.

MAP Mean Average Precision.
MESH Medical Subject Headings.
ML Machine Learning.
MLP Multi-Layer Perceptron.
MRC Machine Reading Comprehension.

NER Named Entity Recognition.
NLP Natural Language Processing.

PMID Pubmed Identification Number.
POS Part-of-Speech.

QA Question & Answering.

RDF Resource Description Framework.
RNN Recurrent Neural Network.

SVMs Support Vector Machines.

Chapter 1

Introduction

This chapter presents the motivation, objectives, general methodology, and contributions of this dissertation.

1.1 Motivation

Natural Language Processing (NLP) can be traced to the 1950's as a field of study focused on interactions between computers and humans in their natural languages such as machine translation, question-answering, information retrieval, summarization, information extraction, topic modelling and sentiment analysis. This kind of language comprehension task can be thought of as an intersection of various areas of knowledge of disciplines such as computational linguistics, information engineering, computer science, and artificial intelligence with the research goal to drive the emergence of systems capable of reasoning or inference over natural language expressions. QA systems have emerged as a layer in the computer and data science knowledge domain precisely to offer models capable of tackling such challenge. These systems are a specialized, advanced form of information retrieval tasks that attempt to output answers in response to queries posed by humans in their natural language. A QA system is, in a nutshell, an artificial system capable of answering human questions. They typically achieve this by using domain-specific structured databases or by perusing a collection of natural language documents. The advancements in efficient and powerful computational hardware, the availability of large quantities of structured and unstructured data, the advent of deep neural networks, all have contributed to the recent rapid pace in research and development of this field.

QA systems date back to the 1960s. One of the very first implementations of these techniques was designed by Simmons et al. [40]. Still using the old punch cards, his system attempted to answer questions of the type 'what do worms eat?'. Two other early examples were the BASEBALL system, which answered questions about a single season of American League baseball games, and LUNAR, that provided answers about the Apollo moon missions. They worked reasonably well, as long as the queries conformed to their narrow domain scope. These systems can be altered to perform different tasks depending on the source of answers such as retrieving information from documents, language learning, online examination systems, human and computer interaction, document management, document classification, document summarization, translation, speech processing [36]. Typically these systems are broadly divided into two groups: open-domain vs closed-domain [1]. An analogy can be drawn with the way humans perceive this task. When confronted with a question, a human might have the answer in their memory collected from experience and their learning through life - this would be a closed-domain setting.

Open systems are typically web based with no restrictions and more general-purposed, whereas closed domain systems are limited in scope, for example, medical, weather forecasting, financial, legal.

Unlike the early open-domain systems that relied on information retrieval to extract answers from unstructured content, modern QA systems build knowledge bases by extracting a rich ontology of entities and relationships from a combination of structured and unstructured content. They take advantage of the latest developments in machine learning, representing text with word embedding and character embedding, and using deep learning — specifically sequence learning methods like Long Short-Term Memory (LSTM) and, more recently, leveraging the potential of the new transformers architecture [7].

The demand for this type of systems increases every day as they are able to produce simple, precise and question-specific answers. This is becoming all the more important as it is combined with the explosion in the digitization of knowledge.

The QA task is, therefore, a benchmark for measuring reasoning and inference ability of systems. The challenges faced by these systems are cumbersome and were summarised by Kodra and Meçe [22] in:

- Lexical gap between questions – differences in language formulation of questions
- Lexical gap between questions and answers – questions and answers can be highly asymmetric in the information they contain
- Deviation from question – phenomenon of answer thread becoming irrelevant to the question. Answers can be given in the form of comments but sometimes users engage in accessory tangent lines of discussion deviating dramatically from the original topic.

Most benchmarks provide a fixed set of corpora and challenge researchers to innovate on the architecture and design of systems. This makes it possible to compare algorithms by running many models on the same dataset in a quest to find the ones that perform best. But the quality of datasets is also of paramount importance. It is now consensual within the data science community that an equally large amount of work is required towards supporting a more data-centric approach where the gathering of high quality datasets plays a major role and is, perhaps, even more determinant in the generation of QA models.

[1] Some authors include a third type – restricted-domain [41]

1.1.1 Datasets for Question-Answering

Historically there was a transition from pure linguistic approaches as well as pattern matching approaches to Machine Learning (ML) and Deep Learning (DL) techniques that have been yielding impressive results when combined with NLP. The growing complexity in algorithms and incredible increase of computational power have been unequivocally determinant in the evolution of QA systems but the success and application of these techniques were only possible due to concomitant advances in development and availability of large training data sets to drive and improve QA systems. This is of fundamental importance. A paramount example demonstrating the consequence of having readily available large free data as critical building blocks for QA is the famous Watson system IBM researchers built to defeat the top 'Jeopardy!' champions in 2011. Watson mined 200 million pages to create a knowledge base, including a full crawl of Wikipedia.

Being mindful that some natural language understanding tasks exist where sources for training are abundant (machine translation or speech recognition are fine examples of this) in QA tasks it becomes substantially much harder to produce corpora suitable for training. Kwiatkowski et al. [23] identify several key problems when elaborating this kind of datasets:

- methods and sources to obtain questions
- methods used to annotate and collect answers
- measurement and quality control of annotations

The reality has been that, for decades, QA research approaches have not been able to achieve its full potential due to the lack of good labeled data. It was not until 2015 that large data sets became readily available, namely, online. It was the emergence of crowd-sourcing and search engines that resulted in an explosion of large-scale (+100K questions) Machine Reading Comprehension (MRC) datasets such as the SQuAD [37], DeepMind CNN/DM [16], MS Marco [30], that allowed highly effective supervised QA systems to gain traction and development momentum. The typical MRC task could be formulated as a supervised learning problem. The goal of a typical MRC task is to learn a predictor f which takes a passage of text p and a corresponding question q as inputs and gives the answer a as output, which could be formulated as a = f (p, q) where it is also necessary that a majority of native speakers would agree that the question q does regard that text p, and the answer a is a correct one that does not contain information irrelevant to that question [26].

Most methodologies that produce such datasets ought to encompass a manual human annotation protocol if they thrive to achieve the status of a gold standard. Even automated methods of *corpus* [2] generation, are expected to provide some form of evaluation performed by human annotators. This is, by its

[2] A corpus of text is defined as a collection of linguistic data with the purpose of verifying a hypothesis about language. It is the equivalent of "dataset" in a general machine learning task but Corpus is the preferred term, as it already existed previous to the machine learning area to refer to a body (collection) of writings. Corpus (pl. corpora) comes from Latin and literally means "body".

nature, resource-consuming. The amount of time and costs inherent to the human supervision of these datasets contributes to the relatively low amount of verified samples found in QA *corpora*.

One such dataset, introduced, in 2019, by Kwiatkowski et al. [23] is the Natural Questions *corpus*. It is an attempt at providing large-scale end-to-end training data for QA and research in natural language understanding. The set results of manual annotation of answers extracted from the Wikipedia pages. Of relevance is the authors conclusion that despite providing a *corpus* with 'high-quality annotations of answers in documents', the idiosyncrasies of natural questions do not yield a comparable performance in their chosen metrics suggesting that significant advances in QA systems are required to tackle natural language understanding.

1.1.2 Biomedical QA systems and Corpora

In the QA biomedical realm, some datasets have been presented. In 2007 the TREC challenge incorporated a genomic *corpus* with a task to retrieve relevant documents out of 38 topic questions. In 2013, the QA4MRE [35] challenge included a QA task regarding Alzheimer's disease. In 2019, Jin et al. [19] presented PubMedQA, a biomedical dataset for answering 'yes/no/maybe' type of questions. This dataset was built with a mixture of manual and automatic labelling where contexts are generated to answer the questions and both are written by the same authors. Also, some automatically collected biomedical QA datasets have been introduced such as emrQA [32] for electronic medical records and BioRead [33].

Nevertheless, within the biomedical domain, one of the main gold standards is the *corpus* developed in the context of the BioASQ challenge. This initiative is an EU-funded action to promote research advances in biomedical semantic indexing and question answering. According to Tsatsaronis et al. [43] it does so by setting up clearly defined tasks and making available realistic, high-quality benchmark datasets and adopting existing evaluation measures. According to the challenge promoters, the methodology used in the making of the dataset samples, involves a team of experts from all sorts of biomedical fields and consists in the construction of questions after achieving a consensus regarding the information contained within abstracts from the PubMed API [3], pertaining a certain biomedical topic.

The BioASQ challenge is split into 2 main groups of tasks: In Task A systems are required to automatically assign Medical Subject Headings (MESH) [4] terms to biomedical articles, thus assisting the indexing of biomedical literature. Task B focuses on obtaining precise and comprehensible answers to biomedical questions. The systems that participate in Task B are given English questions written by biomedical experts that reflect real-life information needs. Task B is divided into two phases. For each question, the experts provided related documents, snippets, concepts and triples, in order to assess the systems that participated in phase A. Furthermore, the experts provided exact and ideal answers for the assessment of phase B. Although most of the current series of challenges in BioASQ have already achieved a high

[3] PubMed is a free resource supporting the search and retrieval of biomedical and life sciences literature with the aim of improving health–both globally and personally. The PubMed database contains more than 32 million citations and abstracts of biomedical literature.

[4] Controlled and hierarchically-organized vocabulary produced by the US National Library of Medicine.

profile status with increasing interest from top-end research teams, the methodologies used to produce current datasets brings to the discussion the question of representativeness of real-world inputs. In other words, one might wonder if these systems have real value if the questions present in their corpora do not bear resemblance to real end-user inputs or queries. The BioASQ *corpus* might, in fact, have this limitation. Given that experts create objective, non-ambiguous, semantically and syntactically correct questions, they may not contain the linguistic richness and variability of real-world human-user queries.

BiQA

Being aware of the difficulty in compiling large high quality datasets for QA tasks, Lamurias et al. [24] attempted to contribute with the developed of an automatic method of corpora generation by gathering queries and answers in the form of PubMed abstracts. Their approach consisted of scraping online Community Question and Answering (CQA) forums in order to extract real life questions from users whilst, at the same time, perusing the corresponding answers from other fellow forum users in search for PubMed articles references. These references are then translated into the corresponding PubMed identifier in order to extract its abstract text, if publicly available. All abstracts retrieved are assumed to contain an answer to the respective query. The premise of BiQA is that QA systems could benefit from the collaborative posting and answering questions from real world users as an approach to *corpus* generation. Their work is a new framework concept of automatically creating a dataset suitable for training of document retrieval systems in QA.

The *corpus* generated is comprised of pairs in the form of question-document posted on Stack Exchange Biology and MedicalSciences forums and Nutrition forum from Reddit.

As the authors acknowledge, this method is pioneer in the field, it allows easy application onto similar online communities and can easily evolve over time. Concurrently it can also be extended, adapted and enriched in order to complement other QA existing *corpora*.

In an attempt to validate their approach, the authors managed to train a deep learning model (one of the contest candidates to the BioASQ challenge) and obtained similar MAP scores to a model trained on the BioASQ *corpus* annotated by experts. They hypothesise this is due to the fact that the BiQA *corpus* not only has more questions but that they actually derive from multiple sources. Furthermore they conducted additional experiments were the BiQA *corpus* was added to the BioASQ training data. This lead to marginal better MAP scores hence concluding their *corpus* might be viable for training and highlighting the importance of more datasets for biomedical QA.

However, both the questions and answers in BiQA, have not been manually inspected by a human annotator in order confirm that they are actually suitable for QA training. Even BiQA authors are aware of the absence of formal review or accountability in these community forums so, evidently, personal biases are very likely abundant in the answers and not possible to verify automatically. Although one assumes that by having an inherent score system within the forum that might provide some form of quality filter for both questions and respective responses, the reality is that given the informal nature of these online communities, the formulation of questions might not be entirely suitable for its purpose and there is also no guarantee that the PubMed abstracts text mentioned in the answers are semantically related to the questions posed. It would be ideal to manually peruse BiQA in order to address these issues in an attempt to, somewhat, characterize the nature and lexicon inherent to the formulations of its questions as well as to establish if there is indeed a semantic relationship between the questions and its abstracts as answers. This characterization could lead to the building of new extra features of BiQA that would, potentially, have the added benefit of contributing to the improvement or refinement of BiQA itself.

1.2 Objectives

The primary objective of this dissertation is to explore the creation of a new, manually annotated, biomedical *corpus* from the existing BiQA dataset by Lamurias et al. [24]. In order to achieve this objective it makes sense to first construct an understanding of the nature, characteristics and intricacies of BiQA. Consequently, this thesis's goals can be seen as two main blocks:

Objective 1 - verification of BiQA *corpus* The starting point of this dissertation is to assess the concept of the BiQA *corpus* as a suitable methodology of automatic creation of a biomedical *corpus*. In reality BiQA does not generate questions or answers but rather gathers them from online community forums. In forums of this nature, it is likely that some questions might be too ambiguous or syntactically poor to a point that even a human reader could not understand its meaning or logic thus rendering a proper answer impossible to formulate. This could eventually impair the performance of models that use BiQA as training data. It it thus an objective of this thesis to attempt to characterize the questions in BiQA in regards to its lexicon, ambiguity and semantic meaning. Additionally it is paramount to also peruse its corresponding gathered abstracts in order to assess if they actually contain a relevant answer to the question. This is performed in a sample of BiQA which includes all three forums.

Objective 2 - creation and validation of a new dataset: BiQA2 In consonance, and concomitantly, with the work performance whilst completing objective 1, a new dataset is created from the perused sample of BiQA - here named BiQA2. The concept for BiQA2 is to use the validated queries and answers (as explained in Objective 1) from BiQA and build upon that information to create additional features, mimicking the features in the gold standard dataset of BioASQ. In BioASQ, a panel of experts gathers abstracts of articles regarding a certain biomedical topic and, from those, selects sentences from which to build a question, 'exact' and 'ideal' answers. Conversely, the starting point in this new dataset - named BiQA2 – are the already existent QA pairs from BiQA. The annotator's work was to search for answers in the abstracts and build the dataset having the same framework of features as in the BioASQ dataset. More concretely, if an abstract does, indeed, contain pieces of information with the ability to answer the query then they would contribute to the extraction of features such as 'snippets', 'type of question' and 'exact answer'. The process of building BiQA2 can be seen as an original exercise in biomedical QA *corpus* creation. The main purposes of BiQA2 are: to serve as an improvement of BiQA by manually curating and selecting/discarding semantically ambiguous or poorly constructed questions and also discarding irrelevant answers; to provide additional data that can be integrated in the BioASQ challenge *corpus*, expanding BiQA beyond the capability of the document re-ranking task.

To test the adequacy, suitability and capability of BiQA2 as a biomedical QA dataset similar to BioASQ, this dissertation delves into the implementation of systems competitors in previous editions of the BioASQ challenge, by using this new *corpus* as input training data of models competing in differentiated tasks and comparing it to the BiQA and BioASQ *corpus* as benchmarks.

1.3 Methodology

The achieve the objective of manual verification of questions and answers in a sample of BiQA - with the three forums equally represented - 400 questions from each forum are validated or excluded as per the annotator's viewpoint. Questions that are poorly formulated or not able to be understood in any way are excluded whilst the validated ones are classified according to their ambiguity, structure and objectivity. The verification of answers (abstracts) also investigates if they actually contain pieces of text capable of responding to its question. Whilst conducting this exercise of perusing the abstracts from the sampled questions, the new *corpus* - BiQA2 - is constructed by selecting the abstracts that contain appropriate answers, extracting those sentences, classify the type of question according to the BioASQ types and propose 'exact answers' to each question. These features are compatible with the biomedical BioASQ *corpus* and could, potentially, serve as additional training data for both Phase A and B of Task B at the BioASQ competition.

To assess the potential of BiQA2 as a biomedical *corpus* capable of contributing to the generation of models in biomedical QA systems, the work in this thesis, also implemented deep learning architectures previously submitted to the BioASQ competition for different tasks. The main system implemented is the same that Lamurias et al. [24] have utilized to validate BiQA. It is the system from the team at the department of informatics from the Athens University of Economics and Business (AUEB) submitted in BioASQ6 [29]. It is a system for both document re-ranking and snippets retrieval sub-tasks and it is based on two distinct Convolutional Neural Networks (CNNs) - one for each sub-task. Several experiments are conducted with different versions of BiQA, BiQA2 and BioASQ in order to assess their performance as training data. In addition, experiments were also conducted by implementing - a now standard - transformers based model (BERT, which is a benchmark for transfer learning ability and task-specific fine-tuning) for the snippets retrieval task. The metric used to assess these models is the Mean Average Precision (MAP) which is the metric of choice in the BioASQ challenge for the task B phase A sub-tasks. The evaluation of all experiments was performed on the 5 test batches of the BioASQ6 competition ,made available online by the AUEB's team. All scripts were written in Python and the DL frameworks utilized were Keras, Tensorflow 1 and Tensorflow 2.

1.4 Contributions

The work developed in the scope of this thesis enabled the assessment of the existing biomedical BiQA corpus [24] by manual inspection and consequent verification of both queries and answers. This is achieved by suggesting a simple methodology of annotations where both the questions and abstracts collected from online biomedical forums are perused and consequently validated or excluded as suitable for the purpose of training a document retrieval QA task.

The annotations performed and the resulting analysis of the queries and abstracts from applying this method to a sample of the highest voted questions in BiQA suggest that the approach taken by its authors' might is an appropriate source of diverse, lexicon-rich, biomedical questions. However, in regards to answers offered by BiQA in form of extracted abstracts from the forums, this thesis identifies a few problems related to the answers gathered from the forums and presents suggestions that could contribute to tackling such issues.

This thesis also explores a methodology of building a new biomedical *corpus* that is derived - or expanded - from the validated questions and answers in the analysis sample of BiQA. Effectively a new biomedical *corpus* was constructed – BiQA2 – having in mind the objective of being used on its own or in conjunction with the BioASQ *corpus* as it incorporates some of its features. The results of using BiQA2 as training data for models submitted to the BioASQ competition seem to indicate that BiQA2 requires some refining and improvement *corpus* but they also show the potential of BiQA2 to contribute to the training of models capable of tackling some tasks in the BioASQ competition such as Document re-ranking and answer to yes/no questions.

Chapter 2

Related Work

2.1 Corpora/Datasets

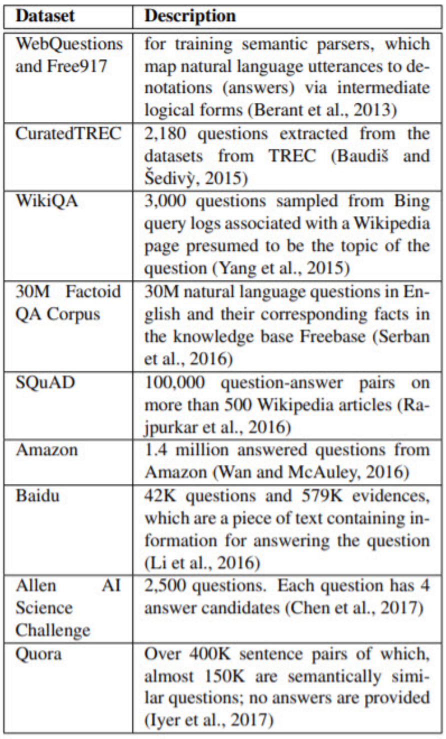
In recent years, numerous datasets have been released in the domain of QA systems to promote new methods that integrate natural language processing, information retrieval, artificial intelligence, and knowledge discovery. According to [9] the majority of these datasets are open domain. Figure 2.1 shows a few examples of such datasets.

The plethora of tasks and usages for which these datasets were built allow for many possible categorizations. There isn't a consensual or standard view on how to classify the various datasets given their particular scopes and idiosyncrasies. Some corpora have been presented as factoid questions only, for example. By contrast, datasets such as the PhotoshopQuiA focus on non-factoid questions with a particular interest in why-questions that are related to causal relationships [9]. Hashemi et al. [14] also understood the importance of non-factoid datasets and presented the ANTIQUE dataset which consists of a collection of open-domain questions and more than 34K manual relevant annotations. Non-factoid questions require complex answers such as descriptions or opinions.

Other authors offer different views on how to systematize this field. Gupta et al. [13], for example, have divided the currently available datasets into 4 main groups:

Open and closed world Question answering - Set of question-answer pairs together with a knowledge database. There is no explicit connection between the QA pair and the knowledge database. An example of this is the SimpleQA dataset that requires simple reasoning over the Freebase database.

Span-Based Answers - Where the answers are multi-word spans from the context. The SQuAD, SQuAD2.0 [37] and HotpotQA [46] are such examples. They contain questions and answers written by annotators who have first read a short text containing the answer. The SQuAD is a triple of question/paragraph/answer from Wikipedia. The SQuAD tasks have been used broadly in the development of all sorts of systems hence helping driving advances in reading comprehension. HotpotQA requires reasoning over multiple Wikipedia pages. This allows the development of systems that can handle longer



Dataset	Description
WebQuestions and Free917	for training semantic parsers, which map natural language utterances to denotations (answers) via intermediate logical forms (Berant et al., 2013)
CuratedTREC	2,180 questions extracted from the datasets from TREC (Baudiš and Šedivỳ, 2015)
WikiQA	3,000 questions sampled from Bing query logs associated with a Wikipedia page presumed to be the topic of the question (Yang et al., 2015)
30M Factoid QA Corpus	30M natural language questions in English and their corresponding facts in the knowledge base Freebase (Serban et al., 2016)
SQuAD	100,000 question-answer pairs on more than 500 Wikipedia articles (Rajpurkar et al., 2016)
Amazon	1.4 million answered questions from Amazon (Wan and McAuley, 2016)
Baidu	42K questions and 579K evidences, which are a piece of text containing information for answering the question (Li et al., 2016)
Allen AI Science Challenge	2,500 questions. Each question has 4 answer candidates (Chen et al., 2017)
Quora	Over 400K sentence pairs of which, almost 150K are semantically similar questions; no answers are provided (Iyer et al., 2017)

Figure 2.1: Open domain Datasets forQA. Obtained from [9].

contexts. SearchQA also presents contexts from more than one document. The questions are not guaranteed to require reasoning because the documents are gathered through IR after the QA pairs are determined.

Free-form Answers - Allow for more flexibility in abstract answers but are more difficult to evaluate on traditional metrics such as the BLEU score. WikiQA [45] and MS Marco [30] contain queries sampled from the BING search engine with human-generated answers. WikiQA has 3047 questions whereas MS Marco contains 100k questions with free-form answers. Another example is the DuReader Chinese language dataset [15] with real-world user queries. This system needs to read entire documents to find answers and it contains 200K questions, 420K answers and 1M documents. The questions and answers come from the Baidu Search and Baidu Zhidao engines.

Community/Opinion Question Answering - There exist a number of datasets that focus on data taken from CQA websites. These data sets contain many 'why?' or 'how?' type of questions. Some examples are Yahoo's Webscope L4 and L6, Qatar Living, and StackExchange.

In particular, for factoid QA, many datasets have been compiled and are characterized for having ample redundant evidence in the text: SQuAD [37]; NewsQA [42]; TriviaQA [20]; Quasar [8]). On the other hand, complex domain-specific MRC datasets such as MCTest [39], BioASQ [43], InsuranceQA [10], have been limited in scale (500-10K samples) due to the complexity of the task or the need for expert annotations that cannot be crowd-sourced or gathered from the web.

Kwiatkowski et al. [23], in particular, argue that the progress in QA systems has been delayed by the lack of appropriate, well-built datasets both for training and testing. They presented the Natural Language Corpus which (according to authors) is the first large publicly available dataset to pair real user queries with high-quality annotations of answers in documents. It is an attempt at providing large-scale end-to-end training data for QA and research in natural language understanding. The set results of manual annotation of answers extracted from the Wikipedia pages to a question that has been issued to the Google search engine by multiple users in a short period of time. In fact, the input of samples in this dataset, into a model, is the question together with an entire Wikipedia page. The output is a long answer and/or a short answer that can be in a Boolean 'yes/no' format. The main objective was to be close to an end-to-end application. Some heuristic rules were applied to filter the questions in order to discard non-questions. The manual annotations are then sampled and some of them validated by other expert annotators. They have tested their *corpus* by implementing a Document-QA approach - by Clark and Gardner, 2018, which performed well in the SQuAD and TriviaQA challenges - for their long-answers, and also the Decomposable Attention model - for their short-answers. The authors conclude that despite providing a *corpus* with 'high-quality annotations of answers in documents', the idiosyncrasies of natural questions do not yield a comparable performance in their chosen metrics suggesting that significant advances in QA systems are still required to tackle natural language understanding.

Biomedical QA

In the realm of biomedical sciences, it is evident the difficulty in providing large-size datasets as the annotation by experts in this domain is limited in its scalability. In 2007 the TREC challenge included a *corpus* based on genomics with a related document retrieval task. Concurrently the QA4MRE challenge also extended its tasks to include a dataset with 40 QA instances. According to the challenge's documentation, Questions are in the form of multiple-choice, each having five options and only one correct answer. The detection of correct answers is specifically designed to require various kinds of inference and the consideration of previously acquired background knowledge from reference documents. Recently Jin et al. [19] published the PubMedQA *corpus* with the goal of having substantial instances with some experts annotations that require reasoning over the contexts to answer the questions. This dataset has manually labelled 1000 articles with question titles and automatically converted statement titles of 211.3K PubMed articles to questions and labelled them with yes/no answers using a simple heuristic. In this dataset, the contexts are generated to answer the questions and both are written by the same author thus ensuring that contexts are absolutely related to the questions making it an optimal benchmark for reasoning. Attempts have been made to automatically collect large-scale datasets for the biomedical domain. Pampari et al. [32] have automatically generated the emrQA dataset for electronic medical records by re-purposing existing annotations. It consists of annotations made by physicians where they pose questions against long time medical records of patients. The resulting *corpus* has 1 million question-logical forms and 400,000+

question-answer evidence pairs. BioRead collected closed-style QA instances by masking biomedical entities in sentences of research articles as context [33]. BioRead was constructed by randomly selecting approx. 90.6k from the +3.4M articles PubMed Central and by then applying MetaMap to each one of the selected articles that recognise words or phrases referring to concepts of the Unified Medical Language System. This yielded a dataset with approximately 16.4 million passage-question instances making it one of the largest MRC in the biomedical domain.

2.2 BioASQ and BiQA

2.2.1 BioASQ challenge and corpus

The BioASQ challenge is an initiative funded by the European Union in the realm of semantic biomedical research. It has been running every year since 2013 as a competition set up with the goal of providing a public evaluation framework of biomedical semantic indexing and QA systems. The structure and components of this challenge are laid out by Tsatsaronis et al. [43]. It tackles the very real difficulties that biomedical professionals encounter when compiling, filtering, and gathering biomedical knowledge from large and exponentially growing databases. These professionals face growing difficulties in keeping up with the rapid increase of research and data. The current search engines are either limited in their resources or, on the other hand, multiple sources of information require the sort of analysis, filtering, and study that is extremely time-consuming. The emergence of QA systems that might help produce answers from a broad body of research is of paramount importance.

The process of annotating documents with well-recognized taxonomies has allowed the matching of questions and answers. However, in the biomedical realm, this process has, for the most part, been done manually, which is a process that, evidently, can't cope with the increasing amount of new information. The BioASQ challenge attempts to contribute to the development of QA systems capable of tackling this issue by setting up very well-defined tasks that can lead to the integration of effective semantic indexing in the biomedical field. In addition, it provides a universal evaluation framework for biomedical indexing and also contributes by making available a high-quality benchmark dataset to support the development of such systems.

According to [43], the BioASQ challenge evaluates the ability of systems to perform:

- large-scale classification of biomedical documents onto ontology concepts;
- classification of biomedical questions on the same concepts;
- integration of relevant document snippets, and information from relevant databases;
- retrieval of information in a concise and user-friendly form.

BioASQ Tasks

The challenge has been split into Task A and Task B. Task A is called 'Large-scale online biomedical semantic indexing. The goal here is to classify documents from the PubMed publicly available library into concepts of the MESH hierarchy. Here, articles that have been recently published and not yet annotated are collected and used as test sets for evaluation of the participant systems.

Task B is named 'Biomedical Semantic Question Answering'. This part of the challenge requires the participating systems to deal with all stages of a QA system by annotation of natural language questions with biomedical concepts and retrieval of documents, snippets of text, triples, exact and ideal answers. One of the objectives of this dissertation is to develop a dataset capable of contributing to the development of models related to task B, therefore justifying a more detailed description.

Task B comprises two phases: In phase A the competition releases batches of 100 questions where the participants can respond with relevant concepts, PubMed articles, snippets of text as well as relevant Resource Description Framework (RDF) triples. It is not mandatory to participate in all these sub-tasks, the competition has the flexibility of evaluating each one of them separately; in phase B, correct (gold) snippets, articles, concepts, and RDF triples are added to the released questions. The purpose is to allow the systems to retrieve 'exact' and 'ideal' answers. Exact answers depend on the question type. For example, the 'yes/no' type of question must be answered with an exact 'yes' or 'no' answer. Similarly, a list type question must be answered with a list of elements. The 'ideal' answer is a paragraph-sized summary that mimics the response one would expect from a fellow scientist.

Task B Dataset format

Since its inception, BioASQ releases new samples of its benchmark dataset every year adding 400-500 questions to the set meant for task B. These datasets have been constructed by teams of biomedical experts as an array of questions in *JSON* format where each question follows the structure in Figure 2.2.

The questions can be of four types: 'yes/no', factoid, list or 'summary'. Each of them has 'ideal' and 'exact' answers except for the 'summary' type which does not have 'exact' answers. The ideal answers are restricted to 200 words and the exact answers also have a limitation of 100 characters.

Metrics

As explained, for each question, the experts produce a correct set of returned concepts, snippets, articles and triples. For each system, the evaluation is performed with the mean average precision (MAP) measure which is a standard measure in the field of information retrieval to evaluate ranked lists of items.

The mean average precision (MAP) of a set of queries is defined by:

$$\text{MAP} = \sum_{q=1}^{Q} \frac{AveP(q)}{Q} \tag{2.1}$$

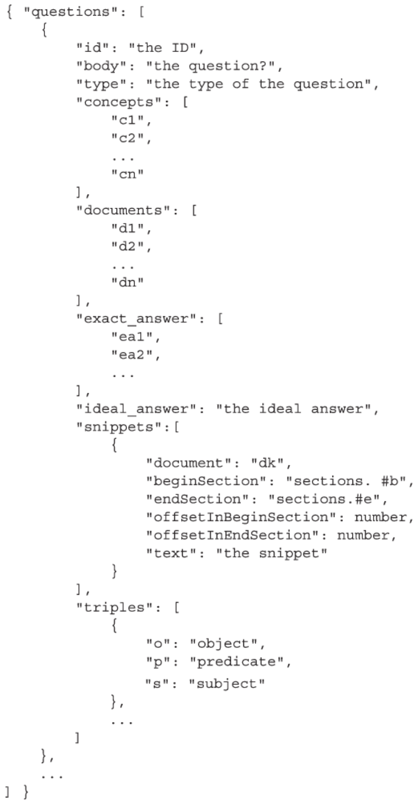
where *Q* is the number of queries in the set and *AveP(q)* is the average precision for of a query, q. For a given query, q, we calculate its corresponding *AveP*, and then the mean of all these scores would give us a single MAP score, which quantifies how good a ranking model performs.

The average precision is a ranking metric, measuring the frequency of relevant recommendations.

$$AveP@N(RetrievedItems_q) = \frac{\sum_{k=1}^{N}(Precision@k(RetrievedItems_q) \cdot relevant(k^{th})}{|RelevantItems_q|} \tag{2.2}$$

$relevant(k^{th})$ is a boolean value, indicating whether the k-th element is relevant, or not.

In the case of evaluation of snippets, the definition of precision was changed to consider a snippet as a set of article-offset pairs. This is because a returned snippet may overlap with one or more golden snippets, without being exactly identical to any of those. In phase B the evaluation of ’exact’ answers is conducted using accuracy for the factoid questions and F1 for ”yes/no”, whilst for list type other measures such as precision, recall and F-measure are used. The ’ideal’ answers are evaluated manually by experts or automatically by measures such as the ROUGE score. The official scores are the result of manual evaluation.



```
{ "questions": [
    {
        "id": "the ID",
        "body": "the question?",
        "type": "the type of the question",
        "concepts": [
            "c1",
            "c2",
            ...
            "cn"
        ],
        "documents": [
            "d1",
            "d2",
            ...
            "dn"
        ],
        "exact_answer": [
            "ea1",
            "ea2",
            ...
        ],
        "ideal_answer": "the ideal answer",
        "snippets":[
            {
                "document": "dk",
                "beginSection": "sections. #b",
                "endSection": "sections.#e",
                "offsetInBeginSection": number,
                "offsetInEndSection": number,
                "text": "the snippet"
            }
        ],
        "triples": [
            {
                "o": "object",
                "p": "predicate",
                "s": "subject"
            },
            ...
        ]
    },
    ...
] }
```

Figure 2.2: Task B dataset Format from Tsatsaronis et al. [43].

An important caveat in the reported evaluations throughout the years is that in some editions of the challenge the denominator in AveP is always 10. Either for documents or snippets retrieved. Furthermore, the final reported official list of relevant documents is updated according to the results of each participant. Hence there might be a discrepancy between the results reported by each individual team on their reported results and the official result by the competition.

2.2.2 BiQA

BiQA is a collection of question-answer pairs generated from three online QA public forums related within the biomedical domain: Biology and Medical Sciences - from StackExchange - and Nutrition - from Reddit.

It represents a methodology for the automatic generation of questions and answers for training and/or evaluation of document retrieval systems. It could be considered a hybrid approach to biomedical *corpora* generation as it automatically extracts questions and respectively associated answers in the form of Pubmed Identification Numbers (PMIDs). Other *corpora* created for competitions such as the BioASQ have their questions and answers curated by biomedical experts from manual interaction with the documentation. Other systems such as PubMedQA [19] attempt to artificially produce questions from documents whilst BiQA leverages the online community input to, rapidly produce, real-life questions and correspondent possible answer documents.

The Queries in BiQA correspond to the question title from the forums' posts. For each question, there is a variable number of answers, which in the case of BiQA, are PMIDs that match a particular article. According to the authors, BiQA has the benefit of providing questions by real-world users with different degrees of expertise or background, therefore providing a higher diversity of formulations. This methodology assumes that the PubMed abstracts referenced inside a forum thread does contain an answer to the question posed.

The process of gathering PMIDs to a question was described by the authors as:

- for each post retrieve all first-level replies hence discarding the answers;
- parse each answer to acquire hyperlinks present in the answer text;
- mapping of URLs from different sources (PMC, doi.org, ScienceDirect, ResearchGate) to PubMed.

Questions that (for some unknown reason) were not mapped to PubMed articles were not included in the *corpus*. The final dataset is composed of more than 7.4K questions and nearly 14k QA pairs.

Table 2.1 shows basic statistics provided by the authors regarding the instances per forum.

The column 'Avg# votes' correspond to the number of votes that the question accumulated in the community. The Medical Sciences set is the smallest in terms of the number of questions but does have the highest number of articles retrieved per answer.

Table 2.1: BiQA samples from Lamurias et al. [24]

Forum	Qs w/PMID	QA pairs	Avg# votes	Avg# PMID
Biology	3961	6925	4.96	1.62
MS	1383	3053	3.91	2.06
Nutrition	2109	4261	6.13	1.79
total	7396	14133	5.01	1.82

The questions posed on forums of this nature - public, non-expert, informal - may be prone to all sorts of misspelling, grammatical mistakes, or constructed in a more levity way without the formality academic research would demand. Whilst such diversity of question constructs might actually be desirable for the training of a natural language model capable of question-answering, it is also true that the input of questions with poor syntax, non-logic or lexically poor, may, in fact, contribute to a less robust result in the modelling of QA systems for a biomedical purpose. Ideally, the questions should be phrased in a readable and understandable format adhering to human speech conventions, basic semantic and grammatical rules. There is limited use of input questions that do not make sense from a standpoint of logic, are too ambiguous, or contain topics non-related to each other yielding a situation where, even for humans, an answer would be extremely difficult or impossible to articulate.

Likewise, the response PMIDs to each question in the forum might not have a corresponding abstract text potentially due to issues such as errors in the API request, the article being updated onto another id number, correct article and title but no abstract text present in the request results or a simple wrong allocation of the PMID in the Pubmed API. Also, a response PMID could be wrongly duplicated if the mentioned article is repeated in the forum answer text.

With the view of validating this methodology, the authors applied the whole BiQA corpus to a deep learning QA system - whose models have been submitted to the document re-ranking sub-task of the BioASQ6 competition [5] - to generate and compare its impact on this type of system, merging the questions of the three subsets into a single corpus, without constrains on the number of votes or PMIDs. Their experiments with this architecture explored generating models with BiQA in conjunction with the BioASQ6b training set and using just the BiQA *corpus*. The results reported by Lamurias et al. [24] are shown in Table 2.2.

Table 2.2: MAP scores from BiQA document re-ranking experiments in Lamurias et al. [24]

Test batch	BioASQ only	+BiQA	Δ	BiQA only	Δ
1	0.2221	0.2235	-0.0014	0.2125	0.0096
2	0.2267	0.2231	0.0035	0.2025	0.0241
3	0.2415	0.2436	-0.0021	0.2279	0.0136
4	0.1686	0.1712	-0.0026	0.1680	0.0006
5	0.1340	0.1355	-0.0015	0.1254	0.0086

These results are, somewhat, inconclusive. On one hand, they show only marginal improvements in MAP scores by adding the BiQA corpus to the BioASQ6b training set, where higher scores would be expected since 14K training samples were added to the original BioASQ set. Conversely, the usage of BiQA alone showed similar scores when compared to BioASQ alone. This could reflect the inability or weakness of this particular system but could also result from idiosyncrasies or characteristics in BiQA that do not meet the expectations of this methodology or a combination of both.

2.3 Question-Answering Systems

QA systems have emerged as frameworks that automatically produce human-readable answers to natural language queries. Soares and Parreiras [41] define the following most common domain-specific terms that one should keep in mind when exploring this topic:

- Question Phrase – what is searched;
- Question type – categorization of the question for its purpose
- Answer type – class of objects sought by the question;
- Question Focus – property or entity being searched;
- Question topic – object or event that the question is about;
- Candidate passage – anything from a sentence to a document retrieved by a search engine in response to a question;
- Candidate answer – text ranked according to its suitability as an answer.

There are many QA systems each with its own application field. There are a variety of variables or features that determine the goal of the system [36]. Notwithstanding, a common framework using NLP and IR techniques can be generalized as seen in Figure 2.3.

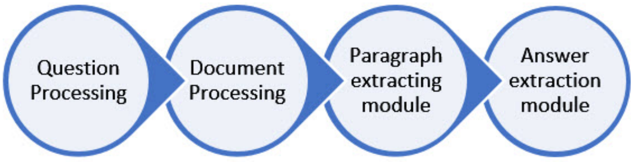


Figure 2.3: QA systems workflow general overview

Question processing module – it receives a question in natural language, analyses the structure for type, meaning, and scope, to avoid ambiguity in the answer (Malik et al., 2013), and compose a meaning formulation compatible with the QA's domain.

Document Processing module – selection and retrieval of documents from the *corpus*. These will be documents likely to contain information pertaining to the user's question. Hence it can integrate a subset of possible documents or generate a neural model which in turn can provide the source for the answer extraction. Usually, the candidate documents are ranked by their potential relevance at this stage.

Paragraph Extraction Module – here the algorithm attempts to extract passages from the documents, rank them and return the ones with the highest score. The system should be abstract enough so that tokenization and passage scoring algorithms can be modularized as well [36].

Answer extraction module – it is the most challenging component of the whole system. The answer produced is, ideally a simple sentence but might also require merging information from different sources, summarization, identify uncertainty, and deal with paradoxes.

Although this general architecture provides a good starting point to understand how these systems work, it is far from consensual how to best characterize or categorize them.

Kodra and Meçe [22] have attempted to summarize the main characteristics of QA systems as:

- System Domain - indicates the domain or specificity of the target domain. They can be open-domain, closed-domain, or even restricted-domain. Usually, open-domain systems are based on implementations on the open world wide web were as restricted and closed domains are subject-specific;

- System Type - They classify the type of existing systems in the literature as being community or non-community based. The vast majority are non-community, where the system is closed, almost like an encyclopedia relying on its own knowledge base. However, a trend has emerged where community-based systems like 'Quora' or 'Yahoo Answers' are gaining significant traction in QA research;

- Question type - One common heuristic to define questions is by their type. Some examples are: Factoids – simple fact that can be answered with a short sentence; List Question – the answer is a list of entities; Definition Question – expect a summary or short sentence; Complex Question – information in a context. Multiple sentences or passages are required to give a meaningful answer. They can simply be joined together or computed by using more complex algorithms as Normalized Row.Scoring, Logistic Regression, Round-Robin or 2-Step RSV;

- Information Source - documents and/or structured knowledge-based;

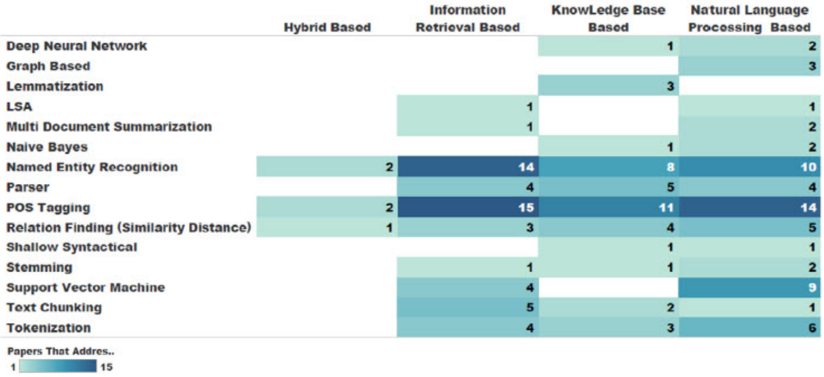
- Information Source Type - single or multiple sources of information.

The very first systems conceived were mostly closed-domain however, in the 1980s and 1990s, researchers shifted their attention to open-domain QA systems. By treating each question as a search query the systems retrieve a set of relevant documents, extracted candidate answers from the results, and then present the best candidate answer to the searcher. The emergence of open-domain QA systems inspired the Text Retrieval Conference (TREC) to establish a question-answering track, which has been running since 1999. Other examples of gain of momentum in QA systems research was in 2009, when Wolfram Alpha launched an "answer engine" based on a collection of curated content, and Siri (Apple) integrated it when it launched in 2011. In 2012, Google embraced QA by launching its Knowledge Graph, leveraging the Freebase knowledge base from its acquisition of Metaweb.

A complementary perspective on how to look at these systems is suggested by Soares and Parreiras [41] and focus on the paradigm they try to implement :

- Information retrieval – where search engines are used to retrieve answers followed by applying filters and ranking scores
- Natural Language Processing – linguistic intuitions and machine learning methods to get answers from retrieved texts
- Knowledge based – having a structured database instead of unstructured. A fine example of this would be an Ontology database describing concepts and their relationships within a certain domain. These are often more sophisticated than simple relational databases and appropriate query languages are developed to perform such queries.
- Hybrid – more modern systems try to make use of as many sources as possible. Integrating a combination of information retrieval, natural language, and knowledge databases. A typical example of this is the well-known IBM Watson system.

Figure 2.4 shows a compilation review of the applied techniques, algorithms, frameworks and tools from Soares and Parreiras [41].



	Hybrid Based	Information Retrieval Based	KnowLedge Base Based	Natural Language Processing Based
Deep Neural Network			1	2
Graph Based				3
Lemmatization			3	
LSA		1		1
Multi Document Summarization		1		2
Naive Bayes			1	2
Named Entity Recognition	2	14	8	10
Parser		4	5	4
POS Tagging	2	15	11	14
Relation Finding (Similarity Distance)	1	3	4	5
Shallow Syntactical			1	1
Stemming		1	1	2
Support Vector Machine		4		9
Text Chunking		5	2	1
Tokenization		4	3	6

Figure 2.4: Techniques, algorithms, frameworks and tools in QA systems. From Soares and Parreiras [41].

The challenges of performing research in this field were also well summarized by Kodra and Meçe [22]. For instance, in community-based online forums, it is common for different users to ask the same questions with different formulations. The authors named this a lexical gap between questions. This results in many questions with the same meaning but with a very different lexicon. Also, it is common to identify grammatical gaps between the questions and answers where the answers could be more elaborate or technically correct whereas the questions could be more informal. Another problem detected in these forums is the deviation from a proper answer. Users can engage in a parallel discussion or simply give unrelated answers. On the other hand, in knowledge-based systems, the most frequent problem is the lexical difference between the question in form of natural language and the structured semantic in the database. Additional reported problems are the fact that some questions involve multiple entities and the identification of an entity and link to a triple in the database.

In addition, one of the most well known issues is called the Question-Understanding problem and is indeed one of the main challenges of QA in the biomedical realm. It occurs when users often phrase questions with long and irrelevant information that contribute to the increase in false positives in answer retrieval. Ben Abacha and Demner-Fushman [3] have been working on this problem and present a possible solution that could be integrated into the pipelines of current systems.

Ishwari et al. [18] review presents a guide through an historical perspective on the techniques, algorithms, and models upon which these systems have been built to deal with unstructured text in QA systems:

Rule based approaches

This was one of the first most promising methods characterized by a selection of rules inferred from grammatical semantics, handcrafted and mostly heuristic in nature, relying on lexical and meaningful context and also based on expert knowledge. It could only solve specific, narrowly defined problems. For example, a system used in the 1954 Georgetown-IBM experiment relied on six grammar rules and 250 lexical items to translate over 60 sentences from Russian to English. Not surprising these representations were akin to decision trees where the rules intent to be linguistic structures mirroring the way humans understand text. To improve these systems matching syntactic, morphological analysis and common knowledge linguistic techniques such as tokenization, Part-of-Speech (POS) tagging and parsing and Named Entity Recognition (NER) [1] were then added. Historically, one of the earliest QA systems was ELIZA, developed in 1964, which had a great successful application with programs such as DOCTOR - where a computer program interacted with users through a text chat interface, answering questions and responding to the users dialog in a manner that mimicked a psychotherapy session between a patient (the user) and their therapist - and BASEBALL [12] system - which was built for answering questions about baseball games played in the American league in one year. Although these systems were quite successful

[1] The task of identifying and categorizing key information (entities) in text. An entity can be any word or series of words that consistently refers to the same thing. Every detected entity is classified into a predetermined category. For example, an NER model might detect the word “super.AI” in a text and classify it as a “Company”.

they rely entirely on the constant update of rigid rules that need to be hard coded into the system. This is time consuming and make the system highly dependent of language specialists to curate the rules. Evidently such systems are not scalable.

Statistical approaches

Statistical methods were the next wave of innovation in QA systems and require the formation of a hypothesis previous to building the model. Contrary to rules-based models these systems can deal with large amounts of data, are domain independent and able to deal with diverse data. In fact, the more data there is the more these approaches thrive. Also the learned statistical method can easily be altered, customized, and be language independent. Different models have been applied depending on the stages of the system. For example, Support Vector Machines (SVMs), Bayesian classifiers, maximum entropy models have all been successfully implemented. These are trained on a *corpus* of questions which have previously been annotated with the respective categories. Stochastic models also haves increased improvement on POS tagging by attempting to handle disambiguation, decoding and smoothing of unknown words utilizing Hidden Markov Chain Models, the Viterbi algorithm ors Linear Interpolation.

Machine Learning approaches

The statistical methods opened the way for introduction of machine learning techniques which bring an entire new perspective on learning to understand linguistic features without explicitly being programmed. Given an annotated *corpus* (dataset) these techniques will, by themselves build a knowledge base. The context is usually processed by means of NER techniques where a taxonomy is built and then acts as the knowledge base. The downside here is that a system like this usually requires large amounts of training data. Evidently these systems are far more scalable than rule based ones. Very frequently an ensemble of machine learning algorithms is compiled for meta-classification tasks and has proven to be very effective [19].

Deep Learning approaches

More recently the advent of deep learning techniques has shown to achieve higher results than machine learning or statistical methods. Not surprisingly Natural Language Processing benefits from the non-linear learning capabilities of neural network systems. The very first ideas to deal with natural language processing in neural networks applied the so called Recurrent Neural Networks (RNNs). This architecture was different from the traditional neural network in the relationships hidden layers maintain with previous values. Conceptually they differ from a standard neural network as the standard input in an RNN is a word sequence instead of the entire sample as in the case of a standard neural network. This gives the flexibility for the network to work with varying lengths of sentences, something which cannot be achieved in a standard neural network due to its fixed structure. It also provides an additional advantage of sharing features learned across different positions of text which can't be obtained in a standard neural network.

However RNNs have limitations such as slow speed in training, the fact that they are capable of capturing the dependencies in only one direction, and the vanishing gradient problem. This usually results in substantial limitations into how many words the network is actually weighting. In other words, it does not work in long sentences, only in very short phrases. One of the solutions proposed to this problem was the well known LSTM architecture. An LSTM is an improvised RNN able to compute longer sequences more efficiently by providing a feature to “remember” the relevant and “forget” the irrelevant parts of the data. LSTMs provide finer control over what is needed by making changes to the internal structure of a neuron.

However, in recent years, the field of NLP has experienced a fast evolution thanks to the development in DL research and the advent of Transfer Learning techniques. Powerful pre-trained NLP models such as OpenAI-GPT, ELMo, BERT and XLNet have been made available by the best researchers of the domain which inevitably contribute to further advances in development of QA systems. In addition, other architectures that have been used in other domains of deep learning research such as convolutional neural networks have also been implemented in QA tasks.

2.4 Deep Learning architectures in QA systems

The following section gives an overview of deep learning architectures implemented in this thesis. Namely, it provides a general guide into CNNs and of transformers based models. It also explores in detail the system submitted by the Department of Informatics of the Athens University Economics and Business Schools (AUEBs) to the BioASQ6 challenge.

2.4.1 Convolutional Neural Networks

CNNs are a variant of neural networks used heavily in the field of Computer Vision aimed at preserving spacial relationships where each layer operates on a small region of the previous layer. These architectures were first introduced in 1989 by LeCun et al. [25] but gained great interest after deeper models achieved amazing results in ImageNet competition in 2012 [2]. The hidden layers of a CNN typically consist of convolutional layers, pooling layers, fully connected layers, and normalization layers. Here it simply means that instead of using the normal activation functions defined above, convolution and pooling functions are used as activation functions and are trained using back-propagation and gradient descent as for standard neural networks [27] (Figure 2.5). Typically, CNNs have two or more fully connected layers at the end of the architecture from where the final outputs are computed.

For a given input image, a convolution is applied based on a receptive field. The convolution process is well-suited for image recognition because it considers locally connected information (neighbour pixels or voxels). These convolutions learn weights in order to extract features such as detection of an edge, a texture, or perhaps a contrast between two colors (Figure 2.6).

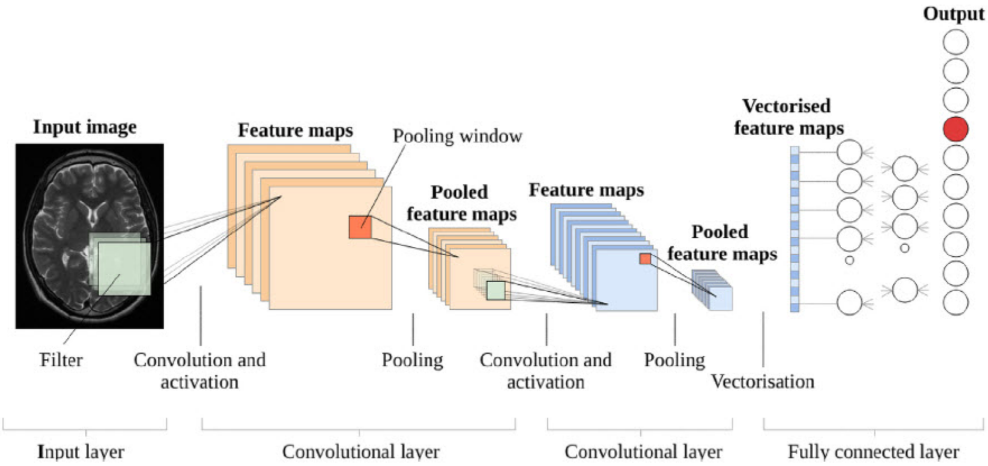


Figure 2.5: Basic CNN structure from Lundervold and Lundervold [27]

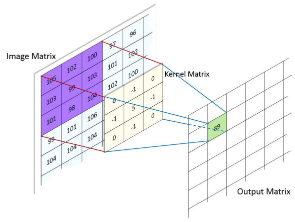


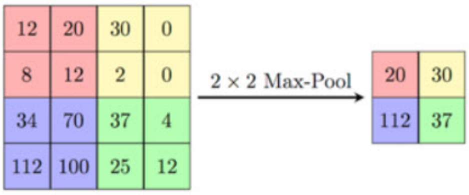
Figure 2.6: CNN kernel. Source: http://machinelearninguru.com/computer_vision/basics/convolution/image_convolution_1.html

Three main components define CNNs:

Convolutional layers: In the convolutional layers the activation from the previous layers are convolved with a set of small parameterized filters, frequently of size 3x3, collected in a tensor *W(j,i)*, where *j* is the filter number and i is the layer number. By having each filter share the exact same weights across the whole input domain, i.e. translational equivariance at each layer, one achieves a drastic reduction in the number of weights that need to be learned. The motivation for this weight-sharing is that features appearing in one part of the image likely also appear in other parts. If you have a filter capable of detecting horizontal lines, say, then it can be used to detect them wherever they appear. Applying all the convolutional filters at all locations of the input to a convolutional layer produces a tensor of feature maps [27].

Activation layer: similar to the basic structure of nodes in Artificial Neural Networks (ANNs) the feature maps resulting from convolutions are fed through non linear activation functions, typically the ReLu function. This produces new tensors also called feature maps.

Pooling: Between convolutional layers pooling layers are introduced to increase the field of view has it takes the convolution outcome as input giving an output that has a lower spatial footprint. These pooling layers reduce over-fitting and computational costs. Pooling is a sample-based discretization process. The objective is to down-sample an input representation (image or hidden-layer output matrix), reducing its dimensionality and allowing for assumptions to be made about features contained in the sub-regions binned. Pooling operations take small grid regions as input and produce single numbers for each region (Figure 2.7). The number is usually computed by using the max function (max-pooling) or the average function (average pooling). Since a small shift of the input image results in small changes in the activation maps, the pooling layers gives the CNN some **translational invariance**. In other words, they lose the information about the exact location of the feature detectors making them unable to acknowledge objects when they are rotated or suffer any other kind of transformation.



12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

20	30
112	37

Figure 2.7: Max pooling schema. Source: https://towardsdatascience.com/understanding-neural-networks-from-neuron-to-rnn-cnn-and-deep-learning-cd88e90e0a90

Typically, after convolutional and pooling layers, CNNs have Fully Connected (FC) layers. They learn the relationship between the features, extracted by previous convolutions and pooling layers and the target. For this to happen, the feature maps from previous layers have to be flatten into 1 dimension prior to input into the first FC. The last FC layer is composed of x number of neurons corresponding to the number of classes. This can be transformed into *n* probabilities using a softmax function to the outputs [11]. CNNs benefit from the fact that weights are shared over the entire input, significantly reducing the computational cost, and allowing the network to extract elementary and higher order local features. The fact that these networks do not need any prior knowledge on the types of the features to extract, has made them popular architectures in medical image processing [38]. Generally speaking, in a CNN with N layers, the output Y(l−1) of layer l − 1, for (2 ≤ l ≤ N), is the input to the layer l resulting in the associated output Yl given by

$$X_{i,j}^{(l)} = \sum_{a=0}^{M-1} \sum_{b=0}^{M-1} \mathbf{W}_{ab} \mathbf{Y}_{i+a,j+b}^{(l-1)} \tag{2.3}$$

$$\mathbf{Y}_{i,j}^{(l)} = \sigma\left(\mathbf{X}_{i,j}^{(l)}\right) \tag{2.4}$$

where X(l) is the pre-activation output, M is the size of kernels, W is the kernel matrix containing the CNN weights to be learned during the back propagation, and σ(•) denotes the activation function [1]. Finally, it is worth to mention that the drawbacks of CNNs are well known. The two main ones being the fact that they can't infer any kind of spatial relationship between objects in the input and the lack of robustness to rotation and affine transformations in images.

2.4.2 AUEB's CNN-based QA system for BioASQ

DL strategies based on CNNs, have been applied to Natural Language tasks in the field of biomedical QA systems. Namely, of particular interest to this thesis is the system developed by the AUEBs team, Department of Informatics, submitted to the document and snippet retrieval tasks of the BioASQ 6 challenge (Task 6B phase A) [5].

In the biomedical domain, document and snippet retrieval have had several Information Retrieval (IR) and ML strategies applied. Most document relevance scoring and re-ranking models are either representation-based or interaction-based. In the models explored by the AUEB's team, they implement an interaction-based approach where the query and documents interaction encoding is induced. According to the authors, this allows for direct modeling of exact or near-matching terms, which, although slower than typical IR models, have been shown to yield better performance.

The AUEB's system is comprised of two interconnected networks. The first tackles the document re-ranking part of the BioASQ challenge Task B Phase A, whilst the subsequent portion attempts to retrieve relevant snippets of text from the re-ranked documents. The document re-ranking component of this system was used by Lamurias et al. [24] to validate their BiQA *corpus*. In this thesis, both components of AUEB's system are employed in an attempt to replicate the results from Lamurias et al. [24] and to validate BiQA2.

The AUEB system's architecture has two separate CNNs, designed to take on a query and retrieve the 10 most relevant pieces of text from abstracts of articles publicly available at the PubMed API. To this effect, at inference time, the system works in the following order: 1) receives a query as input; 2) applies a computationally inexpensive search engine that retrieves an N number of documents related to the query – the BM25 algorithm is the standard choice in this domain.; 3) the N documents are then inputted to the document re-ranker network which, in turn, outputs the 10 most relevant abstracts; 4) after being re-ranked, these abstracts are then passed as input to a second model that retrieves the 10 most relevant sentences from those documents. Figure 2.8 from Brokos et al. [5] illustrates the overall architecture of the system as per the authors.

The reason for the system to run a BM25 algorithm is that DL models are computationally very expensive. Running them on every abstract of PubMed index is not feasible but it is reasonable to do so in 100 or even 200. Hence the initial search engine stage to which the document re-ranker is executed.

Document re-ranking

The AUEB's paper explores different architectures of a so-called term-based interaction way where documents are attributed a score as a proxy to its relation to the query. However in this thesis we only investigate the Position-aware Convolutional Recurrent Relevance (PACRR) model (proposed by Hui

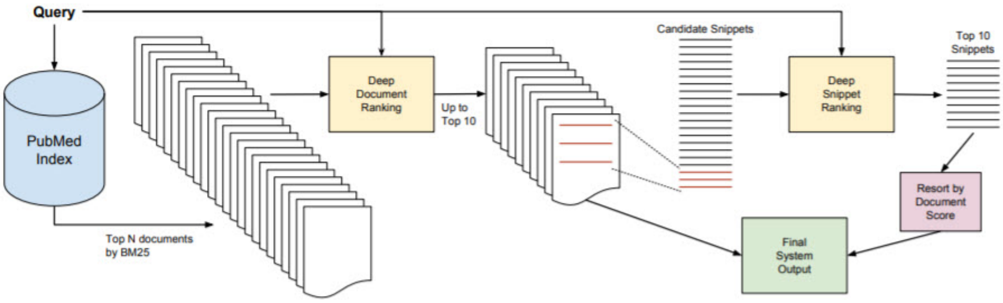


Figure 2.8: AUEB's complete system for Task B Phase B. Source:https://arxiv.org/abs/1809.06366

et al. [17]) since not only it was the best performing model to be submitted to the BioASQ challenge by the AUEB team, but it was also the DL model applied by Lamurias et al. [24] to validate BiQA.

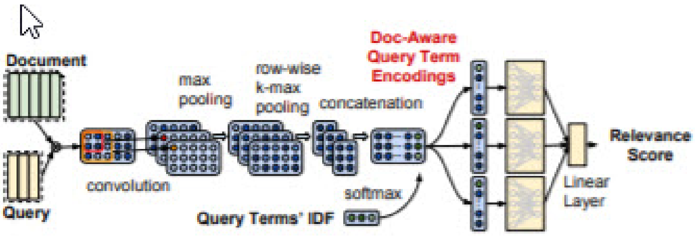


Figure 2.9: TERM-PACRR representation. The Multi-Layer Perceptron (MLP) is applied separately to each document-aware q-term encoding; the resulting scores are combined by a linear layer. Source: https://arxiv.org/abs/1809.06366

This model takes as input both query and document embeddings and computes a cosine similarity matrix between them. In order to keep the dimensions of this matrix fixed - regardless of the size of the query or document – the queries are padded to a maximum number of terms and the abstracts are capped to the first N number of tokens. It is on this similarity matrix that convolutions of variable kernel sizes are applied. For each size, multiple filters can be used. In an attempt to encapsulate the best K signals between each query-document, the authors apply Max Pooling along the dimension of the filters, followed by k-max pooling along the dimension of the documents terms. This results in one matrix per filter which are then concatenated into a single matrix where one row corresponds to a 'document-aware' question term encoding. The authors describe a slight change to the original network by including an individual MLP to score each encoding. These scores are then aggregated in a final linear layer. They called this version TERM-PACRR (Figure 2.9). In addition, other features can be inserted such as Inverse Document Frequency (IDF) of the query terms, which, in this case, is normalized via a softmax function before being appended to the single encodings matrix priot to the MLP.

Snippets retrieval

The second component of AUEB's system is a distinct DL convolutional network named 'Basic CNN' (BCNN). The model receives the output of the document re-ranking, i.e. the top 10 retrieved documents mostly related to the query, splits them into sentences, and scores each sentence by relevance. The input to the network is thus two sequences of tokens - query + sentence from the abstract (snippet), as in Figure 2.10. Snippet sentences are truncated so that they present a constant length. Here a convolutional layer (with the chosen number of kernels) is applied to each query and sentence separately. However, each convolution has the same width for both query and sentence. For each kernel is applied an average pooling layer so that a features map per filter can be computed. By making sure the windowed-average pooling is performed over the same filter width the dimensionality of the feature map is the same for all filters which permits stacking of a random number of convolutions allowing for the extraction of more meaningful features. When using different filters, evidently, results in different feature vectors from each filter. For each one of the vectors average pooling is applied and for each of those similarity scores are computed between the query and sentence features map. This array of similarity scores is then concatenated and passed on to a final linear logistic regression layer.

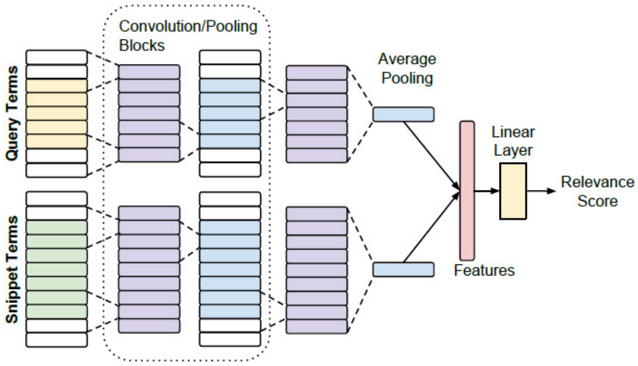


Figure 2.10: BCNN architecture for scoring snippets relative to a query. Source: https://arxiv.org/abs/1809.06366

2.4.3 Transformers/BERT

Pre-trained word vectors have been tremendously useful in NLP as an approximation to language modelling at a time when hardware was very slow and deep learning models were not widely implemented. These models found adoption through their efficiency and ease of use. Since then, the standard way of conducting NLP projects has largely remained unchanged: word embeddings pre-trained on large amounts of unlabeled data via algorithms such as word2vec [28] and GloVe [34] are used to initialize the first layer of a neural network, the rest of which is trained on data of a particular task. On most tasks with limited amounts of training data, this led to a boost of two to three percentage points in metrics gains.

These methods are shallow approaches that trade expressivity for efficiency not being able to capture higher-level information that might be more useful. A model initialized with word embeddings needs to learn from scratch not only to disambiguate words but also to derive meaning from a sequence of words. This is the core aspect of language understanding, and it requires modelling complex language phenomena such as compositionality, polysemy, anaphora, long-term dependencies, agreement, negation, and many more. It should thus come as no surprise that NLP models initialized with these shallow representations still require much larger amounts of data — they see major improvements when trained on millions, or billions, of annotated training examples. Interestingly, pre-training entire models to learn both low and high-level features has been practiced for years by the computer vision (CV) community. Most often, this is done by learning to classify images on the large ImageNet dataset. ULMFiT, Elmo, and the OpenAI transformer have now brought the NLP community close to having an ”ImageNet for language”, i.e. a task that enables models to learn higher-level nuances of language, similarly to how ImageNet has enabled training of CV models that learn general-purpose features of images. Researchers have developed various techniques for training general-purpose language representation models using the enormous quantities of unannotated text on the web (this is known as pre-training). These general-purpose pre-trained models can then be fine-tuned on smaller task-specific datasets, e.g., when working with problems like question answering and sentiment analysis. This approach results in great accuracy improvements compared to training on the smaller task-specific datasets from scratch. One such recent model that made a breakthrough in the NLP community, taking over the top score spots across multiple tasks is the Bidirectional Encoder Representations from Transformers (BERT) – presented by Devlin et al. [7] in 2018. It took the DL community by storm as it presented state-of-the-art results in a wide variety of NLP tasks, like QA.

BERT relies on the Transformer model architecture, instead of LSTMs. A Transformer works by performing a constant number of steps applying an attention mechanism to understand relationships between all words in a sentence, regardless of their respective position. It lets go of the recurrence in RNNs and LSTMs and instead relies entirely on an attention mechanism to draw global dependencies between input and output. Instead of predicting the next word in a sequence, BERT makes use of a novel technique called Masked Language Model (MLM): it randomly masks words in the sentence and then it tries to predict them. Masking means that hat a fraction of the words of a *corpus* are masked (hidden) and the model looks in both directions and it uses the full context of the sentence, both left and right surroundings, in order to predict the masked word. Unlike the previous language models, it takes both the previous and next tokens into account at the same time. The existing combined left-to-right and right-to-left LSTMs based models were missing this “same-time part”.

A basic Transformer consists of an encoder to read the text input and a decoder to produce a prediction for the task. Since BERT’s goal is to generate a language representation model, it only needs the encoder part. Essentially, the Transformer stacks a layer that maps sequences to sequences, so the output is also a sequence of vectors with a 1:1 correspondence between input and output tokens at the same index (Figure 2.11). BERT does not try to predict the next word in the sentence. Instead, it leverages multihead attention between all term pairs in the text sequence — but makes it very deep. Its main version, BERT-Large, includes 24 Transformer layers, each with 1024 hidden dimensions and 16 attention heads. It in total has 340 million learned parameters, much bigger than typical neural networks-based IR models.

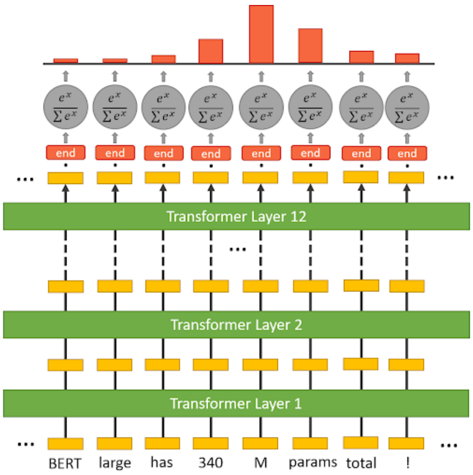


Figure 2.11: BERT architecture overview. Source: http://www.mccormickml.com/

A Pre-trained BERT can be viewed as a black box that provides 768-dimensional vectors for each input token in a sequence. Here, the sequence can be a single sentence or a pair of sentences separated by the separator [SEP] and starting with a token [CLS].

According to Devlin et al. [7] the input to the encoder for BERT is a sequence of tokens, which are first converted into vectors and then processed by the neural network. The authors of BERT use three distinct embeddings to represent the input token (Figure 2.12): 1. Token embeddings: A [CLS] token is added to the input word tokens at the beginning of the first sentence and a [SEP] token is inserted at the end of each sentence. 2. Segment embeddings: A marker indicating Sentence A or Sentence B is added to each token. This allows the encoder to distinguish between sentences. 3. Positional embeddings: A positional embedding is added to each token to indicate its position in the sentence.

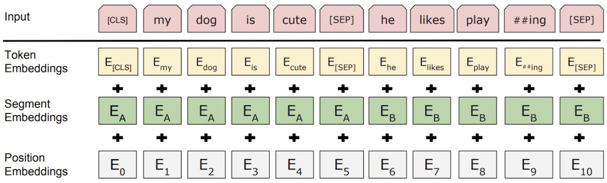

Input
[CLS]
my
dog
is
cute
[SEP]
he
likes
play
##ing
[SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

Figure 2.12: BERTs embeddings inputs. Source: Devlin et al. [7]

Fine-tuning tasks

The authors of BERT added the “[CLS]” token at the start of the sequence, whose embeddings are treated as the representation of the text sequence(s), and suggests to add task-specific layers on the “[CLS]” embedding in fine-tuning.

If we want to fine-tune the original model based on our own dataset, we can do so by just adding a few extra layers on top of the core model. For example, if we want to perform a sentence or sentence pair classification task we use the output from the [CLS] token and connect it one or more neural dense layers or to a Logistic regression classifier or take an average of all the outputs and then run a logistic regression on top. There are many possibilities, and what works best will depend on the data for the task. We can also tune for Single Sentence Tagging Task where we need to predict some tags for each token rather than the word itself. For example, for a POS Tagging task like predicting Noun, Verb, or Adjective, we will just add a Linear layer of size (768 x n_outputs) and add a softmax layer on top to predict one of the POS classes. As another example of fine tuning, we may want to create a question answering application which is merely a prediction task — on receiving a question as input, the goal of the application is to identify the right answer from some *corpus*. Given a question and a context paragraph, the model predicts a start and an end token from the paragraph that most likely answers the question. This means that using BERT a model for our application can be trained by learning two extra vectors that mark the beginning and the end of the answer.

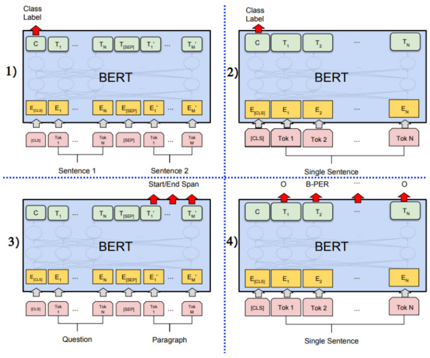


Figure 2.13: BERT applications: 1. Sentence Pair Classification tasks 2. Single Sentence Classification Task 3. Question Answering Tasks 4. Single Sentence Tagging Task. Source:https://arxiv.org/abs/1810.04805

Chapter 3

BiQA Verification

3.1 Verification methodology

The following section describes the methodology applied to the manual verification and characterization of a sample of BiQA through the perusal of its questions and corresponding answers/abstracts.

The annotations for verification of the BiQA set - and construction of BiQA2 - were performed in a sample of 400 questions from each of the three forums adding to a total of 1200 questions. The queries were sorted by the highest score in voting by the forum participants. The objective of verifying BiQA was, in practice, translated into the assessment of certain basic lexical and grammatical features of the Queries and the ability of the provided abstracts to actually provide an answer to its respective query.

3.1.1 Queries verification

Primarily it is important to ascertain if the queries in BiQA are constructed with semantic and/or grammatical sense thus allowing a clear human interpretation. If it is not possible to ascertain a question's meaning or intent because it is just too ambiguous or it does not make sense in any way then that particular question was considered invalid and no further consideration was given that question or to the answers gathered from it. This information could be taken into consideration at a later stage, for example, to predict the ability of a QA system to determine if the questions are answerable. On the other hand, if it was considered as having a minimum standard of understanding of what it being asked then it would then proceed to be characterized by the features detailed below. Its correspondent answers would also be analysed.

Syntax, objectivity and acuity

In order to attempt a characterization of the lexicon within the questions in BiQA, each question formulation was classified according to its syntax, objectivity and acuity. These grammatical features should provide insights into the nature of the *corpus* questions. A question is considered to have good syntax when (from the annotator's viewpoint) has a reasonable grammatical construction. The difference between subjectivity and acuity might not be obvious at first glance. However, they are a distinct aspect of language. A sentence is usually considered ambiguous when it can refer to more than one state of affairs or object whereas it is usually considered subjective when the subject matter requires some person's point of view in order to make the reference. Ambiguity generally has to do with some unclarity in the reference due to multiple possible referents. Subjectivity generally has to do with some personal point of view with respect to how the referencing is being made. Both primarily have to do with linguistic meaning.

Example of a subjective but clear question in BiQA: 'Are artificial sweeteners safe?'. It is clearly stated without ambiguity but there is subjectivity from the point of view of what an individual might consider safe. On the other hand the sentence 'How do artificial sweeteners affect weight loss?' is an objective question but it is ambiguous in its formulation.

A sentence is usually considered subjective when the subject matter requires some person's point of view in order to make the reference.

Structure

Another feature incorporated in the analysis of BiQA is the type of phrase structure. It could take the set of values $\{simple, double, triple, quadruple, compound, sentence\}$. One of the objectives of BiQA is to provide a diverse lexicon of questions, leveraging the input of non-formal or non-expert users. This is evident in the formulation of some questions in the forums. This feature attempts to categorize such formulation diversity. Some examples are:

- 'Why did the process of sleep evolve in many animals? What is its evolutionary advantage?'
- 'Why are there no organisms with metal body parts, like weapons, bones, and armour? (Or are there?)
- 'Can humans ever directly see a few photons at a time? Can a human see a single photon?'
- 'Right now, what is the most widely accepted/proposed way to heal your gut? What are your experiences with healing your gut with change in your nutrition? What are other, uncertain more controversial ways to do it? How did you heal your gut?'

3.1.2 Abstracts verification

A second layer of verification lies within the content of the abstracts. The BiQA *corpus* can be used for supervised learning tasks where it assumes that all abstracts retrieved will contain an answer to the query hence labelled as positive samples for training purposes. In the annotation performed, the abstracts gathered as answers in BiQA, are thoroughly read in order to understand if, as per the annotator judgment, contain information capable of answering the corresponding question taking into consideration the scope and intent of all the themes/topics articulated in the query. It will then be classified as being relevant or, in other words, a positive sample for training purposes. On the other hand, an abstract considered as negative, or not related to the query, either does not contain any information related to any of the topics in the questions or that information might only address a minority, but never all, of the topics in that question. In other words, if the abstract does not contain information addressing the full intent of the question but merely has sparse pieces of information - for example, glancing only at one of the concepts presented in the question - it would be considered a negative example. In addition, a negative sample could also be an abstract that despite clearly stating or complying with the scope of the question - for example, if it is referring to a systematic review or meta-analysis research paper where an answer is most likely contained within the full article - do not contain an answer in the abstract text itself.

Finally, it was also registered the rare cases where the retrieved PMIDs did not have a corresponding abstract text in the English language.

3.2 BiQA characterization

The analysis performed in this thesis was based on a sample of 400 highest scoring questions for each forum in BiQA. It translates into 1200 questions overall, yielding a total of 2716 QA pairs analysed which corresponds to a sample of 19.69% from BiQA.

Table 3.1 shows the breakdown of valid questions and PMIDs per forum.

Table 3.1: Validation of BiQA QA sampled pairs.

	Biology		Medical		Nutrition		All	
	#	%	#	%	#	%	#	%
Q-A pairs analysed | % BiQA	858	12.84	1000	33.18	858	20.95	2716	19.69
Questions analysed | % BiQA	400	10.47	400	29.18	400	19.59	1200	16.73
Valid Questions | % BiQA Q analysed	366	91.5	377	94.25	333	83.25	1076	89.67
Valid PMIDs | % BiQA QA analysed	809	94.29	947	94.7	820	95.57	2576	94.85
Valid Questions with no valid abstracts	15	3.75	15	3.75	12	3	40	3.33

Given the methodology described, around 10% of questions were deemed as not valid. These are questions that might not be well understood by the annotator, that are open-ended questions with a broad spectrum of possible answers or questions that refer to some sort of other information in the forum post (such as an article or tweet or an image) that BiQA does not capture.

Some examples of queries considered invalid:

’What insect is this? (Central Africa)’

’Contact Inhibition of Cell Division: Signaling Pathway’

’50:50 sugar/fat mixture’
’Superhuman eyesight’
’On the genetics behind caste marriages’
’remo’
‘Which of these things is needed for first time pregnancy check-ups?’
‘Is this study linking RFR and Cancer valid?’
‘Are minerals in salts overrated?’
‘Breakthrough bleeding and antibiotics’
‘Can’t poop. Please send help.’

The perusal of abstracts revealed only a small portion identified as being retracted, repeated or had the wrong identification code. These were deemed invalid and amount to little more than 5%. Approximately 3% of questions had no abstract text and this remains even across the three forums. However, there seems to be an observable difference in regards to the valid questions in the Nutrition forum when compared to the other two, since nearly 17% of its queries were considered invalid.

The global distribution of abstracts relevance to the queries is presented in Figure 3.1.

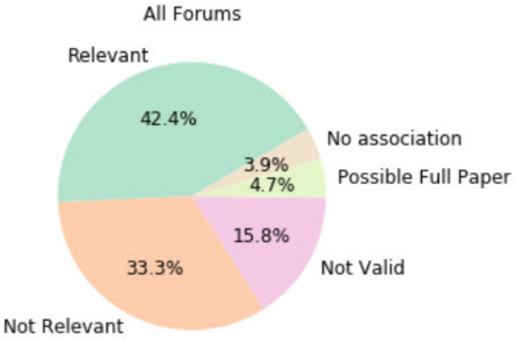


Figure 3.1: Distribution of Abstracts relevance to respective queries in a sample of 1200 queries from BiQA.

Overall, 42.4% of abstracts were considered as having information, i.e. snippets of text, capable of answering their respective queries as per the annotator’s view. 33.3% of abstracts were considered to not containing enough information to answer the respective queries. In reality, these abstracts considered ’Not Relevant’, may address or mention at least one of the topics in the query but fail to provide a concrete objective answer (according to the respective type of question) to the query posed. On the other hand, less than 4% of abstracts were deemed as not related to the queries in any way (’No Association’).

The ’Not Valid’ label corresponds to the abstracts whose questions were considered not valid or where the PMIDs didn’t exist, was duplicated, retracted or simply where no abstract was available through the Pubmed API.

Of note is the proportion (4.7% in total) of abstracts that were considered as relevant to the query but that do not contain retrievable snippets. The description in both title and abstract text in these papers points towards the full text of the article itself most certainly containing information capable of answering the query in a comprehensive manner but such information is not part of the abstract text. An example of this is the case of abstracts corresponding to meta-analysis or systematic reviews that clearly state the scope and intent of the query topics in the abstract but fail to provide any insight into the research's methods or findings.

If we break down the relevance of the abstracts between the three subsets it becomes clear that the medical forum yields nearly 50% of QA positive samples on the corresponding forum whereas the nutrition forum is only 38%.

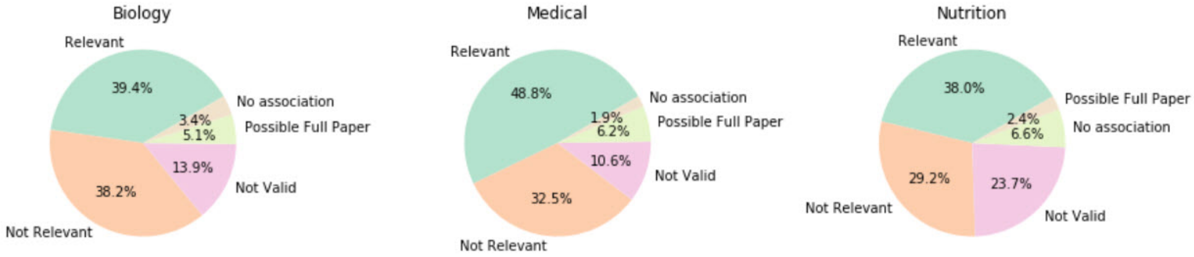


Figure 3.2: Distribution of Abstracts relevance to respective queries per forum

Objectivity, Syntax, Structure, Acuity

The valid queries were also classified by their syntax, semantic nature, ambiguity and complexity. Overall we find that more than half of all valid queries are still somewhat ambiguous and a third contain some level of subjectivity. The overall syntax structure was considered simple but there still exist more than 10% of queries formulated in a compound or complex manner.

Breaking down these features per forum some differences become evident particularly between the Nutrition forum and the others. In fact, both Biology and Medical forums are very similar in acuity (around 50-60% being unambiguous) whereas 65% of questions in Nutrition are considered ambiguous. 70-80% of questions in Biology or Medical were marked as being objective whereas the nutrition forum has 45% of questions as subjective.

The syntax was also similar between Biology and Medical, with 81% and 79% respectively, presenting with good syntax, whereas in Nutrition it is slightly lower at 69%.

Figures 3.3 and 3.4 give an overall perspective of the major differences in the grammatical nature of the question throughout BiQA2.

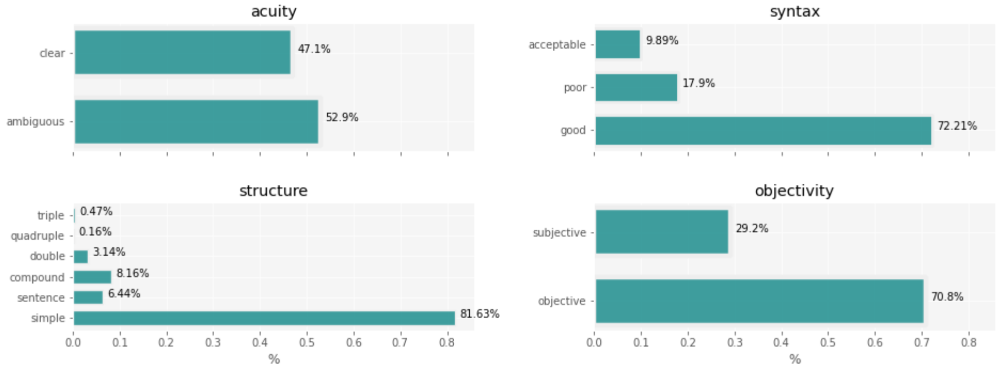


Figure 3.3: Queries grammatical features

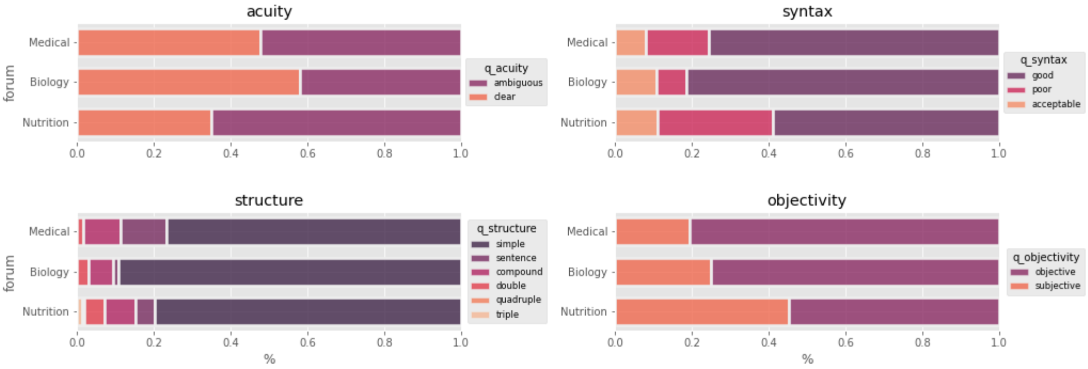


Figure 3.4: Queries grammatical features per forum.

Chapter 4

BiQA2

The following chapter describes in detail the rationale that presided to the concept of building a new, refined and extended dataset from a sample of BiQA. Such a dataset is named BiQA2[1] in the context of this thesis and it is an attempt at creating training examples with features similar to those observed in the BioASQ dataset gold standard set. BiQA2 ought to contribute as an addition to the array of existing biomedical datasets for QA tasks such as the BioASQ dataset.

This section also presents a series of experiments to test or validate the potential of the BiQA2 as training data for the development of models based on deep learning architectures that competed in the BioASQ challenge in recent years. Namely, the training makes use of the system presented by the AUEB team in BioASQ6 for task B phase A for both document re-ranking and snippets retrieval. For the task of snippets retrieval, it is also presented an implementation with a model based on BERT, a recent DL architecture with impressive results on various NLP tasks.

4.1 Methodology

BiQA2 is an attempt to explore the concept of building a biomedical manually annotated *corpus* having BiQA as starting point. As explained previously, BiQA can be used in the training, or generation, of models capable of competing in the BioASQ challenge sub-task of document re-ranking. It does so by assuming that the retrieved abstracts are answers to a certain query. In other words, it considers all its pairs as positive samples. In this dissertation, the main objective is to try to extend the scope of BiQA into BiQA2 by including or adding features to the validated QA pairs from BiQA with the view of potentially use them as a training data source for models competing in both phases of Task B of the BioASQ challenge. BiQA2 effectively explores the possibility of expanding BiQA by adding features such as ’snippets’, ’exact answer’, ’question type’ (similar to the BioASQ dataset). At the same time, because we exclude the abstracts that do not contain text capable of answering its respective question, BiQA2 ought to be a refined version of BiQA in respect to the generation of models related to the document

[1] BiQA2 can be found at https://github.com/MACVLX/BiQA2

re-ranking sub-task .

Snippets of text

In Task B Phase A, the challenge is set up in a way that the participating systems must retrieve, at most, 10 sentences from a list of 10 of the re-ranked relevant documents/abstracts provided by the ranking model. Evidently, these sentences contain answers to the queries.

The first step in building BiQA2 is thus to select snippets of text that can adequately answer the queries. These portions of text come from the abstracts in BiQA that were considered relevant. If an abstract contains sentences that can effectively answer the correspondent query then they were considered as being effective positive Question-Abstract pair. Consequently, the feature ’snippets’ in BiQA2 provides sentences of text as potential data for training or testing of models capable of identifying candidate sentences.

As explained in section 2.2.1, in the construction of the BioASQ dataset, human experts are asked to capture one or more consecutive relevant full sentences from the abstracts. In the case of BiQA2, it was decided that the snippets do not have to adhere to this strict criterion of being full sentences. Partial sentences were allowed, particularly in the case of very long sentences where only a small section is, in fact, relevant. In the case of abstracts containing multiple relevant sentences although not consecutive, the snippets were also captured but considered as separate/independent samples for model training purposes and not concatenated into a single sentence (this is analogous to the BioASQ dataset methodology of snippet selection).

In addition to extracting sentences from relevant abstracts, BiQA2 attempts to emulate the BioASQ *corpus* structure so that it can also potentially be integrated into the development of models related to Phase B Task B of the BioASQ competition. To that end, the QA pairs were annotated with the type of question and the type of answer, following the BioASQ set representations whenever possible.

Features ’question type’ and ’exact answer’

In BioASQ the feature ’question type’ takes into consideration the nature of the question: yes/no, factoid, list, summary. For each of these, there can be both ”ideal” and “exact” answers - with the exception of summary questions, where ”exact” answers are not allowed. According to Tsatsaronis et al. [43], ”ideal” answers (i.e., paragraph-sized summaries) are what a human would expect as an answer by a peer biomedical scientist. In BiQA2, it was decided to not include the ”ideal answer” feature as per the BioASQ dataset, for reasons presented in the Discussion section.

In BiQA2, the BioASQ definition of the question types and corresponding ”exact” answer was followed yielding questions of the types yes/no, factoid, list, and summary. However, in BiQA, there are lexical idiosyncrasies and challenges both in the formulation questions and in possible ”exact” answers contained within the various snippets collect for a same query. It’s not always clear which values to attribute in each question type and ”exact answer” if we were to adhere strictly to the BioASQ categories. Both features ”question type” and ”exact answer” can’t be thought of as separate entities but rather reasoned as dependent on one another when annotating their question type and answer. One of such challenges is related to the yes/no type of question. In BioASQ, the dataset is built by a panel of experts whose task is to curate questions that can effectively be answered by agreed snippets within abstracts from PubMed. By reaching a consensus regarding the information contained within the analyzed abstracts, the experts avoid ambiguity or potential antitheses and contradictions that could occur in abstracts that address a certain topic. In other words, some abstracts convey contradictory information regarding the same research topic but the experts agree on what is the most correct information and discard the contradictory ones. Conversely, in BiQA, a conflict of interpretations and contradictory snippets might exist between different articles expressed in the forum answers for a certain question.

A simple solution would be to eliminate the questions with conflicting abstracts. However, this could result in the removal of considerable amount of samples/abstracts from the *corpus*. Instead, it was decided to also attribute an ”exact answer” value to each individual abstract, consequently, there are two exact answers:

- **question exact answer** - the global ’exact answer’ taking into account the contributions of all snippets. It corresponds to the feature ‘exact answer’ in BioASQ.

- **snippet exact answer** - the ’exact answer’ contained within the snippet of a certain abstract, i.e., one exact answer per abstract. This effectively is an extra feature to the BioASQ format. All the positive pairs with ’question type’ values will have a corresponding ’exact answer’ except for the summary type questions.

For example, in the particular case of yes/no questions, a ’yes’ or ’no’ global ”exact answer” as an aggregate consensual response from its various abstracts might not be possible as they might express conflicting views or conclusions. Because the QA pairs of a query might have different answers in the case of yes/no questions, the value ”uncertain” was introduced in BiQA2 as the third class to this type of question. Thus the ’yes/no’ question type is now converted to a yes/no/uncertain type. If all abstracts do not point towards a ’yes’ or ’no’ global answer then the answer to the question is ”uncertain”. Additionally, if the information within an abstract is also not a clear ‘yes’ or ‘no’ then the ”exact” answer for that particular abstract is also ”uncertain”. Figure 4.1 provides some examples taken from the Medical forum.

Also found in the questions of BiQA was the presence of formulations that require a particular choice between given options. These options are, usually, stated on the query itself, thus, once again, reflecting the nature of the questions in these non-academic forums. Typically these types of questions presented with 2 options plus - implicitly - both ”none” and ”all” options. Therefore BiQA2 introduces an extra type of question: choice. Examples of each question type can be found in Table 4.2.

Table 4.1: Medical forum examples of uncertain type questions

Query	PMID	snippet	snip exact answer
Are organic foods healthier than conventional foods?	24968103	the concentrations of a range of antioxidants such as polyphenolics were found to be substantially higher in organic crops/crop-based foods, with those of phenolic acids, flavanones, stilbenes, flavones, flavonols and anthocyanins	yes
	11833635	With the possible exception of nitrate content, there is no strong evidence that organic and conventional foods differ in concentrations of various nutrients.	no
	19640946	there is no evidence of a difference in nutrient quality between organically and conventionally produced foodstuffs	no
	11327522	There appear to be genuine differences in the nutrient content of organic and conventional crops.	yes
Are there any side effects to cracking knuckles?	1130029	A survey of a geriatric patient population with a history of knuckle cracking failed to show a correlation between knuckle cracking and degenerative changes of the metacarpal phalangeal joints.	no
	10067714	acute injuries can result from the forceful manipulation needed to achieve the audible pop of cracking knuckles and that patients should be counseled accordingly.	yes
Are artificial sweeteners safe?	17828671	The weight of existing evidence is that aspartame is safe at current levels of consumption as a nonnutritive sweetener.	yes
	16507461	indicate that APM is a multipotential carcinogenic agent, even at a daily dose of 20 mg/kg body weight, much less than the current acceptable daily intake.	no

Table 4.2: Examples of BiQA2 types of questions.

yes/no/uncertain	If food prepared in a microwave oven less healthy?
	Are egg yolks actually bad for you?
	Do vaccines cause autims?
choice	What would you say is worse? Consuming excess saturated fats daily, or excess sugar daily?
	Is eating spicy hot (pungent) food (hot chilli, pepper, etc) healthy or harmful?
factoid	How much protein do we really need?
	How many whole eggs per day would be considered safe?
list	What are some ways to stop and prevent acne?
	What senses are active while sleeping?
summary (simple)	Effect of the common cold on the immune system
	Why do Humans not produce Vitamin C like other mammals?
summary (double)	Do drinks like coke zero, pepsi black, diet coke etc actually have 0 sugar and 0 cals? If yes, then what is the downside to such drinks?
	What (actually) is cholesterol and why does it matter?
summary (triple)	What is the deal with apple cider vinegar? Why is it part of most ”cleanse” or ”detox” diets? I’ve read that it stabilizes blood sugar. Should people drink it everyday?

4.2 Characterization and features

Snippets

BiQA2 was built by using BiQA's validated queries and matching abstracts that were considered relevant. Snippets of text that contain information capable of answering the corresponding query were selected from such abstracts and are included in the feature 'snippets' for that query, similar to the format in BioASQ.

Taking into account the three forums, the work developed in this thesis yielded 1157 abstracts containing relevant snippets of text. These correspond to 640 questions (53.33%) from BiQA, i.e., questions for which exists, at least, 1 snippet capable of producing an answer. Therefore, questions that did not have any snippets are not included in BiQA2. The same abstract can contain multiple snippets, hence the number of resulting snippets can exceed the total number of abstracts considered relevant. Almost 1500 snippets of text were extracted from the abstracts. Table 4.3 shows the breakdown per forum regarding the number of snippets in BiQA2.

Table 4.3: BiQA2 Questions and Abstracts per forum.

	Biology		Medical		Nutrition		Total	
	#	%	#	%	#	%	#	%
Abstracts with snippets | % All Abstracts analyzed from BiQA	340	39.63	490	49.00	327	38.07	1157	42.58
Questions | % All Queries analyzed from BiQA	204	51	238	59.5	198	49.5	640	53.33
Snippets extracted from documents	388	-	705	-	400	-	1493	-

The average number of snippets per question is two. There is only one significant outlier associated with the question 'Effect of cigarettes on passive smokers' with 48 snippets. The medical forum has substantially more snippets than the other forums. The distribution of the number of snippets per question and the approximate number of words per snippet is shown in Figure 4.1.

Question Type

The type of question was also a recorded feature (according to the categories in BioASQ) but with the modifications described in section 4.1. Summary and yes/no/uncertain type questions account for around 90% of all the validated questions, followed by residual amounts of factoid (5.7%), list (4%) and choice (1.4%). These proportions are very similar across the three forums (Figure 4.2).

Question exact answer + Snippet exact answer

The BioASQ dataset contains an ”exact answer” per question (except for the summary type), which means that every question that contains at least one snippet will have an ”exact answer” as per the BioASQ definition. In BiQA2, however, there was an attempt to be more flexible by also making available the

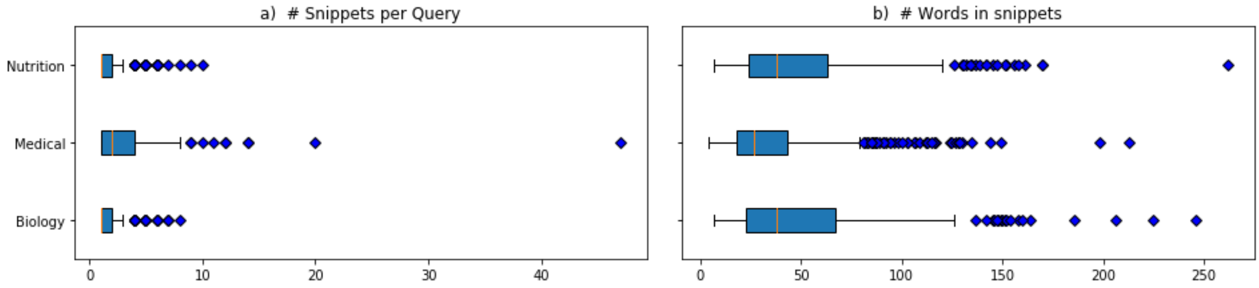


Figure 4.1: Snippets statistics. a) distribution of number of snippets per question; b) distribution of number of words in snippets

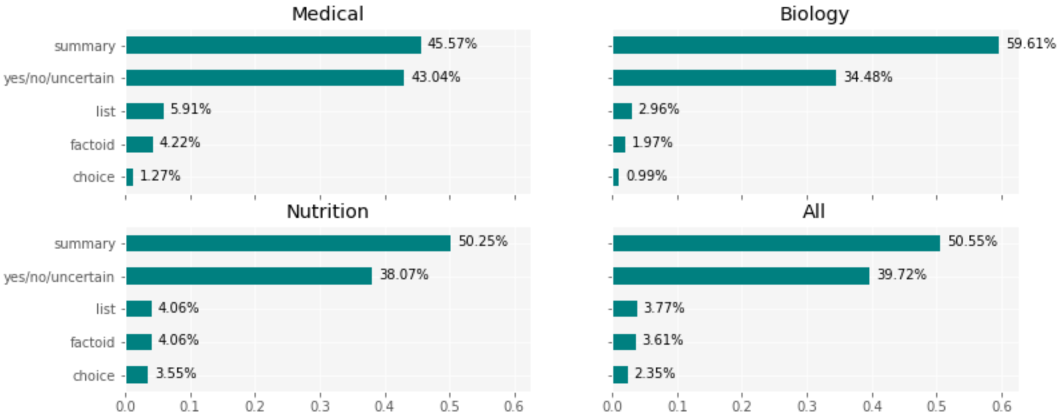


Figure 4.2: BiQA2 question type per forum

exact answer of the specific snippet in the case of yes/no/uncertain and choice type of questions, which does not happen in BioASQ. This is necessary as some abstracts point towards one answer whilst others (referring to the same query) suggest an opposite or contradictory view. Furthermore, it was also acknowledged the existence of abstracts where the authors conclude that there isn't enough information to reach a definitive conclusion regarding the query topic. For this reason, particularly in yes/no questions, there will be some questions where the 'global exact answer' is uncertain but we can still find abstracts - corresponding to that same query - containing snippets that can provide within them a yes or no or even uncertain "exact" snippet answer. Table 4.4 discriminates the amount of yes/no/uncertain classification attributed to the query as a whole and to individual snippets.

Table 4.4: yes/no/uncertain question type samples

type	queries	snippets
yes	282	392
no	106	139
uncertain	87	75

4.3 Validation with Deep Learning models in BioASQ tasks

In order to assess the potential of BiQA2 as a valid *corpus* for training of biomedical QA systems, it was decided to generate DL models by implementing architectures that have previously competed at Task B Phase A of the BioASQ challenge.

The implemented experiments, described in the section below, address the tasks of Document re-ranking and Snippets Retrieval where different combinations and versions of the BioASQ, BiQA and BiQA2 are utilised in order to explore BiQA2's suitability to train models that fulfil the objectives of such tasks. For every experiment, there were independent train, validation and test sets in an attempt to prevent data leakage. For each task, the data utilized will be explained in detail.

From a standpoint of implementation, Python was the chosen programming language throughout the thesis given its versatility and widespread usage in NLP research field where an abundance of domain optimized libraries are available.

4.3.1 Document Re-ranking with AUEB's system

4.3.1.1 TERM-PACRR implementation and modelling details

To compare BiQA2 with both BioASQ and BiQA, it was decided to apply BiQA2 into the same system used by Lamurias et al. [24] to validate and test BiQA as a training set for a document re-ranking model.

The referred system has been described in detail in section 2.4.2. It was developed by the AUEB team as part of the 6^{th} edition of the BioASQ competition in 2018 [5]. Their full system consists of two distinct convolutional based neural networks that tackle both sub-tasks of re-ranking a set of documents and extraction of snippets from said documents. These sub-tasks are part of Phase A, Task B in the BioASQ challenge.

The experiments conducted with the AUEB's system were based on the team's publicly available code at GitHub [2]. For the document re-ranking stage, their best performing model was an arquitecture named TERM-PACRR. This was the neural network implemented by Lamurias et al. [24] to generate a document re-ranking model using BiQA.

[2] https://github.com/nlpaueb/aueb-bioasq6

It consists of a supervised learning convolutional network-based system that performs features extraction on a similarity matrix computed between the query and the abstract text which are then passed to a final linear layer that attributes a score on the abstract as a proxy of relatedness to the query.

This step is called re-ranking because, at inference time, the relevance score is used to attribute a new rank to previously retrieved documents that have been associated with the query by a faster or computationally cheaper algorithm in a search engine. In practice, the AUEB system very first step consists of retrieving 100 documents per query, using the BM25 scoring algorithm. Only then the document re-ranking trained CNN will score each abstract text leading to a re-ranking of the 100 documents in their relatedness to the query. The system then retains the top N documents most relevant to the query, which in the case of the BioASQ challenge, is 10 documents.

In summary, the input data into TERM-PACRR is a ranked list of 100 documents that results from a BM25 algorithm search on each query.

Training data

BioASQ, BiQA and BiQA2 can all be a source of training data for the specific task of re-ranking a set of documents according to its relevance to a certain query. As training data for this task, we performed experiments with all these *corpora*. However, to understand the nuances of the training experiments, it is necessary to first understand how the AUEB's team used the BioASQ *corpus* to build the positive and negative samples of labelled training data to the network as this will have an impact on the number of samples in each version of the experiments.

The AUEB's team GitHub repository contains data extracted from the BioASQ6 dataset which was converted to an input format compatible to their model. More specifically, they used BioASQ training data from years 1-5 and batch 5 of year 5 as a development set which was used to tune the models when selecting the best epochs of the training loops. Their available data for training consists of BioASQ6 queries. For each query 100 documents/abstracts were retrieved from PubMed using the BM25 algorithm. From this set of ranked documents, training positive and negative QA pairs are then built as input to the network.

For each query in the dataset, the AUEB team considered as positive samples the abstracts that are simultaneous present in the retrieved BM25 ranked list and in the BioASQ dataset for that query. The negative samples for the query are then randomly picked from the remaining documents in the initial 100 documents BM25 retrieved list. This means that, for each positive abstract, there is a negative one making it a balanced set approach for training. Also to note is the fact that the title of the abstract is concatenated to the abstract text as input of the document to the network.

Lamurias et al. [24] also applied the BM25 algorithm search but on a local copy of PubMed to retrieve 100 documents on each BiQA set query.

The AUEB authors claim that setting 100 documents for the initial retrieval is enough for the BM25 system to return the majority of documents the BioASQ dataset identifies as relevant for that query. However, this might not the case in BiQA or BiQA2. A substantial amount of relevant documents in both BiQA and BiQA2 might be left out as samples for training due to this mismatch.

Test data

The test data used to assess the performance of models trained with different datasets are the batches made available at the AUEB's repository which corresponds to the BioASQ test batches 1-5 of the BioASQ6 challenge. These data are used to test all the resulting trained models of this thesis. Each batch has 100 questions from BioASQ. The results of each trained model on the test sets will be reported per batch.

Experiments

The first models generated used the architecture named TERM-PACRR, and trained it with the positives and negatives sampling approach detailed above. This was applied to BiQA2 and also to both BiQA and BioASQ datasets in order to benchmark their performance to BiQA2.

Different versions of the datasets were compiled as training data for the TERM-PACRR model with distinct amounts of training samples for each version - this is expressed in Table 4.5. The details of each experiment is as follows:

- BioASQ - using only the provided BioASQ6 set from AUEB GitHub repository as a baseline comparison to the other *corpus*. This set is a list of 100 Pubmed abstracts retrieved by the BM25 algorithm for each question in the BioASQ set.

- BiQA - also used as a baseline and in an attempt to replicate the results by Lamurias et al. [24]. In this version, only the original BiQA dataset queries are used to retrieve 100 related abstracts by using the BM25 algorithm in a local copy of Galago[3]. Every document is scored by the BM25 and each document is then labelled as positive if it was both retrieved by BM25 and is associated with the corresponding query in BiQA. This means that if there is no match between the abstracts retrieved and the ones present in BiQA, this question will not be used given that no positive abstract can be fed to the network. The negative samples were picked randomly from the remaining abstracts of the retrieved.

- BiQA2 - for the document re-ranking sub-task, BiQA2 offers an attempted to refine which abstracts can be considered more related (i.e. effective positive samples) to train this type of network. The same approach was carried out in terms of the Galago documents retrieval search as for the BiQA set.

[3] Galago is a search engine toolkit written in Java, specifically designed for research by Croft et al. [6]. It consists of various search engine components which are pluggable for indexing and retrieval.

- BioASQ + BiQA - it is yet another implementation of a baseline benchmark trying to replicate the results of Lamurias et al. [24] by joining both sets.
- BioASQ + BiQA2 - the merge of both BioASQ and BiQA2 datasets to try to improve the model performance compared to each dataset separately.

Table 4.5: Baseline experiments for TERM-PACRR training. 'pos samples': number of positive samples yielded by each *corpus* according to the methodology in TERM-PACRR; 'queries in *corpus*': number of total available queries in each *corpus*; 'queries in training': number of queries from which the positive samples are computed.

training data	pos samples	queries in corpus	queries in training
BioASQ	15017	2151	1997
BiQA	1229	7218	998
BiQA2	168	627	131
BioASQ + BiQA	16246	9369	2995
BioASQ + BiQA2	15185	2778	2128

Note that the samples expressed in Table 4.5 are only the positive ones so the total positives and negatives are double that value. But still, it is very clear that the sampling approach taken by the AUEB group yields very few samples to both BiQA and BiQA2 compared to their potential positive examples. The number of queries that actually made it to training is substantially less than the ones present in the *corpus* (both in the case of BiQA and BiQA2). Given this reduced number of samples, further models were generated using the entire potential of both sets for positive samples in training. This was achieved by making sure all abstracts in BiQA or BiQA2 are added into the returned BM25 ranked list of documents, in case they were not returned by BM25 search. For the purposes of this thesis, these models are refereed to as models with 'ideal' inputs. The difference in this approach - of producing the samples data - to the one implemented by Lamurias et al. [24] and the AUEB team, is that the documents in BiQA that were left out - because of an absence of a match between the list retrieved from the Galago search and the ones in BiQA - are now included as input positive abstracts to train the model. Note that they are added to the list of scored retrieved documents after the Galago search, which results in lists of more than 100 documents, i.e. no search engine retrieved document is excluded. Also, it was necessary to artificially provide high BM25 scores to the 'ideal' input documents as these are taken as extra features added into the final scoring layers. These artificial scores were computed by taking the maximum score in all retrieved documents and increasing that value on each added relevant document. The additional experiments and respective samples can be seen in Table 4.6.

Furthermore, it was decided to run additional experiments for a direct comparison between BiQA and BiQA2 as described below:

- Q$_{BiQA2}$ + A$_{BiQA}$- The training set is made exclusively of the questions common to BiQA and

Table 4.6: Document re-rank model training datasets versions with 'ideal' positive samples

training data	pos samples	queries in corpus	queries training
BiQA ideal	12841	7225	6854
BiQA2 ideal	1155	639	637
BioASQ + BiQA ideal	27858	9376	8851
BioASQ + BiQA2 ideal	16172	2790	2634

BiQA2 (Q_{BiQA2}) but with the answers/abstracts from BiQA (A_{BiQA}). The goal is to assess if the abstracts selected as relevant to the questions curated in BiQA2 actually perform better when compared to the original set of abstracts present in the same questions in BiQA. We conduct experiments applying the BM25 approach above but also the methodology of 'ideal' inputs, thus forcing all samples of both datasets to be included.

- Q_{BiQA} + A_{BiQA2} - This version is a compilation of the BiQA *corpus* where the answers/abstracts of the questions that are common to BiQA and BiQA2, are replaced by the ones present only in BiQA2. In other words, all the abstracts from the questions used in BiQA to create BiQA2, are replaced with the correspondent abstracts/QA pairs from BiQA2. Once again, both versions of the BM25 method only and 'ideal' sampling were applied.

Table 4.7: Additional Document re-rank models

training data	pos samples	queries in corpus	queries training
Q_{BiQA2} + A_{BiQA}	195	706	150
Q_{BiQA2} + A_{BiQA} ideal	1721	706	699
Q_{BiQA} + A_{BiQA2}	1220	7218	992
Q_{BiQA} + A_{BiQA2} ideal	12270	7225	6807

These experiments should give some insights regarding the strength and validation of the annotations in BiQA2 as a refinement methodology for the relevant documents in BiQA.

TERM-PACRR training implementation, hyperparameters and epochs

TERM-PACRR used the GenSim implementation of the word2vec model to build the word embeddings on 28 million articles of PubMed collection. These embeddings were comprised of 200 dimensions and were not updated when training our datasets. The tokenization of queries and abstracts was performed by the 'bioclean' tool made available by the BioASQ challenge.

In all experiments, the tokenized queries were truncated to a length of 30 tokens in the queries and 300 tokens in the abstract texts. Padding was applied to both queries and abstracts in order to produce a uniform similarity matrix as input to the convolutional network. The number of convolutional filters was 16 in all training sets, with a 3x3 window. The final dense neural layer was comprised of 50 units (neurons).

As per the original paper by the AUEB team, additional traditional IR features were inputted into the final linear layer of TERM-PACRR which combines the q-term scores. These additional features are the BM25 score, uni-gram and bi-gram word overlap count between query and document) and IDF representation of the query. In case any IDF value is missing it assumes the maximum IDF value available. It was decided to use the IDF scores already computed and made available by the AUEB group since this was also the approach by Lamurias et al. [24].

The training of TERM-PACRR was performed in each experiment for 100 epochs. The provided BioASQ development set was used at the end of each epoch to evaluate the resulting model using the MAP scoring which always assumes 10 relevant documents - the maximum permitted by the BioASQ competition. The procedure described in the AUEB paper (and simultaneously followed by Lamurias et al. [24]) consisted of training a combination of 10 runs of each model.

4.3.1.2 TERM-PACRR results

This section presents the results from training the document re-ranking model TERM-PACRR of the AUEB's team presented to tackle task B Phase A of BioASQ 6 competition. The evaluation measure is the competition's own version of Mean Average Precision (MAP) of the 10 most relevant documents. As explained above, the resulting models are evaluated in the 5 test batches made available in BioASQ6.

Table 4.8: TERM-PACRR model for Document re-ranking MAP results per test batch

training set	batch 1	batch 2	batch 3	batch 4	batch 5	mean
BioASQ	0.123	**0.112**	0.11	0.096	**0.064**	0.101
BiQA	0.11	0.098	0.097	0.09	0.051	0.089
BiQA2	0.106	0.092	0.094	0.086	0.053	0.086
BioASQ + BiQA	**0.124**	**0.112**	0.111	**0.098**	0.063	0.101
BioASQ + BiQA2	**0.124**	**0.112**	**0.113**	**0.098**	0.063	**0.102**
BiQA ideal	0.105	0.088	0.088	0.081	0.049	0.082
BiQA2 ideal	0.104	0.089	0.09	0.082	0.052	0.084
BioASQ +BiQA ideal	0.104	0.089	0.090	0.086	0.054	0.085
BioASQ + BiQA2 ideal	0.115	0.100	0.096	0.087	0.056	0.091
$Q_{BiQA2} + A_{BiQA}$ ideal	0.104	0.092	0.091	0.085	0.052	0.085
$Q_{BiQA2} + A_{BiQA}$	0.110	0.097	0.095	0.089	0.055	0.089
$Q_{BiQA} + A_{BiQA2}$ ideal	0.105	0.089	0.088	0.081	0.049	0.082
$Q_{BiQA} + A_{BiQA2}$	0.112	0.100	0.094	0.090	0.052	0.090

The results relative to training with BioASQ only are analogous to the ones submitted by the AUEB's team for BioASQ6 [5]. However it was not possible to replicate the results reported by Lamurias et al. [24] in regards to the performance of BiQA as training data. The most likely reason for this is the fact that the test sets used by Lamurias et al. [24] were slightly different than the ones provided by the AUEB's repository. Lamurias et al. [24] used the enriched version of the test batches in BioASQ6 which is a version compiled after the submission of models from all teams. The experts panel takes into consideration the documents retrieved by the different competitors and this results in an extension of the list of documents for certain queries making it an enriched set compared to the original test sets for that particular year of the challenge.

The average MAP scores for all batches of the batch sets was computed and used to observe the percentage difference in the mean scores between all experiments (figure 4.3.

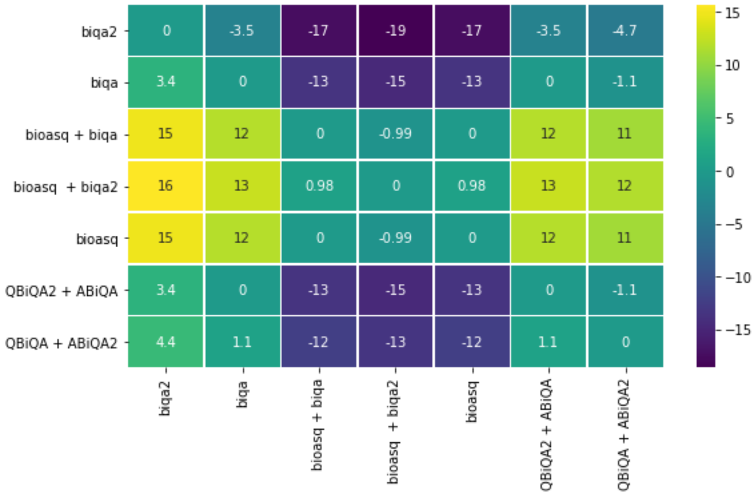


Figure 4.3: % difference in mean MAP Scores

Overall it is clear that - when used alone using the BM25 filter methodology from AUEB's team system - both BiQA and BiQA2 alone demonstrate lower performance than the BioASQ dataset. However, MAP scores are higher when BiQA or BiQA2 are used in combination with the BioASQ set. In fact, the best MAP score is achieved with the combination of BioASQ and BiQA2. The results from BiQA and BiQA2 are very similar, with BiQA being only 3% above BiQA2 but with a substantially less amount of training samples. Surprisingly, in the experiments where all the potential positives from BiQA and BiQA2 (training sets named 'ideal'), the performance actually was lower when compared to the BM25 approach of AUEB's, despite the increased availability of training samples of both sets. 'BiQA', 'BiQA2', 'BioASQ + BiQA', 'BioASQ + BiQA2' yielded higher MAP scores than their respective 'ideal' counterpart experiments.

The results with 'Q_{BiQA2} + A_{BiQA}' should only be directly compared with the results from BiQA2 because they are both using the exact same queries but with the refined answers in BiQA2. In this case, BiQA2 shows slightly better performance on batches 2 and 3 but lower on the others. This is also verified in the 'ideal' implementations. Similarly, the results of 'Q_{BiQA} + A_{BiQA2}' should be analyzed in the context of training with documents from BiQA only. Here the scores are marginally higher for 'Q_{BiQA} + A_{BiQA2}' (0.6%) when compared to BiQA.

4.3.2 Snippets retrieval with AUEB's system

4.3.2.1 TERM-PACRR + BCNN implementation and modelling details

BiQA2 tries to expand on BiQA by also providing sentences for training of models associated with the snippets retrieval stage of the BioASQ phase A task B.

As detailed in section 2.4.2, in association with the TERM-PACRR model, the AUEB team designed an additional model capable of extracting sentences from the re-ranked documents/abstracts. After re-ranking and retrieving the 10 most associated documents using TERM-PACRR, the system uses a secondary and differentiated convolutional-based neural network model to retrieve sentences that are most associated with the query - architecture named Basic Convolutional Neural Network (BCNN). The system's architecture does this by framing the task also as a ranking problem. It captures all the sentences in the 10 re-ranked documents and attributes a score to each of them as a proxy of relevance. As per the rules in the BioASQ competition, the 10 most likely sentences related to a query are retrieved. In this thesis, this system was also implemented to test the capability of BiQA2 as training data for snippets retrieval systems.

Training Data

The AUEB GitHub repository contains a train set built from the BioASQ6 *corpus* converted into a format for input into their BCNN model. This dataset was used as a training benchmark with which to compare BiQA2 training potential for BCNN. According to the authors in [5] the sentences contained in this training, set are the output of splitting all the relevant documents re-ranked by the document re-ranking model into sentences - having '.' as a separator. They then consider the sentences that overlap with BioASQ gold snippets as the positive samples and the other ones as negative. Although the authors do not mention if the resulting samples are balanced or non-balanced, by perusing their code, it becomes clear that for each positive sample there is only one negative sample sentence, randomly picked from the remaining abstract sentences that do not match the gold snippets. Of note is the fact that the BioASQ guidelines demand the panel of experts to consider as gold snippet one or more consecutive sentences from the abstracts. If more than one non-consecutive sentences exist in the same abstract they are considered as independent gold snippets.

The BiQA2 dataset was converted into the same format for input into BCNN and respective training. This was achieved by taking the snippets in BiQA2 and respective queries as positive samples. The negative sentences were randomly picked from the abstracts from which the positive samples originated.

In case there was no text remaining from the abstract then the negative sentence is a random sentence from any other abstract in the dataset.

In summary, BCNN was trained separately with the available BioASQ data (for benchmark), BiQA2 and both joined together. The number of samples resulting from the methodology described above can be seen in Table 4.9 :

Table 4.9: Number of samples in BCNN training

training data	pos samples	questions	total samples
BioASQ	21875	1615	43750
BiQA2	1490	643	2980
BioASQ + BiQA2	23365	2258	46730

The input format of a sample into BCNN is two sequences of terms: the query and the sentence from the document (positive or negative).

Training implementation

The implementation of training also followed the code implementation in AUEB's GitHub repository[4]. The overall architecture was described in section 2.4.2. It uses Python and TensorFlow version 1. The word embeddings also used embeddings already available which were pre-trained by applying word2vec to 28 million MEDLINE/PubMed article abstracts. These embeddings consist of 200 dimensions. The tokenization of queries and snippets is performed with NLTK's English splitterlibrary [4] where the queries were allowed to have any number of tokens, however, snippets are truncated to a maximum of 40 tokens counted from the beginning of the sentence. Training used a binary log-loss and the AdaGrad gradient descent optimization algorithm with a fixed learning rate of 0.08 and an L2 regularization method with a λ of 0.0004. 50 convolutional filters were set with windows of 4 for each convolutional layer. During each epoch, the similarity scores feature retrieved by the neural network and accumulated in a list which, at the end of each epoch, is concatenated and applied to a Logistic Regression layer.

The batches size was 32 for the three training datasets and 50 epochs were computed.

Similarly to TERM-PACRR, additional features are included in the final layer of the model. In BCNN these features are the binary word overlap and bi-gram overlap between the query and snippet.

The authors of BCNN improved the performance of the model by applying a post-processing method where they retain only the best scored Ks snippets for each query and then re-rank those snippets by the relevance scores of the documents they came from. Since the objective was not to achieve a better model - but solely to compare the performance of datasets onto the same model trained with the same criteria and parameters - it was decided not to implement this approach which might impair the ability to compare our training with the BioASQ dataset with the results from their paper.

[4] https://github.com/nlpaueb/aueb-bioasq6/tree/master/models/snippets/ABCNN

Furthermore, in neural networks training, it is standard practice to have a validation set with which an evaluation for the network loss and analysis of metrics are computed to guide the training process. However, in the particular experiments with this system, there was no validation data as the AUEB GitHub repository did not show any code for this purpose. In fact, the only mention of this matter, in the original paper, was merely a statement that the best epoch in training was later selected in the test data.

Testing

The testing of BCNN training is done taking into consideration the whole AUEB system built for task B, phase A of the BioASQ competition. This means that the trained BCNN performs snippet extraction from the 10 documents yielded by the Document re-ranking stage, i.e. from the output of the TERM-PACRR model. In BCNN, at inference time, documents returned by TERM-PACRR are split into sentences, ranked, and sorted by the system. To that end, the testing performed for each of the 3 trained models (BiQA2, BioASQ, and BiQA2 + BioASQ) is done taking as input the resulting documents from our experiments with BioASQ, BiQA, and BiQA2 from the TERM-PACRR model. The test batches 1-5 passed through the document re-ranking stage of the system are the input testing data for this second stage of the snippet retrieval. The resulting snippets are then evaluated by the BioASQ tool which uses MAP scoring as a metric according to the gold snippets.

4.3.2.2 TERM-PACRR + BCNN results

Table 4.10 shows the MAP scores results of training the BCNN model- from the AUEB's team - using BiQA2, BioASQ, and BioASQ+BiQA2 available snippets.

The test batches first pass through the trained TERM-PACRR models (column 'Doc_rerank output' indicates the training data used) and only then have snippets extracted from the documents with the BCNN model trained in either BiQA2, BioASQ, or a merge of both (column 'BCNN train Dataset').
The MAP scores achieved by using the BioASQ dataset as training for both TERM-PACRR and BCNN replicate the ones obtained by the AUEB team in their submission to the snippet retrieval stage of the BioASQ competition [5].

These results show that training with BiQA2 alone produces the lowest scores in all batches of test data. The best MAP scores are overwhelmingly achieved when training BCNN with the BioASQ set alone. Despite the occasional increase in performance in, for instance, batch 2 (when using BiQA2 output from the document re-rank) it is very clear that, globally, the performance does decrease when joining both BiQA2 and BioASQ.

Table 4.10: MAP scores for BCNN model

Doc_rerank output	BCNN training data	batch 1	batch 2	batch 3	batch 4	batch 5	Mean
BioASQ	BiQA2	0.054	0.037	0.066	0.044	0.018	0.044
BioASQ	BioASQ	**0.084**	0.064	0.102	**0.071**	**0.037**	**0.071**
BioASQ	BiQA2+BioASQ	0.068	**0.072**	0.106	0.064	0.031	0.068
BiQA2_ideal	BiQA2	0.048	0.029	0.056	0.043	0.013	0.038
BiQA2_ideal	BioASQ	0.080	0.063	0.095	0.064	0.025	0.066
BiQA2_ideal	BiQA2+BioASQ	0.070	0.061	0.106	0.062	0.023	0.064
BiQA2	BiQA2	0.050	0.033	0.068	0.042	0.016	0.042
BiQA2	BioASQ	0.079	0.059	0.097	0.065	0.031	0.066
BiQA2	BiQA2+BioASQ	0.063	0.062	**0.111**	0.061	0.030	0.066
BiQA_ideal	BiQA2	0.054	0.028	0.055	0.040	0.014	0.038
BiQA_ideal	BioASQ	0.083	0.061	0.096	0.064	0.026	0.066
BiQA_ideal	BiQA2+BioASQ	0.073	0.058	0.111	0.060	0.024	0.065
BiQA	BiQA2	0.051	0.031	0.064	0.044	0.015	0.041
BiQA	BioASQ	0.079	0.062	0.096	0.065	0.029	0.066
BiQA	BiQA2+BioASQ	0.064	0.067	0.110	0.065	0.026	0.067

4.3.3 Snippets retrieval with TERM-PACRR + BERT

4.3.3.1 TERM-PACRR + BERT implementation and modelling details

The system by AUEB's team, submitted to BioASQ6, is based on CNNs. Given the current state-of-the-art transformers-based approach to NLP in general and QA in particular, we also experimented with BERT-based models by replacing the BCNN model in the AUEB's team pipeline with pre-trained biomedical BERT models from the HuggingFace library [44]. The Document re-ranking TERM-PACRR model was still used to feed the resulting documents into BERT-based models to retrieve snippets to assess if the results are in line with the BCNN approach. Given that the most current self-attention state-of-the-art techniques and their application into transformers models for natural language processing has been achieving tremendous success it was decided to experiment with BiQA2 with the expectation of achieving better results than the ones observed with the BCNN architecture for Snippet retrieval.

The most recent BioASQ competition (BioASQ 8) has few submissions to task B phase A with the application of transformers. One of them is presented in the work by Kazaryan et al. [21]. The authors describe several experiments with the well-know Bidirectional Encoder Representations from Transformers (BERT) architecture both for document re-ranking as well as for snippets retrieving confirming that the BioASQ dataset can be effectively used to train a transformer-based natural language model. They achieve very high MAP scores in both tasks of phase A task B. However there was no implementation made available by the authors and their descriptions in the paper isn't clear on most of the parameters and lacks the necessary details to attempt replicating the model.

Therefore, in this thesis, the implementation of a model of this nature, to be trained with BiQA2, attempts to mimic the work by Nogueira and Cho [31]. They have 're-purposed BERT as a passage re-ranker and achieved state-of-the-art results on the MS MARCO passage re-ranking task'.

BERT Snippets retrieval training implementation

Similar to the BCNN system, the snippets retrieval task was framed as a re-ranking problem of the sentences contained in the documents also re-ranked from an N list of related documents retrieved by a standard search engine mechanism such as BM25. Here BERT is applied to the concatenated question-snippet sequence[5].

The objective of the re-ranking algorithm is to estimate a score of how relevant a candidate passage is to a certain query. The input is the query as sentence A and the passage text as sentence B. The text is truncated such that the concatenation of query, passage and separator tokens have the maximum length of 256 tokens. We use a BERT as a binary classification model, that is, we use the '[CLS]' vector as input to a single layer neural network consisting of 512 neurons to obtain the probability of the passage being relevant. This probability is computed for each passage independently where a final list of passages is gathered by ranking them with respect to these probabilities. The 10 most relevant sentences to a query are selected. Training starts from a pre-trained BERT model and fine-tunes it to the re-ranking task using the cross-entropy loss. ADAM gradient descent optimizer is used with a learning rate of $1x - 5$. This fine-tuning of BERT is performed throughout 6 epoch iterations with a batch size of 5 sentences on a single GPU with 8 GB of memory.

The BERT models used in these experiments are freely available from the HuggingFace library which is an open-source repository of transformers models readily and easily available for public usage. This is now a standard library where all kinds of pre-trained BERT models can be downloaded and fine-tuned for specific tasks. In our particular case, given the hardware limitations, it was decided to use the BERT-base uncased model. It consists of 12 layers (transformer blocks), 12 attention heads, and 110 million parameters. According to the information on Devlin et al. [7] it was pre-trained on a large *corpus* of English data in a self-supervised fashion, i.e. on the raw texts only, with no humans labelling them in any way and is primarily aimed at being fine-tuned on tasks that use the whole sentence (potentially masked) to make decisions, such as sequence classification, token classification or question answering.

In addition to BERT-base, we also experimented with the biomedical domain-specific model PubMedBERT_base_uncased for abstracts from Microsoft. This model was pre-trained on biomedical text from scratch and, according to pre-training *corpus* comprises 14 million PubMed abstracts with 3 billion words (21 GB), after filtering empty or short abstracts.

These models at HuggingFace offer their own tokenization process and embeddings. BERT uses WordPiece tokenization. The vocabulary is initialized with all the individual characters in the language,

[5] This thesis implemented its own version of BERT's architecture into snippets retrieval based on the TensorFlow 2 approach by https://github.com/airKlizz/MsMarco.

and then the most frequent/likely combinations of the existing words in the vocabulary are iteratively added. For the implementation in this thesis, the 'encode_plus' method was utilized which contains the encoded sequence and also the mask for sequence classification and the overflowing elements given that a maximum number of tokens is specified.

No layers were frozen during fine-tuning. All the pre-trained layers along with the task-specific parameters are trained simultaneously.

Training and validation data

The exact same data utilized for training of BCNN was used for these experiments (table 4.9). BiQA2, BioASQ and both joined together were independently used to train the BERT models. However, contrary to the training with the AUEB's system, a validation dataset was built from data available at the AUEB's GitHub repository. This validation set is comprised of 2216 query-sentence pairs. By the end of each epoch the performance of the models' weights was measured by computing the accuracy in this set. Evidently the best performing epoch at the validation set will be the one to use when applying it to the test batches.

4.3.3.2 TERM-PACRR + BERT results

The overall MAP results related to the snippets retrieval task maintain the same pattern of scores in both BCNN and BERT models. In the vast majority of batches and regardless of the input documents from the Document re-rank stage, the scores are higher when the models are trained on the BioASQ *corpus* alone. Merging BiQA2 with BioASQ for training consistently decreases the resulting model's performance. Not surprisingly, the results are indeed better with a BERT model when compared to the convolutional network approach with the best results being achieved by the fine-tuning of the domain specific pubmedBERT model pre-trained in PubMed abstracts.

The overall MAP results related to the snippets retrieval task maintain the same pattern of scores in both BCNN and BERT models. In the vast majority of batches and regardless of the input documents from the Doc re-rank stage, the scores are higher when the models are trained on BioASQ alone. Merging BiQA2 with BioASQ for training consistently decreases the resulting model's performance. Not surprisingly, the results are indeed better with a BERT model when compared to the convolutional network approach with the best results being achieved by the fine-tuning of the domain specific pubmedBERT model pre-trained in PubMed abstracts.

4.3.4 'yes/no' BioASQ phase B

Additionally to the tasks of document re-ranking and snippets retrieval, BiQA2 has also been used as training data as part of a master thesis work by a colleague that is currently working in the research

Table 4.11: Snippets retrieval MAP scores for training with Bert_base_uncased

Doc_rerank output	BERT_base training data	batch 1	batch 2	batch 3	batch 4	batch 5	Mean
BioASQ	BiQA2	0.044	0.050	0.092	0.047	0.021	0.051
BioASQ	BioASQ	**0.082**	**0.080**	**0.107**	**0.065**	**0.044**	**0.076**
BioASQ	BiQA2+BioASQ	0.062	0.075	0.100	0.059	**0.044**	0.068
BiQA	BiQA2	0.042	0.043	0.090	0.046	0.018	0.048
BiQA	BioASQ	0.073	0.079	0.106	0.061	0.033	0.071
BiQA	BiQA2+BioASQ	0.059	0.068	0.098	0.054	0.034	0.063
BiQA_ideal	BiQA2	0.046	0.042	0.093	0.043	0.021	0.049
BiQA_ideal	BioASQ	**0.082**	0.071	0.104	0.059	0.033	0.070
BiQA_ideal	BiQA2+BioASQ	0.067	0.067	0.092	0.056	0.032	0.063
BiQA2	BiQA2	0.041	0.047	0.091	0.044	0.019	0.049
BiQA2	BioASQ	0.067	0.074	0.110	0.060	0.037	0.070
BiQA2	BiQA2+BioASQ	0.056	0.063	0.105	0.056	0.036	0.063
BiQA2_ideal	BiQA2	0.043	0.043	0.092	0.044	0.020	0.049
BiQA2_ideal	BioASQ	0.072	0.077	0.105	0.060	0.034	0.070
BiQA2_ideal	BiQA2+BioASQ	0.062	0.068	0.095	0.054	0.034	0.063

Table 4.12: Snippets retrieval MAP scores for training with pubmedBert_base_uncased

Doc_rerank output	PubmedBERT training data	batch 1	batch 2	batch 3	batch 4	batch 5	Mean
BioASQ	BiQA2	0.064	0.068	0.094	0.072	0.043	0.068
BioASQ	BioASQ	**0.079**	**0.086**	0.101	**0.075**	**0.045**	**0.077**
BioASQ	BiQA2+BioASQ	0.074	0.079	0.089	0.071	0.041	0.071
BiQA	BiQA2	0.058	0.061	0.093	0.071	0.041	0.065
BiQA	BioASQ	0.074	0.082	0.102	0.069	0.037	0.073
BiQA	BiQA2+BioASQ	0.075	0.074	0.085	0.066	0.032	0.066
BiQA_ideal	BiQA2	0.063	0.053	0.093	0.066	0.035	0.062
BiQA_ideal	BioASQ	0.089	0.073	0.102	0.069	0.032	0.073
BiQA_ideal	BiQA2+BioASQ	0.078	0.076	0.088	0.065	0.028	0.067
BiQA2	BiQA2	0.054	0.057	0.094	0.071	0.038	0.063
BiQA2	BioASQ	0.078	0.074	0.103	0.068	0.039	0.072
BiQA2	BiQA2+BioASQ	0.071	0.069	0.093	0.064	0.036	0.067
BiQA2_ideal	BiQA2	0.059	0.057	0.097	0.069	0.036	0.064
BiQA2_ideal	BioASQ	0.078	0.077	**0.104**	0.069	0.033	0.072
BiQA2_ideal	BiQA2+BioASQ	0.072	0.077	0.091	0.064	0.028	0.067

group LASIGE [6]. This group is a frequent contributor and participant in the BioASQ competition.

[6] LASIGE is a research and development (R&D) unit at the Faculty of Sciences of the University of Lisbon (FCUL), in the field of Computer Science and Engineering.

Three models - named ’model direct’, ’model 256’ and ’model 512’ - were developed in the context of the ’yes/no’ sub task of Task B phase B at the BioASQ competition. In this task the models should generate a ’yes’ or ’no’ exact answer as per the competition’s rules. All models developed by our colleague are based on the BERT architecture but innovate on the design of layers that are added to the original BERT’s layers output. In ’model direct’, the output of BERT’s [CLS] token goes directly to a softmax function for classification of ’yes’ or ’no’, whereas in ’models 256’ and ’512’ the output of BERT is fed to a 256 or 512 fully connected neural layer, respectively. The generation of models was conducted having as baseline for the BioASQ9 challenge training set for the ’yes/no’ snippets classification task from Task B Phase B. Further experiments were conducted by adding the ’yes/no’ snippets - available in BiQA2 - to the BioASQ set in order to assess if the added number of samples could increase the performance of the proposed models. Table 4.13 shows that, in both ’256’ and ’512’ models, BioASQ combined with data from BiQA2 achieved the highest mean Macro F1 scores - which is the competition’s metric for this particular task.

Table 4.13: F1 scores for ’yes/no’ exact answer models

model	training data	batch 1	batch 2	batch 3	batch 4	batch 5	mean
256	BioASQ	0.561	**0.653**	0.832	0.594	0.674	0.663
	BioASQ+BiQA2	**0.688**	0.63	**0.903**	**0.73**	0.674	**0.725**
512	BioASQ	**0.709**	**0.704**	**0.896**	0.519	0.652	0.696
	BioASQ+BiQA2	0.636	0.671	0.805	**0.641**	**0.764**	**0.703**
direct	BioASQ	**0.561**	**0.734**	**0.748**	0.519	0.573	**0.627**
	BioASQ+BiQA2	0.405	0.41	0.734	**0.578**	**0.597**	0.545

Chapter 5

Discussion

The main objectives in this dissertation were the manual verification of a sample of questions and PMIDs retrieved automatically from online public forums – BiQA - as a reliable source of training data for information retrieval and natural language models, and the making of a new dataset – BiQA2 – from such validated QA pairs.

5.1 Queries-Abstracts verification in BiQA and challenges in BiQA2

A sample of 400 queries was selected from each forum using as a criterion the highest total score given to the query by the forum users. These scores – can be assumed - reflect how good, relevant or popular the question is to the forum's community hence higher scores could serve as a proxy for more thoughtful, elaborate, well-constructed, rich, or simply more viewed questions and answers. If these reveal to be of low standard for a dataset of this nature then one should expect the lower scored questions to have even lower quality. The choice of 400 samples per forum was based on the fact that the BioASQ team adds between 400 and 500 samples to their training set every year. 400 questions correspond to nearly 20% of questions in all forums, which should provide enough representativeness of the whole corpus.

Of the questions analyzed only slightly more than 10% were not valid, which, in the context of this thesis, means they are poorly constructed, too ambiguous or with very poor syntax making them not suitable for systems training as they might actually reduce the performance of certain language models. These are queries formulated in such a way that even a human person would struggle to understand the meaning and intent and not consider the attempt at an answer without first asking for clarification. These questions were not included in BiQA2 and were marked as not valid for purposes of verification of the original BiQA. The abstracts corresponding to these queries in BiQA were, consequently, not analyzed and discarded for BiQA2. Nevertheless, even the ones considered valid constituted somewhat a challenge from a grammatical viewpoint hence the characterization of some of those grammatical features. Seventy per cent of queries were considered objective and 52% ambiguous. This was particularly notorious in the Nutrition subset of the corpus where more than 65% of queries were considered ambiguous. The same trend was observed in terms of syntax and structure of queries. From a lexical and grammatical perspective, the Nutrition Forum seems to yield more challenging queries than the Medical or Biology forums. Although the vast majority of questions are simple in their formulation (similar to BioASQ), in BiQA, between 10-15% are built in a compound way or where the query itself contains more than one question.

The PMIDs and corresponding abstracts checked manually, confirm that the BiQA methodology does extract articles currently existing in PubMed. Nearly 95% of PMIDs, present in BiQA, are accounted for in the PubMed API. Inherent to the nature of these types of forums is the more informal or speech-like composition of queries.

The abstracts were classified in their relatedness or ability to provide concrete, direct answers to the respective queries. One of the main insights collected from perusing the abstracts was that, with the exception of the 4% classified as having ’no association’, there were 33% of them considered as having ‘No relevance’. These are abstracts that are not completely unrelated to the query but, for the most part, offer very limited information regarding the query’s full scope or intent. Only about 42% of abstracts offer a comprehensive, direct, clear answer to the complexity of the question posed.

Our findings throughout this exercise identified a few reasons that justify an abstract not being considered relevant for its query:

- The abstract may be related to only a specific topic in the query but not the whole scope or intent of the query. It seems that the abstracts being referenced in the forums answers are, for the most part, used as complementary or auxiliary material for the construction of a rationale by the users answering the questions and not necessarily have a direct quote or sentence with a comprehensive answer. The users that take the time to provide an elaborate answer, in the forum, could be trying to construct a certain argument around the query without responding directly to it, maybe because there might not be an objective or consensual answer anyway. if this is the case they might be using PubMed articles only marginally related to the topics to justify or enrich certain viewpoints expressed about their rationale as a response to the query. This might be happening particularly in queries that are very complex or formulated in a compound manner or that are simply subjective or ambiguous. This translates into a BiQA set which seems mostly populated with abstract texts that are somewhat related to the queries although they might not address its full range of topics or full scope/intent.

- The full article could contain answers but the abstract itself may not. The most paradigmatic case where this happens is in meta-analysis or review articles where the information in the abstract or title hints that the article clearly addresses the query but fails to provide any insights in the abstract itself.

- Not be related in any way to the queries hence having as its theme an entirely different biomedical topic.

The approach in making BiQA2 is very distinct from the BioASQ process of creating samples for their dataset where PubMed articles are gathered first and the information within them is analyzed at a second stage. Independently of being a list, a factoid, a yes/no type of question, their starting point is by perusing abstracts and compiling pieces of text containing scientific facts. Objective questions are then constructed with the goal of addressing directly those facts. The approach in BiQA2 is different. It is not a 'blank sheet' approach as it is conditioned to the existing questions and abstracts. At its core, it is an exercise attempting to curate or select abstracts that would serve best as input training for QA modelling systems. One of the challenges in BiQA2 was the degree of difficulty in understanding domain-specific terms present in the abstracts. It took considerably more time to analyze abstracts related to the question in the Biology forum, for instance. Not seldom, it was doubtful if the abstracts contained answers to the correspondent question in this forum. This was not the case in the Medical set, where doubts were not common. This might not be surprising given that the annotator has a professional background in the medical field. Indeed, it could actually be argued, that for this particular task annotators are required to have some scientific background knowledge in these domains. It was clear that the absence of familiarity with certain topics delays and impairs the ability to reach conclusions in certain abstracts.

5.2 BiQA2 Question types

The attempt to mimic the BioASQ dataset and provide further training samples to the tasks in phase B of task B lead to the classification of the questions (and subsequent differentiated 'exact answer') according to the BioASQ corpus guidelines. The nature and formulation of questions resulted in an overwhelming majority being marked as type 'Summary' or 'yes/no' (+90%). Although a new type of question ('Choice') was identified, the reality is that it is scarcely represented, alongside the 'list' and 'factoid' types, impairing their usage as training data if using BiQA2 alone. However, perhaps they can be included in existing datasets as a small contribution of more samples. Another option would be to transform or classify these questions into a different type, for example, 'Summary'. In this thesis, it was decided not to include the feature 'ideal answer' at BiQA2. In BioASQ this feature is the result of the work of a panel of experts that reach a consensus about the specific text serving as 'golden answers' or ground truth. Given the idiosyncrasies of abstracts in BiQA, namely contradictory, uncertain answers in some queries, and ambiguous questions, it does not make sense to pursue this avenue. Particularly after the poor results, yielded from the DL models implemented at the snippets retrieval subtask, it seems that the snippets of text within the abstracts in BiQA2 require further refining before considering constructing 'ideal answers'.

yes/no/uncertain questions

In BioASQ all the snippets of text compiled for a given question are classified as either ‘yes’ or ‘no’. The abstracts in BiQA, on the other hand, can contain contradictory information between themselves. Furthermore, in some cases, the abstracts authors conclude that a ‘yes’ or ‘no’ binary answer can’t be reached with the current data available. For this reason, the value ‘uncertain’ was introduced not only for each snippet of text but also as the main response to the question. This is why BiQA2 not only contains an answer for each question (as in BioASQ) but also the answer contained within each snippet. The introduction of both the new value ‘uncertain’ and the answer each snippet contains is believed to extend the potential of the dataset.

5.3 Document re-ranking task

The experiments conducted, to access the ability of a dataset to train a model capable of re-ranking a determined number of abstracts, were done with both BioASQ6 and BiQA datasets in order to replicate the results in Lamurias et al. [24] and Brokos et al. [5], and used as a benchmark to compare the potential of BiQA2 as training data capable of generating models to tackle some tasks in the BioASQ competition. By making use of the data provided at the AUEB’s team GitHub repository it was possible to achieve MAP scores very close to the results reported in their paper for document re-ranking. However, it was not possible to replicate the results published by Lamurias et al. [24]. One possible reason for this could be the usage of different ‘ground truth’ samples provided by the BioASQ challenge to compute the MAP scores. Quite often the competition adds or ‘enriches’ (official name) the dataset and that can have an extreme effect on the performance metrics. In this thesis, the gold (ground truth) dataset corresponding to the test batches used was the one provided by the AUEB’s team repository. This might explain why the results obtained with TERM-PACRR do not match the ones in Lamurias et al. [24] both for the BioASQ set and for BiQA alone.

Nevertheless, it was still possible to use BioASQ and BiQA as benchmarks enabling a comparison with BiQA2. In Table 4.8, it is clear that both BiQA and BiQA2 - when used alone and following the BM25 search algorithm as stated by Brokos et al. [5] and used by Lamurias et al. [24] - achieve lower MAP scores compared to the usage of BioASQ alone in TERM-PACRR. However, when combined with the BioASQ dataset, BiQA2 in conjunction with BioASQ actually yields the best mean results. The MAP gain as an average of all batches is 1.1% for BioASQ + BiQA2 when compared to BioASQ alone.

Conversely, the results for BiQA alone were better than the results for BiQA2 alone (3% better on the batches average) with the exception of Batch 5. Despite the differences in MAP scores being very small, this is somewhat surprising given the better joint performance of BioASQ and BiQA2. However, these results need to take into consideration the small number of samples. Following the sampling method from AUEB’s team, BiQA2 only yields 336 samples (balanced) for training whereas BiQA (following the same methodology) yields 2458 which is about 7 times more. One might speculate that this is the reason why BiQA2 alone performs poorly compared with the others but when combined with an already sound dataset such as BioASQ it helps to increase the performance.

5.3.1 'Ideal' inputs

As explained in the methodology section, the number of samples yielded by the sampling method of AUEB's team leaves out the vast majority of possible positive samples in both BiQA and BiQA2. For this reason, it was decided to explore all their positive documents by adding them to the BM25 retrieved list of 100 documents. The number of balanced samples indeed increased to +2K in BiQA and BiQA to +25K. However, despite this nearly 10-fold increase in samples, the results are actually worse when compared to the initial experiments where the abstracts are left out if they are not both in the retrieved BM25 search and in the BiQA or BiQA2. This was not expected. The major increase in the number of samples was expected to yield better results. We speculate on two possible reasons for this:

- The model itself may not be suitable for this task. Its ability to extract meaningful features from the similarity matrix computed between the query and the documents can be limited by the number of kernels (filters), size and stride of the convolutional filters, and the number of convolutions. The similarity distance computed is the standard cosine similarity that could also be a limiting factor. Additionally one must consider that these types of models - based in a convolutional network for computing a natural language task such as document classification of text to a query – might not be the best option altogether. Indeed the MAP results in this thesis, as well as the ones reported by Brokos et al. [5] are low.

- The BM25 search algorithm could, in reality, be acting as a good filter for positive samples in BiQA and BiQA2. The experiment in series A only allows as positive samples the documents that are both in the datasets and have been retrieved by the BM25 search. If the results are better when this filter is applied despite a substantial amount of fewer samples, then it stands to reason that a substantial amount of the documents considered as related to the query in both BiQA and BiQA2 are actually not good positive samples for this kind of model.

Expanding on this last point, one observes that, when used fully ('ideal' versions), BiQA2 demonstrates better mean scores in the test batches than BiQA. This could be an indication that - despite the filtering provided in the BM25 algorithm – the refinement provided by BiQA2 by excluding documents from BiQA is meaningful.

5.3.2 BiQA and BiQA2 combinations

The experiments run with Q_{BiQA2} and A_{BiQA}, Q_{BiQA} + A_{BiQA2} were meant to further evaluate BiQA2s refinement of answers when compared to the original BiQA. Table 5.1 shows some experiments already displayed in their respective results sections but paired here for easier visualization and discussion.

training data	batch_1	batch_2	batch_3	batch_4	batch_5	Mean	pos samples
BiQA2	0.106	0.092	0.094	0.086	0.053	0.086	168
QBiQA2 + ABiQA	0.110	0.097	0.095	0.089	0.055	0.089	195
BiQA2 ideal	0.104	0.089	0.090	0.082	0.052	0.084	1155
QBiQA2 + ABiQA ideal	0.104	0.092	0.091	0.085	0.052	0.085	1721
BiQA	0.110	0.098	0.097	0.090	0.051	0.089	1229
QBiQA + ABiQA2	0.112	0.100	0.094	0.090	0.052	0.090	1220
BiQA ideal	0.105	0.088	0.088	0.081	0.049	0.082	12841
QBiQA + ABiQA2 ideal	0.105	0.089	0.088	0.081	0.049	0.082	12270

Table 5.1: Comparison between pairs of MAP results.

The MAP scores are, once again, very marginally different, which is surprising given the disparity in the number of samples between experiments. At first glance when comparing BiQA2 (BM25) and Q$_{BiQA2}$ + A$_{BiQA}$ it seems that the BiQA2 selected abstracts aren't better related to the queries than in BiQA. However, not only Q$_{BiQA}$ + A$_{BiQA2}$ yields higher scores than BiQA alone, but also Q$_{BiQA}$ + A$_{BiQA2}$ ”ideal” also perform better than BiQA2 ”ideal”. Both these findings might indicate that the abstracts selected for BiQA2 could contain information more related to the queries than in BiQA.

5.4 Snippets

Using the system provided by the AUEB team the task of snippet retrieval was performed on the resulting list of relevant documents re-ranked by TERM-PACRR. The evaluation of MAP scores in the test batches is, thus, conditional to the performance with various training experiments as described above. The snippets retrieval experiments were performed with AUEB's BCNN model and transformers-based models. The results yielded by training BCNN with BioASQ only were similar to the results submitted by AUEB's team to BioASQ6. The transformers-based models implemented here performed better than the BCNN which is not surprising since these type of models are now state-of-the-art in dealing with natural language tasks. Furthermore, the domain-specific BERT model trained on PubMed abstracts (pubmedBERT) yielded the best results (although being just marginal improvements) confirming these models ability to transfer learning and task-specific fine-tuning with extra data. Overall the results on the three snippets retrieval models show the same tendency when comparing BioASQ and BiQA2 as potential datasets for the task of snippets extraction: when used alone for training, BiQA2 achieves lower MAP scores than BioASQ alone, which makes sense since the number of samples is smaller, however, when combined, BioASQ and BiQA2 yield lower scores on all models. This is somewhat surprising – since it happened in both CNN and transformers-based models - and indicates that the snippets extracted by the annotator, when building BiQA2, contribute to hinder the performance in BioASQ6 test batches. One could speculate on possible reasons for BiQA2's performance:

- The complexity and compounded nature of some questions might impair the ability of these types of models to extract semantic meaning from the questions and respective answers. The overwhelming majority of question types are Summary and yes/no/uncertain. In the case of summary questions, they can be very complex and formatted in a way that the current models are simply not equipped to deal with yet. This is analogous to the conclusions reported by Kwiatkowski et al. [23] when building their Natural Questions corpus where it is argued that current methods cannot yield high results in *the corpus* containing real users queries when compared to corpus where highly curated and objective questions are built. In addition, the 'yes/no/uncertain' type of question, can also impair a model's ability to extract meaning from particular snippets of text if there is a substantial amount of conflicting text for the same question. In the case of 'yes/no/uncertain' questions, around 18% of them are considered uncertain.

- The subjectivity and ambiguity inherent to the questions contribute to the poor selection of snippets by the annotator. Particularly in the case of summary questions, it is often the case that the question can have different valid interpretations. This might lead to the selection of abstracts and snippets that can be considered a valid approach to the question but actually being referring to distinct interpretations. Here is an example:

  Query: ***How to avoid fatigue if I foresee irregular sleeping time?***

  Snippet 1: *melatonin PR 2 mg 1-2 h before bedtime was associated with significant improvements relative to placebo in many sleep and daytime parameters, including sleep quality and latency, morning alertness and health-related quality of life. Melatonin PR 2 mg was very well tolerated in clinical trials in older patients, with a tolerability profile that was similar to that of placebo. Short- or longer-term treatment with melatonin PR 2 mg was not associated with dependence, tolerance, rebound insomnia or withdrawal symptoms.*

  Snippet 2: *These results demonstrate that even an ultra short period of sleep is sufficient to enhance memory processing.*

5.5 'yes/no' BioASQ Phase B

Conversely to the topics discussed above in regards to the quality of snippets as a whole, the preliminary results achieved by our colleague at LASIGE when adding yes/no type of questions does seem to show the potential of increasing the performance in these models - from mean of 0.663 (BioASQ) to 0.725 (BioASQ+BiQA2) in model "256" and from 0.696 (BioASQ) to 0.703 (BioASQ+BiQA2). The snippets of text used were only the ones that objectively contained a yes or a no response. Question or snippets that were considered uncertain were not used in training. The fact that the Macro F1 score actually increased in some BioASQ9 test batches - on all 3 models implemented - could indicate that the snippets associated with the yes/no/uncertain questions have the potential to contribute as an additional source of training data for these types of models.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

BiQA's verification

The first component of the work developed in this thesis focused on perusing a sample of both questions and abstracts in BiQA to attest their suitability as a biomedical corpus for QA tasks such as document re-ranking. From the QA pairs analysed there was about 10% whose questions were considered as not understandable, not even questions but rather statements or questions that missed crucial elements of information to merit an answer. Of the remaining questions, not even 50% were considered as having enough information - within their abstracts – to yield an objective response to the query. The Nutrition forum stands out from the Biology and Medical forums as having more ambiguous, more complex and a broader scope of topics within the queries, making it challenging to find an appropriate answer within the abstracts. The characterization of BiQA quantified in the annotations presented in this dissertation suggest that – based on our sample - the method explored in BiQA does fulfil its objective of gathering biomedical questions in a more informal, more natural-speaking formulated manner but it has a lot of room to improve in providing the high standard answers the authors expected to acquire from the forums. In other words, we conclude that the majority of PubMed abstracts gathered from the forums treads do not contain objective information capable of answering its respective queries. This also seems to be corroborated by the results achieved in this thesis upon implementation of the AUEB's team DL system for Document re-ranking and Snippets retrieval. The fact that the trained models yield lower MAP scores with the 'ideal' implementation does suggest that there might be some problems within BiQA that limit its ability to contribute to better performance in the training of models for document re-ranking tasks as set out in Lamurias et al. [24]. Our primary objective is thus accomplished given that we were able to provide a rationale for an analysis of BiQA that allows identification and understanding of potential difficulties or pitfalls that might hinder its applicability as training data for document re-ranking models.

BiQA2

In addition, this thesis explores building a new corpus - BiQA2 - derived from the validated questions and answers in a sample of BiQA. Effectively a new corpus was constructed having in mind the objective of being added to the BioASQ corpus incorporating some features of the BioASQ corpus. The results achieved in this thesis, by applying the features in BiQA2 as training data into models that have competed in the BioASQ challenge seem to indicate that the curating methodology adopted by the annotator's actually contributed to help increase the performance of a CNN model capable of attributing a score to the documents as a proxy of its relatedness to a given query. Conversely, the results obtained in the experiments regarding the snippets retrieval task, are somewhat, disappointing as given a decrease in performance was observed when added to the BioASQ corpus. Although these results are discouraging regarding the quality of annotations, it also points to the difficulty of extracting high standard answers from PubMed abstracts gathered in the forums from which BiQA is built. In regards to BiQA2, the current potential as training data for tasks in Phase B seems to be limited to the yes/no types of questions. This is because almost the entire corpus is composed of summary and yes/no questions, however, at this stage in BiQA2, the summary questions do not have an ”exact” answer according to the BioASQ challenge. This means that, as it stands, less than 40% of BiQA2 can actually be utilised for this task in the BioASQ challenge. Nevertheless, it is possible to gather nearly 400 queries as samples for this yes/no sub-task as a possible contribution to the BioASQ set, which is not a negligible number considering that it is similar to the number of new questions curated by the BioASQ experts for each yearly competition.

The process of making BiQA2 from BiQA was cumbersome and very time-consuming. The perusal of abstracts searching for potential snippets of text that effectively answered the queries was, most of the time, replete with doubts and uncertainty. Given the nature of the queries, one person alone as an annotator will inevitably introduce their own biases when approaching a subjective or ambiguous question. It is, therefore, of paramount importance to have a group of annotators working in tandem to embrace this kind of challenge in order to increase confidence in the resulting corpus. The objective of constructing an enriched corpus from BiQA seems to have been only partially accomplished given the results achieved. In order to fulfil the objective of becoming a valid complement to the BioASQ corpus, further refinement and improvement would need to be done in BiQA2.

6.2 Future Work

The work performed in this dissertation suggests that the current workflow or methodology employed by the makers of BiQA might not produce the expected quality on abstracts hence it requires further investigation on methods that aid in refining effective relevant abstracts from the forums answers. This could be achieved by implementing simple existing search algorithms such as the BM25 which would act as filters for the abstracts or implement more advanced methods comprised of machine learning or deep learning techniques. However, the formulation of the queries in BiQA, is, in fact, representative of the natural, informal manner a human user poses questions when not in an academic or scientific setting. Therefore these questions might be a good starting point for the development of models capable of dealing with their ambiguity and subjectivity. For this reason, and in regards to BiQA2, it seems to be worth continuing the work on this thesis by exploring the remaining QA pairs in BiQA to expand BiQA2. Ideally, this should be conducted by a team of expert annotators that could pursue the following suggestions:

- review and refine the selection abstracts, snippets and exact answers in BiQA2;
- explore new abstracts by manually conducting PubMed searches with BiQA's queries;
- attempt to construct ideal answers in summary type questions so that BiQA2 can be used in phase B of Task B in the BioASQ challenge;
- perhaps to integrate a step in BiQA method in order to simplify or summarize the queries that are formulated in a double, triple or even quadruple manner into a simpler formulation. A solution akin to the one proposed by Ben Abacha and Demner-Fushman [3] to this Question-Understanding problem could potentially be easily integrated in the pipeline of BiQA's methodology.

Finally, the models used to test BiQA2 in this thesis were not trained in each forum separately given the low amount of data each separate forum yields. However, this is something that can be explored in the future.

References

[1] P. Afshar, A. Mohammadi, and K. N. Plataniotis. Brain tumor type classification via capsule networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3129–3133. IEEE, 2018. 26

[2] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson. Deep learning for brain mri segmentation: state of the art and future directions. *Journal of digital imaging*, 30(4):449–459, 2017. 24

[3] A. Ben Abacha and D. Demner-Fushman. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1215. URL `https://www.aclweb.org/anthology/P19-1215`. 22, 71

[4] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009. 53

[5] G.-I. Brokos, P. Liosis, R. McDonald, D. Pappas, and I. Androutsopoulos. AUEB at BioASQ 6: document and snippet retrieval. *arXiv preprint arXiv:1809.06366*, 2018. 18, 27, 45, 51, 52, 54, 64, 65

[6] W. B. Croft, D. Metzler, and T. Strohman. *Search engines: Information retrieval in practice*, volume 520. Addison-Wesley Reading, 2010. 47

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4, 30, 31, 56

[8] B. Dhingra, K. Mazaitis, and W. W. Cohen. Quasar: Datasets for question answering by search and reading. *arXiv preprint arXiv:1707.03904*, 2017. 12

[9] A. Dulceanu, T. Le Dinh, W. Chang, T. Bui, D. S. Kim, M. C. Vu, and S. Kim. PhotoshopQuiA: A corpus of non-factoid questions and answers for why-question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018. 11, 12

[10] M. Feng, B. Xiang, M. R. Glass, L. Wang, and B. Zhou. Applying deep learning to answer selection: A study and an open task. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 813–820. IEEE, 2015. 12

[11] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. 26

[12] B. F. Green Jr, A. K. Wolf, C. Chomsky, and K. Laughery. BASEBALL: an automatic question-answerer. In *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, pages 219–224, 1961. 22

[13] 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. 11

[14] H. Hashemi, M. Aliannejadi, H. Zamani, and W. B. Croft. ANTIQUE: A non-factoid question answering benchmark. In *European Conference on Information Retrieval*, pages 166–173. Springer, 2020. 11

[15] W. He, K. Liu, J. Liu, Y. Lyu, S. Zhao, X. Xiao, Y. Liu, Y. Wang, H. Wu, Q. She, X. Liu, T. Wu, and H. Wang. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2605. URL `https://www.aclweb.org/anthology/W18-2605`. 12

[16] K. M. Hermann, T. Kočiskỳ, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. *arXiv preprint arXiv:1506.03340*, 2015. 5

[17] K. Hui, A. Yates, K. Berberich, and G. De Melo. PACRR: A position-aware neural ir model for relevance matching. *arXiv preprint arXiv:1704.03940*, 2017. 28

[18] K. Ishwari, A. Aneeze, S. Sudheesan, H. Karunaratne, A. Nugaliyadde, and Y. Mallawarrachchi. Advances in natural language question answering: A review. *arXiv preprint arXiv:1904.05276*, 2019. 22

[19] 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. 6, 13, 17, 23

[20] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017. 12

[21] A. Kazaryan, U. Sazanovich, and V. Belyaev. Transformer-based open domain biomedical question answering at BioASQ8 challenge. 2020. 55

[22] L. Kodra and E. K. Meçe. Question answering systems: A review on present developments, challenges and trends. *International Journal of Advanced Computer Science and Applications*, 8(9): 217–224, 2017. 4, 20, 21

[23] 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*, 7:453–466, 2019. 5, 6, 13, 67

[24] A. Lamurias, D. Sousa, and F. M. Couto. Generating scientific question answering corpora from Q&A forums. *arXiv preprint arXiv:2002.02375*, 2020. 7, 8, 9, 10, 18, 27, 28, 45, 46, 47, 48, 50, 51, 64, 69

[25] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer, 1999. 24

[26] S. Liu, X. Zhang, S. Zhang, H. Wang, and W. Zhang. Neural machine reading comprehension: Methods and trends. *Applied Sciences*, 9(18):3698, 2019. 5

[27] A. S. Lundervold and A. Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019. 24, 25

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 29

[29] A. Nentidis, A. Krithara, K. Bougiatiotis, G. Paliouras, and I. Kakadiaris. Results of the sixth edition of the BioASQ challenge. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, pages 1–10, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5301. URL `https://www.aclweb.org/anthology/W18-5301`. 9

[30] 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. 5, 12

[31] R. Nogueira and K. Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019. 56

[32] A. Pampari, P. Raghavan, J. Liang, and J. Peng. emrQA: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*, 2018. 6, 13

[33] 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. 6, 14

[34] 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)*, pages 1532–1543, 2014. 29

[35] A. Peñas, E. Hovy, P. Forner, A. Rodrigo, R. Sutcliffe, and R. Morante. QA4MRE 2011-2013: Overview of question answering for machine reading evaluation. pages 303–320, 09 2013. ISBN 978-3-642-40801-4. doi: 10.1007/978-3-642-40802-1_29. 6

[36] A. M. Pundge, S. Khillare, and C. N. Mahender. Question answering system, approaches and techniques: a review. *International Journal of Computer Applications*, 141(3):0975–8887, 2016. 4, 19, 20

[37] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. 5, 11, 12

[38] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21, 2016. 26

[39] 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*, pages 193–203, 2013. 12

[40] R. F. Simmons, S. Klein, and K. McConlogue. Indexing and dependency logic for answering english questions. *American Documentation*, 15(3):196–204, 1964. 3

[41] M. A. C. Soares and F. S. Parreiras. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences*, 32(6):635–646, 2020. 4, 19, 21

[42] 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. 12

[43] 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*, 16(1):138, 2015. 6, 12, 14, 16, 40

[44] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-art natural language

processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`. 55

[45] Y. Yang, W.-t. Yih, and C. Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018, 2015. 12

[46] 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. 11