A Deep Learning Question Answering Method Over Mixed Closed Domain Information Sources

Andrés Enrique Rosso Mateus

Universidad Nacional de Colombia
Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2020

A Deep Learning Question Answering Method Over Mixed Closed Domain Information Sources

Andrés Enrique Rosso Mateus

Submitted to the Engineering School of the Universidad Nacional de Colombia, in partial fulfillment of the requirements for the degree of:

Doctor of Engineering

Systems and Computer Engineering

Advisor:
Fabio A. González Ph.D.

Co-Advisor:
Manuel Montes-Y-Gómez Ph.D.
INAOE (Instituto Nacional de Astrofísica, Óptica y Electrónica)

Research area:
Computer Science

Research Group:
MindLab Research Group

Universidad Nacional de Colombia
Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia

2020

Dedication

To my parents Carmen and Luis Enrique

Acknowledgements

This thesis would not have been possible without the help and support of many people. First and foremost, I would like to thank my thesis advisors, Dr. Fabio Gonzalez, Dr. Manuel Montes-y-Gomez, and Dr. Paolo Rosso, for their invaluable guidance, constant encouragement, understanding, and patience. Also, I would like to express my sincere gratitude to all the people who are part of the research group MindLab, who have become not only colleagues but also good friends. Special thanks to my colleagues Sebastian Sierra, Johan Rodriguez, John Arevalo, Oscar Perdomo and Victor Contreras, who contributed to my research in several ways. I would like to thank Paolo Fidanza who allowed me to apply some of the knowledge acquired in his company.

I greatly appreciate the support received through the collaborative work undertaken with the Laboratory of Language Technologies of the Computational Sciences Department at the National Institute of Astrophysics, Optics, and Electronics (INAOE).

I also want to thank COLCIENCIAS, the institution that mainly funded this thesis by the doctoral fellowship grant 727/2016. Also, my thanks go to Universidad Nacional de Colombia, their professors, infrastructure, and funding support for the doctoral thesis project and international mobility.

Finally, I am grateful to my wife Carolina Suárez, to my parents, Carmen Mateus and Luis Rosso as well as my sister Sandra Rosso, who have always given me their unconditional support, encouragement, and advice.

Abstract

Question Answering (QA) is an active research area due to its usefulness in accessing the ever increasing amount of data. Information needs have led to the emergence of new information retrieval paradigms in which the user can easily access accurate information.
QA methods allow to solve queries submitted by the user in natural language concisely and effectively, reducing the need for manual validation of large documents. In closed domains, such the biomedical one, these methods are relevant due to the large amount of specialized documents that make difficult the task of finding specific information as well as the usefulness of this information to support practice and research.
In this research work, passage retrieval, which is often the final step in a question-answer system, was particularly addressed. This task evaluates the text fragments that make up the documents that may contain the answer to the question submitted by the user. This evaluation carries out semantic and sometimes syntactic checks that allow to deduce if the text passage is a valid answer, to finally return a ranked list of passages that have a higher probability of being an answer.
In a closed domain, such as the biomedical domain, passage retrieval is particularly challenging due to the complexity of biomedical terminology and the heterogeneity of information sources. These challenges, along with others that will be detailed throughout the document, make it necessary to use other sources of information, such as semantic ones, which, when used in combination with textual sources, help to manage the complexity of language.
On the other hand, the use of deep learning in this field has great interest and recently it has become increasingly popular as an important tool to solve the task of passage retrieval, however there are very few methods that merge the different modalities of information that in a domain like biomedicine offer obvious advantages.
In this research work, different deep learning techniques were explored. In addition, several methods of information fusion were evaluated to take advantage of the complementarity of the modalities. The proposed methods were systematically evaluated in different open and closed domain data sets. Particularly in the biomedical domain the results were outstanding, surpassing the state of the art and demonstrating their effectiveness in the biggest global challenge for this particular task, BioASQ.

Keywords:
Question answering, Passage retrieval, Deep Learning, Machine Learning, Biomedical information retrieval, BioASQ.

Esta tesis de doctorado se sustentó el 22 de 04 de 2021 a las 5:00 p.m., y fue evaluada por los siguientes jurados:

Elizabeth León Guzmán, Phd.
Profesora Asociada
Coordinadora Curricular Ingeniería de Sistemas y Computación
Departamento de Ingeniería de Sistemas e Industrial
Universidad Nacional de Colombia
Bogotá D.C., Colombia

Thamar Solorio, Phd.
Associate Professor
Computer Science department at University of Houston
University of Houston
Houston, Texas, United States

Diego Mollá-Aliod, Phd.
Department of Computing
Macquarie University
Sydney, Australia

Sergio G. Jiménez Vargas, Phd.
Grupo de investigación de Lingüística Computacional
Instituto Caro y Cuervo
Bogotá D.C., Colombia

Content

List of Figures

List of Tables

1 Introduction

The exponential growth of information has also shaped the way it is searched and accessed [148]. Every two years the volume of the textual content produced over the course of history doubles [106], which makes the traditional ways of accessing information obsolete.

In the biomedical field, which is a closed domain example, more than 3,000 medical articles are published every day [124], making it into one of the fields where alternatives to the traditional paradigms of information access become more necessary. A study showed that, with the current rate of publication, a physician would need to check over 130 scientific journals as well as read 27 articles per day to stay updated on breast cancer alone [7].

Among the information retrieval paradigms that may contribute to alleviate this evident need is Question Answering (QA), where a user submits a question in natural language so that the system accurately returns the most likely answer. QA consists of many stages, ranging from the formulation of the question, through the recovery of the documents and finally the extraction of the answer. The extraction can occur in two ways, first returning the text fragment that answers the question (known as answer extraction or passage retrieval) or generating the answer directly from the fragments that support it.

Methods for passage retrieval have mainly explored textual sources which are also the most numerous. However, in closed domains, there have been efforts to standardize the language and to alleviate its ambiguity, which has resulted in the development of the semantic information sources or knowledge bases. Such resources unequivocally represent the domain concepts and its relationships. Another advantageous use of semantic information resources is when the question is too short or does not contain the most relevant terms in the target corpus, it can be alleviated by the use of synonyms of related terms that can match the desired passages.

Therefore, the use of these sources of information is valuable for passage retrieval task, but it poses some challenges such as the computational representation and later fusion of each modality to take advantage of the benefits that each one of them offers. This thesis work has focused on the use of these complementary information sources for closed domain passage retrieval. Through the joint use of semantic and textual information sources, we have been able to obtain encouraging results in biomedical passage retrieval.

The chapter firstly presents the research problem together with the research questions to later establish the objectives. At the end of the chapter, the contributions are enumerated and the published works are listed.

1.1 Problem Statement

Passage retrieval methods employ different approaches to identify relevant passages to a particular question. Traditional methods address the task by adapting document retrieval methods such as vector space model to the passage retrieval task, this is not completely adequate since the length of the passages is much less in comparison to document length, to mention only one of other differences between the two tasks.

Recently, as a result of the increased availability on large volumes of information, deep learning methods are becoming more predominant, which leads to better performance. Regardless of the passage retrieval approach, most of them base their operation on textual sources exclusively, as they represent 80% of the total amount of information available for the specific task [87]. The large volume of data makes it easier to implement accurate methods, but textual content has well-known limitations, such as ambiguity of the terms, where according to the context, the term adopts a certain meaning.

In an effort to reduce the inherent drawbacks of language, alternative forms of knowledge representation began to be promoted, allowing it to be unambiguous and more precise. The initiatives to create a semantic web gave a special impulse to these projects [11, 4], which made possible the rapid development and adoption of representation standards such as the ontologies or knowledge graphs, among others.

Despite the fact that the semantic language representation is the remaining 20%, the advantage is that the vocabulary is controlled and the relationships between the concepts are explicit and in a hierarchical structure. These latter properties are useful for addressing challenges such as the lexical gap or language uncertainty. The semantic representation resources are more common in the closed domains, and many efforts have been made recently to improve the completeness.

In order to make use of the different modalities of information in a complementary way, two important sub-tasks must be tackled: information representation and information fusion. In the first one, different approaches have been proposed, ranging from frequency-based representations, language models or learned distributional representations. All of them capture semantic features of the term sequence and are a starting point, however, an aggregated representation of the question and the passage is desired for this particular task. In the second sub-task it is necessary to consider how to exploit the complementarities of the information modalities. In this case it is necessary to define if the modalities will be exploited separately and eventually combined, or if they will be merged at an earlier stage.

This thesis addresses the use of different information sources to recover text passages from a huge data corpus. Likewise, different fusion approaches are explored in order to take advantage of the data complementarity and, by means of the use of deep learning models extract the features and analyze them to effectively solve the related task. The biomedical domain was taken as a use case due to its complexity and the impact of the research in an area that requires better approaches in information retrieval.

This thesis has explored the following research questions:

1. How to learn a similarity measure between the queries and candidate passages that take advantage of semantic and textual information sources?

2. How to learn a fused information representation that benefits from complementary multimodal information sources?

3. Does the involvement of semantic knowledge enhance the closed domain question answering performance?

4. How to combine deep learning text representations with structured domain knowledge?

1.2 Objectives

1.2.1 General objective

To develop a closed domain deep learning question answering method that takes advantage of textual and semantic information sources.

1.2.2 Specific objectives

1. To design and apply different representational approaches for questions and passages, considering both textual and semantic information sources.

2. To design and implement deep learning methods that extract question-passages similarity features, based on multi-modal information fusion.

3. To systematically evaluate the performance of passage retrieval methods on closed domain datasets.

1.3 Main Contributions

- **Contribution 1: Use of semantic knowledge representation as a complement to the textual one**

  Passage retrieval methods meet many of the challenges that NLP faces. Sometimes it is hard to disambiguate textual information, fortunately semantic information sources can provide a solution for alleviating such common language-related issues. Although semantic information have been used over many years and these resources are extensive in the closed domain such as the medical domain, most passage retrieval approaches do not make use of them.

In this work, semantic information sources were exploited to enrich the textual representation with the identified biomedical concepts from the question and candidate passages, in addition to expanding the key concepts of the question to increase coverage. The proposed approach was helpful in two main respects:

1. The disambiguation of the textual source is achieved through the use of identified unique medical concepts using semantic structured sources such as ontologies and terminology databases.
2. The coverage is increased by using the semantic representation and its hierarchical relationship offered by this modality.

The contribution are described in the Chapter 4.

- **Contribution 2: Enriched information representation from multimodal sources**

  The language representation is an important factor on which the task of passage retrieval relies to be solved effectively. In most of the approaches, question-passage sequences are encoded separately.

  Our approach is quite different, based on the hypothesis that the answer-passage sequences have a stronger semantic correlation if the passage is a valid answer than if it is not. We address the representation by means of the semantic interactions rather than independently. We explored similarity-based representations on vector representations, statistical co-occurrences covered in the Chapter 5 and transformer's attention layers based on similarity encoding discussed in the Chapter 7 .

- **Contribution 3: A Multi-modal information fusion strategy based on deep learning**

  To take advantage of textual and semantic representations it is essential to create a combined representation that captures the most relevant patterns of each modality and merges them in a complementary way.

  For this purpose three fusion strategies were proposed: early fusion approach covered in the Chapter 5, intermediate fusion and late fusion discussed in the Chapter 4. The fusion strategies make use of different deep learning building blocks, such as: convolutional and recurrent layers among others.

- **Contribution 4: A novel deep metric learning approach for closed domain passage retrieval**

  Metric learning has been widely used for image processing tasks, whereas it has not been fully explored in passage retrieval. Inspired on the use in image-processing where the correspondences of the images are projected in a metric space, we proposed a deep metric learning model which encodes question-passage interactions using a siamese

architecture but taking as in the case of triplet network, a triple input (question, positive passage and negative passage) which enables the model to produce a metric space where interactions are well represented.

Another important factor in the success of the model is the sampling strategy. The proposed informative sampling strategy selects first easy samples, which refers to those passages that are not semantically related to the question and that for the same reason have to be located spatially far away, opposite are the hard samples that despite not being a valid answer they are semantically related and would be located over the border that separates positive and negative samples in the metric space. This sampling strategy considerably improves the learning process and leads to an improvement in the overall performance of the model, this model is proposed in Chapter 6.

- **Contribution 5: State-of-the-art performance improvement in passage retrieval by merging information from multiple sources**

  The proposed models for passage retrieval as well as for document retrieval in closed domain were systematically validated. The results obtained mainly in the biomedical domain showed that the passage retrieval task is effectively solved by merging semantic and textual sources. When evaluating the models in the BioASQ competition and comparing the results with the best competition's models, an improvement of around 20% in the official ranking metric (Mean Average Precision - MAP) was achieved [103].

Following is the list of papers that has been published during the development of this research:

1. Rosso-Mateus, Andrés, Manuel Montes-y-Gómez and Fabio A. González. "A Deep Metric Learning Method for Biomedical Passage Retrieval" In Proceedings of the 28th International Conference on Computational Linguistics (COLING2020), In Press. 2020. [103]

2. Rosso-Mateus, Andrés, Manuel Montes-y-Gómez, Paolo Rosso, and Fabio A. González. "Deep fusion of multiple term-similarity measures for biomedical passage retrieval." Journal of Intelligent & Fuzzy Systems Preprint (2020): 1-10. [99]

3. Rosso-Mateus, Andrés, Fabio A. González, and Manuel Montes. "Mindlab neural network approach at bioasq 6b." In Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering, pp. 40-46. 2018. [102]

4. Rosso-Mateus, Andrés, Fabio A. González, and Manuel Montes-y-Gómez. "A Two-Step Neural Network Approach to Passage Retrieval for Open Domain Question Answering." In Iberoamerican Congress on Pattern Recognition, pp. 566-574. Springer, Cham, 2017. [101]

5. Rosso-Mateus, Andrés, Fabio A. González, and Manuel Montes-y-Gómez. "A Shallow Convolutional Neural Network Architecture for Open Domain Question Answering." In Colombian Conference on Computing, pp. 485-494. Springer, Cham, 2017. [100]

Also there is a collaboration work which was presented at BioASQ 7 challenge (2019):

1. Pineda-Vargas, Mónica, Andrés Rosso-Mateus, Fabio A. González, and Manuel Montes-y-Gómez. "A Mixed Information Source Approach for Biomedical Question Answering: MindLab at BioASQ 7B." In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 595-606. Springer, Cham, 2019. [91]

1.4 Thesis Outline

The thesis addresses three broad topics: deep representation learning, information fusion and similarity/metric learning. The chapters relate to this topics as follows: chapter 3 is focuses on deep representation learning, chapter 4, 5, and 7 present methods for information fusion and chapter 6 addresses the similarity/metric learning problem. The details of each chapter content is presented next.

- **Chapter 1: Introduction** —— This chapter covers the introduction of this thesis. It includes problem statement, research objectives and motivation and significant contributions of our work.

- **Chapter 2: Background and Related Work** —— This chapter presents the literature review in question answering (QA) field. Each stage in the QA pipeline is described, although a more rigorous description of passage retrieval task is provided.

- **Chapter 3: Pseudo-relevance feedback deep learning method for open domain passage retrieval** —— We present our work exploring deep learning approaches for open domain passage retrieval.

- **Chapter 4: Multimodal Fusion Strategy for Biomedical Passage Retrieval** —— This chapter presents a model for biomedical passage retrieval that explores different information fusion alternatives. It also describes our participation in the BioASQ challenge and the results obtained.

- **Chapter 5: Deep Fusion of Multiple Term-Similarity Measures For Biomedical Passage Retrieval** —— This chapter presents a novel approach for biomedical passage retrieval which is able to combine different information sources using a similarity matrix fusion strategy.

- **Chapter 6: A Deep Metric Learning Method For Biomedical Passage Retrieval** —— This chapter describes a novel approach for metric learning which is able to map the interactions between the question and the answer on a metric space built with the semantic interactions of the text and the semantic information.
- **Chapter 7: Transformers based representation for Biomedical Passage Retrieval** —— This chapter presents a work in progress that, using Bert's attention layers representation capacity, takes advantage to extract features that allow to discriminate if a passage is related to a question.
- **Chapter 8: Conclusions and Future Work** —— In this chapter some of the research conclusions are shared in addition to the most interesting future steps in passage retrieval task.

2 Background and Related Work

This chapter briefly discusses some of the most important definitions and concepts related to the field of question answering, these concepts are needed for the following chapters. As was mentioned earlier, the focus of this research work is the passage retrieval QA sub-task, which is the most relevant and commonly assumed as the final step in question answering systems. For this reason, most of the state-of-the-art research is dedicated to this task in particular.
In the first part of this chapter, the question-answer problem and its component tasks are described in detail. Furthermore, we examine open and closed domain passage retrieval in addition to the differentiating properties they have. Since the study of passage retrieval task on closed domains is one of the research objectives of this work, the biomedical domain is taken as a use case which is presented along with the challenges it. Some of the most important datasets are presented at the end of the chapter, including the relevant challenges and the used metrics.

2.1 Question Answering

Question answering is a rapidly growing information retrieval paradigm that aims to find short and concrete answers analyzing thousands of documents where such answers can be found. In this paradigm, instead of returning the document that may be related to the question, it fulfill the need for information by returning a sentence, a paragraph, a fragment of text or even a word that is the answer to the question asked [46].
Question answering systems have to mine vast volumes of textual information in order to gather the relevant evidence to produce an answer that satisfies the information requirement. Most of those systems have in common a high level architecture [2] depicted in the Figure **2-1**.
The standard Question answering system architecture is a pipeline of information that flows through chained tasks, the first being **1) question processing**, in this task a question posed in natural language is transformed into a query that is used to retrieve the documents in the following step, **2) document retrieval** in this phase an information retrieval algorithm retrieves the most related documents for the compound query, **3) passage retrieval**, the retrieved documents are analyzed in detail where each piece of text (passage) may become a possible answer to the question posed, the result of this step is a set of passages that can be transformed or enriched in the final step **4) answer extraction**.

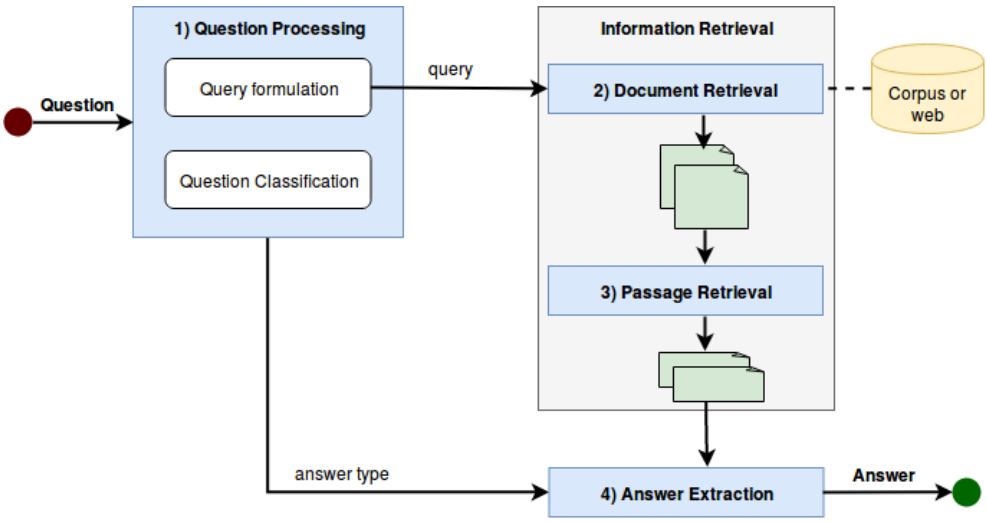


Figure 2-1: Question answering stages

A detailed QA phases description is presented in the following section.

2.1.1 Question Answering Phases

Phase 1: Question Processing

Question processing is the phase in which the objective of the question is identified, the question type is classified, and the question is reformulated into a semantically equivalent query that is simpler and more appropriate for the document search engine.
This stage is very important because by formulating a wrong query, documents that can answer the expected question will not be retrieved. Some sub-tasks are listed below.

- Key-word Extraction: It is important to define the key-words that will be used to identify the relevant documents. These key-words are extracted by matching with named entities as the case of Srihari et al. approach [115], or using a markov model through sequence labelling as is described in Veyseh work [129].

- Query Formulation: In this task, the key-words related to the question are used to compose a query that can be sent to the document retrieval system. There are approaches which make use of semantic information sources to expand the query, as the approach proposed by Yang et al. [137].

- Question Classification: Question classification methods identify what is the expected entity type for a given question, for example, people, location, time, medical concept, etc. Question classifiers can be built by hand-writing rules [58], by supervised machine learning [150, 41, 38], symbolic [112], or with some hybrid approach [8, 127]. A very

exhaustive work made by the Facebook team [133] presents a taxonomy for question type.

- Semantic Parsing: Is a task related to the mapping of natural language sentences into a structured representation. In this task, there are supervised machine learning methods as Berant et al propose. [10], rule-based approaches [136, 126] or machine learning as proposed by Krishnamurthy et al. based on distant supervision relation learning [47].

Phase 2: Document Retrieval

Document retrieval is defined as the matching of a query built on a set of large free text records. The objective is to find the documents which most closely coincide with the search terms in the query. Question Answering Systems rely on document retrieval to provide the subset of information that will be analyzed in depth, in which the answer to the question is most likely to be found. For this reason, the effectiveness of the system is strongly influenced by this sub-task [72].
Most frequent approaches in document retrieval are those based on TF-IDF, where the relevance ranking function is based on the frequency of the query terms and documents. The best-known implementation of this family of methods is BM25, this method ignores the proximity of the terms, it only takes into account the occurrence, similar to a bag of words approach [98]. Also, there is an important drawback related to the vocabulary used in the query and the document, related documents with different words are not identified although they carry the same information. Other Approaches based on Latent low dimensional document representations like Latent Semantic Indexing (LSI) [28], Latent Dirichlet Allocation (LDA) [15] are more robust to vocabulary differences using a semantical representation where the document search is focused on concepts and not in words.
Despite the outstanding performance of neural networks (NNs) in fields like computer vision or automatic language translation, the use of these approaches in Document Retrieval (DR) tasks had relatively less attention. Further on, there would appear approximations based on language models, [42, 149], which make it possible to know the probability that a document occurs given query terms. Most of those models were based on word embeddings [33, 70, 151]. A key advantage of neural networks based models are their ability to work from raw input data, so there is no need to extract features from the text, also those methods are not extremely affected by vocabulary differences or terms order as classic methods do. Many different architectures and approaches have been proposed, such as auto-encoders [49, 111], recursive neural networks [113], recurrent neural networks [88], convolutional neural networks [57], various embedding methods [33, 69], and deep reinforcement learning [107, 73].
This stage was not a central part of our research work, as Passage Retrieval is. However, during the research it was noted that some datasets that were intended to be used did not provide the relevant documents for the query, so it was necessary to explore methods to resolve this dependency as well.

Phase 3: Passage Retrieval

The passage retrieval task is responsible for analyzing in depth each text fragment in the established candidate documents with the objective of finding the text sequence that answers the question.
The passages are usually paragraphs or sentences that are semantically compared with the question, filtered and ranked according to their semantic matching. In many QA systems the process ends by returning these ordered passages, while other systems use this list of passages to compose a single answer.
There are several approaches for passage retrieval, some of them employ deep learning approaches[84, 125], symbolic approaches [142, 135] and passage filtering using question matching keywords [85, 22]. This task was the main objective of this research work, hence a comprehensive analysis of the most relevant methods is made at the end of the chapter.

Phase 4: Answer Extraction

Given the question Q and a subset made up of passages that probably answer the question S(1), ..., S(N), the objective in answer extraction is to find a term or combinations of contiguous terms that are the precise answer to Q. If the question is looking for an entity (factual question) this entity is identified and returned as the answer.
The main sub-tasks in this stage are:

- Answer Processing: In this subtask, the goal is to process the information from the passages to identify important pieces of information mainly based on two approaches: 1) pattern extraction using regular expression and 2) n-gram clustering sometimes called the redundancy-based approach [19, 54], where all the unigrams, bigrams and trigrams that appear in the snippet are extracted, weighed and finally concatenated to produce a response fragment.

- Answer Summarizing: The answer to a given question may be supported by many passages found in different documents. In this task, the goal is to produce a concise representation from a fragment or a set of fragments to compose a complete answer. Approaches to achieving that task include knowledge-based methods [65], statistical-based [77] and so on.

Question Answering example

To motivate a brief example, consider a user trying to know the name of Lionel Messi's agent. With this intention in mind, let review the phases that the system must accomplish to retrieve an answer.

- **User Question**: *Who is Lionel Messi's agent?*

- **Phase 1. Question Processing**: In this stage, the system identifies that the user is looking for a person entity, extracts the keywords as Lionel Messi, agent, etc, and if we want to include structured knowledge in the process, the query must be translated into an RDF triplet (subject, predicate, object) that should be as (Lionel Messi, agent, X?)

- **Phase 2. Document Retrieval**: In this stage, we must compose the query that is going to be submitted to the document retrieval method. The query should include keywords and query expansions terms. Then, the query is processed and the system returns an ordered list of related documents, that can be news-sites that talk about Lionel Messi or Messi's Wikipedia page.

- **Phase 3. Passage Retrieval**: In this stage, the documents are analyzed in deep to find the passages that could contain the desired answer. If we retrieve Messi's Wikipedia page, the query related passages are in the following list and the passage with the highest score is highlighted in bold:
  - *Since 2008, when he was 20, Messi has been in a relationship with Antonella Roccuzzo, a fellow native of Rosario.*
  - *Messi and Roccuzzo have three sons: Thiago (born in 2012), Mateo (born in 2015), and Ciro (born in 2018).*
  - *Messi enjoys a close relationship with his immediate family members, particularly his mother, Celia, whose face he has tattooed on his left shoulder.*
  - ***His professional affairs are largely run as a family business: his father, Jorge, has been his agent since he was 14, and his oldest brother, Rodrigo, handles his daily schedule and publicity.***
  - *Since leaving for Spain at age 13, Messi has maintained close ties to his hometown of Rosario, even preserving his distinct Rosarino accent.*

- **Phase 4. Answer Extraction**: In this example, we knew in advance that the answer was related to a person entity that was the name of his agent. So, for our example the system's answer should be: ***His father, Jorge, has been his agent since he was 14.***

2.2 Passage Retrieval as Question Answering Core Task

Passage retrieval is the main focus of this thesis, this task is also known as answer selection or snippet retrieval, and is usually the final step in a question answering system. Given the input question q and a set of passages $p = p_0, ...p_m$, the objective is to return those passages that are a valid answer to the question q. It can be seen as the problem of learning

a prediction function $f(q, p) \rightarrow a$ from a training data set, where a is close to zero 0 when the passage is not a a valid answer and close to 1, when it does. Traditional models for passage retrieval were based on Information Retrieval approaches, such as the BM25 algorithm estimates the passage relevance measuring similarities in a sparse representation [98], other including indexing meta-information on terms [26], or Query Likelihood [93] showing similar performance in passage retrieval task using term occurrence as main representation. A key issue in the traditional approach is that an exact match of the most important question-answer terms is required, otherwise, a low score will be achieved. A less critical factor that impacts the overall performance is the order in which the terms are arranged; traditional methods do not take into account the order and it is shown to be significant to the task [13]. Lets consider the following example:

- Question: *What are the symptoms of the flu?*
- Answer: *For most people, influenza begins with a fever and a cough.*

In the related example, there are very few term coincidences because the answer is using different synonyms for the flu and also the word symptoms is not present.
Subsequently, with the growth of the NLP field, researchers begin to use lexical, syntactical, and linguistic relations as input features [109]. Pizzato et al. [92] research the indexing and retrieval of annotated text using a representation based on semantic role labeling. Guo et al. proposed a model that combines seven different features drawn from the relationships within the texts, such as the lexical matching and page link [37]. Chen [24] employed a language model to evaluate if two sentences have a question-answer relationship, and Surdeanu [118] made use of semantic role tagging, syntactic dependency strings, and so on to enhance retrieval.
NLP approaches are commonly combined with machine learning (ML) to take advantage of the linguistic features. The work of Othman et al. combines lexical, syntactic, and semantic features which are analyzed by a support vector machine model to predict whether the input passage is relevant to the posed question [85]. Another ML model was proposed by Khalifa et al. which uses a lexical-based hybrid method with a naive Bayes classifier that takes advantage of domain knowledge by exploiting the auxiliary information (thesaurus) [45].
Most ML and NLP strategies require a heavy pre-processing stage to extract the handmade features, in order to feed the discriminatory model. This requirement is not present in deep learning (DL) based models, where the DL ability for automatic extraction of these features from the representation is exploited.
Recently, researchers have been studying deep learning approaches to automatically extract features and learn semantic correlations between questions and answers. Yu et al. [145] present a Bigram model that using a one-dimensional CNN model and as input representation a pretrained semantic word embeddings (bag-of-words or bigram model). The work presented by Severyb et al. [109] make use of a Siamese Convolutional Neural Network

for learning question-answer representations patterns, in this approach QA pairs are concatenated together with the TF information and passed across a feed-forward network to produce a relevance score. Wang et al. [131] propose a bidirectional neural model to encode question-answer sentence terms, then an output layer learns the similarity patterns in the encoded pairs. The model proposed by Tan et al. [119] implements two bidirectional long short-term memory networks to encode questions and passages separately, and measures their closeness by cosine similarity. Cohen et al. [27] have proposed an Hybrid CNN and BiLSTM approach, where the input layer processes the query and candidate answer characters before feeding a convolutional layer with different kernel lengths. This convolutional layer creates abstract features that are processed by a Bidirectional Short Term Memory (BiLSTM) layer to capture time dependencies and determine pair relevance. Recently attention models become very popular in passage retrieval, the Yang et al. work was some of the first models [138], the approach is based on a value-shared weighting scheme, they also combine different matching signals weighted by the importance learned from an attention mechanism.

In the biomedical field the predominant approach is deep learning. For example, [32] proposed to apply a word embedding representation for question-passage sequences and then to compute their semantic relationship employing a weighted cosine distance. Another relevant approach, which obtained the best results in the 2018 BioASQ edition, was presented by the auen-nlp team [20]. This approach is based on an ABCNN architecture [143], which models pair of sentences with a convolutional neural model and an attention mechanism, and uses a linear classification layer to produce an output relevance score. The model proposed by Pappas et al. computes context-sensitive term embeddings with multiple CCN filter which capture context and similarity relevance information [89]. Finally, [121] used Bert contextual word embeddings [51] to represent question and passage pairs, and fine-tuned the model to produce a ranking score, most recent works -not have employed pre-trained transformers language models that are fine-tuned on the downstream classification task.

It is remarkable that although the biomedical domain is plenty of semantic knowledge as biomedical terminology databases and ontologies, most of the approaches have not made use of these resources [61]. An exception is the work presented by [12], where ontologies are used to expand query terms. Semantic resources offer information that is complementary to textual information and that can be used to alleviate problems as polysemy or synonymy disambiguation.

Table **2-1** present some remarkable approaches to passage retrieval. As was mention most of them are based on Deep Learning and the preferred text sequence representations are LSTM and CNN. The latest ones are mainly based on Bert which is expected because the remarkable performance in many NLP tasks. As was mentioned, most of these are based on Deep Learning and the preferred text sequence representations are LSTM and CNN. More recent models are mainly based on BERT, which is expected due to the outstanding performance in several NLP tasks. It is predominant the use of passage retrieval methods for

open domain, in which there is a greater number of datasets that will be enunciated later, whereas for biomedical domain the most prominent is BioASQ.

Author (year)	Approach	Dataset	Domain
Yu (2014) [145]	DL (CNN + Bigram)	TrecQA	Open
Severyn (2015) [109]	DL (CNN)	TrecQA	Open
Wang (2015) [131]	DL (BiLSTM)	TrecQA	Open
Tan (2015) [119]	DL (BiLSTM + Cosine)	TrecQA	Open
Yang (2016) [138]	DL (CNN + Attention)	TrecQA	Open
Miller (2016) [68]	DL (Key-Value Memory Network)	WikiQA	Open
He (2016) [39]	DL(CNN pairwise)	TrecQA / WikiQA	Open
Cohen (2018) [27]	DL (CNN + BiLSTM)	TrecQA	Open
Tay (2018) [120]	DL (Hyperbolic NN)	TrecQA	Open
Ma (2018) [60]	Statistical (Noise Contrastive Estimation)	WikiQA	Open
Yang (2019) [139]	DL (Residual encoder NN)	WikiQA	Open
Galko (2018) [32]	DL (CNN + Bigram)	BioASQ	Biomedical
Brokos (2018) [20]	DL (ABCNN)	BioASQ	Biomedical
Yoon (2019) [144]	DL (CNN + Latent-cluster)	TrecQA / WikiQA	Open
Pappas (2019) [89]	DL (CNN + context-embedding)	BioASQ	Biomedical
Telukuntla (2020) [121]	DL (Bert)	BioASQ	Biomedical
Gard (2020) [60]	DL (Bert fine-tuning)	TrecQA / WikiQA	Open

Table 2-1: List of most prominent passage retrieval approaches

2.2.1 Challenges in Passage Retrieval

Passage retrieval have many linguistic challenges that could affect the method precision. Some of the most important are presented as follows.

- **Vocabulary Gap**: This challenge is related to the use of different words or combination of words in the query and the related document or passage. Although queries and passages are using different words, the information contained in both elements must be

highly correlated. Consequently, there is a vocabulary gap between concepts expressed in different words, as for example:

  - *Where can I buy cheap laptops?*
  - *Where can I purchase affordable notebooks?*

- **Polysemy**: In the natural language there are cases in which a word is used to express quite different meanings. The challenge is known as Polysemy, and it is also present in the Biomedical Domain as the following example shows:
  - *Cold (temperature): having a lower than usual temperature.*
  - *Cold (disease): viral infectious disease of the upper respiratory tract that primarily affects the nose. The throat, sinuses, and larynx.*
  - *Cold (acronym): Chronic Obstructive Lung Disease.*

- **Complex and elaborated queries**: This challenge is especially evident in a closed domain such as biomedical science, in which vocabulary is complex, acronyms are constantly used, and language is highly specialized. An example of a biomedical query is the following:
  - *Which markers are screened with the triple test for the detection of syndromes in fetus?*

- **Use of Abbreviations**: It is very common in both open and closed domains. There are problems with the use of abbreviations, such as ambiguity when the same abbreviation is used in two different contexts, for example:
  - *TCF can refer to T Cell Factor or Tissue Culture Fluid. It causes a conflict in the identification of the desired entity.*

The use of semantic data to disambiguate textual information can mitigate some of these challenges that prevent a better performance in QA models.

2.3 Open and Closed Domain Passage Retrieval

Open-domain passage retrieval, following the setting of the annual TREC competitions [130], is defined as the task to find passages that answer a question in a large collection of textual documents, for example, documents, blogs, reports, news, articles, e-mails, logs, large web pages, and recently including large-scale structured knowledge bases, such as Freebase [17], DBPedia [5], Google's Knowledge Graph, YAGO2 [43], etc.
Conversely, closed domain passage retrieval refers to the restriction made over the knowledge field to extract the answer, some example domains are music, mathematics, biology, medicine, etc.

In the closed domain the terminology is controlled and several semantic resources are available, e.g. ontologies, thesaurus and other terminology databases, which offers some advantages such as:

1. Sharing a common understanding of the structure of information.
2. Enabling reuse of domain knowledge.
3. Making domain assumptions explicit.
4. Separating domain knowledge from the operational knowledge.

A use case of closed domain is the Biomedical one, it is explored through this research work.

2.4 Biomedical Passage Retrieval

Every day, more than 3000 new articles are published in biomedical journals [63], which means around 2 articles are published every minute. In biomedical field, passage retrieval plays an important role based based on the premise that clinical decision making is supported by experience and research literature. Finding useful information in the enormous amount of biomedical articles represents a challenge for expert users, even more so when the user is actually a patient [62].
The biomedical domain presents the aforementioned challenges of language, such as highly specialized queries that become more noticeable in this field, vocabulary gap, use of abbreviations, among others that are listed below.

- **Biomedical terminology evolution**: The biomedical literature employs thousands of entity names, and every day new names are added to the list, which makes it a challenge to keep the dictionaries and lexicons up to date [110]. For example, if we consider only humans, the fly, the mouse and the worm, there are about 70,000 genes. Those genes comprise more than 100,000 proteins. Furthermore, there are over a million species, cell lines and molecules. The possible concepts that can exist is huge.

- **Synonymy**: This is more likely to occur in gene names. For example, the yeast gene UBC6 is also known as DOA2. If multiple text passages coming from different documents refer to the same gene with different names, it is hard to determine if the passages are referring to the same entity [110].

- **Variability in spelling**: It happens when a term have different spelling in the same language based on the country or region where is used. For example haematoma in British English versus hematoma in American English.

- **Biomedical Polysemy**: In biomedical science, the context has a fundamental influence on the meaning of a particular term. For example, these gene names may also refer to protein names according to the context [110]:
  - CAT1
  - LacZ
  - MAP kinase
  - Sonic hedgehog

The use of semantic data in the biomedical domain has a great potential. Such data can reduce uncertainty through the use of terminological or taxonomic databases that can either disambiguate the language or reduce the mentioned lexical gap. Some of the most important resources explored in this research work are presented below.

- **Ontologies**: An ontology is defined as a formal explicit description of concepts in the domain of discourse, together with their attributes, roles, restrictions, and other defining features [83]. In biomedicine, an abundance of ontologies has been developed for different purposes an now they play a central role in integrating the information coming from different fields. The overwhelming importance of ontologies to biomedical research and to clinical practice, has pull serveral organizations, professional societies, and individual laboratories to create their own onotologies but the effort has been mostly uncoordinated, OPEN BIOMEDICAL ONTOLOGIES (OBO) is a project created with the aim of create controlled shared vocabularies across different biological and medical domains [104]. Some of the most known biomedical ontologies are: MESH, Gene Ontology (GO), Sequence Ontology (SO), Generic Model Organism Project (GMOD), Functional Genomics Data (FGED), Ontology for Biomedical Investigations (OBI), Plant Ontology Consortium (POC), an extended description of the more relevant as follows.
  - **MESH**: The National Library of Medicine (NLM) designed a medical subject taxonomy to index PubMed and Medline documents. Each article is indexed with one or more MeSH terms. The use of mesh concept allow to improve the information retrieval by limiting the search space.
  - **GO**: The Genetic Ontology (GO) is the most comprehensive knowledge base of genetic field in the world. Such knowledge can be read by humans and machines, providing a basis for computer analysis of large-scale genetic and molecular biology research.
- **Biomedical Terminology Databases and Systems**: Terminology databases are intended to facilitate access to medical literature as well as to support the development of computer systems that understand biomedical language.

To achieve these goals, the language is translated into a standardized form, i.e., the terms used to express the same concept are unified and the relationships between concepts are established to make navigation easier, even between resources coming from different fields. Two are the resources most used by the scientific and academic community: SNOMED and UMLS, which are described below.

- **SNOMED CT**: SNOMED Clinical Terms is a systematically organized collection of terms, synonyms and definitions used in clinical documentation and medical reporting. It is considered to be the most complete clinical care terminology base in the world. The primary purpose of SNOMED CT is to encode the medical concepts used in diagnosis and patient care. The comprehensive coverage of SNOMED CT includes: clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices and samples.
- **The Unified Medical Language System (UMLS)** : is a compendium of many controlled vocabularies in the biomedical sciences (created 1986) [16]. It provides a mapping structure among these vocabularies and thus allows one to translate among the various terminology systems; it may also be viewed as a comprehensive thesaurus of biomedical concepts. The three main componentes os UMLS are:
  - **Metathesaurus**: Comprises over 1 million biomedical concepts and 5 million concept names from many vocabularies, including CPT, ICD-10-CM, LOINC, MeSH, RxNorm, and SNOMED CT. Also includes Hierarchies, definitions, and other relationships and attributes.
  - **Semantic Network**: Broad categories (semantic types) and their relationships (semantic relations).
  - **SPECIALIST Lexicon and Lexical Tools**: A large syntactic lexicon of biomedical and general English and tools for normalizing strings, generating lexical variants, and creating indexes.

The focus of this research work will be biomedical passage retrieval, as a use case of closed domain.

2.5 Passage Retrieval Evaluation Campaigns and Datasets

There are several passages retrieval benchmark datasets for the open domain and closed domain task evaluation of factoid question answering. Some of the most relevant are presented as follows.

2.5.1 TREC QA

TREC QA was a seminal campaign, and its main purpose is to perform domain-independent answer retrieval evaluation over large and unstructured corpora [130]. Wang et al. [132] developed a benchmark collection using the Text REtrieval Conference (TREC) 8-13 QA data. They used the questions in TREC 8-12 for training and set aside TREC 13 questions for development (84 questions) and testing (100 questions), the complete statistics for the related dataset are in Table **2-2**.

Split	#Questions	#Pairs
TRAIN ALL	1,229	53,417
TRAIN	94	4,718
DEV	82	1,148
TEST	95	1,517

Table 2-2: TrecQA dataset statistics

TREC QA data set has become one of the most widely used benchmarks for open-domain passage retrieval. Some examples of the questions and passages contained in the dataset are presented in the related Table **2-3**

Question: *How many members were in the crew of the Challenger?*
Answer 1: *More than 200 safety modifications followed the loss of Challenger, most notably the redesign of the solid-fuel booster that triggered the disaster.*
Answer 2: *These changes, among thousands of major and minor design modifications that were made since the shuttles started flying, included redesigned solid-fuel booster rockets, more spacecraft sensors, and a parachute escape system for the crew.*
Answer 3 (correct): *On Jan 28, 1986, the space shuttle Challenger exploded 73 seconds after liftoff from Cape Canaveral, killing all seven crew members.*

Table 2-3: Sample Questions and Answer Passages in TrecQA Dataset

2.5.2 Wiki QA

This dataset was released in 2015 by Microsoft Research Group [141], which contains Question-Answer pairs for an open domain. The Microsoft research group collected Bing Search Engine query logs and extracts the questions the user submit from May of 2010 to July of 2011, and the answers are sentences of Wikipedia summary page. WikiQA dataset is larger than the previous TREC QA filtered dataset, see Table **2-4**.
Examples for the dataset are in Table **2-5**.

Split	#Questions	#Pairs
TRAIN	2,118	20,358
DEV	296	2,716
TEST	633	6,156

Table 2-4: WikiQA dataset statistics

Question: *How many players on a side for a football game ?*
Answer 1: *American football , known in the United States as football , is a team sport*
Answer 2 (correct): *It is played by two teams, eleven players to a side, who advance an oval ball over a rectangular field that is 120 yards long by 53.3 yards wide and has goalposts at both ends.*
Answer 3: *The team in possession of the ball (the offense) attempts to advance down the field by running with the ball, or passing it.*

Table 2-5: Sample Questions and Answer Passages in WikiQA Dataset

2.5.3 BioASQ

BioASQ challenge is focused on indexing and question answering tasks over biomedical articles [122]. BioASQ information retrieval challenge is composed of two phases, Phase A and B.

- Phase A: Given a question the system must return relevant documents (from PubMed articles baseline [79]), relevant snippets (extracted from articles).
- Phase B: Given a question and a set of relevant articles and snippets. The system must provide an exact answer (e.g., named entities) and ideal answers (summaries) [122].

The available dataset consists of biomedical articles published in MEDLINE. For each article in the dataset, title, abstract, and MESH terms are provided. The total number of articles is close to 30 million.
Throughout this research, the BioASQ dataset will be used to test the approaches implemented on biomedical passage retrieval. Because the data set consists of both questions and positive answers only, we have gathered the negative passages which are in the same document where the positive passages are. The obtained dataset was very unbalanced, only 18% of the total number of pairs is positive, the statistics for BioASQ dataset are presented in Table **2-6**, a sample of the dataset is in Table **2-7**.

Split	#Questions	#Pairs
TRAIN	2747	27,600
DEV	500	6,345
TEST	500	6,156

Table 2-6: BioASQ dataset statistics

Question: *Which viruses are best known to cause myocarditis?*
Answer 1: *Myocarditis can occasionally lead to sudden death and in up to 10% of patients may progress to dilated cardiomyopathy.*
Answer 2: *Because the initial onset is difficult to recognize clinically, and the diagnostic tools available are unsatisfactory, new strategies to diagnose myocarditis are needed.*
Answer 3 (correct): *Enteroviruses (EV) are an important cause of neonatal disease, including hepatitis, meningoencephalitis, and myocarditis that can lead to death or severe long-term sequelae.*

Table 2-7: Sample Questions and Answer Passages in BioASQ Dataset

2.5.4 SQuAD

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset. It consists of questions and answers posed by human curators on a set of Wikipedia segments of text that answers the questions. The dataset consists of 100,000+ question-answer pairs on 500+ articles, up to now is the biggest dataset [95], complete statistics and a sample data in Table **2-8** and **2-9** respectively.

Split	#Questions	#Pairs
TRAIN	78,713	27,600
DEV	8,886	6,345
TEST	10,570	6,156

Table 2-8: SQuAD dataset statistics

Question: *What is Nigeria's official language?*
Answer 1: *Nigeria has one of the largest populations of youth in the world.*
Answer 2: *The country is viewed as a multinational state, as it is inhabited by over 500 ethnic groups, of which the three largest are the Hausa, Igbo, and Yoruba.*
Answer 3: *It is played by two teams , eleven players to a side , who advance an oval ball over a rectangular field that is 120 yards long by 53.3 yards wide and has goalposts at both ends.*
Answer 4: *These ethnic groups speak over 500 different languages and are identified with a wide variety of cultures.*
Answer 5 (correct): *The official language is English.*

Table 2-9: Sample Questions and Answer Passages in SQuAD Dataset

2.6 Performance Metrics for Document and Passage Retrieval

This section describes the evaluation metrics used in the passage retrieval task, these are the same ones used in information retrieval. Models are then evaluated on their ability to correctly retrieve and rank answers for a given questions.

For a given set of relevant documents 'gold standard' and a set of documents retrieved by the system, precision and recall are defined as Equation 2-1 and 2-2 shows respectively.

$$Precision = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \tag{2-1}$$

$$Recall = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \tag{2-2}$$

The measure $F1$ is mainly a weighted harmonic mean of recall and precision, as follows in Equation 2-3.

$$F_1 = 2\frac{precision \cdot recall}{precision + recall} \tag{2-3}$$

The aforementioned measures do not take into account the order of the recovered passages, but this property is very important in the task to be solved. Given that it is required to return the list of ordered passages the metrics presented below model the effectiveness of the model by evaluating the returned passages ranking versus the gold standard ranking, the Equation 2-4 describe the metric.

$$Average\ Precision(AP) = \frac{\sum_{k=1}^{n}(P(k) \times R(k))}{\text{RD}} \tag{2-4}$$

Where n is the total number of returned passages and RD denotes the number of relevant passsages in the gold standard. $P(k)$ is the precision of system when retrieved list considers only first k relevant items and $R(k)$ is an indicator function which is equal to 1 if the $k-th$ item belongs to the gold standard item set otherwise its equal to 0.
Once the $Average Precision(AP)$ is calculated over a set of queries, it is possible to calculate the $Mean\ Average\ Precision\ (MAP)$, which is defined as follows the Equation 2-5:

$$Mean\ Average\ Precision\ (MAP) = \frac{1}{|Q|}\sum AP(q_i) \tag{2-5}$$

Where $AP(q_i)$ denotes the average precision for a given query qi. The precision of the geometric mean is equivalent to that of the MAP; with the only difference being that the MAP uses the arithmetic mean and the $GMAP$ uses the geometric mean.
The equation for calculating GMAP is as follows in Equation 2-6:

$$GMAP = \sqrt[n]{\prod_{n}^{i=1}(AP_i + \epsilon)} \tag{2-6}$$

The mean reciprocal rank is a statistic measure that evaluates the ranking order over a golden raking for a given set of queries, the order is commonly a probability of correctness. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer, as is described in Equation 2-7

$$Mean\ Reciprocal\ Rank\ (MRR) = \frac{1}{n}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{2-7}$$

Where n is the total number of queries and $rank_i$ is the position of gold standard element in the retrieved list.

3 Pseudo-Relevance Feedback for Open Domain Passage Retrieval

The matching of questions and passages relies largely on the efficiency of the representation chosen. Thus, the representation used should be able to capture intricate semantic associations between sequences (of questions and passages). This task is related to **representational learning**, which can be defined as the learning of an informative representation that is adapted to the goal task through the extraction of higher-level features. In this chapter, we present a representation based on the interactions between the terms of the question and passage sequences. This approach is different from the more traditional ones [67, 138], where the question and passage sequences are separately represented to further identify in a later step the patterns that are relevant whenever a passage becomes a valid answer employing a discriminatory model.
The method combines a term by term cosine similarity matrix with a convolutional neural network. The method was evaluated in an open-domain question-answering task. The work presented in this chapter was published in the following papers.

- Rosso-Mateus, A., González, F. A., & Montes-y-Gómez, M. (2017, September). A Shallow Convolutional Neural Network Architecture for Open Domain Question Answering. In Colombian Conference on Computing (pp. 485-494). Springer, Cham.
- Rosso-Mateus, A., González, F. A., & Montes-y-Gómez, M. (2017, November). A Two-Step Neural Network Approach to Passage Retrieval for Open Domain Question Answering. In Iberoamerican Congress on Pattern Recognition (pp. 566-574). Springer, Cham.

The chapter is organized as follows. Section 1 introduce the proposed method, Section 2 describes the method and the proposed architecture. Section 3 depicts the experimental setup in detail. Section 4 discusses the results achieved by the method in the two datasets selected for evaluation, and finally, Section 5 presents the conclusions and future work directions.

3.1 Introduction

Most of the state-of-the-art approaches exhibit good performance in ranking the first candidate passage. That is, the first-ranked candidate passage frequently contains a valid answer

to the posed question. Based on this observation, this model proposes a two-stage ranking approach. In the first stage, the passages are ranked according to their similarity with the question. This initial ranking is generated by a convolutional neural network, which is applied to a matrix encoding question−passage term similarities, and returns a score that indicates the degree of similarity between the question and the candidate passage. In the second stage, passages are re-ranked based on their similarity with the first passage in the initial ranking. To generate this new ranking, a convolutional neural network is also applied, but at this time the matrix is made by first_passage−other_passage term to term similarities. This strategy is analogous to the pseudo-relevance feedback method used in information retrieval, where the highest-ranked results are used to expand the question-based query.

3.2 Model Description

The method proposed in this chapter is presented in Figure **3-1**, each of the steps will be detailed in the next section. The whole process consists of two stages: the training phase where the similarity model is obtained, and the testing phase, where the calculated model is used to rank the question-passage pairs. During training: (1) question-passage pairs (pairs) are pre-processed, (2) the similarity matrix between pairs terms is calculated, (3) a convolutional neural network model is trained to predict the relevance of the answer to the question. Once the model is built it can be used to predict the rank order of candidate passages. At testing time, for a particular question, the model is applied to predict the relevance score of the set of candidate passages: (4) passages are ranked according to their scores, (5) passages are re-ranked according to their similarity with the highest-ranked passage at step (4), producing a new ranking of the passages.

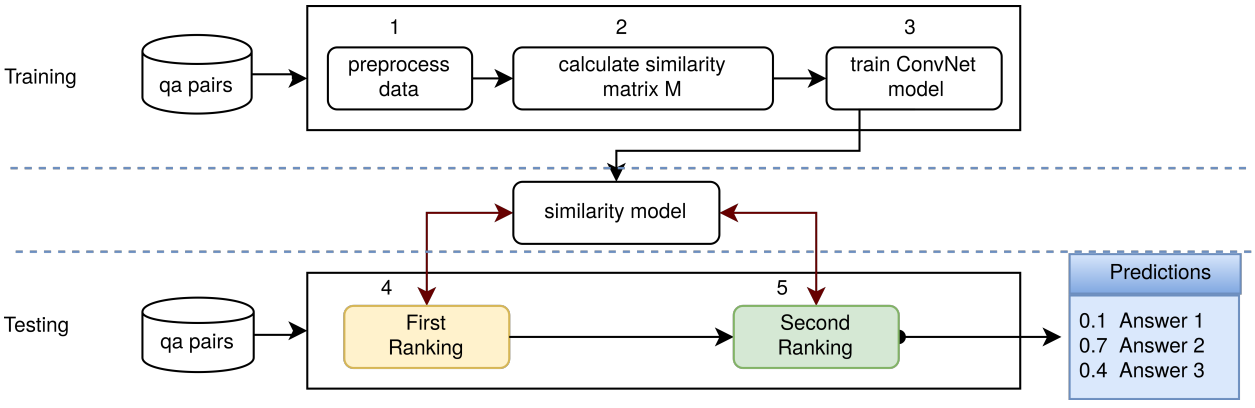


Figure 3-1: Process of ranking and re-ranking qa pairs

3.2.1 Step 1. Pre-process Data

Questions and candidate passages are processed using: tokenization to delimit terms; lowercasing to standardize the terms; pos-tagging, using the nltk pos-tagger [14], to extract syntactical information that will be used in salience weighting; and transforming terms to a word2vec vector representation [66], to make possible their semantic similarity comparison.

3.2.2 Step 2. Calculate Similarity Matrix

The similarity matrix M represents the semantic relatedness of the i-th question term and the j-th passage term according to a similarity measure. Each element $M_{i,j}$ of this matrix is a composition of a similarity score and a salience score as described by the Eq. 3-1.

$$M_{i,j} = scos(q_i, a_j) * sal(q_i, a_j) \tag{3-1}$$

Similarity Score.

The similarity score for a question-passage pair terms (q_i, a_j) is calculated by means of the cosine distance between their word2vec vectors as indicated by Formula 3-2.

$$scos(q_i, a_j) = 0.5 + \frac{q_i \cdot a_j}{2 \left\|q_i\right\|_2 \left\|a_j\right\|_2} \tag{3-2}$$

In the case that it does not exist the word2vec representation for one of the terms, their similarity is measured based on their distance in Wordnet [134]. In particular, we use a similarity measure for the edge distance between the first common concept related to q_i and a_j. If there is not a common concept between the terms, then we calculate the Levenshtein distance between the words [52], defined as the number of operations (insertions and eliminations of characters) needed to transform q_i to a_j.

Salience Weighting.

As not all terms are equally informative for measuring text similarities [56, 30], we consider weighting the terms from the question and the answer based on part of speech functions: verbs, nouns, and adjectives are considered to be the most relevant. We model this information through a salience score.

The salience score is calculated as follows. If both terms are relevant then their score is 1. If only one of the terms is important then the score is 0.6, in case none of them is relevant the score is 0.3. The salience function is defined in the Formula 3-3.

$$sal(q_i, a_j) = \begin{cases} 1 & if\ imp(q_i) + imp(a_j) = 2 \\ 0.6 & if\ imp(q_i) + imp(a_j) = 1 \\ 0.3 & if\ imp(q_i) + imp(a_j) = 0 \end{cases} \quad (3\text{-}3)$$

Where $imp(q_i)$ and $imp(a_j)$ are the evaluation of importance weighting function for every question and passage term. The related function returns 1 if the term is a verb, noun, or adjective, otherwise, a 0 is returned.
Finally, we sort the calculated matrix M leaving the most related terms in the top-left cell, and if the number of rows or columns exceeds 40, the remaining data is truncated. This step provides an invariable representation of the similarity patterns that can be exploited by the convolutional network.

3.2.3 Step 3. Convolutional Model

Convolutional neural networks (CNN) are a popular method for image analysis, due to their ability to capture spatial invariant patterns. In the proposed method, they play a similar role, but instead of receiving an input image, the CNN receives the similarity matrix M. The hypothesis is that it will be able to identify term-similarity patterns that help to determine the relevance of a question-answer pair. Patterns identified by CNN are sub-sampled by a pooling layer. The output of the pooling layer feeds a fully-connected layer. Finally, the output of the model is generated by a sigmoid unit. This output corresponds to a score, $\mathbf{simScore}(q, a)$, that can be interpreted as a degree of relatedness between the question q and the answer a.
The architecture of the convolutional model is presented in Figure **3-2**.

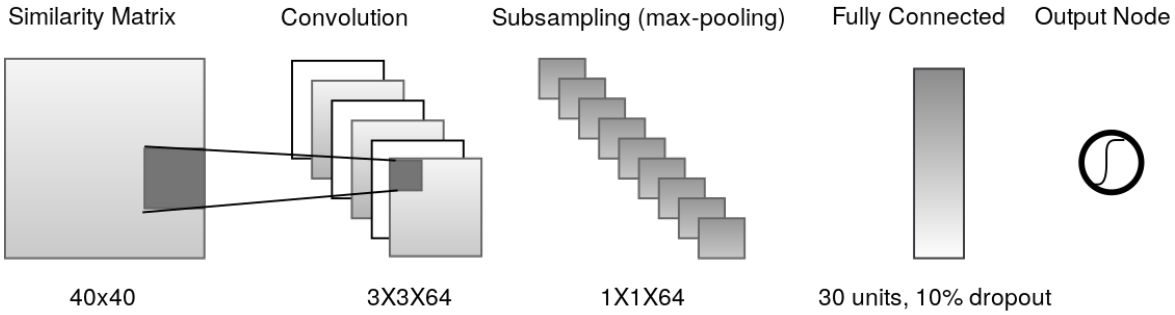


Figure 3-2: Convolutional neural network model architecture.

3.2.4 Step 4 and 5. Two Ranking Stages

During the testing phase, a new query, along with candidate passages, are submitted to the method. The candidate passages $(a_1, a_2, ..., a_k)$ are ranked using the CNN model producing

their first rank. Based on the premise that the first candidate passage, a^*, is expected to be highly correlated with the question q, a second score, $simScore(a^*, a_k)$, is calculated by comparing each candidate passage with the highest-ranked passage. A new ranking is calculated by using a new score corresponding to a linear combination of the first and second scores as is shown in Eq. 3-4.

$$finalScore(q, a_k) = (1 - \alpha) * simScore(q, a_k) + \alpha * simScore(a^*, a_k) \quad (3\text{-}4)$$

This strategy promotes candidate passage which share similar terms with the highest-ranked answer. This is a strategy analogous to pseudo-relevance feedback in information retrieval [97], where the original question-based query is extended with terms from the highest-ranked documents.

3.3 Experimental Setup

3.3.1 Test Datasets

The proposed method was compared to baseline and state-of-the-art methods using two information retrieval performance measures Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP), the performance metrics were described in Chapter2. To evaluate the method, TrecQA and WikiQA datasets were used, the description of the.

- **TrecQA**: The dataset has two partitions. In TRAIN partition the correctness of answer was carried out manually while in TRAIN-ALL the correctness of candidate answer sentences was identified by regular expressions against the answer, this can induce noise in the data, the statics of the related dataset are presented in Table **3-1**.

Table 3-1: TrecQA dataset

Split	#Questions	#Pairs
TRAIN ALL	1,229	53,417
TRAIN	94	4,718
DEV	82	1,148
TEST	95	1,517

- **Wiki QA**: The Microsoft research group collected Bing Search Engine query logs and extract the questions that the user submitted from May of 2010 to July of 2011, the answers are the sentences of Wikipedia summary page and were manually labeled by experts, Table **3-2** presents the statistics.

Table 3-2: WikiQA dataset

Split	#Questions	#Pairs
TRAIN	2,118	20,358
DEV	296	2,716
TEST	633	6,156

3.3.2 Baseline Models

Three baseline models were implemented to evaluate the performance of the proposed method. 1) Word Count, which is a word matching method that counts the number of non-stop words that occur both in the question and in the answer sentences. 2) Weighted Word Count, a modified approach that weighs the word counts using semantical information [141]. 3) DeepMind model [146], a semantic parsing method based on similarity metric learning and latent representations.

The list of comparative methods are the following: Word Count, Weighted Word Count, DeepMind model [146], Paragraph Vector (PV) [49], Attention-Based Model (aNMM) [138], Convolutional Neural Network Method (CNN) [109], Pairwise Word Interaction Model (Pairwise CNN) [39], and the proposed model without rerank (this work) and with rerank (This Work Rerank).

3.4 Results

Table **3-3** summarizes the results of all the evaluated methods applied to both TrecQA and WikiQA datasets. In the case of TrecQA two configurations were evaluated: the TRAIN partition and the TRAIN ALL partition, which were described in Subsection 2.5.

In the TrecQA dataset, the proposed method presents the best performance of all the evaluated methods. This is consistent in both configurations. Also, we can observe that the use of re-ranking improves the method performance in terms of MAP. The main reason is that, in most cases, the first ranked passage is relevant; this can be evidenced by the high value of the MRR measure.

In the WikiQA dataset, the best result is obtained by the Pairwise CNN method [39], however, the proposed method has a competitive performance that outperforms the other evaluated methods. This can be evidenced by the overall performance of all the methods, this dataset seems to be more challenging. One difficulty with this dataset is that it contains several questions without a valid answer in the dataset. The re-ranking strategy produces an important improvement for the TrecQA dataset, while with the WikiQA dataset it did not improve the performance. It can be concluded that a lower MRR in this dataset means that the top ranked answer is less likely to be relevant and thus it has less probability of improving the ranking of relevant answers. As we are introducing a weighting term α to

Table 3-3: Overview of results QA answer selection task datasets. We also include the results of the baseline models. ('-' is Not Reported)

	TREC TRAIN ALL		**TREC TRAIN**		**WikiQA**	
Method	MAP	MRR	MAP	MRR	MAP	MRR
Baselines						
Word Count	0.6402	0.7021	0.6402	0.7021	0.4891	0.4924
Weighted Word Count	0.6512	0.7223	0.6512	0.7223	0.5099	0.5132
DeepMind model	0.6531	0.6885	0.6689	0.7091	0.5908	0.5951
aNMM [138]	0.7385	0.7995	0.7334	0.8020	-	-
CNN [109]	0.7459	0.8078	0.7329	0.7962	-	-
Pairwise CNN [39]	0.7588	0.8219	-	-	**0.7090**	**0.7234**
PV [49]	-	-	-	-	0.5110	0.5160
This Work	0.7644	**0.8414**	0.7605	0.8344	0.6368	0.6614
This Work (Rerank)	**0.7737**	0.8403	**0.7750**	**0.8350**	0.6351	0.6583

scale the second score, we calculated this term based on the exploration with the validation partition, which gives 0.32 as the optimal value.
In general, we can say that the proposed method exhibits a very competitive performance when compared to state-of-the-art methods. However, its main strength is the fact that it is simpler than the other methods. This can be objectively measured by counting the number of parameters that the learning algorithm has to adjust during training. Table **3-4** shows the number of parameters for some of the evaluated methods. The proposed method has fewer parameters than the other methods, in orders of magnitude. This has a positive impact on the number of computational resources that are required during training and testing.

Table 3-4: Number of Parameters

Split	# Of Parameters
aNMM [138]	14,000
CNN [109]	100,000
Pairwise CNN (2016)[39]	1.7 million
This Work	**3,198**

3.5 Conclusion

This chapter presents an novel method for open domain passage retrieval based on convolutional neural networks and a pseudo-relevance-feedback-inspired re-ranking strategy.
The experimental results show that the proposed method is competitive when compared

with state-of-the-art methods by the its publishing time, despite being a simple model with a reduced set of parameters. The second ranking improves the first one in about 2% in the MAP metric. This observation validated our hypothesis that the first ranked passage contains information that can help to re-rank the subsequent answers.

4 Multimodal Fusion Strategy for Biomedical Passage Retrieval

The use of different modalities to represent question-passage interactions can bring several advantages.

- It increases coverage when the representation is missing or noisy in some of the modalities.
- It offers a natural way to model the importance of a pair of terms when similarities are strong in several modalities.
- It provides a complementary view of how the terms in the passage and the response interact.

In the biomedical domain, there are several semantic information sources that can be used as additional modalities to the textual similarity representation. In this chapter, we present a method to efficiently combine information coming from textual and semantic sources. The approach was tested in the biomedical domain using the largest available dataset for biomedical passage retrieval (BioASQ). Also, we show the results of our participation in the BioASQ 6 (2018) and BioASQ 7 (2019) challenge edition.
The methods described in this chapter were published in the following papers:

- Rosso-Mateus, A., González, F. A., & Montes-y-Gómez, M. (2018). MindLab Neural Network Approach at BioASQ 6B. In Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering (pp. 40-46). http:www.aclweb.organthologyW18-5305
- Pineda-Vargas, M., Rosso-Mateus, A., González, F. A., & Montes-y-Gómez, M. (2019, September). A Mixed Information Source Approach for Biomedical Question Answering: MindLab at BioASQ 7B. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 595-606). Springer, Cham.

The remaining sections of this chapter are organized as follows: Section 1 method introduction, Section 2 describes the model architecture and the strategies used for document and passage retrieval, Section 3 discusses the results as well as the conclusions, finally a summary of our participation in the related challenge is presented.

4.1 Introduction

The use of semantic information can alleviate some of the issues with the textual information. In this chapter we present a method for biomedical passage retrieval that effectively combines textual and semantic sources.

Our passage retrieval model is based on the hypothesis that the questions and the relevant passages share semantic properties in each of the modalities involved (textual and semantic), which means that by combining them we can achieve a more effective outcome. In order to obtain the semantic information representation, we have identified the biomedical concepts listed in the Unified Medical Language System (UMLS) thesaurus [9] and then a vector representation for biomedical concepts is used to encode them. Finally two fusion methods are proposed: Mixed Data Representation Intermediate Method (MIF) and Mixed Data Representation Late Fusion (MLF) that exploit the convolutional network architecture that was previously used.

As was mentioned earlier in BioASQ challenge is mandatory to retrieve also the relevant documents, for this reason a document retrieval method is proposed with a re-ranking strategy. The first stage involves retrieving N most relevant documents using BM25, while the second stage consists of re-rank the N document using two methods: word mover's distance and document centroid.

4.2 Methods

4.2.1 Model Architecture

The model for the task is composed of two main modules as shown in Figure **4-1**. A document retrieval module searches the PubMed Baseline Repository (MBR) [79] for relevant documents, and a fine-grained information retrieval model to identify the 10 most relevant snippets.

We used Elastic Search (ES) engine [35] for document indexing and BM25 as relevance ranking function [1]. With ES we retrieve the top n relevant documents given a question. Subsequent to this, we re-rank the top n relevant documents using Word Mover's Distance (WMD) and Document Centroid Rerank, obtains the 10 most relevant documents.

Most related documents are analyzed in depth. We split the documents into sentences and those sentences feed the snippet retrieval stage. We process the snippets with a Convolutional Neural Network (CNN) to obtain a semantic similarity relevance score.

Finally, the scored snippets are sorted in descending order and the 10 with the highest scores are selected. The documents are re-ranked based on a standardized linear combination between Elastic Search score and the average of their snippets scores. The 10 most related documents and snippets were submitted to BioASQ server.

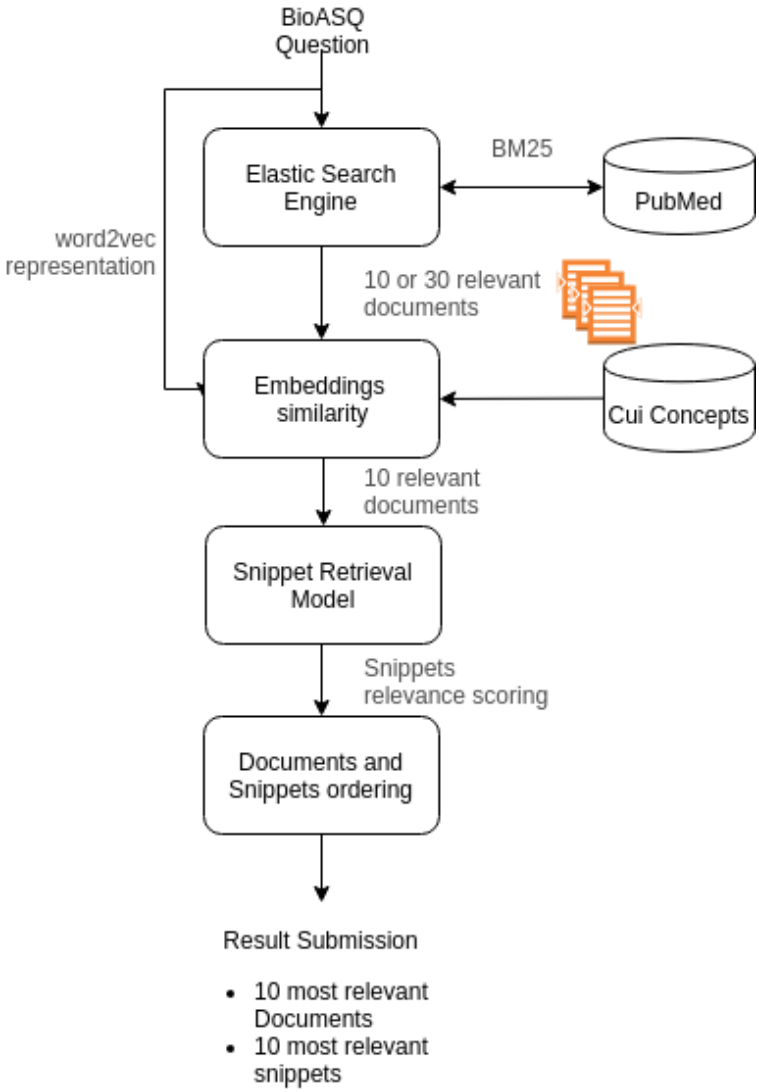


Figure 4-1: BioASQ Model Diagram

A detailed description of the model will be presented in the following sections.

4.2.2 Document Retrieval

One of the first tasks in question answering is retrieving the documents that could contain the answer to a user's question. This task affects the question answering task because if a retrieval system finds irrelevant documents for a question, the later stages as snippets retrieval, will inevitably fail.

We used Elastic Search (ES) [35] to index approximately 27 millions medical articles, using information like the title, abstract and keywords, applying standard text preprocessing operations such as tokenization, remove stopwords and stemming.

Step 1. Get the top n relevant documents

Okapi-BM25 is used as relevant ranking function that involves different factors including: inverse document frequencies, term frequencies, and the length of the document and the query. With this strategy, we return the n most relevant documents (we used 10 and 30 as n in experimentation). These documents are analyzed using two embeddings similarity strategies that compare the question with the title and the abstract of the document.
We use multi-match query with type cross-fields [1] for the search, that first analyzes the query and produce a set of terms, then it searches for each term in the fields that have been specified, for this case, "abstract", "title" and "mesh-term" as shown in Figure **4-2**.

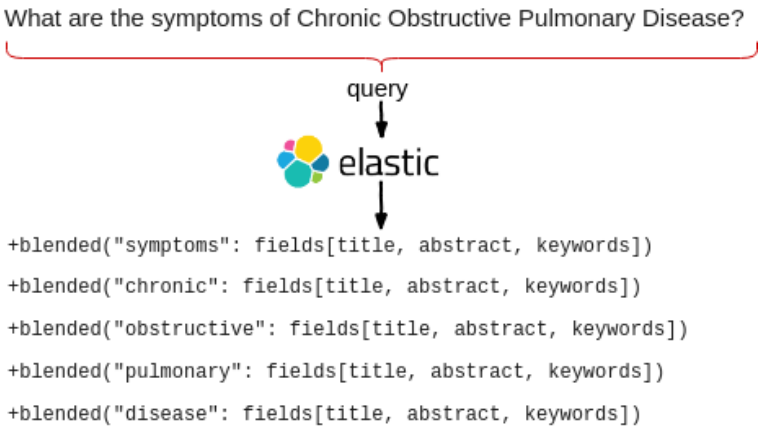


Figure 4-2: Multi-match, cross-fields ES search

Step 2. Re-rank with embedding similarity

After obtaining the n most relevant documents, we re-rank using two embedding similarity methods. As was mentioned before first ranking is based on BM-25 which only takes into account exact term matching, to improve the results we propose a Word Mover's Distance and Document Centroid re-ranking strategy.

Word Mover's Distance: The first was Word Mover's Distance (WMD) (CITAR), a special case of the Earth Mover's Distance (CITAR). Is a metric for the distance between two documents, calculating the similarity in the Word2Vec embedding space. The query and each document are represented as a weighted point cloud of embedded words as follows in in Figure **4-3**. The distance between them is the minimum cumulative distance that words from query need to travel to match exactly the point cloud of document.

[1] cross-fields Elastic Search `https://www.elastic.co/guide/en/elasticsearch/reference/7.1/query-dsl-multi-match-query.html#type-cross-fields`.

Figure 4-3: Word Mover's Distance between two documents.

Let q be the query user , $d \in D$ where D is a set of n relevant documents, and $|q|, |d|$ the number of distinct tokens in q and d respectively. Let $\boldsymbol{T}$ be a flow matrix where $\boldsymbol{T}_{ww'}$ denotes how much the word w in q travels to word w' in d and C is the transportation cost with $C_{w,w'} := dist(\boldsymbol{v}_{q_w}, \boldsymbol{v}_{d'_w})$ normally provided by their Euclidean distance in the word2vec embedding space. Finally, we can define the WMD between the two documents as the minimum cumulative cost required to move all words from d to q.

$$\min_{\boldsymbol{T} \geq 0} \sum_{w,w'}^{n} \boldsymbol{T}_{ww'} C(w, w') \tag{4-1}$$

Finally this module returns the 10 documents with less Word Mover's distance.

Doc Centroid Rerank: The second approach method is based in word2vec, computing for a given query q and each document d in the top n relevant documents, the mean of its words vectors. Then we compute the cosine similarity between the mean of the word's vectors of query and the mean of the word's vectors of documents. Then, we reorder the documents by similarity and returns the top 10.

4.2.3 Passage Retrieval

Our passage retrieval model is based on two main hypothesis: first, that question and answer passages are semantically correlated term by term and concept by concept; second, that structured and unstructured information are complementary modalities that can jointly represent, in a better way, the semantic content of questions and passages.

The proposed method has two stages as Figure **4-4** shows. The first one (training phase) has the objective to learn the similarity patterns for question-answer pairs. In the second stage (testing), the trained similarity model is used to obtain the ranking scores of a set of candidate answers (snippets) for a particular question. The method uses two representations schemes, textual and structured, for both answers and questions. Both representations are learned from data using a convolutional neural network architecture. The representations are combined using an intermediate fusion strategy.

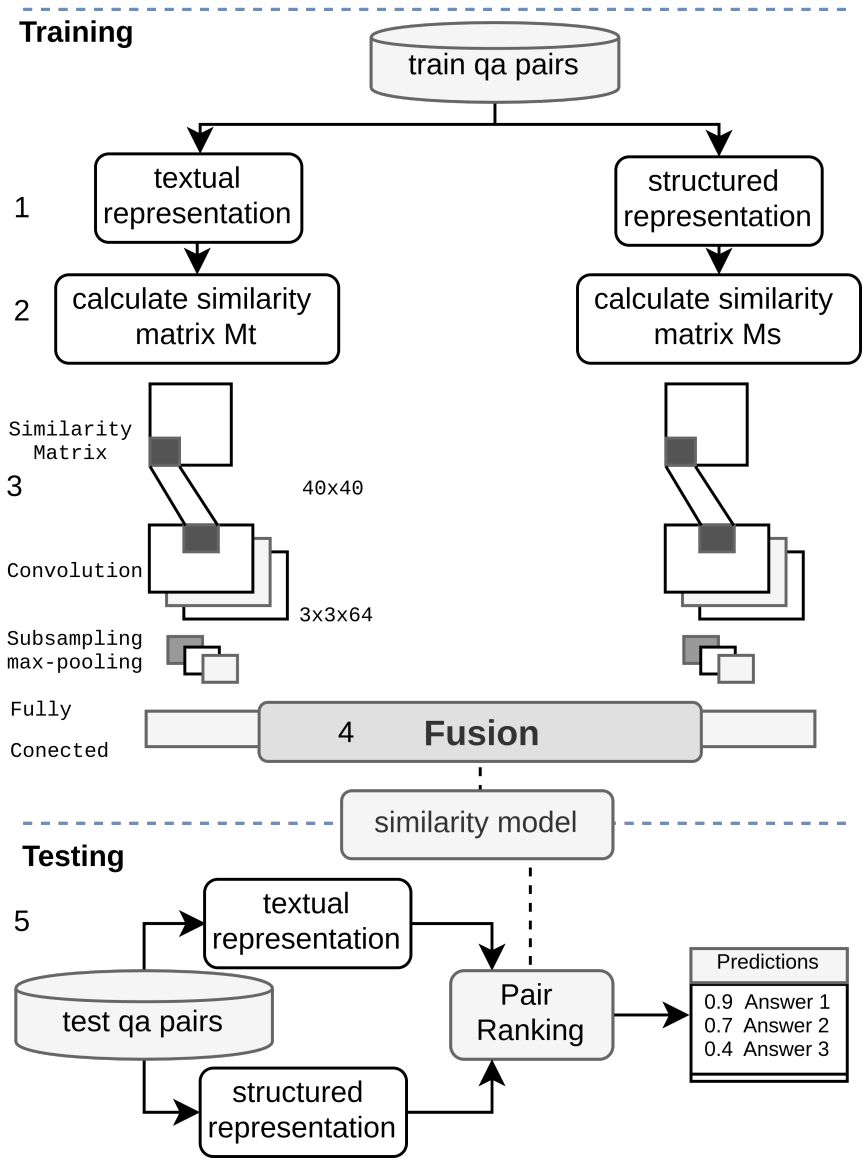


Figure 4-4: Passage Retrieval Process

The details of the process are presented next. For a given question and candidate answer pair (q_i, a_j), its textual representation is denoted by (qt_i, at_j), and for structured representation by (qs_i, as_j).

- **Step 1 - Extract Representation**: The question and answer pairs are needed to be transformed to feed the neural network, the process is different for each modality.
  - **Textual Representation**: First the text is cleaned and tokenized, the grammatical tagging is carried out with NLTK POS-tagger to extract syntactical information that will be used in salience weighting; each term is transformed later in a vector embedding using a pre-trained word2vec model provided by NLPLab,

which is trained on Wikipedia and PubMed documents [2].

– **Semantic Representation**: To identify medical concepts we have to use QuickUMLS [114] which is an unsupervised biomedical concept extraction. The identified concepts are then transformed into a continuous vector representation using a cui2vec embedding. This embedding maps medical concepts instead of words. Concepts are referred by their concept unique identifier (CUI) from the Unified Medical Language System (UMLS) thesaurus [9]. In contrast with the textual representation, there are less words in the text fragments that can be embedded in the structured representation as it is shown in Figure **4-5**. This has to do with the reduced size of the cui2vec vocabulary. To overcome this restriction (4 concepts in average per question) we applied expansion to question Cui embeddings. The followed approach was the centroid method proposed by Kuzi et al. [48].

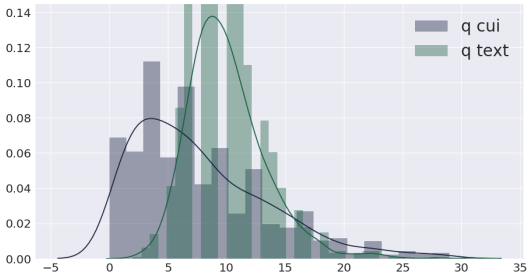


Figure 4-5: Question Terms and Cuis distribution

- **Step 2 - Calculate Similarity Matrix**: Each i, j-entry of the similarity matrices M_t and M_s, represents the semantic relatedness of the i-th question term (or concept) and the j-th answer term (or concept) according to the embedding (nlplab or cui2vec).

  – **Textual Similarity Matrix** M_t: In the case of textual representation the cosine similarity between terms is weighted based on the grammatical function of the term pair, this grammatical weighting is called a salience score $sal(qt_i, at_j)$. The similarity between element i-th and j-th is calculated as Eq. 4-2 shows.

$$M_{i,j} = scos(qt_i, at_j) * sal(qt_i, at_j) \tag{4-2}$$

$$scos(qt_i, at_j) = 0.5 + \frac{qt_i \cdot at_j}{2 \left\| qt_i \right\|_2 \left\| at_j \right\|_2} \tag{4-3}$$

[2]BioNLP word vector representation, trained with biomedical and general-domain texts http://bio.nlplab.org

$$sal(qt_i, at_j) = \begin{cases} 1 & if\ imp(qt_i) + imp(at_j) = 2 \\ 0.6 & if\ imp(qt_i) + imp(at_j) = 1 \\ 0.3 & if\ imp(qt_i) + imp(at_j) = 0 \end{cases} \quad (4\text{-}4)$$

Where $imp(qt_i)$ and $imp(at_j)$ are the importance weighting for every question and answer term. The related function returns 1 if the term is a verb, noun or adjective, otherwise, returns 0.

– **Structured Similarity Matrix** M_s: In the case of structured information we calculate just the cosine similarity between cui2vec concept vectors.

Finally, both matrices (M_t, M_s) are sorted according to the similarity score to facilitate the similarity pattern identification.

- **Step 3. Convolutional Model**: The architecture of the convolutional model is shown in Figure **4-4** step 3. A convolutional layer is fed with both similarity matrices M_t and M_s, CNN layer will identify element-similarity patterns to rank the relevance of a question-answer pair using both knowledge representations. Patterns identified by each CNN filter are sub-sampled by a pooling layer. The pooling layer for all the filters is merged with two fully connected layers with 10% of dropout.

- **Step 4. Multimodal fusion**: The dense outputs of the modalities are merged in a unique dense layer, which feeds another dense layer. Finally, the output score of the model is generated by a sigmoid unit on top of the last dense layer.

- **Step 5. Pair Ranking**: Candidate answers $(a_1, a_2, ..., a_k)$ are ranked against the query q using the trained similarity model. The model produces the final similarity score taking into account information from both modalities.

4.3 Experimental Setup

4.3.1 Datasets

For training and testing of the proposed models, BioASQ challenge datasets have been used as follows.

- **Training dataset**: We have used the BioASQ challenge dataset from versions 5 and 6 for training [123]. It consists of 2747 questions annotated manually with relevant documents and passages by an expert panel. The relevant documents and passages for each question are taken from the document baseline published by PubMed for 2018 [80]. By the year of BioASQ version 6, the total number of documents was 26,759,399.

The documents are composed by the title and the abstract for the biomedical paper. Since the dataset provides only positive passages in original, the negative passages have to be collected by participating teams. We collected negative passages by using these two approaches, first taking the passages which are in the relevant document but does not answer the question and then taking the same number of passages that were collected previously but this time they come from unrelated documents. The final statistics for the training data set are in Table **4-1**.

#Questions	#Pairs	#Positives	#Negatives
2747	345,247	25,621	319,626

Table 4-1: BioASQ 5 & 6 training dataset with negative samples

- **Test dataset**: For testing, we used the test data set that was provided in the 2019 challenge Bioasq 7 version. The dataset is comprised of 5 batches, each containing 100 questions and variable candidate passages. The statistics for testing dataset are presented in Table **4-2**.

Batch	**# of questions**	**Avg. relevant documents**	**Avg. Relevant passages**
Batch 1	100	9.4	5.5
Batch 2	100	12.4	7.5
Batch 3	100	17.3	11.28
Batch 4	100	13	8.6
Batch 5	100	8.9	4.9

Table 4-2: BioASQ 7 test dataset statistics

The experimentation process is divided in two phases, the first one focused on document retrieval process and the second one for snippets retrieval.

Document Retrieval

We indexed the full data of 2018 PubMed baseline in ElasticSearch engine (ES) version 6.2.2 with the default configuration, this is our baseline. The number of processed files were 928 and the total number of medical articles was 26,759,399. For each article, we extracted the title, MESH concepts and abstract to be indexed. The indexing time was around 18 hours in an Intel Xeon processor Intel(R) at 2.60GHz with 82 GB RAM and GeForce GTX TITAN X.

Other index that we used was *Index v3*; also generated in ElasticSearch engine but using the parameters b and $k1$ for BM25.

We proposed three experiments for document retrieval: Retrieve 10 most relevant documents with *BM25 index-v2* (our BioASQ 6 strategy), the second one is retrieve documents with *BM25 index-v3* and re-rank using Word Mover's Distance. The last one is retrieve documents with *BM25 index-v3* and re-rank using Doc Centroid Rerank.
The experiments were implemented with the BioASQ 7 data and the results are presented in Tables **4-3** and **4-4**

Model	**Mean precision**	**Recall**	**F-Measure**	**MAP**	**GMAP**
BM25_v2_10d	0.20784	0.47294	0.22974	0.13196	0.02574
BM25_v3_WMD	0.21204	0.481	0.23484	0.1138	0.0186
BM25_v3_centroid	0.20184	0.44906	0.22186	0.12038	0.01588

Table 4-3: Document Retrieval results for BioASQ 6 (summarized)

Batch	**System**	**Mean precision**	**Recall**	**F-Measure**	**MAP**	**GMAP**
6b1	BM25_v2_10d	0.212	0.5061	0.2449	0.1408	0.0284
	BM25_v3_WMD	0.232	0.5322	0.2653	0.1336	0.0256
	BM25_v3_centroid	0.181	0.3725	0.2004	0.1145	0.0041
6b2	BM25_v2_10d	0.2301	0.5286	0.2549	0.1569	0.0341
	BM25_v3_WMD	0.2301	0.5328	0.2564	0.1218	0.0266
	BM25_v3_centroid	0.2301	0.5328	0.2564	0.1334	0.0282
6b3	BM25_v2_10d	0.2551	0.5245	0.2603	0.1806	0.0576
	BM25_v3_WMD	0.2571	0.5286	0.2622	0.1511	0.0326
	BM25_v3_centroid	0.2571	0.5286	0.2622	0.1593	0.0369
6b4	BM25_v2_10d	0.183	0.4983	0.2127	0.0903	0.0049
	BM25_v3_WMD	0.183	0.4983	0.2127	0.0903	0.0049
	BM25_v3_centroid	0.183	0.4983	0.2127	0.1111	0.0061
6b5	BM25_v2_10d	0.159	0.3072	0.1759	0.0912	0.0037
	BM25_v3_WMD	0.158	0.3131	0.1776	0.0722	0.0033
	BM25_v3_centroid	0.158	0.3131	0.1776	0.0836	0.0041

Table 4-4: Document Retrieval results for BioASQ 6

Snippet Retrieval

For training it was observed that dataset was very unbalanced, only 8% of the total number of pairs are labeled as a relevant answer. To balance the dataset, the sampling is carried out using the same number of positives and negative examples, this strategy is also applied in the validation phase.
The model training was done using RMSprop optimization algorithm with 32 samples in mini-batch and the defined loss function is binary cross entropy. The number of maximum

epochs was set to 50. In each epoch, we evaluate MAP and MRR, and after 5 epochs without any improvement in MAP metric, we apply early stopping to avoid over-fitting.

Information Fusion Approaches

As we have used information that comes from textual representation and structured representation the combination of those modalities is also a model parameter to explore. In that way we have evaluated four different configurations to measure the performance involving different information representation approaches:

- Approach 1: Only Textual Representation. Questions and candidate answers are represented using only the textual embedding.
- Approach 2: Only Structured Representation. Questions and candidate answers are represented using only the concept embedding.
- Approach 3: Mixed Data Representation Intermediate Method –MIF. In this model the fusion of textual and structured representations is carried out in an intermediate dense layer after the textual and structured patterns are identified by the CNNs layers **4-4**. The merged layer is then connected to the sigmoidal output unit with dropout as regularization strategy.
- Approach 4: Mixed Data Representation Late Fusion –MLF. In this approach each model (textual and structured) independently calculates a score for each question-answer pair, $score\_t$ for textual representation, and $score\_s$ for structured representation. Lastly, a linear combination produces the final score f_score, as shown in Equation 4-5. The *alpha* value was found using cross validation with the validation partition; it was set to 0.73.

$$f\_score(q, a_k) = (1 - \alpha) * score_t(q, a_t) + \alpha * score_s(q, a_s) \tag{4-5}$$

Model parameters

The model hyper-parameters were tuned using hyper-parameter exploration. The parameters chosen are listed next.

- **Convolution Parameters:** The number of convolutional filters used are 64, width 3 and length 3, the stride used is 1 without padding.
- **Convolution Activation Function:** After a convolutional layer, it is useful to apply a nonlinear layer [34]. We tested different activation functions and RELU gave us the best performance.

- **Pooling Layers:** For the pooling layer, we used max pooling.

- **Dropout Layer:** We add a dropout layer as a regularization strategy [116], setting the parameter in 10%.

Finally, the number of parameters to learn in our model is not very high (5,192) compared with other Convolution Neural approaches used in similar tasks (Question Answering) which are in order of millions and hundreds of thousands [109, 39]

4.3.2 Model Tuning

In this section, we will describe the strategy used to improve the overall performance of our system. The metrics were calculated over the training dataset released by BioASQ for the 6th version.

- Mesh concept indexing: Document retrieval is mainly based on Elastic Search key-word matching evaluation with BM25 ranking function. We used a cross-fields query approach which looks for each term in the title, abstract and concepts indexed fields. Considering the retrieval of 10 most related documents, the performance using cross-fields approach were (Recall = 0.24, MAP = 0.19) while not using this were (Recall=0.278, MAP= 0.221).

- Word representation: The choice of a good word representation is important to generate a semantically good model where relations between terms or sentences are more easy to establish. We tested our system using different pre-trained word2vec models and the best representation was the skip-gram model provided by NLPLab, which is trained on Wikipedia and PubMed abstracts [71]. The MAP score in the snippet retrieval sub-task improved from 0.126 to 0.142.

- Training dataset generation: The training corpus was generated with questions and answer passages extracted from 2016, 2017 and 2018 BioASQ training datasets. We tested different rates of negative samples (passages in related documents that does contain the answer) in order to increase the negative sample coverage. This assumption is based on the hypothesis that it is not easy to determine that a related snippet does not contain the answer. With a higher negative sample generation, these cases are more common, and the method can learn a better discriminant function. The rate that experimentally achieved the best results considers using 10 negative samples per 1 positive sample. The MAP score in snippet retrieval sub-task, improved using 6b training partition from 0.142 to 0.151.

4.4 Results and Discussion

In this section, we present the results for the sixth version of BioASQ challenge in task B phase A. The first sub-task is to retrieve the most related articles based on a question posed in natural language. The second one is to retrieve the snippets that have more correlation with the question in order to use them to compose an answer. The answer composition is carried out in phase B, which was not the scope of our participation.

4.4.1 Document Retrieval

The results shown in the Table **4-5** reveal, that our ES document retrieval implementation did not have a good performance, the recall obtained is low in all the batches. In the first batch, we had a technical issue that corrupted the results, it also happened for snippet retrieval. The best result was obtained in batch 3 (Recall = 0.49), the team leader in this batch reached 0.56, an important difference. As it was mentioned before, document retrieval is very important for snippet retrieval, it is the first information filter and it feeds the method to rank their snippets. Despite the low recall in this step, we will see in the next section that snippet retrieval scores are very promising.

Batch	Document Retrieval	
	Mean precision	Recall
	F-Measure	MAP
1	-	-
	-	-
2	0.1150	0.4685
	0.1621	0.0709
3	0.1320	0.4984
	0.1782	0.0891
4	0.1240	0.4467
	0.1717	0.0846
5	0.0890	0.2961
	0.1260	0.0540

Table 4-5: Document retrieval results

4.4.2 Snippet Retrieval

In this stage, we analyzed in depth the returned set of documents from the previous method, and identify the text snippets that can answer the posed question.

Based on the evidence shown in Table **4-6**, the snippet retrieval approach obtained a good performance. We could have had a better performance in snippet retrieval with a higher score in document retrieval, but it was enough to reach the second position in all the batches except the first one (due to the technical issue).
We can state that the proposed method exhibits a very competitive performance compared with other methods.

Batch	Snippet Retrieval	
	Mean precision	Recall
	F-Measure	MAP
1	-	-
	-	-
2	0.1111	0.2426
	0.1416	0.0938
3	0.1614	0.2657
	0.1877	0.1344
4	0.1043	0.2180
	0.1306	0.0980
5	0.0404	0.1134
	0.0542	0.0475

Table 4-6: Snippet retrieval results

4.5 BioASQ 6 and 7 Participation Overview

In this section we are going to give a quick report of our participation in BioASQ 6 challenge (2018) [76], and the edition 7 [74], in the 8 version we have not participate in all the batches.

4.5.1 BioASQ 6 Participation

For the related edition, passage retrieval task was tackled by 50 different systems, developed by 15 teams. In this task the winner team was a collaboration between Google and University of Athens "AUEB" [20]. They have used novel extensions of deep learning models for retrieving question-relevant snippets, using a self-trained biomedical word embedding and a DRMM model [36].
The proposed model that is described in detail in the following paper [102], achieved the second position despite of the low performance in document retrieval step. In the figure **4-6** we can observe the averaged scores in all the five batches. The obtained results are very competitive, regarding to official ranking metric (MAP), we are in the highest quartile.

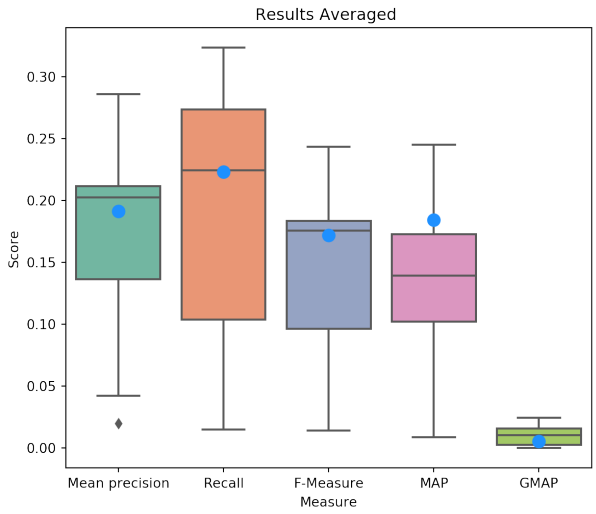


Figure 4-6: Snippet retrieval results in BioASQ 6 (2018), blue point is our model

4.5.2 BioASQ 7 Participation

In this edition of the shared task the number of systems were 73 different systems, developed by 18 teams [74]. Most of the teams used Deep Learning approaches and the winner team was the same of the previous edition, it used a different deep learning approach based on BCNN model [143] which reach the highest MAP scores [89].
The method that was described in this chapter reached the first position in the first batch, and the second in the remaining batches. As in version six, our proposed document retrieval method was not as competitive as the winning team's despite improvements made with different re-ranking approaches described in this chapter. It also impacted the outcome in passage retrieval in which we were beaten by a small margin by the opposing team. In the figure **4-7** we can observe the averaged scores in all the five batches.

4.6 Conclusion

In this chapter we presented a passage retrieval method for biomedical domain that takes advantage of multi-modal information coming from textual and semantic information sources such as UMLS. The method takes information represented in each modality and fuses it together using two different strategies (late and intermediate fusion), thus improving the performance of the system compared to the single modality. The method is competitive with state-of-the-art models for biomedical passage retrieval.
To test the effectiveness of the proposed model, we participated in the BioASQ 6 and 7 challenge, which evaluates the tasks of document and passage retrieval. The results obtained in the 7th version for the passage retrieval were competitive, reaching the second place in the batches (2, 3, 4, 5) and the first place in the batch 1. If we consider that the results in

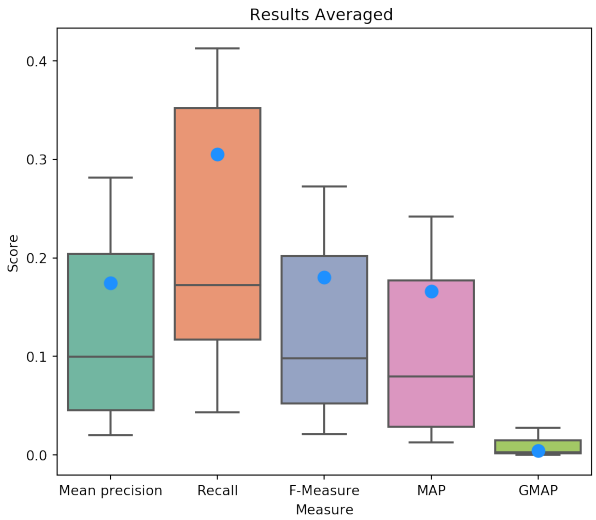


Figure 4-7: Snippet retrieval results in BioASQ 7 (2019), blue point is our model

the phase of document retrieval were not among the best, it is even more remarkable this result in passage retrieval task.

5 Deep Fusion of Multiple Term-Similarity Measures

The fusion method from the previous chapter combines the textual and semantic information sources using a deep neural network architecture. Since each modality is incorporated separately in the deep neural network architecture, any relationships which can potentially be present between modalities are not fully exploited. The approach presented in this chapter is essentially different, although textual and semantic sources are still involved, their representations are combined into a single representation to be jointly exploited so that the modalities complement each other. In addition, considering that the word representation does not provide full coverage for the possible sequence terms, we enhance the representation by adding a term-to-term co-occurrence-based similarity. The experimental evaluation results obtained are, by a wide margin, better than those obtained with the previous methods.
The methods presented in this chapter were published in the following paper:

- Rosso-Mateus, A., Montes-y-Gómez, M., Rosso, P., & González, F. A. (2020). Deep fusion of multiple term-similarity measures for biomedical passage retrieval. Journal of Intelligent & Fuzzy Systems, 39-2, 2239-2248.

The remainder of the chapter is organized as follows: Section 1 presents the introduction and motivation for the proposed passage retrieval model; Section 2 shows the model architecture and implementation details; Section 3 presents a systematic evaluation of the method; finally, Section 4 exposes some conclusions and discusses our future work.

5.1 Introduction

Almost all passage retrieval methods calculate some sort of similarity between the query and the passage. Some similarities are based on term-term similarities and others involve more semantic information. Semantic similarity measures are mainly based on large corpora where important relational patterns are extracted. Some of the approaches, as for example probabilistic hyperspace analog to language (HAL) [6], propose a semantic window of length K which is moved across the corpus of text.Terms contained in the window co-occur with a strength inversely proportional to term by term distance. They reported that when window size increases (K greater than 5), there was a diminishing on performance in information retrieval task.

Other approaches take into consideration the semantic and ontological relationships that exist between words. Thus, based on this knowledge, semantic similarity can be calculated following the minimal path between two nodes [117]. Ramage et al. have proposed a random walk algorithm [96] that compares the random walk graph generated between two terms to measure the semantic relatedness. They used WordNet and corpus statistics. These approaches are efficient when the coverage of the ontology is wide; in the biomedical domain, it is hard to have a 100% coverage.

Apart from ontological text representations, recently, authors have been working with word embeddings. These models represent each word as an n-dimensional vector, with the property that semantically related vectors are close to each other. Cosine similarity is one of the similarity measures that can be applied when text is represented as vectors. Other measures include Euclidean distance, soft-cosine similarity, and so on. Based on that, it can be said that the similarity measure election will guarantee the success of the model.

In Mikolov's model [67], the semantic relation strength between a pair of terms is given by the occurrence in context windows. This parameter choice will punish distant terms that can give important information, e.g., the following snippet of a biomedical article has two highly related entities "**calcitonin**" and "**migraine**" with 20 terms separation between them:

Calcitonin *gene-related peptide, the most abundant neuropeptide in primary afferent sensory neurons, is strongly implicated in the pathophysiology of* ***migraine headache****, but its role in* ***migraine*** *is still equivocal.*

The consequence will be a low spatial correlation in the semantic vector space. However, in some domains (such as biomedical), it is important to capture also more 'topical' relationships [55].

In this work, we propose a passage retrieval method that takes advantage of different resources to build similarity measures. The obtained representation fits a deep learning model to extract similarity patterns in order to improve the performance on the passage retrieval task. The proposed approach combines three different similarity representations: 1) word2vec embedding cosine similarity, 2) term co-occurrence and 3) concepts co-occurrence. These similarities, extracted from large corpora, contribute with local and topical relatedness. The way to exploit these similarity patterns is based on a convolutional neural network.

5.2 Method

5.2.1 Overall architecture

The overall architecture is depicted in Figure **5-1**. Figure swim lines indicate different stages which are explained in the following sections.

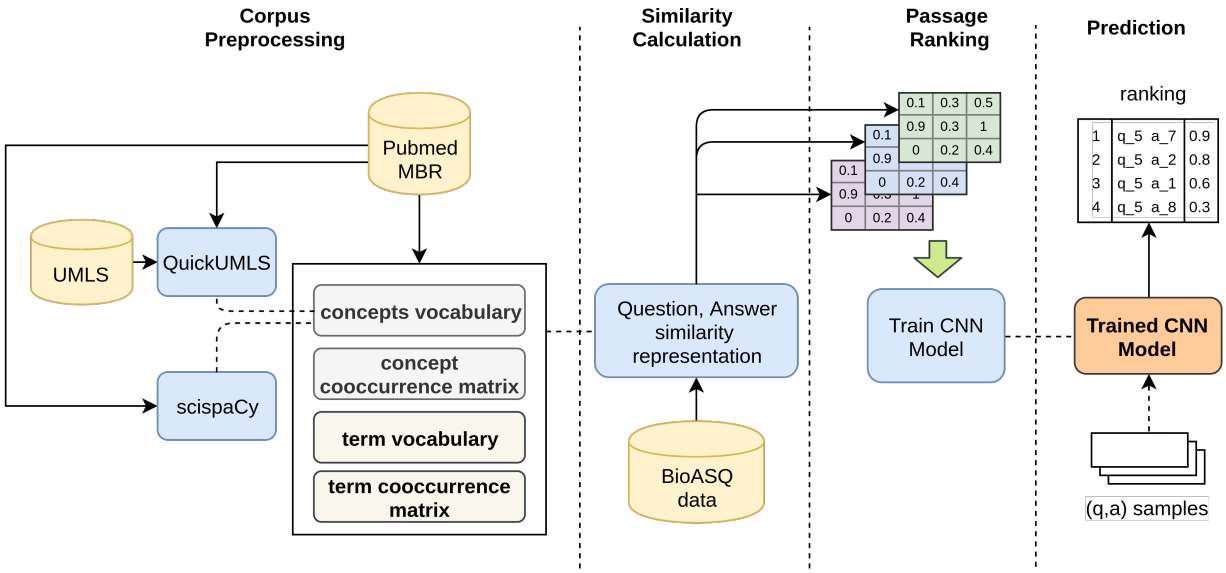


Figure 5-1: Passage retrieval overall architecture

Corpus Preprocessing

The first part of the process is to calculate the co-occurrence between pairs of terms and pairs of biomedical concepts. In this stage, we take a random sample of 30.000 biomedical documents from PubMed Baseline Repository (MBR) document set [80]. The objective is to build the vocabulary and to calculate the co-occurrences for both terms and concepts. For the later, we need to identify the biomedical concepts. For this task, we have used the terminology data source UMLS Meta-thesaurus[1] which contains information about over 1 million biomedical concepts and 5 million concept names. As the process to match every term to a concept is computationally expensive, we take advantage of the QuickUMLS tool provided by Soldani et al. that has a good performance identifying concepts in large texts [114].

Experimentally, we have determined that the coverage of UMLS is not 100%. To overcome this limitation, a second check is performed with the Scispacy tool [78]. This Spacy model provides biomedical named entity recognition which increases the biomedical concept identification coverage. Once the vocabularies of terms and concepts were built, we filter out frequent terms and concepts which provide less information. Also, very rare terms and concepts are not taken into account. Figure **5-2** shows the count frequency of term and concepts.

Now we have to indicate if a word appears in a given document and if keep it in a binary vector. The resulting matrix will have a dimension $NtimesM$, where N is the number of documents and M is the vocabulary size, with value 1 when the vocabulary word appears in the given n-th document.

[1]UMLS Meta-thesaurus http:umlsks.nlm.nih.gov

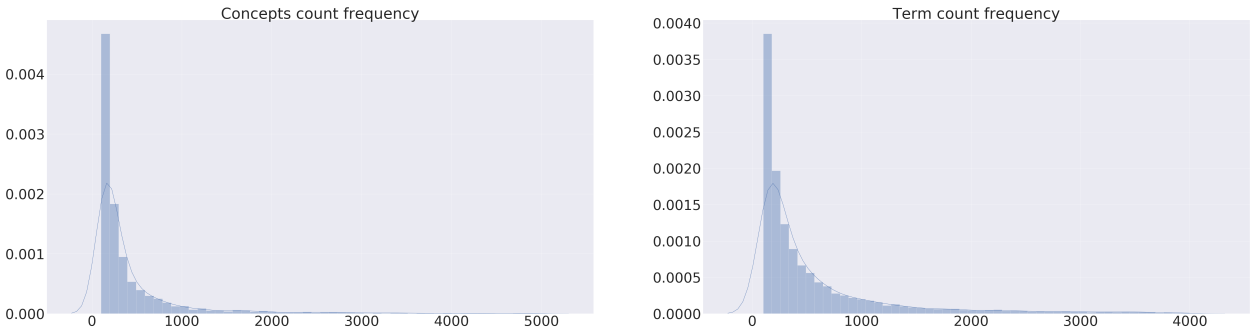


Figure 5-2: Term and concept count frequency

With the document-word appearance matrix X calculated, we have to calculate the word by word normalized co-occurrence matrix to achieve that we apply the Equation 5-1.

$$Tc\_norm = (XX^T)(1/diag(XX^T) \tag{5-1}$$

The produced information in this step is:

- Term vocabulary
- Term co-ocurrence matrix
- Concept vocabulary
- Concept co-occurrence matrix

The process was also applied to sentence level co-occurrence, but instead of calculating the co-occurrence in documents, we split them into sentences and continued with the same process. Empirically, we have stated that document level similarity matrix achieves higher scores. Henceforth, in this paper we will understand co-occurrence similarity as document level similarity. Once co-occurrence matrices are calculated for terms and concepts, it is time to represent the model input data in the similarity matrices that the CNN model expects (co-occurrence term similarity; co-occurrence concept similarity and cosine similarity).

Co-occurrence similarity

The co-occurrence-based representation offers an additional perspective on the semantic term-to-term relationships and therefore a representation for the question-answer pair (q, a). To obtain this representation we take the value of the pre-calculated co-occurrence for the pair (q, a), if q or a are not in the vocabulary then the value is equal to 0, this allows us to align the representation of the similarity matrices in the three tensor dimensions. This process is also followed for medical concepts identified as such.

Cosine similarity

Cosine similarity is another question and passage data representation. Each pair (q, a), is defined as a weighted cosine similarity score between question and passage pair words, as described below.

- **Step 1: Pre-processing**: Question and answer sentences are cleaned and tokenized; a grammatical tagging is carried out with NLTK POS-tagger to extract syntactical information that will be used for the salience weighting; each term is transformed later in a vector embedding using a pre-trained word2vec model provided by NLPLab, which was trained on Wikipedia and PubMed documents [2].

- **Step 2: Calculate similarity matrix** (qt_i, at_j): Each i, j-entry of the similarity matrix M_t, represents the semantic relatedness of the i-th question term and the j-th answer term according to their word embedding.

- **Step 3: Matrix weighting** M_t: as not all terms are equally informative for measuring text similarities [56, 30], we have applied a term weighting based on the grammatical function of the term pair "salience score" $sal(qt_i, at_j)$.

  The term pair similarity (qt_i, at_j) is calculated as Eq. 5-2 shows.

$$M_{i,j} = scos(qt_i, at_j) * sal(qt_i, at_j) \tag{5-2}$$

$$scos(qt_i, at_j) = 0.5 + \frac{qt_i \cdot at_j}{2 \left\|qt_i\right\|_2 \left\|at_j\right\|_2} \tag{5-3}$$

$$sal(qt_i, at_j) = \begin{cases} 1 & if\ imp(qt_i) + imp(at_j) = 2 \\ 0.6 & if\ imp(qt_i) + imp(at_j) = 1 \\ 0.3 & if\ imp(qt_i) + imp(at_j) = 0 \end{cases} \tag{5-4}$$

  The value of $imp(x)$ function is based on the POS-tagging label. We consider verbs, nouns, and adjectives to be "important" [56, 30]. As a consequence $imp(x)$ is 1 for important label and 0 for the others, if both terms are important the $imp(qt_i)+imp(at_j)$ would be 2, and therefore the weighted will be 1.

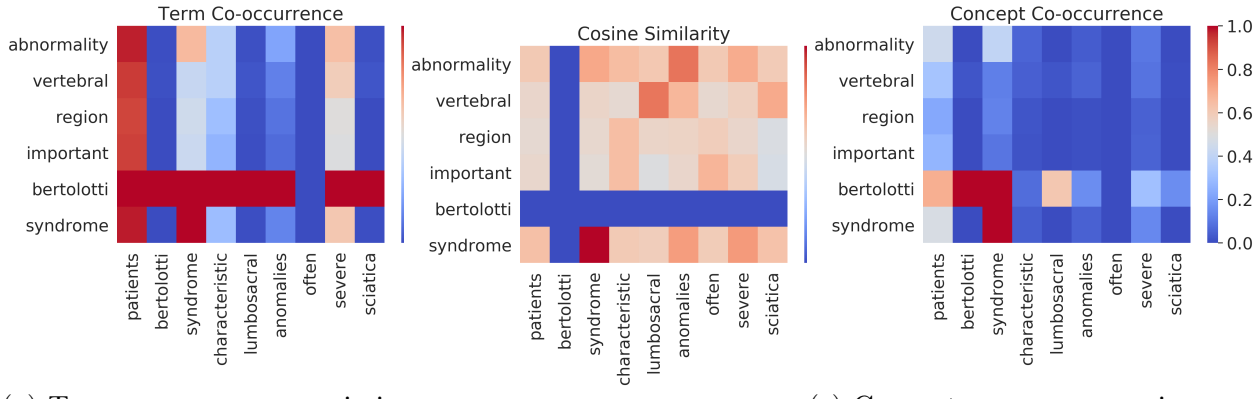
[2] BioNLP word vector representation, trained with biomedical and general-domain texts http://bio.nlplab.org

Here we can observe that the three similarity measures used in the proposed approach capture different aspects of semantic relatedness. When only using one similarity measure, the method may fail to capture all the important aspects of the semantic relatedness. This can be seen in the following example:

Q: Abnormality in which vertebral region is important in Bertolotti's syndrome?

A: Patients with Bertolotti's syndrome have characteristic lumbosacral anomalies and often have severe sciatica.

The three similarity matrices visualisation is represented with the following heat maps, see Figure **5-3**. It can be observed that the cosine similarity matrix does not have a high value for the "Bertolotti" term. It is because there is no vector representation for the term, but the co-occurrence matrices for term or concept have the highest values in the related cell values. In the same way, the "Bertolotti" concept is highly correlated with "syndrome" and "lumbosacral" in the concept co-occurrence matrix which are important concepts to answer the question. In the case of the term co-occurrence matrix, the similarity is less precise but gives a high score for the related term "sciatica". As the similarity matrices show, they are complementary to each other and they produce important patterns to rank a set of candidate answers.



(a) Term co-occurrence similarity matrix

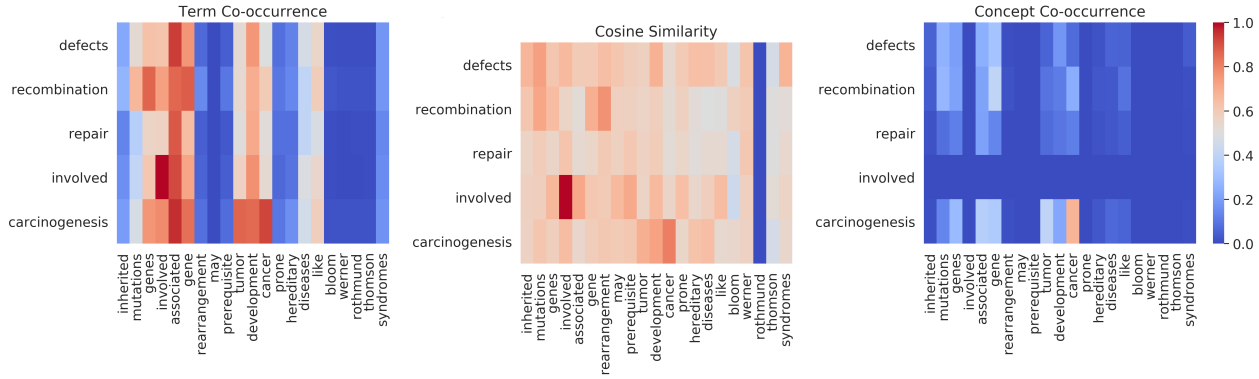
(b) W2V cosine similarity

(c) Concept co-occurrence similarity matrix

Figure 5-3: Example 1. Similarity matrices

Another example of how similarity measures contribute to an improved representation of the query-passage relationship is presented in the following example:

Q: Are defects in recombination repair involved in carcinogenesis?



(a) Term co-occurrence similarity matrix

(b) W2V cosine similarity

(c) Concept co-occurrence similarity matrix

Figure 5-4: Example 2. similarity matrices

A: Inherited mutations in genes involved in HR are associated with gene rearrangement and may be a prerequisite for tumor development in some cancer-prone hereditary diseases like Bloom, Werner, and Rothmund-Thomson syndromes.

We can see similarity matrices as heat maps in Figure **5-4**. For this case, cosine matrix has a high similarity score between "recombination" and "rearrangement", while co-occurrence representation score is low. All three matrices have a high score for "carcinogenesis" in the question and "tumor", "development" and "cancer" in the answer.

The objective with the incorporation of additional and complementary information is to feed the neural model with meaningful features that allow the model to identify when a question and answer pair are highly correlated. During the training phase, the CNN model has to determine those similarities patterns that we hypothetical highlight.

Passage ranking

Convolutional Neural Nets (CNN) were originally developed for image processing, where the important information may appear on arbitrary regions of the image, represented frequently as a 3 channel RGB matrix. The same assumption can be applied to our similarity matrices.

Once the (q, a) pairs are represented as the three similarity matrices, we feed them to the CNN model presented in Figure **5-5**. The CNN layer will identify word-similarity patterns in each of the three channels. The patterns are captured for the 64 filters to be then subsampled by a pooling layer. The pooling layer for all the filters is merged with a fully connected layer. Finally, an output sigmoid unit produces a similarity score based on the evidence coded by the neural networks units activation values.

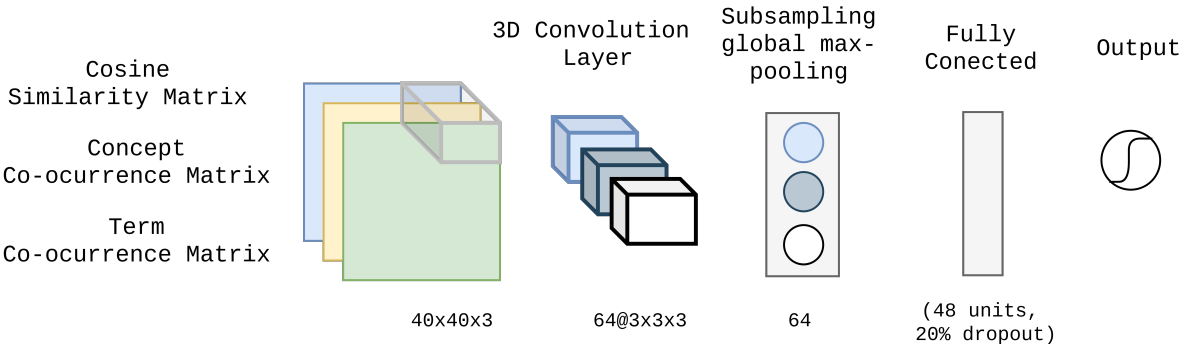


Figure 5-5: Multiple channel convolutional neural network

5.2.2 Prediction

Once the training phase has been completed we obtain a similarity discrimination model that is capable of measuring the semantic correlation between question and answer pairs and produce a final score.
The next step is to use the model to rank candidate answers $(a_1, a_2, ..., a_k)$ against a given query q. The candidate answers are retrieved based on the highest scores.

5.3 Experimental Evaluation

The experimentation was carried out over the BioASQ 6 challenge dataset. We evaluate different method combinations in order to measure how important is each of the similarity measures for the passage retrieval task. Finally, we will combine all three similarity matrices to validate the complementary information hypothesis.

5.3.1 Data set

The training was conducted with the question and answer pairs from the 2016, 2017, and 2018 BioASQ Task B training datasets. As previously mentioned, the BioASQ dataset does not include negative samples; we collected negative samples from both related and unrelated documents. The complete statistics of the training dataset are described in the table **5-1**.
The obtained dataset was very unbalanced, only 7% of the total number of pairs are labeled as a relevant answer. To balance the dataset, the sample extraction in the training phase was done with the same number of positives and negative samples, this strategy is also applied in the validation phase.

5.3.2 Experimentation models

In order to compare the discriminative power of the proposed model using the related three similarity feature matrices, we introduce the following model configurations: 1) using just

#Questions	#Pairs	#Positives	#Negatives
3295	500,248	32,944	467,304

Table 5-1: BioASQ dataset with negative samples

the term co-occurrence matrix as input to the CNN (term); 2) using just the concept co-occurrence matrix as input to the CNN (concept); 3) combine term and concept co-occurrence (term + concept); 4) using the cosine similarity matrix (w2v), 5) combine cosine similarity with term co-occurrence (w2v + term); 6) combine cosine similarity with concept co-occurrence (w2v + concept); 7) combining all three similarity measures (w2v + term + concept). Besides these methods, we will compare the latest configuration (w2v + term + concept) against the proposed baseline models: a self-trained finetuning BERT model (BERT) and the winner model from last year BioASQ challenge (aueb-nlp-5). To give a broad definition of BERT model we are going to detail the process followed to finetune BioBert.

Bert finetunned model baseline

Language pre-trained models have proven to be useful for universal textual representations. One of the last pretrained models is BERT (Bidirectional Encoder Representations from Transformers) which has achieved an important result for different NLP tasks. Recently a pretrained BERT model over biomedical and open domain data was released by Lee et al [50].
In order to validate state-of-the-art methods, we have finetuned BioBert to achieve the passage retrieval task. We have followed the approach for sentence pair classification task. The data used to finetune the model was the same to train the proposed model.

5.3.3 Results and Discussion

The results for different model configurations are reported in Table **5-2**.
Results show that the most informative individual similarity measure is the cosine similarity (w2v) with the proposed POS-tagging salience weighting. Term and concept co-occurrences have very close scores in all the batches when used separately. The combination (term + concept) improves significantly the scores as expected. Combining (w2v + term) and (w2v + concept) is quite similar, the scores are close, but when all three similarity measures are jointly used there are important improvements in the MAP metric across all batches.
We present the following question (Q) and answer candidate (A) example extracted from experimental data-set to show the model contribution.

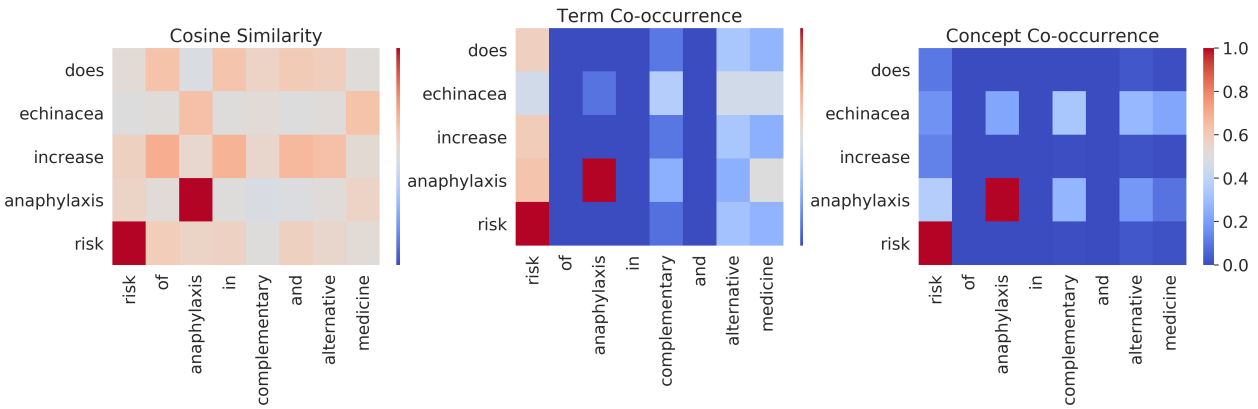
Q: Does echinacea increase anaphylaxis risk?

DFMTS (this model)	B1 MAP	B2 MAP	B3 MAP	B4 MAP	B5 MAP
term	0.1979	0.2842	0.2626	0.1629	0.0857
concept	0.2076	0.2828	0.2617	0.1537	0.0861
term + concept	0.2106	0.3329	0.3008	0.2178	0.0987
w2v	0.1942	0.2946	0.2671	0.1581	0.0914
w2v + term	0.2145	0.3612	0.3289	0.2210	0.1019
w2v + concept	0.2191	0.3547	0.3178	0.2281	0.1101
w2v + term + concept	**0.2322**	**0.3838**	**0.3571**	**0.2409**	**0.1163**

Table 5-2: Snippet retrieval results combining similarity matrices

A: Risk of anaphylaxis in complementary and alternative medicine.

The produced similarity matrices are depicted as heat maps in order to visualize the similarity strength between terms and concepts, see Figure **5-6**. In this example, the concept similarity matrix offers higher values for co-occurrence similarity between echinacea and anaphylaxis allergic reaction. Verifying in the medical literature, there are documented adverse reactions associated with echinacea which support our observations.



(a) cosine similarity matrix (b) concept co-occurrence matrix (c) term co-occurrence matrix

Figure 5-6: Similarity matrices example where concept co-occurrence have a better performance over the others

Model results against baseline

We have conducted our experimentation with the test batches released for BioASQ 6b. In order to compare our results with state-of-the-art methods, we have included last year winner team (Athens University and Google [20]) results. Since snippet retrieval highly depends on document retrieval, and with the objective to make a fair comparison of our proposed method, we asked the winner team to share with us the documents obtained in the document retrieval step. They shared the submitted files and, therefore, a snippet retrieval isolate comparison was possible to carry out.

The scores presented in Table **5-3** for aueb-nlp-5 [20] were extracted from the BioASQ results leader board table. This is the system that reached the highest scores. In the same way, we reported the scores that our Bert fine-tuned model and our fusion model with three similarity measures obtained when using the same set of documents from aueb-nlp-5.

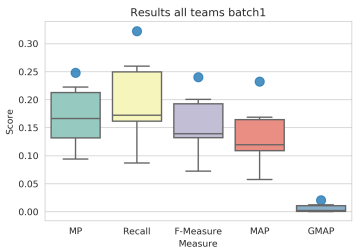
Model	B1 MAP	B2 MAP	B3 MAP	B4 MAP	B5 MAP
aueb-nlp-5	0.1684	0.3187	0.332	0.2138	0.1147
bert	0.106	0.1389	0.2021	0.1223	0.063
DFMTS (concept + term + w2v)	**0.2322**	**0.3838**	**0.3571**	**0.2409**	**0.1163**

Table 5-3: Snippet retrieval results using the documents provided by AUEB [20]

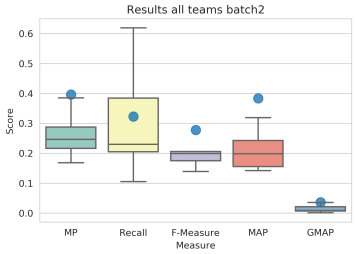
The proposed model scores are consistent in the five batches and the difference against the best model from last year (aueb-nlp-5) is 3.5 percent points on average, across all the batches. We can also see that the BERT based model is competitive, although their scores are below those from the other two models.

The next comparison was carried out against the 15 best models from the 2018 BioASQ challenge. In order to visualize the scores in a more friendly way, we have consolidated the results from the leader board table in a box plot, see Figure **5-7**. There is one box plot for each batch, and the X-axis corresponds to the reported metrics in BioASQ 6 (mean precision, F-score, recall, MAP, GMAP). The blue point is the score obtained with our model using the documents supplied by [20].

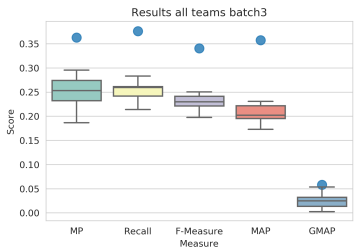
In batch1, batch3 and batch4 we reached the best results as ilustrated in the boxplot. In batch2, the only measure where the model is not the best is recall. Still, they are in the highest quartile. In the last batch, the scores for all the teams are lower than in previous batches. The result of our model is the highest in MAP and competitive according to the other metrics.



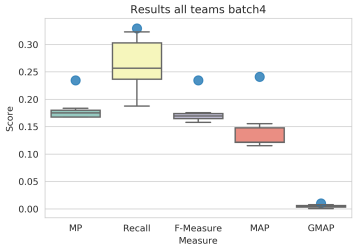
(a) Results batch 1



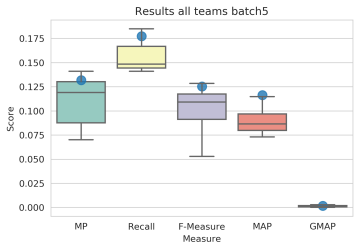
(b) Results batch 2



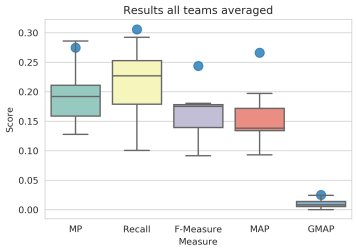
(c) Results batch 3



(d) Results batch 4



(e) Results batch 5



(f) Results averaged for all batches

Figure 5-7: 15 best systems results for task 6b, blue points correspond to the proposed model

5.4 Conclusion

In this chapter, we have presented a novel approach for biomedical passage retrieval. The proposed method is based on different similarity measures which offer complementary information in order to semantically match question and answer passages.

The proposed similarity measures come from concepts and term co-occurrence, in addition to a word-embedding cosine similarity. Concepts are extracted using the UMLS terminology data source and a biomedical-trained Scispacy model. The multiple similarity representation is exploited by a convolutional neural network which extracts similarity-based patterns and produces a semantic relatedness score, which is further used to rank the answer candidate passages. We have tested different combinations of similarity measures, and the most accurate was the one in which we used all three similarity measures, which validate the hypothesis that the similarities are complementary to each other.

The proposed model was tested within BioASQ 6b dataset and the scores obtained were compared against the best models reported for the 2018 challenge. The obtained results showed that the proposed model outperformed all the methods used in BioASQ challenge with a substantial difference. Motivated by the obtained results, future work will be focused on extending the similarity representation and exploiting it with more sophisticated neural models that better use the multiple information.

6 A Deep Metric Learning Method For Biomedical Passage Retrieval

Deep learning-based passage retrieval methods usually approach the problem as a classification problem that attempts to discriminate relevant passages from non-relevant ones. Training is performed by presenting random samples (positive or negative), although some negative samples are semantically related to the answer and others are completely different. This semantic relationship between question and passage can be modeled more naturally using a deep-metric-learning approach, where relevant passages have a distance close to zero to the question and non-relevant passages have a large distance.
This chapter presents a deep-metric-learning approach that employs a triplet input consisting of (question, positive and negative passage), but unlike commonly used architectures in this type of approach, we propose a siamese architecture instead of a triplet network, where each sub-network captures the interactions between the question and the passage of the positive and negative pairs respectively. In addition, a suitable sampling strategy is presented that allows to improve the model performance by presenting first easy negative training samples and then more difficult ones.
The method discussed in this chapter were presented in the following paper:

- Rosso-Mateus, A., González, F. A., & Montes, M. (2020, December). A Deep Metric Learning Method for Biomedical Passage Retrieval. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 6229-6239).

6.1 Introduction

Metric learning has been broadly used in face identification and other image processing tasks. This approach has a powerful and simple mathematical formulation that allows to produce a compact representation in a metric space that can be used to identify image correspondences. The same idea can be applied to the passage retrieval task where answer passages should share semantic patterns with the question and this can be measured by a metric in an appropriate metric space. This idea has not been explored in depth in the context of passage retrieval, except for the work of [18], where a siamese network was used for learning a metric between questions and candidate answers in an open-domain question answering task on a proprietary dataset.

This chapter presents a novel deep metric learning method that learns a metric between question and passages bringing close semantically related pairs. Most of the metric learning approaches learn to embed samples in a latent space where a metric (usually Euclidean) captures relationships between samples. The proposed approach directly learns the metric fusing different similarity measures through a siamese convolutional deep learning architecture. Also, the chapter presents a sampling strategy that chooses easy and then hard negative samples in the training phase, improving the overall model performance. The experimental results show that the method is able to induce a metric between questions and passages that helps to discriminate relevant passages from non-relevant passages.
The proposed architecture is similar to a triplet network (because of the three inputs: question, answer passage, non-answer passage) and also to a siamese architecture because it is composed of two convolutional neural networks with shared weights. However, different from these, it allows to extract important semantic features from several question-passage internal similarity measures that provide a complementary view of their relatedness. The similarity measures include a structured view of the question and passage, incorporating valuable information that is usually available in close domain problems.
To validate the model performance we carried out a systematic evaluation considering a widely used domain-specific collection, the BioASQ dataset [124], and comparing it against state-of-the-art models. The results show that the performance of the proposed model outperforms previous approaches with a wide margin. The main contributions are the following:

- We formulate a novel deep metric learning architecture which encodes question-passage semantic interactions improving state-of-the-art performance in biomedical passage retrieval.
- We develop an informative sample filtering method that helps to identify easy and hard negative samples to be used during training leading to faster convergence and better performance.

It is important to highlight that the proposed model could be easily implemented, and the number of its parameters is much less than in the state-of-the-art models [20], which have in the order of millions while ours in the order of thousands.
The rest of chapter is organized as follows: Section 2 shows the details of the proposed metric learning method; Section 3 present the sampling strategy; Section 4 presents a systematic evaluation of the method; Section 5 discusses the results against the state of the art models; finally, Section 6 exposes some conclusions and discusses our future work ideas.

6.2 Deep Metric Learning For Passage Retrieval (DMLPR)

The traditional deep metric learning approach is composed of two steps. First, a deep neural model is trained to learn a mapping from a given data representation (commonly images) to

an Euclidean space, then Euclidean distances in the learned spaces are expected to measure the dissimilarity between objects [108, 59]. The first deep metric learning approaches used a siamese architecture, where the model receives a pair question-answer and each component is mapped to the Euclidean space by the same neural network. An evolution of this architecture was the triplet network, where the model receives triplets instead of pairs. The triplets consist of two matching examples (positive and anchor) and one non-matching sample (negative). For both siamese and triplet networks, each sample is individually mapped to to the embedding space.
In contrast to the classic metric learning approach, which learns a metric embedding space for individual samples, our approach learns a combined question-passage embedding that codifies the pair relatedness. The proposed architecture is describe in detail in the following sections.

6.2.1 Model Architecture

Our model architecture is presented in Figure **6-1**. The model accepts three text sequences: the question, a passage that answers the posed question (referred as positive), and a passage that does not contain a valid answer (referred as negative). In the first step of the model, the relatedness of question and passages is calculated using different term-level question-passage similarity measures. This similarities are represented as matrices for the positive (q, p_+) and negative (q, p_-) pairs. These matrices feed a siamese convolutional model which identifies the internal patterns of the interactions between question and passages. The internal patterns are then used to calculate a measure of semantic relatedness, these are noted as $dis_{(q,p_+)}$ and $dis_{(q,p_-)}$ for the positive and negative pairs respectively. The model is trained by minimizing the loss function from Equation 6-1, the distances for positive pairs are encouraged to be close to 0, while negatives pairs should have a distance greater than a margin α.

$$\frac{1}{N}\sum_{i}^{N}[dis(q,p_+) - dis(q,p_-) + \alpha] \tag{6-1}$$

The two main blocks of this model, the input layer and convolutional layer, are described in the following subsections. The model implementation is publicity available with downloadable source code in Github [1].

6.2.2 Input layer: Similarity Measures Calculation

Input training samples are composed of a question and two passages, one positive and the other negative. A question-passage pair is represented by its internal semantic interactions, which are extracted analyzing the term-by-term semantic similarity using three different

[1] DMLPR source code https://github.com/****/***

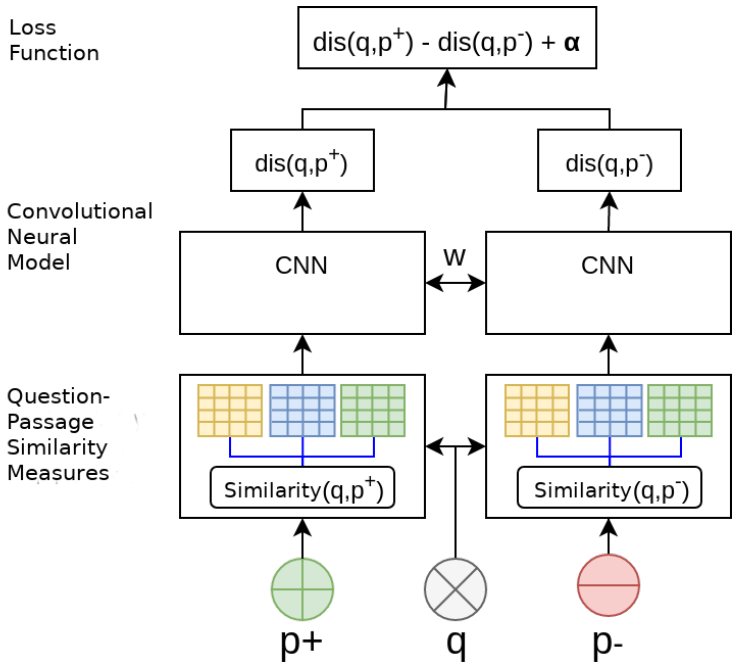


Figure 6-1: Overall model architecture; the input is composed of a question and a positive and negative passages, it includes a convolutional layer and a loss function that compares the distances between the positive and negative pairs.

similarity measures: 1) a word embedding cosine similarity, 2) a term co-occurrence measure, and 3) a concept co-occurrence measure. This representation was presented in a previous work [99], where the internal interactions are defined by three similarity matrices comparing each term in the question q_i against each term in the candidate passage p_j. A brief description of these matrices is presented below.

Cosine similarity: it captures the relatedness of terms using the BioNLP pre-trained word embeddings[2]. After representing terms in the embedded space, their cosine similarity is measured $cos\_sim(\vec{q_i}, \vec{p_j})$ and weighted by its grammatical importance, giving emphasis to verbs, nouns, and adjectives [56, 30].

Term and concept co-occurrence measures: they capture statistical term by term coincidences at sentence level. Concept co-occurrence gives special attention to biomedical concepts discarding common words. In both cases co-occurrence matrices are pre-calculated extracting sentences from 30,000 PubMed biomedical documents[3]. In the case of concept identification, each term is compared against UMLS Meta-thesaurus[4] using the QuickUMLS tool [114]. To increase the concept identification coverage, a second check was done with the Scispacy tool [78].

[2]The BioNLP word vector representation was trained with biomedical and general-domain texts `http://bio.nlplab.org`

[3]NIH PubMed Baseline Repository `https://mbr.nlm.nih.gov/Download/Baselines/2018`

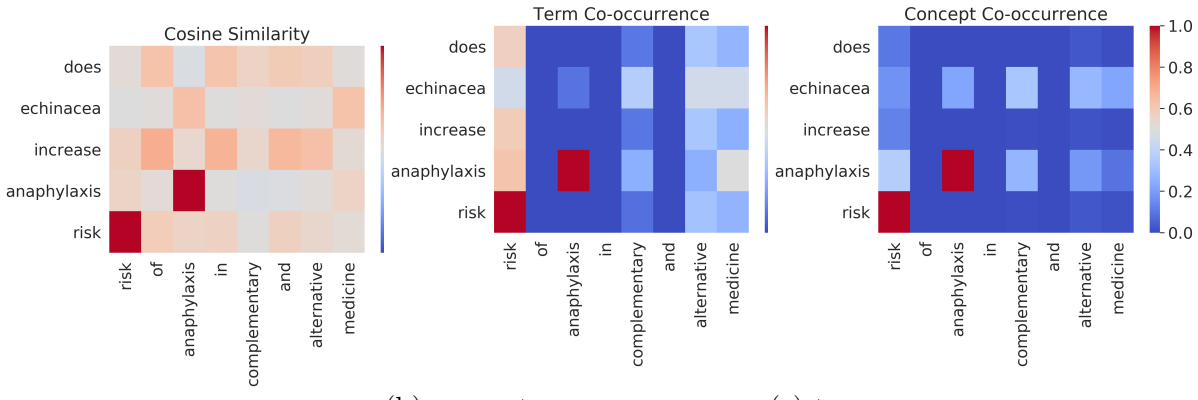
[4]UMLS Meta-thesaurus `http://umlsks.nlm.nih.gov`

To visualize the information captured with the three similarity matrices and to emphasize their complementariness, Figure **6-2** shows some heat maps that indicate the different interactions between a question and a related passage.

Q: Does echinacea increase anaphylaxis risk?

A: Risk of anaphylaxis in complementary and alternative medicine.

In the presented example, the concept similarity matrix offers higher semantic similarity values for question row term 'echinacea' and the related answer passages 'complementary', 'alternative', 'medicine', and 'anaphylaxis' highlighting important relationships. Cosine similarity gives higher values to 'increase' question term and its related row. Term co-occurrence has a similar behaviour to concept co-occurrence, but the last has more focus over important terms. The more informative modality in this example is concept co-occurrence highlighting an important relationship between 'echinace' and the set of terms: 'anaphylaxis', 'alternative' and 'medicine'. This relationships reveal that echinacea has adverse anaphylaxis allergic reactions associated, as is documented in medical literature.



(a) cosine similarity matrix (b) concept co-occurrence matrix (c) term co-occurrence matrix

Figure 6-2: An example of the similarity matrices for a given question (rows) and passage (columns), aiming to visualize the sequences internal interactions.

6.2.3 Convolutional Neural Model

The result of the question-passage similarity calculation is a tensor with three similarity channels. This bi-dimensional multi-channel representation is analogous to that used with images. Convolutional neural networks (CNN) are an effective way of extracting patterns from this kind of representation, and, therefore, we employed a CNN to learn an enhanced representation of the question-passage interactions.

The proposed model has a siamese architecture; each subnet processes a negative or positive input sample pair respectively. The weights of the subnets are shared as it is usual in this kind of architectures. The output of each subnet corresponds to an estimation of the distance for the corresponding input pair as it is depicted in Figure **6-3**.

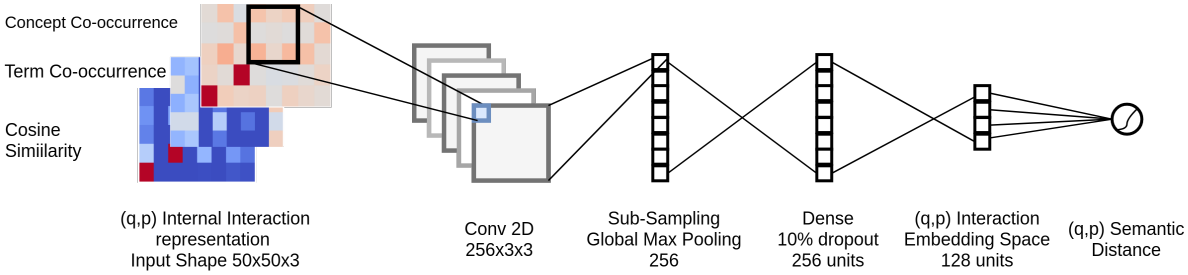


Figure 6-3: Convolutional model used in siamese architecture, each sub-net employ this architecture

The first layer of each subnet is composed of 256 3x3 convolutional filters with a Relu activation function. This layer acts as a feature extraction layer analyzing similarity patters in three dimensions. The identified patterns are then summarized by a global max-pooling layer which is connected to a fully connected layer with 128 units and Relu activation. Finally, a sigmoid unit outputs the estimated distance measure.

6.3 Informative Negative Passage Identification

Selecting informative training samples is very important in deep metric learning, as it is described in previous works [21, 44]. Our approach discriminate hard negative samples based on the semantic relatedness of question and passage pairs using the cosine similarity over BiosentVec sentence embeddings [23]. During training, we first feed the model with easy negative samples, and then with hard negative samples that are more challenging to classify. The process to filter hard and easy training samples is as follows:

1. **Represent samples in an embedded space**: question and passage text sequences, q_i and p_j, are transformed to its BioSentVec embedding representation [23]; the vectors $(\vec{q_i}, \vec{p_j})$ are obtained.

2. **Calculate the similarity between question and passage**: we employed the cosine similarity to measure the semantic relatedness between each question and candidate passage, $cos\_sim(\vec{q_i}, \vec{p_j})$.

3. **Estimate the densities for negative and positive samples**: based on the obtained similarity scores, we calculated the density for positive and negative samples; refer to Figure **6-4**.

4. **Filter hard negative samples**: for each negative sample x, we determined whether it is hard or easy by comparing $p(x \in positive)$ and $p(x \in negative)$; if the sample is more likely to be positive, then it is considered 'hard', otherwise it is labeled as 'easy'.

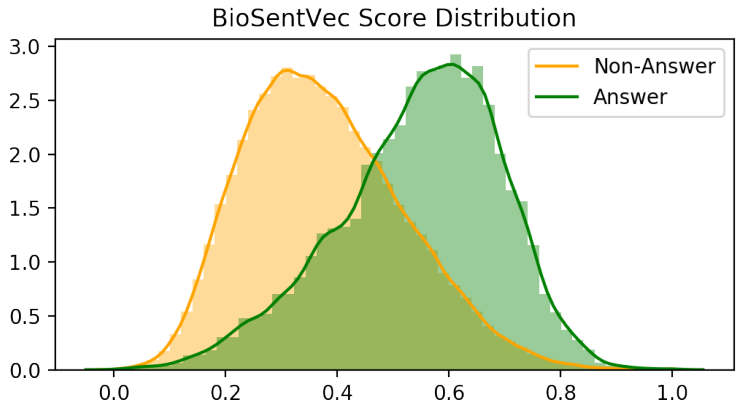


Figure 6-4: Cosine similarity density distribution for BioASQ negative and positive sample pairs

6.4 Experimental Evaluation

6.4.1 Experimental Setup

We evaluated the proposed metric learning model on the BioASQ biomedical challenge dataset; the description of the dataset, as well as the implementation details are presented below.

BioASQ Challenge Dataset

The BioASQ challenge dataset only provides positive passages, while negative examples should be individually collected by the participating teams.
For our experiments, we took the BioASQ training sets from the 2016, 2017 and 2018 editions. From them, we filtered out positive passages and selected negative passages from the relevant documents taking into account the following conditions:

1. **Removal of repeated positive passages**: As there are a significant number of repeated passages, duplicated passages were removed based on the Levenshtein Distance [147], as implemented in the FuzzyWuzzy tool[5].

[5]FuzzyWuzzy approximate string match library https://github.com/seatgeek/fuzzywuzzy

2. **Removal of outliers**: Few passages contain 1 or more than 400 words. To have a more homogeneous training dataset, we removed outliers using the Median Absolute Deviation (MAD) robust statistic [53].

3. **Selection of homogeneous negative passages**: Positive and negative passages should have similar lengths. We have identified that 95% of the positive passages have length between 13 and 55 terms, therefore, we selected the negative passages that allowed a distribution similar to that of the positive ones.

Table **6-1** presents the statistics of the BioASQ training dataset after filtering out positive and adding negatives examples using the strategy discussed in Section 6.3 [6].

#Questions	#Pairs	#Positives	#Negatives	#Hard Neg.	#Easy Neg
3295	500,248	32,944	467,304	108,130	359,174

Table 6-1: BioASQ dataset with negative samples

For testing, we used the test dataset provided in the 2018 version of the challenge. This dataset is composed of 5 batches each one with 100 questions and different number of candidate answer passages [7]

Baselines

- **Bert fine-tuned model**: We used Bert model pretrained on biomedical texts (BioBert, [50]) and it was fine-tuned using question-passage pairs. It was trained with the same training set as the proposed model.

- **Siamese model**: This is vanilla siamese model that receives a question and a passage [31]. Both text sequences were represented with BioNLP word embeddings [8].

- **Triplet network w2v-rep**: This is a conventional triplet network [108] that receives three sequences a question (the anchor), a positive passage and a negative passage. The input sequences are represented with BioNLP word embeddings.

- **Triplet network sim-rep**: This combines a conventional triplet network with the multi-similarity representation proposed in this paper. Instead of sequences, the model receives three tensors representing the similarities between three different question-answer pairs. The purpose of this method was to explore whether the gains obtained

[6] The derived training dataset is publicly-available at https://github.com/****/****

[7] The number of candidate passages per batch in the BioASQ 6b test dataset are 957, 1137, 1283, 789 and 895 respectively.

[8] BioNLP word vector representation, trained with biomedical and general-domain texts http://bio.nlplab.org

by the DMLPR could by matched by a conventional triplet network using the same representation.

Implementation Details

The proposed model was developed in TensorFlow v.2 within the Keras framework. The number of epochs was set to a maximum of 10, with a batch size of 32 samples. It was observed that a balanced sample batch has an important effect on the method's convergence, hence training samples were equally balanced between positive and negative. The number of parameters for the DMLPR model was 40,193, which is much lower than in other deep learning approaches, for example, Aueb-nlp5 has 1.5 million of parameters [20].

6.4.2 Experimental Results

Ablation Study

The following results aim to evaluate and compare the different model configurations, varying the sampling method and input representation. The reported results correspond to the Mean Average Precision (MAP) averaged over the five batches of the BioASQ 6b test dataset.
Table **6-3** presents the analysis of the contribution of the different similarity measures. It shows the results using each of the similarity representations separately and together (i.e., word2vec cosine similarity, term co-occurrence and concept co-occurrence). The Word2vec cosine similarity is the most informative single representation, nevertheless, the combination of the three representations considerably improves the isolated representation. It can be concluded that these three representations are complementary to each other.
Regarding the negative sampling strategy, we evaluated four different scenarios: **hard**, only hard negative samples are used for training; **easy**, only easy negative samples are used; **easy-hard** the model is first trained with easy negative samples and after this with hard negative samples; and **random**, were there is not distinction between easy and hard negative samples.
Table **6-2** presents the results for the four sampling strategies. As it can be observed random sampling produces higher scores than only **easy** or **hard** sampling. However, the best results were obtained in the **easy-hard** scenario, were the model is warmed-up with the easy negative samples, which prepares it better to take advantage of the hard negative samples.
To further understand the contribution of the negative sampling strategy, we visualized the space of characteristics that is generated in the dense layer of 128 units of the proposed architecture. Figure **6-5** shows a two-dimension projection of the the positive, easy negative, and hard negative samples generated by tSNE. As it can be observed, a geometrical distribution based on semantic relatedness is kept in the feature space; hard negative samples are closer to positive passages than easy negative samples.

Sampling	MAP
easy-hard	**0.294**
random	0.238
hard	0.227
easy	0.098

Table 6-2: MAP score averaged over 5 batches with different sampling strategies.

Modality	MAP
all	**0.294**
w2vcos	0.146
terms	0.138
concepts	0.129

Table 6-3: MAP score averaged over 5 batches using different representation modalities.

BioASQ Challenge Results

The results of the passage retrieval task largely depends on the performance obtained in the document retrieval stage. To have a fair comparison of the different passage retrieval approaches, we used in all experiments the same set of documents, which were retrieved by AUEB-NLP, the winning document-retrieval strategy of BioASQ 6 [20]. Thanks to the fact that the winning team of version 6 AUEB-NLP [20], shared the documents retrieved by them, we can make a fair comparison using the same set of documents. BioASQ ranks winner teams using Mean Average Precision metric, We report results averaging official metrics over the 5 batches, the reported metrics are: Mean Average Precision (MAP), Mean Precision, Recall, F-Measure, and G-MAP.

Table **6-4** presents the obtained results. The proposed method outperformed all baselines methods according to the averaged MAP score. With respect to the winning method of the BioASQ version 6 (**AUEB-NLP**), an average increase of 25% in MAP was observed, while a 10% improvement was achieved with regard to the **Triplet loss metric sim-rep**. It is also notable that the representation using multiple similarities as input is considerably better than using the sequences without interaction between them, since it exceeds the Siamese model and **Triplet loss metric w2v-rep** by about 65%. The Bert model has moderate performance scores, and the margin with respect to the proposed model is wide.

We also compared the results of the DLMPR method against the top 15 models in the BioASQ 2018 challenge. Their results were taken from the BioASQ 6b leader board[9] and averaged over the five batches. Figure **6-6** shows a boxplot with these results. The x-axis corresponds to reported metrics in BioASQ 6 (mean precision, recall, f-score, MAP, GMAP), the bluepoint indicates the average results of DMLPR in the five batches. It is noticed that DMLPR improved the recall, f-score, MAP, and GMAP of all participating teams by a wide margin. The Mean Precision score is in the higher quartile close to the best result.

[9]BioASQ portal https://www.bioasq.org

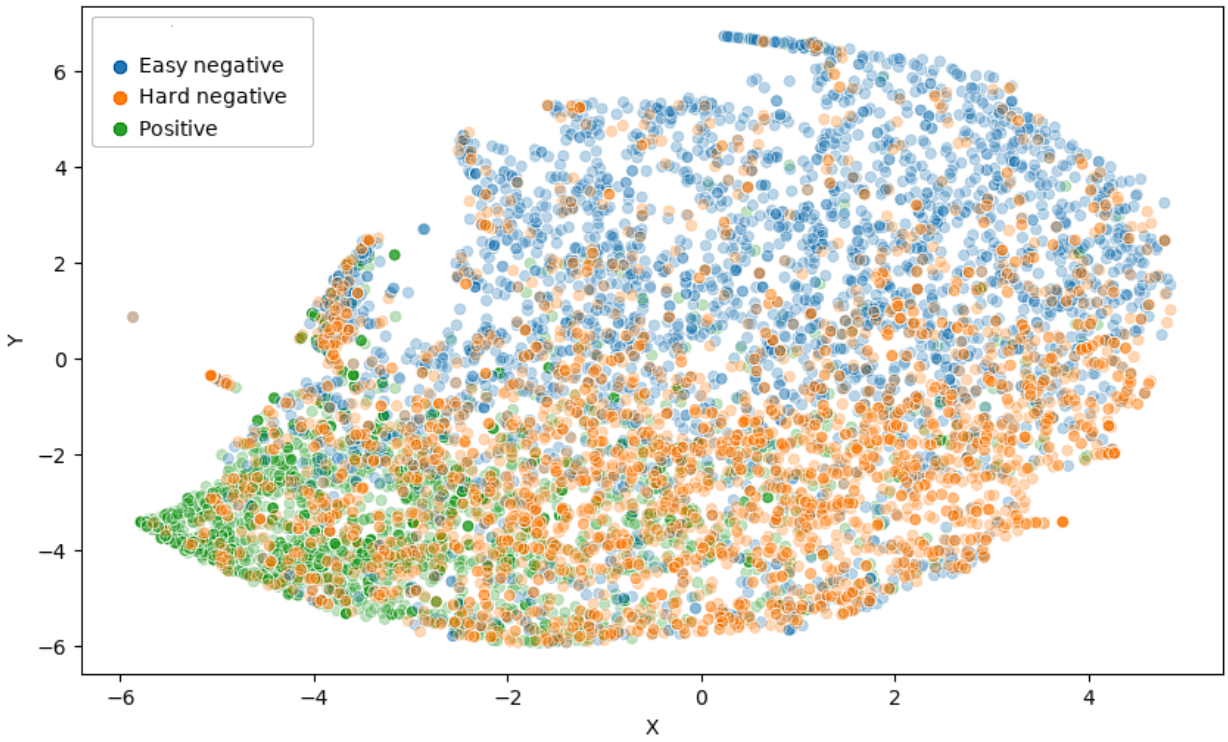


Figure 6-5: Visualization for the generated metric space using 2D tSNE dimensional reduction, points are BioASQ positive, hard-negative and easy-negative test-partition samples.

6.4.3 Results Discussion

The results obtained show that the proposed method has a significant improvement over the state-of-the-art methods as well as over the baselines. The good performance of the DMLPR model depends on different factors.

The representation based on the three similarity matrices is, by a wide margin, more effective to capture the semantic relatedness of the question and answer sequences than taking independent representations. Most of the current state-of-the-art works exclusively used learned representation for text. The results of the ablation study show that using domain knowledge to identify important concepts in the text and using them to calculate a complementary similarity enriched the question-passage representation.

Another factor, and a distinctive characteristic of this work, is the combination of a metric learning approach with a CNN applied over text-similarity matrices. The results show that it successfully captures the question-passage interactions. Finally, the negative sampling strategy that identify easy and hard negative samples was very important for successfully train the model. This is not a common strategy in passage retrieval methods, and the present work shows that it could have a very positive impact.

Method	Mean precision	Recall	F-Measure	MAP	GMAP
Bert	0.172	0.191	0.186	0.144	0.010
Siamese	0.119	0.156	0.131	0.129	0.002
Triplet loss sim-rep	0.226	0.262	0.241	0.266	0.021
Triplet loss w2v-rep	0.107	0.169	0.122	0.131	0.001
AUEB-NLP	0.215	0.229	0.180	0.231	0.015
USTB	0.188	0.292	0.178	0.138	0.011
DMLPR	**0.243**	**0.358**	**0.231**	**0.294**	**0.030**

Table 6-4: Passage retrieval results for the proposed baselines and the best models in BioASQ challenge 6b task [75]

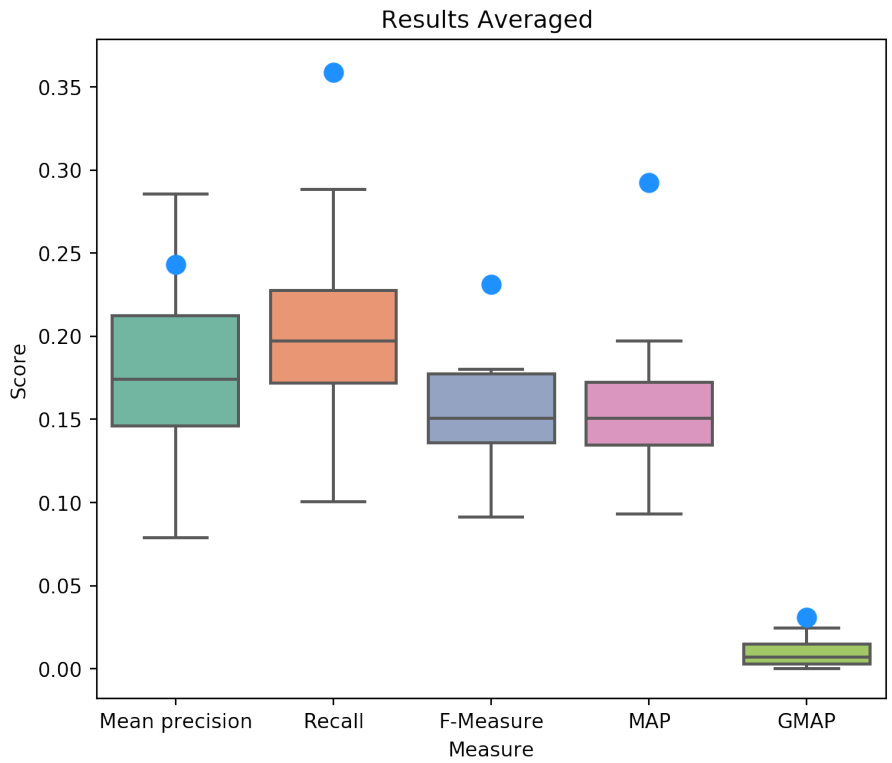


Figure 6-6: 15 best systems results for BioASQ task 6b, blue points correspond to the DMLPR model.

6.5 Conclusion

We present a novel deep-metric learning approach for biomedical passages retrieval that surpasses previous approaches evaluated in the BioASQ dataset. The model presents innovations in terms of the architecture that combines multi-similarity representation, a CNN,

and a siamese design, as well as in terms of the training strategy that identify hard and easy negative samples which are used to gradually train the model.

7 BERT Attention-based representation for Biomedical Passage Retrieval

In the previous chapters, we have introduced representations for question-answer sequences based on different similarity representations, such as the cosine or the co-occurrence similarity matrix. Those representations have been shown to be effective in capturing semantic correlations between candidate passages and the question. In an earlier chapter, we exploited these representations using a new deep metric learning architecture, which was shown to be more effective at analyzing the input representation than previously explored models. In this chapter, we present a novel input representation based on transformers-based attention similarity. Large-scale pre-trained neural networks such as BERT are successful in NLP tasks, most of the proposed implementations of such models focus on using the outputs of the network as input features for a classification layer [82, 140].
In this chapter, we propose a different approach, since, as noted in previous work, the head attention layers encode important semantic and syntactic relations [25], therefore we have used the attention weights as the input representation of the question-passage, in addition to previously explored representations. Bert's attention-based representation offers complementary views of co-occurrence and cosine similarity, providing enriched features that can be exploited by the deep metric learning model. We structure up this chapter as follows: the first section presents a quick overview of Bert's model, then a description of Bert's attention mechanism is provided to further explore the properties that Bert's attention-based representation can offer to address the task of passage retrieval. This is followed by a description of the architecture of the proposed model for experimentation and results. Some conclusions are discussed at the end of the chapter.

7.1 Introduction

Neural pre-trained language models such as ELMo [90], OpenAI GPT [94], and BERT [29] have achieved stunning results in NLP tasks ranging from natural language inference to question answering. One of those popular model, BERT, has recently been applied to document retrieval and question answering tasks in several published works, most of them work over open domain data [81, 82, 64, 140]. Furthermore, for the biomedical domain, Bert's alternatives have been proposed, which have been trained on large volumes of data composed of scientific papers and clinical notes, some of which are BioBert [51] and Clinical BERT [3].

For text classification tasks with BERT, the common approach is to use the hidden state h of the final layer over the special token [CLS], which would be the full representation of the input sequences. On top of this, a classifier (softmax) is placed to predict the probability given the hidden state representation. This is known as fine tuning, in which it is possible to leave the weights of the BERT model fixed or allow certain layers to be adjusted along with the classification ones. Passage-retrieval can me modeled as a classification task that can be approached in this way, some of the most relevant papers that have implemented the related approach are [105, 40, 86, 121].
However, we propose a different approach, where BERT's pre-trained attention maps are used as a semantic representation of the question-passage pair which is then used by a deep metric learning-based model in conjunction with other representations (cosine and co-occurrences) to discriminate between answer and non-answer passages. The use of this representation is therefore appropriate since attention maps can be seen as the weight that a term has when calculating the representation of any other term, that is, as the more semantically correlated this weight is greater.
We will experimentally test later the use of this attention-based representation for the problem of passage retrieval. Each attention layer will be evaluated separately, as well as the combination of them. The representation will be combined with other representations that have been presented in previous chapters.
The results obtained demonstrate that Bert's attention layers are a rich representation mode that can be exploited for passage retrieval tasks. According to the results, this representation combined with others improves the performance of the models already discussed.

7.2 Background: Bert Attention Mechanism

In his seminal work "Attention is all you need" Vaswani et al. [128] successfully reproduce the attention mechanism without employing a recurrent network which incorporates context. In order to do so, they propose the "transformer" model that is totally based on self-attention mechanisms. The attention in deep learning can be interpreted in general terms as a weighting vector of significance for predicting or inferring an element. The attention vector is employed to estimate the strength of the current element's correlation or " attendance " to other elements.
BERT comprise multiple layers where each one contains multiple attentions heads. While our methods are applicable for any model which employs an attention mechanism, here we use BERT [29]. According to Devlin et al. description of Bert's attention mechanism, each of the heads in each layer will calculate the weight of "attention" between two word pairs, using standardized softmax point products. The output of the attention head is a weighted sum of the vector values. On the last of these layers the output is connected to the '[CLS]' token that is used to perform the classification task typically. Therefore, as you will see below, the model goes layer by layer focusing its attention on the answer terms that are

most correlated with the key question terms.
Acording to Clark et al. [25] the first layers capture relatively simple linguistic features-such as the syntactic function of a word in a sentence-and the upper levels help classify more complex features, such as the way words combine to make up meaning. BERT's attention layers encode the semantic relationships that exist in the text sequences, with this in mind we propose the hypothesis that BERT's attention mechanisms can effectively represent the interactions between the question and the passages terms, these interactions are a valuable source of information that added to the previous representation ways are useful for passage retrieval task.

7.3 Bert Attention as Similarity Representation

In the proposed model the question q is used as the first sequence in Bert and the passage p is the second sequence separated by the special token '[SEP]'. Since Bert encodes terms as tokens, where a word can be composed of several tokens, the maximum token sequence size of the question and the passage is specified as 512. Once codified the input sequences in Bert we extract the attention layer values that will feed the deep metric learning model. As Clark et Al. explains [25], it is highly likely that attention heads in the same layer tend to have similar behavior. This would enable us to reduce the 12 attention heads per layer to 1 per layer, a fair statistic to achieve this is the average. In order to see the representation properties offered by BERT's attention layers, the existing attention weights between the question and the passage will be visualized. Since the terms were split in tokens, we also need to recombine the tokens to have term attention values. Therefore, we collapsed the tokens that make up a term as suggested by [25]: for the attention of a split word, we summed up the attention weights on its tokens. For the attention of a split word, we take the average of the attention weights on its tokens. The model that we have employ to extract the attention values is BioBert, which is a biomedical pretrained bert trained with scientific papers and clinical notes [50]). This model was fine-tuned for passage retrieval task as is sugested by Devlin et al. [29]
Suppose you have the following question-passage pair:

- **Question**: What causes Katayama Fever?
- **Answer**: Schistosomiasis is a helminthic infection that is endemic in tropical and subtropical regions.

In the visualization **7-1**, **7-2**, **7-3**, the attention maps layer by layer is presented. The darkness of a line indicates the strength of the attention weight.

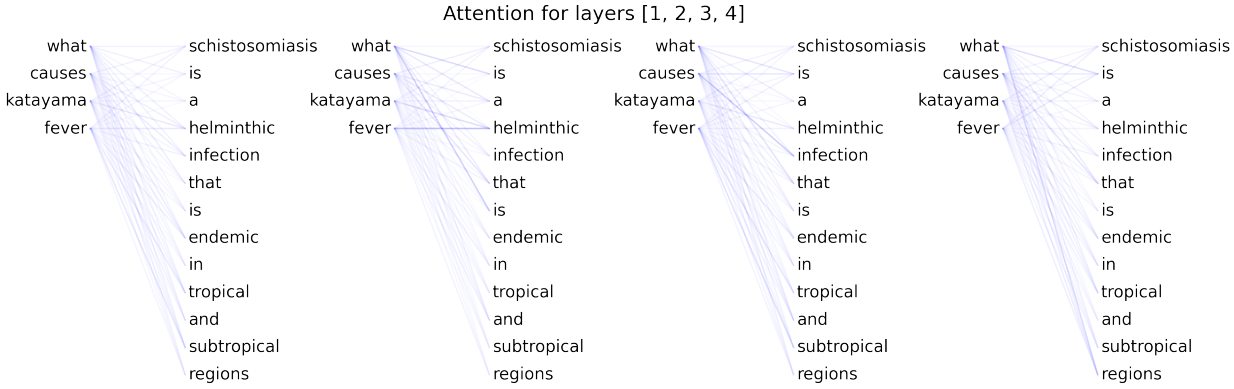


Figure 7-1: Attention between question and passage terms for 1, 2, 3, 4 layers

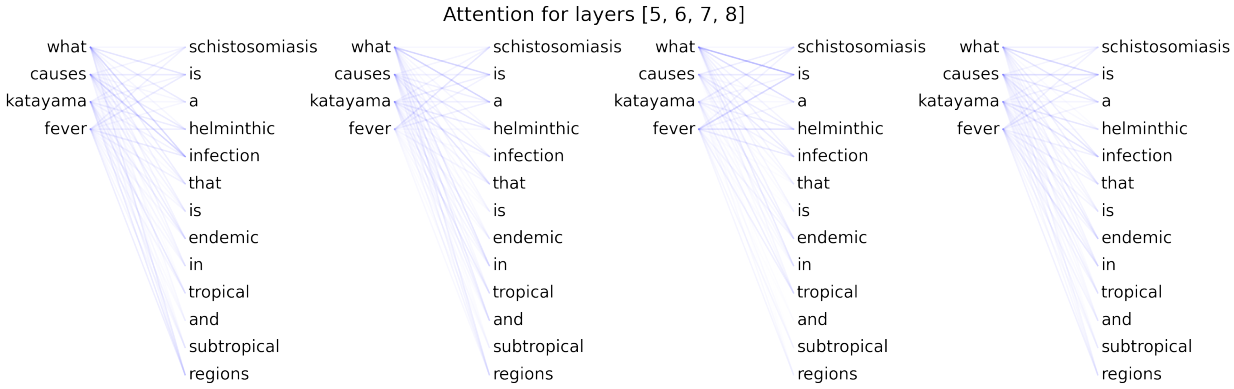


Figure 7-2: Attention between question and passage terms for 5, 6, 7, 8 layers

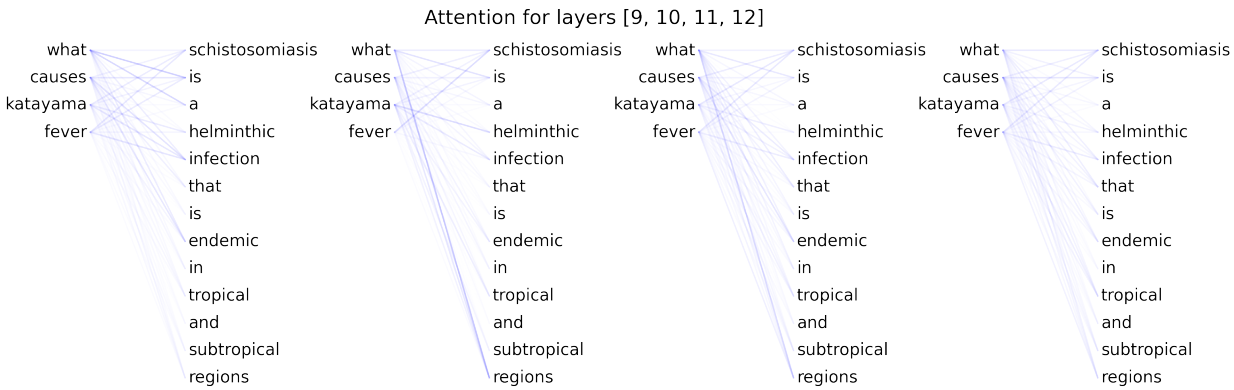


Figure 7-3: Attention between question and passage terms for 9, 10, 11, 12 layers

As it can be observed in the first layers (1,2,3,4) the attention between the terms is more uniform compared to the last layer (9,10,11,12), the last ones have strong relations of attention

between the following terms: fever-Schistosomiasis, Katayama-helminthiasis and Katayama-Schistosomiasis among others. This behavior has also been reported by Clark et al. [25], who found that the lower layers of attention have a very broad attention span. Whereas the latter layers are a kind of aggregation of attention that focuses more on those relevant terms related to the end results.
To simplify the analysis of which term in the attention mechanism is focusing layer by layer, the figure **7-4** presents the average of the attention received by each term in each Bert layer. It is interesting to note that the term that has the highest semantic relationship with the question "Schistosomiasis" receives the most attention in the last 3 layers, this term is a key term to answer the question.

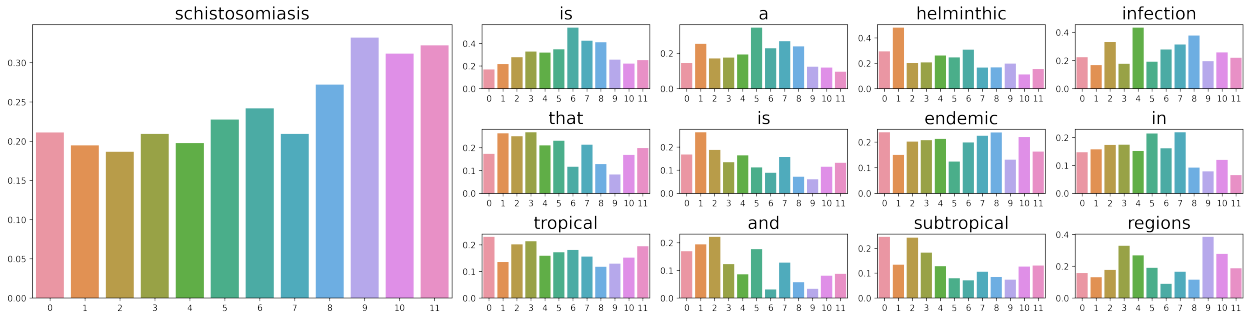


Figure 7-4: Average received attention in each BERT for all answer terms

7.4 Methods

Our model architecture is presented in Figure **7-5**. The diagram presents the training and testing phases.

- **Training**: In the training phase the first step is to obtain the input representation for the question-answer pairs. As the proposed deep metric learning model expects a positive (q, p_+) and a negative (q, p_-) pair for the same question, a tensor calculation using three different term-level similarity measures will be carried out S. For BERT attention-based representation, we calculated the head-averaged attention for the 12 BioBert layers. For the other representations: cosine similarity, term co-occurrence and concept co-occurrence are obtained in the same way as was presented in the previous chapter, finally the 15 similarity matrices are aligned and appended in a similarity tensor $S \in \mathbb{R}^{15\times40\times40}$, where 40 is the maximum sequence length for question and passages. Once the representation S is calculated, it feeds the deep metric learning model presented also in the previous chapter. The siamese CNN model is used to calculate a measure of semantic distance between question and the passage, these are noted as $dis_{(q,p_+)}$ and $dis_{(q,p_-)}$ for the positive and negative pairs respectively. The model is trained by minimizing the loss function from Equation 7-1, the distances for

positive pairs are encouraged to be close to 0, while negatives pairs should have a distance greater than a margin α, N is the batch size.

- **Testing**: Once the model has been trained, it is used to produce predictions for the incoming question-passage pairs. First the pair is represented as expected in a tensor shape and then is evaluated by the model. The model output is a semantic distance between question and passage, if the distance is greater than the margin α then the passage is not considered as valid answer.

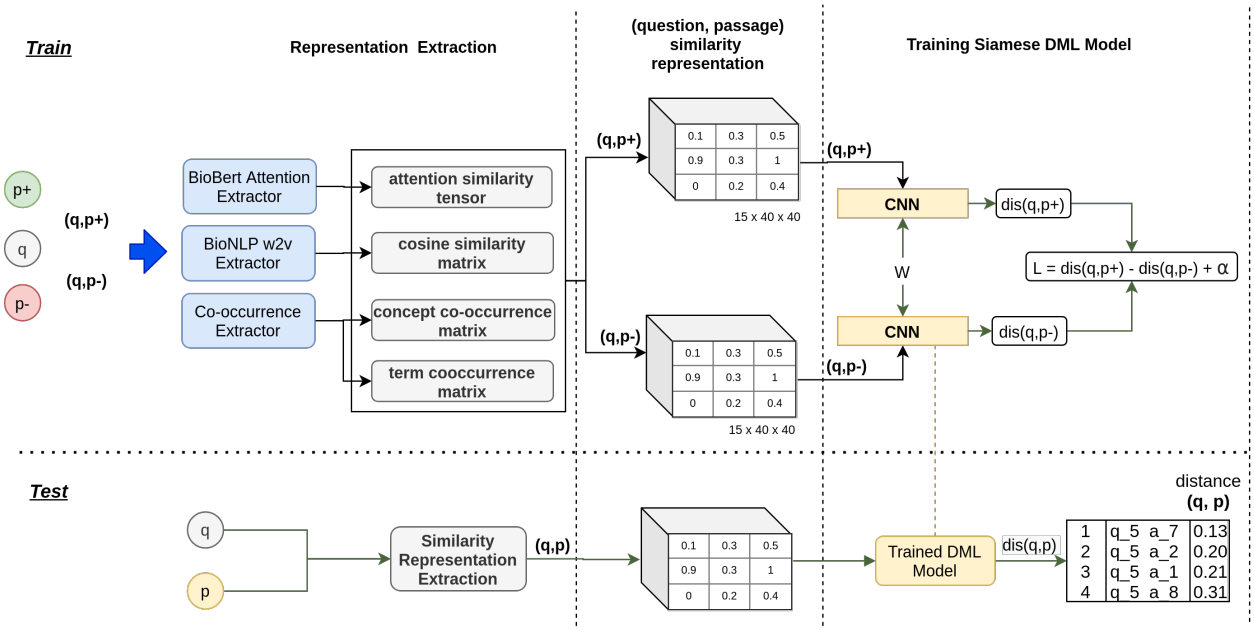


Figure 7-5: Passage retrieval model architecture

$$\frac{1}{N}\sum_{i}^{N}[dis(q,p_{+}) - dis(q,p_{-}) + \alpha] \tag{7-1}$$

The details for cosine similarity, term co-occurrence and concept term co-occurrence question-passage representation calculation in addition to the convolutional model and the sampling strategy was presented in the Chapter 6. For this reason we are not going to extend the related discussion.

7.5 Experimental Evaluation

7.5.1 Data-set and Training

Model training was conducted with the question-and-answer pairs from the 2016, 2017, and 2018 BioASQ Task B training datasets. In this dataset the samples were labeled as

hard or easy negatives based on the semantic relatedness between the question and passage using cosine similarity over the BiosentVec sentence embedding. As explained in the earlier chapter, we first fed the model with easy negative samples, and then with hard negative samples that are more difficult to classify. With this informative sampling the model is warmed-up with the easy negative samples, which better prepares it to take advantage of the hard negative samples.

The resulting statistics of the dataset are presented in Table **7-1**.

#Questions	#Pairs	#Positives	#Negatives	#Hard Neg.	#Easy Neg
3295	500,248	32,944	467,304	108,130	359,174

Table 7-1: BioASQ dataset with negative samples

The proposed model was developed in TensorFlow v.2 within the Keras framework. The number of epochs was set to a maximum of 10, with a batch size of 32 samples. For testing purpose we have used the test batches released for BioASQ 6b as was done in the previous chapters, as testing dataset is composed of five batches score metrics are averaged.

Ablation Study

To compare the discriminative power in the proposed model as well as in each of the attention layers, the following configurations of the model will be evaluated, 1) using only the head-averaged attention for individual layers 2) averaging across the heads and layers 3) using the 12 head-averaged attention layers.

Figure **7-6** presents the related scenarios, each bar shows the MAP score over the five batches using only a specific layer, averaging them and using all layers. The results support our previous observations related with the effectiveness of higher layers to capture semantic properties. The score obtained for the last layer compared with the first one, is 37% higher, a similar improvement is observed in layers 8, 9, 10 and 11. In the same way averaging the head attention is less effective that use all the 12 attention layers as input representation.

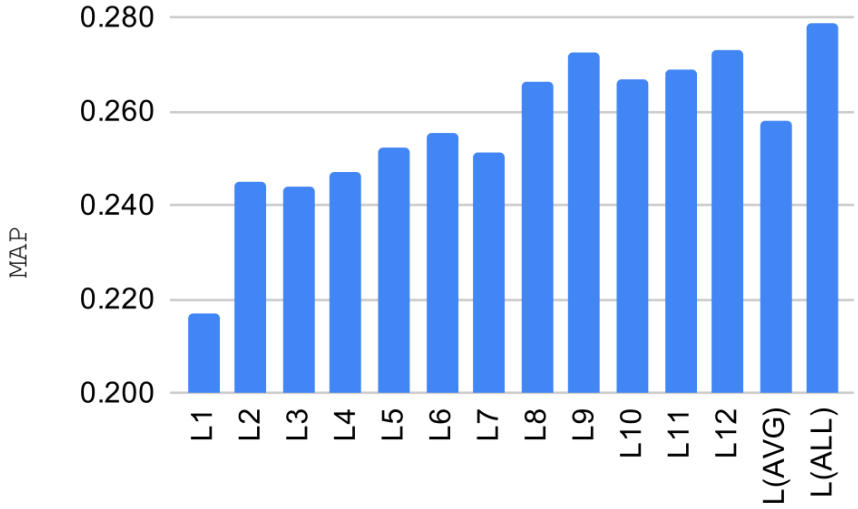


Figure 7-6: MAP averaged scores over 5 batches

Model results against baselines

In order to compare and contrast our method we have conducted the experimentation using as baselines previous chapters models. The set of documents used to retrieve the passages is the same for all the methods in the evaluation. A description of the baseline models is presented as follows.

- **Bert fine-tuned model**: We used Bert model pretrained on biomedical texts (BioBERT, [50]) and it was fine-tuned using question-passage pairs. It was trained with the same training set as the proposed model.
- **DFMTS**: This approach was the one presented in Chapter 5, that presents a deep fusion strategy for multiple term similarity measure.
- **DMLPR**: This is the deep metric learning presented in Chapter 6, the model has three similarity representation matrices (cosine, term co-occurence, concept co-ocurrence).

7.5.2 Results

In addition to the previous presented baselines we have included the results from the the winning team of BioASQ in the same challenge edition [20], this method use also the same set of relevant documents.
As can be observed in Table **7-2**, the proposed method **DMLPR(BERT+W+TC+CC)** outperformed all baselines methods according to the averaged MAP score. Two versions are presented DMLPR(BERT+W+TC+CC) and DMLPR(BERT), using only attention-based

representation and combining it with the previously obtained representations (cosine similarity, term co-occurrence and concept co-occurrence) respectively. The last one is remarkably more effective than the first which use only attention-base representation, with an improvement of 20% in MAP metric. With respect to the winning method of the BioASQ version 6 (**AUEB-NLP**), an average increase of 52% in MAP was observed.

Method	Mean prec.	Recall	F-Measure	MAP	GMAP
AUEB-NLP	0.215	0.229	0.181	0.231	0.015
DFMTS	0.237	0.259	0.219	0.276	0.024
DMLPR (W+TC+CC)	0.243	0.358	0.231	0.294	0.033
DMLPR (BERT)	0.209	0.342	0.213	0.279	0.021
DMLPR (BERT+W+TC+CC)	**0.279**	**0.370**	**0.251**	**0.355**	**0.034**

Table 7-2: Passage retrieval results for the proposed model DMLPR(Bert) and baselines in BioASQ challenge 6b task [75]

7.6 Conclusion

This chapter presented a new model that uses BERT's attention layers as a representation of the semantic interactions between a question and a passage. This representation proved to be effective in capturing relevant patterns that enable a proper discrimination for a passage that conforms a valid answer to the question. This new representation was combined with the other representations introduced in previous chapters and is demonstrated to offer an alternative and complementary perspective to the other involved representations. The method proposed was compared against the previously described methods, as well as against the state of the art models. By far, this model surpasses the state of the art models and by about 20% to the previously proposed models.

8 Conclusions

We have proposed several representations for questions and passages based on their interactions that have proved to be highly effective in passage retrieval task. The first was based on the semantic similarity between the terms of the question and the answer given their vector representation. Later on, the representation based on co-occurrences was proposed, which is complementary to the previous one. This is easier to see when there are terms out of the vocabulary which have no representation in the embedding space and also when the terms are not well represented due to the training parameters, eg. the size of the window used in word2vec. The subsequent representation was based on the Bert's attention mechanism, which allows to know the significance of a term in the context of another term, establishing a semantic relationship by means of attention values. The last two representations are novel for the task of passage retrieval; there is no similar work that uses Bert's attention layers or co-occurrences as a method of representation.

The present thesis has proposed several information fusion methods, such as early, intermediate and late fusion. Depending on the approach, the fusion models took textual and semantic information that allowed discriminating the passages that were a valid answer to the question and thus considerably improved the final result of the task.

The use of mixed information sources (textual and semantic) fused into a single representation, provides the following benefits: 1) the representation produced is unambiguous thanks to the semantic information sources that have a single representation and meaning, 2) the representation based on question-passage interaction comprises multiple similarity dimensions that provides more supporting arguments to identify the correct answer to a particular question. The captured semantic interactions are complementary, as demonstrated in the ablation studies conducted.

One approach that was effective by far in modeling the passage retrieval task was the metric learning approach, in which a metric is induced to model the semantic interactions of question-passage pairs. The described method processes a triple input composed of (question, negative and positive passage), but, unlike triple networks, the architecture is based on a siamese neural model. This approach allows to process simultaneously the pairs of questions and negative and positive answers obtained from the same question. The objective is to have a metric space that closely locates the question-passage pairs that are a valid answer and keep away the ones that are not. Another factor that influences model performance is the sampling strategy, which selects easy and hard negative samples based on semantic similarity to the question.

8.1 Future Research

This research answered several research questions but open new ones that will guide our future research efforts.

- **What additional mechanisms different to the ones explored in this thesis can be used to represent the interactions between question and passages?**

  The semantic correlation between question and passage terms does not always occur directly or is not efficiently captured by the defined representation. This drawback can be alleviated by including new representation strategies that enable a complementary view of the question and the passage interactions.

  Complementarity in the representation has driven the incorporation of different similarity measures such as word and concept co-occurrences or BERT's attention maps similarity. However, there are others that can be useful such as: diffusion kernels, robust distributional word similarities or knowledge graphs.

- **Can other deep neural model approach or architecture discriminate better if a passage is a valid answer or not?**

  We have mainly explored convolutional neural networks to take advantage of similarity-based representation. But this decision is motivated by the mechanism of representation in which question-passage interactions are summarized in a term-by-term similarity tensor. However, other models or architectures can take advantage of such representation or any other.

  Instead of representing the question-passage by its term-level interactions, we can use the complete text sequence and employ a BiLSTM model to capture information at context level. Another option is to explore Deep Adversarial Metric Learning, where the negative samples are generated automatically and following the negative statistical data distribution.

- **Can additional steps appended to passage retrieval pipeline improve the overall performance?**

  As it was possible to verify, the approaches proposed are effective in finding the most relevant passages, likewise the final ranking is completely based on the similarity score provided by the output layer of the deep model. However, even if there is a high semantic similarity between the question-passage pair, this passage may not be a valid answer to the question.

  There are two additional tasks in the passage retrieval pipeline that may alleviate this challenge. The first one is to filter the candidate passages based on the expected entities or medical concepts expected as answers. Second, to perform a reordering of the passages based on a model trained over a subset of positive and negative passages,

but the latter being difficult to discriminate as such, that would allow to have a fine discrimination model that being combined with the base model would improve the total performance of the system.

Bibliography

[1] Agichtein, Eugene ; Brill, Eric ; Dumais, Susan: Improving web search ranking by incorporating user behavior information. In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* ACM, 2006, S. 19–26

[2] Al-Chalabi, Hani ; Ray, Santosh ; Shaalan, Khaled: Semantic Based Query Expansion for Arabic Question Answering Systems. In: *Arabic Computational Linguistics (ACLing), 2015 First International Conference on* IEEE, 2015, S. 127–132

[3] Alsentzer, Emily ; Murphy, John ; Boag, William ; Weng, Wei-Hung ; Jindi, Di ; Naumann, Tristan ; McDermott, Matthew: Publicly Available Clinical BERT Embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019, S. 72–78

[4] Androutsopoulos, Ion: A challenge on large-scale biomedical semantic indexing and question answering. In: *BioNLP workshop (part of the ACL Conference)* (2013), S. 92–98

[5] Auer, Sören ; Bizer, Christian ; Kobilarov, Georgi ; Lehmann, Jens ; Cyganiak, Richard ; Ives, Zachary: Dbpedia: A nucleus for a web of open data. In: *The semantic web*. Springer, 2007, S. 722–735

[6] Azzopardi, Leif ; Girolami, Mark ; Crowe, Malcolm: Probabilistic hyperspace analogue to language. In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* ACM, 2005, S. 575–576

[7] Baasiri, Rudeina A. ; Glasser, Stanley R. ; Steffen, David L. ; Wheeler, David A.: The breast cancer gene database: a collaborative information resource. In: *Oncogene* 18 (1999), Nr. 56, S. 7958–7965

[8] Banerjee, Somnath ; Bandyopadhyay, Sivaji: Ensemble approach for fine-grained question classification in bengali. In: *27th Pacific Asia Conference on Language, Information, and Computation*, 2013, S. 75–84

[9] Beam, Andrew L. ; Kompa, Benjamin ; Fried, Inbar ; Palmer, Nathan P. ; Shi, Xu ; Cai, Tianxi ; Kohane, Isaac S.: Clinical Concept Embeddings Learned from Massive Sources of Medical Data. In: *arXiv preprint arXiv:1804.01486* (2018)

[10] Berant, Jonathan ; Chou, Andrew ; Frostig, Roy ; Liang, Percy: Semantic parsing on freebase from question-answer pairs. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, S. 1533–1544

[11] Berners-Lee, Tim ; Hendler, James ; Lassila, Ora: The semantic web. In: *Scientific american* 284 (2001), Nr. 5, S. 34–43

[12] Bhogal, Jagdev ; MacFarlane, Andrew ; Smith, Peter: A review of ontology based query expansion. In: *Information processing & management* 43 (2007), Nr. 4, S. 866–886

[13] Bilotti, Matthew W. ; Elsas, Jonathan ; Carbonell, Jaime ; Nyberg, Eric: Rank learning for factoid question answering with linguistic and semantic constraints. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, S. 459–468

[14] Bird, Steven: NLTK: the natural language toolkit. In: *Proceedings of the COLING/ACL on Interactive presentation sessions* Bd. 1 ACL, 2006, S. 69–72

[15] Blei, David M. ; Ng, Andrew Y. ; Jordan, Michael I.: Latent dirichlet allocation. In: *Journal of machine Learning research* 3 (2003), Nr. Jan, S. 993–1022

[16] Bodenreider, Olivier: The unified medical language system (UMLS): integrating biomedical terminology. In: *Nucleic acids research* 32 (2004), Nr. suppl_1, S. D267–D270

[17] Bollacker, Kurt ; Evans, Colin ; Paritosh, Praveen ; Sturge, Tim ; Taylor, Jamie: Freebase: a collaboratively created graph database for structuring human knowledge. In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* AcM, 2008, S. 1247–1250

[18] Bonadiman, Daniele ; Kumar, Anjishnu ; Mittal, Arpit: Large Scale Question Paraphrase Retrieval with Smoothed Deep Metric Learning. In: *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, 2019, S. 68–75

[19] Brill, Eric ; Dumais, Susan ; Banko, Michele: An analysis of the AskMSR question-answering system. In: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* Association for Computational Linguistics, 2002, S. 257–264

[20] Brokos, Georgios-Ioannis ; Liosis, Polyvios ; McDonald, Ryan ; Pappas, Dimitris ; Androutsopoulos, Ion: AUEB at BioASQ 6: Document and Snippet Retrieval. In: *arXiv preprint arXiv:1809.06366* (2018)

[21] Bucher, Maxime ; Herbin, Stéphane ; Jurie, Frédéric: Hard negative mining for metric learning based zero-shot classification. In: *European Conference on Computer Vision* Springer, 2016, S. 524–531

[22] Buscaldi, Davide ; Flores, Jorge G. ; Meza, Ivan V. ; Rodriguez, Isaac: Sopa: Random forests regression for the semantic textual similarity task. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, S. 132–137

[23] Chen, Qingyu ; Peng, Yifan ; Lu, Zhiyong: BioSentVec: creating sentence embeddings for biomedical texts. In: *2019 IEEE International Conference on Healthcare Informatics (ICHI)* IEEE, 2019, S. 1–5

[24] Chen, Yueguo ; Gao, Lexi ; Shi, Shuming ; Du, Xiaoyong ; Wen, Ji-Rong: Improving context and category matching for entity search. In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014, S. 16–22

[25] Clark, Kevin ; Khandelwal, Urvashi ; Levy, Omer ; Manning, Christopher D.: What does bert look at? an analysis of bert's attention. In: *arXiv preprint arXiv:1906.04341* (2019)

[26] Clarke, Charles L. ; Cormack, Gordon V. ; Lynam, Thomas R. ; Terra, Egidio L.: Question answering by passage selection. In: *Advances in Open Domain Question Answering*. Springer, 2008, S. 259–283

[27] Cohen, Daniel ; Croft, W B.: A Hybrid Embedding Approach to Noisy Answer Passage Retrieval. In: *European Conference on Information Retrieval* Springer, 2018, S. 127–140

[28] Deerwester, Scott ; Dumais, Susan T. ; Furnas, George W. ; Landauer, Thomas K. ; Harshman, Richard: Indexing by latent semantic analysis. In: *Journal of the American society for information science* 41 (1990), Nr. 6, S. 391

[29] Devlin, Jacob ; Chang, Ming-Wei ; Lee, Kenton ; Toutanova, Kristina: Bert: Pre-training of deep bidirectional transformers for language understanding. In: *arXiv preprint arXiv:1810.04805* (2018)

[30] Dong, Li ; Wei, Furu ; Zhou, Ming ; Xu, Ke: Question Answering over Freebase with Multi-Column Convolutional Neural Networks. In: *ACL* Bd. 1, 2015, S. 260–269

[31] Feng, Minwei ; Xiang, Bing ; Glass, Michael R. ; Wang, Lidan ; Zhou, Bowen: Applying deep learning to answer selection: A study and an open task. In: *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* IEEE, 2015, S. 813–820

[32] Galkó, Ferenc ; Eickhoff, Carsten: Biomedical question answering via weighted neural network passage retrieval. In: *European Conference on Information Retrieval* Springer, 2018, S. 523–528

[33] Ganguly, Debasis ; Roy, Dwaipayan ; Mitra, Mandar ; Jones, Gareth J.: Word embedding based generalized language model for information retrieval. In: *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* ACM, 2015, S. 795–798

[34] Goodfellow, Ian ; Bengio, Yoshua ; Courville, Aaron ; Bengio, Yoshua: *Deep learning*. 1. MIT press Cambridge, 2016

[35] Gormley, Clinton ; Tong, Zachary: *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*. 1. " O'Reilly Media, Inc.", 2015

[36] Guo, Jiafeng ; Fan, Yixing ; Ai, Qingyao ; Croft, W B.: A deep relevance matching model for ad-hoc retrieval. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* ACM, 2016, S. 55–64

[37] Guo, Shangmin ; Zeng, Xiangrong ; He, Shizhu ; Liu, Kang ; Zhao, Jun: Which is the effective way for gaokao: Information retrieval or neural networks? In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, S. 111–120

[38] Hardy, H ; Cheah, Yu-N: Question classification using extreme learning machine on semantic features. In: *Journal of ICT Research and Applications* 7 (2013), Nr. 1, S. 36–58

[39] He, Hua ; Lin, Jimmy J.: Pairwise Word Interaction Modeling with Deep Neural Networks for Semantic Similarity Measurement. In: *HLT-NAACL* Bd. 1, 2016, S. 937–948

[40] He, Yun ; Zhu, Ziwei ; Zhang, Yin ; Chen, Qin ; Caverlee, James: Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, S. 4604–4614

[41] Hermjakob, Ulf: Parsing and question classification for question answering. In: *Proceedings of the workshop on Open-domain question answering-Volume 12* Association for Computational Linguistics, 2001, S. 1–6

[42] Hiemstra, Djoerd: *Using language models for information retrieval*. 1. Taaluitgeverij Neslia Paniculata, 2001

[43] Hoffart, Johannes ; Suchanek, Fabian M. ; Berberich, Klaus ; Weikum, Gerhard: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. In: *Artificial Intelligence* 194 (2013), S. 28–61

[44] Kaya, Mahmut ; Bilge, Hasan S.: Deep metric learning: a survey. In: *Symmetry* 11 (2019), Nr. 9, S. 1066

[45] Khalifa, Khalid ; Omar, Nazlia: A hybrid method using lexicon-based approach and Naive Bayes classifier for Arabic opinion question answering. In: *J. Comput. Sci.* 10 (2014), Nr. 10, S. 1961–1968

[46] Kolomiyets, Oleksandr ; Moens, Marie-Francine: A survey on question answering technology from an information retrieval perspective. In: *Information Sciences* 181 (2011), Nr. 24, S. 5412–5434

[47] Krishnamurthy, Jayant ; Kollar, Thomas: Jointly learning to parse and perceive: Connecting natural language to the physical world. In: *Transactions of the Association for Computational Linguistics* 1 (2013), S. 193–206

[48] Kuzi, Saar ; Shtok, Anna ; Kurland, Oren: Query expansion using word embeddings. In: *Proceedings of the 25th ACM international on conference on information and knowledge management* ACM, 2016, S. 1929–1932

[49] Le, Quoc V. ; Mikolov, Tomas: Distributed Representations of Sentences and Documents. In: *ICML* Bd. 14, 2014, S. 1188–1196

[50] Lee, Jinhyuk ; Yoon, Wonjin ; Kim, Sungdong ; Kim, Donghyeon ; Kim, Sunkyu ; So, Chan H. ; Kang, Jaewoo: BioBERT: pre-trained biomedical language representation model for biomedical text mining. In: *arXiv preprint arXiv:1901.08746* (2019)

[51] Lee, Jinhyuk ; Yoon, Wonjin ; Kim, Sungdong ; Kim, Donghyeon ; Kim, Sunkyu ; So, Chan H. ; Kang, Jaewoo: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. In: *Bioinformatics* 36 (2020), Nr. 4, S. 1234–1240

[52] Levenshtein, Vladimir I.: Binary codes capable of correcting deletions, insertions, and reversals. In: *Soviet physics doklady* Bd. 10, 1966, S. 707–710

[53] Leys, Christophe ; Ley, Christophe ; Klein, Olivier ; Bernard, Philippe ; Licata, Laurent: Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. In: *Journal of Experimental Social Psychology* 49 (2013), Nr. 4, S. 764–766

[54] Lin, Jimmy: An exploration of the principles underlying redundancy-based factoid question answering. In: *ACM Transactions on Information Systems (TOIS)* 25 (2007), Nr. 2, S. 6

[55] Lison, Pierre ; Kutuzov, Andrey: Redefining context windows for word embedding models: An experimental study. In: *arXiv preprint arXiv:1704.05781* (2017)

[56] Liu, Feifan ; Pennell, Deana ; Liu, Fei ; Liu, Yang: Unsupervised approaches for automatic keyword extraction using meeting transcripts. In: *Proceedings of human language technologies.* Bd. 1 ACL, 2009, S. 620–628

[57] Liu, Xiaodong ; Gao, Jianfeng ; He, Xiaodong ; Deng, Li ; Duh, Kevin ; Wang, Ye-Yi: Representation learning using multi-task deep neural networks for semantic classification and information retrieval. (2015)

[58] Loni, Babak: A survey of state-of-the-art methods on question classification. (2011)

[59] Lu, Jiwen ; Hu, Junlin ; Zhou, Jie: Deep metric learning for visual understanding: An overview of recent advances. In: *IEEE Signal Processing Magazine* 34 (2017), Nr. 6, S. 76–84

[60] Ma, Zhuang ; Collins, Michael: Noise Contrastive Estimation and Negative Sampling for Conditional Models: Consistency and Statistical Efficiency. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, S. 3698–3707

[61] Majdoubi, Jihen ; Tmar, Mohamed ; Gargouri, Faiez: Using the MeSH thesaurus to index a medical article: combination of content, structure and semantics. In: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* Springer, 2009, S. 277–284

[62] Malakasiotis, Prodromos ; Androutsopoulos, Ion ; Bernadou, Agiatis ; Chatzidiakou, Nephelie ; Papaki, Eliza ; Constantopoulos, Panos ; Pavlopoulos, Ioannis ; Krithara, Anastasia ; Almyrantis, Yannis ; Polychronopoulos, Dimitris [u. a.]: Challenge Evaluation Report 2 and Roadmap. In: *BioASQ deliverable D* 5 (2014)

[63] Mao, Yuqing ; Wei, Chih-Hsuan ; Lu, Zhiyong: NCBI at the 2014 BioASQ Challenge Task: Large-scale Biomedical Semantic Indexing and Question Answering. In: *CLEF (Working Notes)*, 2014, S. 1319–1327

[64] Mass, Yosi ; Roitman, Haggai ; Erera, Shai ; Rivlin, Or ; Weiner, Bar ; Konopnicki, David: A Study of BERT for Non-Factoid Question-Answering under Passage Length Constraints. In: *arXiv preprint arXiv:1908.06780* (2019)

[65] McKeown, Kathleen ; Radev, Dragomir R.: Generating summaries of multiple news articles. In: *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* ACM, 1995, S. 74–82

[66] Mikolov, Tomas ; Chen, Kai ; Corrado, Greg ; Dean, Jeffrey: Efficient estimation of word representations in vector space. In: *arXiv preprint arXiv:1301.3781* (2013)

[67] Mikolov, Tomas ; Corrado, Greg ; Chen, Kai ; Dean, Jeffrey: Efficient Estimation of Word Representations in Vector Space. In: *Proceedings of the International Conference on Learning Representations (ICLR 2013)* (2013), S. 1–12. – ISBN 1532–4435

[68] Miller, Alexander ; Fisch, Adam ; Dodge, Jesse ; Karimi, Amir-Hossein ; Bordes, Antoine ; Weston, Jason: Key-Value Memory Networks for Directly Reading Documents. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, S. 1400–1409

[69] Mitra, Bhaskar ; Craswell, Nick: Neural text embeddings for information retrieval. In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* ACM, 2017, S. 813–814

[70] Mitra, Bhaskar ; Nalisnick, Eric ; Craswell, Nick ; Caruana, Rich: A dual embedding space model for document ranking. In: *arXiv preprint arXiv:1602.01137* (2016)

[71] Moen, SPFGH ; Ananiadou, Tapio Salakoski2 S.: Distributional semantics resources for biomedical text processing. In: *Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan*, 2013, S. 39–43

[72] Monz, Christof: Document retrieval in the context of question answering. In: *European Conference on Information Retrieval* Springer, 2003, S. 571–579

[73] Narasimhan, Karthik ; Yala, Adam ; Barzilay, Regina: Improving information extraction by acquiring external evidence with reinforcement learning. In: *arXiv preprint arXiv:1603.07954* (2016)

[74] Nentidis, Anastasios ; Bougiatiotis, Konstantinos ; Krithara, Anastasia ; Paliouras, Georgios: Results of the Seventh Edition of the BioASQ Challenge. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* Springer, 2019, S. 553–568

[75] Nentidis, Anastasios ; Krithara, Anastasia ; Bougiatiotis, Konstantinos ; Paliouras, Georgios ; Kakadiaris, Ioannis: 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. Brussels, Belgium : Association for Computational Linguistics, November 2018, S. 1–10

[76] Nentidis, Anastasios ; Krithara, Anastasia ; Bougiatiotis, Konstantinos ; Paliouras, Georgios ; Kakadiaris, Ioannis: 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*. Brussels, Belgium : Association for Computational Linguistics, November 2018, S. 1–10

[77] Neto, Joel L. ; Santos, Alexandre D. ; Kaestner, Celso A. ; Alexandre, Neto ; Santos, D [u. a.]: Document clustering and text summarization. (2000)

[78] Neumann, Mark ; King, Daniel ; Beltagy, Iz ; Ammar, Waleed: ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In: *arXiv preprint arXiv:1902.07669* (2019)

[79] NIH. *PubMed Baseline Repository*. 2017

[80] NIH. *PubMed Baseline Repository*. 2018

[81] Nogueira, Rodrigo ; Cho, Kyunghyun: Passage Re-ranking with BERT. In: *arXiv preprint arXiv:1901.04085* (2019)

[82] Nogueira, Rodrigo ; Yang, Wei ; Cho, Kyunghyun ; Lin, Jimmy: Multi-stage document ranking with BERT. In: *arXiv preprint arXiv:1910.14424* (2019)

[83] Noy, Natalya F. ; McGuinness, Deborah L. [u. a.]. *Ontology development 101: A guide to creating your first ontology*. 2001

[84] Oh, Jong-Hoon ; Torisawa, Kentaro ; Kruengkrai, Canasai ; Iida, Ryu ; Kloetzer, Julien: Multi-column convolutional neural networks with causality-attention for why-question answering. In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* ACM, 2017, S. 415–424

[85] Othman, Nouha ; Faiz, Rim: Question answering passage retrieval and re-ranking using n-grams and SVM. In: *Computación y Sistemas* 20 (2016), Nr. 3, S. 483–494

[86] Ozyurt, Ibrahim B. ; Bandrowski, Anita ; Grethe, Jeffrey S.: Bio-AnswerFinder: a system to find answers to questions from biomedical texts. In: *Database* 2020 (2020)

[87] Pakray, Partha ; Bhaskar, Pinaki ; Banerjee, Somnath ; Pal, Bidhan C. ; Bandyopadhyay, Sivaji ; Gelbukh, Alexander F.: A Hybrid Question Answering System based on Information Retrieval and Answer Validation. In: *CLEF (Notebook Papers/Labs/Workshop)*, 2011

[88] Palangi, Hamid ; Deng, Li ; Shen, Yelong ; Gao, Jianfeng ; He, Xiaodong ; Chen, Jianshu ; Song, Xinying ; Ward, Rabab: Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 24 (2016), Nr. 4, S. 694–707

[89] Pappas, Dimitris ; McDonald, Ryan ; Brokos, Georgios-Ioannis ; Androutsopoulos, Ion: AUEB at BioASQ 7: document and snippet retrieval. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* Springer, 2019, S. 607–623

[90] Peters, Matthew E. ; Neumann, Mark ; Iyyer, Mohit ; Gardner, Matt ; Clark, Christopher ; Lee, Kenton ; Zettlemoyer, Luke: Deep contextualized word representations. In: *arXiv preprint arXiv:1802.05365* (2018)

[91] Pineda-Vargas, Mónica ; Rosso-Mateus, Andrés ; González, Fabio A. ; Montes-y Gómez, Manuel: A Mixed Information Source Approach for Biomedical Question Answering: MindLab at BioASQ 7B. In: Cellier, Peggy (Hrsg.) ; Driessens, Kurt (Hrsg.): *Machine Learning and Knowledge Discovery in Databases*. Cham : Springer International Publishing, 2020, S. 595–606

[92] Pizzato, Luiz A. ; Mollá, Diego: Indexing on semantic roles for question answering. In: *Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering*, 2008, S. 74–81

[93] Ponte, Jay M. ; Croft, W B.: A language modeling approach to information retrieval. In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, S. 275–281

[94] Radford, Alec ; Narasimhan, Karthik ; Salimans, Tim ; Sutskever, Ilya. *Improving language understanding by generative pre-training.* 2018

[95] Rajpurkar, Pranav ; Zhang, Jian ; Lopyrev, Konstantin ; Liang, Percy: SQuAD: 100,000+ Questions for Machine Comprehension of Text. (2016), S. 2383–2392

[96] Ramage, Daniel ; Rafferty, Anna N. ; Manning, Christopher D.: Random walks for text semantic similarity. In: *Proceedings of the 2009 workshop on graph-based methods for natural language processing* Association for Computational Linguistics, 2009, S. 23–31

[97] Riezler, Stefan ; Vasserman, Alexander ; Tsochantaridis, Ioannis ; Mittal, Vibhu ; Liu, Yi: Statistical machine translation for query expansion in answer retrieval. In: *ACL*, 2007

[98] Robertson, Stephen ; Zaragoza, Hugo: *The probabilistic relevance framework: BM25 and beyond*. 1. Now Publishers Inc, 2009

[99] Rosso-Mateus, Andres ; Montes-y Gomez, Manuel ; Rosso, Paolo ; Gonzalez, Fabio A.: Deep fusion of multiple term-similarity measures for biomedical passage retrieval. In: *Journal of Intelligent and Fuzzy Systems* 39, Nr. 2, S. 2239–2248

[100] Rosso-Mateus, Andrés ; González, Fabio A. ; Montes-y Gómez, Manuel: A Shallow Convolutional Neural Network Architecture for Open Domain Question Answering. In: *Colombian Conference on Computing* Springer, 2017, S. 485–494

[101] Rosso-Mateus, Andrés ; González, Fabio A. ; Montes-y Gómez, Manuel: A Two-Step Neural Network Approach to Passage Retrieval for Open Domain Question Answering. In: *Iberoamerican Congress on Pattern Recognition* Springer, 2017, S. 566–574

[102] Rosso-Mateus, Andrés ; González, Fabio A. ; Montes, Manuel: Mindlab neural network approach at bioasq 6b. In: *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, 2018, S. 40–46

[103] Rosso-Mateus Andrés, Fabio A. G.: A Deep Metric Learning Method for Biomedical Passage Retrieval. (2020)

[104] Rubin, Daniel L. ; Lewis, Suzanna E. ; Mungall, Chris J. ; Misra, Sima ; Westerfield, Monte ; Ashburner, Michael ; Sim, Ida ; Chute, Christopher G. ; Storey, Margaret-Anne ; Smith, Barry [u. a.]: National center for biomedical ontology: advancing biomedicine through structured organization of scientific knowledge. In: *Omics: a journal of integrative biology* 10 (2006), Nr. 2, S. 185–198

[105] Rybinski, Maciej ; Xu, Jerry ; Karimi, Sarvnaz: Clinical trial search: Using biomedical language understanding models for re-ranking. In: *Journal of Biomedical Informatics* 109 (2020), S. 103530

[106] Rydning, David Reinsel-John Gantz-John: The digitization of the world from edge to core. In: *Framingham: International Data Corporation* (2018)

[107] Salakhutdinov, Ruslan ; Hinton, Geoffrey: Semantic hashing. In: *International Journal of Approximate Reasoning* 50 (2009), Nr. 7, S. 969–978

[108] Schroff, Florian ; Kalenichenko, Dmitry ; Philbin, James: Facenet: A unified embedding for face recognition and clustering. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, S. 815–823

[109] Severyn, Aliaksei ; Moschitti, Alessandro: Learning to rank short text pairs with convolutional deep neural networks. In: *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015, S. 373–382

[110] SHATKAY, HAGIT ; FELDMAN, RONEN: Mining the Biomedical Literature in the Genomic Era: An Overview. In: *JOURNAL OF COMPUTATIONAL BIOLOGY* 10 (2003), Nr. 6

[111] Shen, Yelong ; He, Xiaodong ; Gao, Jianfeng ; Deng, Li ; Mesnil, Grégoire: A latent semantic model with convolutional-pooling structure for information retrieval. In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* ACM, 2014, S. 101–110

[112] Silva, Joao ; Coheur, Luísa ; Mendes, Ana C. ; Wichert, Andreas: From symbolic to sub-symbolic information in question classification. In: *Artificial Intelligence Review* 35 (2011), Nr. 2, S. 137–154

[113] Socher, Richard ; Manning, Christopher D. ; Ng, Andrew Y.: Learning continuous phrase representations and syntactic parsing with recursive neural networks. In: *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop* Bd. 2010, 2010, S. 1–9

[114] Soldaini, Luca ; Goharian, Nazli: Quickumls: a fast, unsupervised approach for medical concept extraction. In: *MedIR workshop, sigir*, 2016

[115] Srihari, Rohini ; Li, Wei: Information Extraction Supported Question Answering. (1999)

[116] Srivastava, Nitish ; Hinton, Geoffrey E. ; Krizhevsky, Alex ; Sutskever, Ilya ; Salakhutdinov, Ruslan: Dropout: a simple way to prevent neural networks from overfitting. In: *Journal of Machine Learning Research* 15 (2014), Nr. 1, S. 1929–1958

[117] Steinberger, Josef ; Jezek, Karel: Using latent semantic analysis in text summarization and summary evaluation. In: *Proc. ISIM* 4 (2004), S. 93–100

[118] Surdeanu, Mihai ; Ciaramita, Massimiliano ; Zaragoza, Hugo: Learning to rank answers to non-factoid questions from web collections. In: *Computational linguistics* 37 (2011), Nr. 2, S. 351–383

[119] Tan, Ming ; Xiang, Bing ; Zhou, Bowen: LSTM-based Deep Learning Models for non-factoid answer selection. CoRR abs/1511.04108 (2015). In: *arXiv preprint arXiv:1511.04108* (2015)

[120] Tay, Yi ; Tuan, Luu A. ; Hui, Siu C.: Hyperbolic representation learning for fast and efficient neural question answering. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, S. 583–591

[121] Telukuntla, Sai K. ; Kapri, Aditya ; Zadrozny, Wlodek: UNCC biomedical semantic question answering systems. BioASQ: Task-7B, Phase-B. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* Springer, 2019, S. 695–710

[122] Tsatsaronis, George ; Balikas, Georgios ; Malakasiotis, Prodromos ; Partalas, Ioannis ; Zschunke, Matthias ; Alvers, Michael R. ; Weissenborn, Dirk ; Krithara, Anastasia ; Petridis, Sergios ; Polychronopoulos, Dimitris ; Almirantis, Yannis ; Pavlopoulos, John ; Baskiotis, Nicolas ; Gallinari, Patrick ; Artiéres, Thierry ; Ngomo, Axel-Cyrille N. ; Heino, Norman ; Gaussier, Eric ; Barrio-Alvers, Liliana ; Schroeder, Michael ; Androutsopoulos, Ion ; Paliouras, Georgios: An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. In: *BMC Bioinformatics* 16 (2015), Nr. 1, S. 138. – ISBN 14712105 (Electronic)

[123] Tsatsaronis, George ; Balikas, Georgios ; Malakasiotis, Prodromos ; Partalas, Ioannis ; Zschunke, Matthias ; Alvers, Michael R. ; Weissenborn, Dirk ; Krithara, Anastasia ; Petridis, Sergios ; Polychronopoulos, Dimitris ; Almirantis, Yannis ; Pavlopoulos, John ; Baskiotis, Nicolas ; Gallinari, Patrick ; Artiéres, Thierry ; Ngomo, Axel-Cyrille N. ; Heino, Norman ; Gaussier, Eric ; Barrio-Alvers, Liliana ; Schroeder, Michael ; Androutsopoulos, Ion ; Paliouras, Georgios: An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. In: *BMC Bioinformatics* 16 (2015), Nr. 1, S. 138. – ISBN 14712105 (Electronic)

[124] Tsatsaronis, George ; Schroeder, Michael ; Paliouras, Georgios ; Almirantis, Yannis ; Androutsopoulos, Ion ; Gaussier, Eric ; Gallinari, Patrick ; Artieres, Thierry ; Alvers, Michael R. ; Zschunke, Matthias [u. a.]: BioASQ: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In: *AAAI fall symposium: Information retrieval and knowledge discovery in biomedical text*, 2012

[125] Tymoshenko, Kateryna ; Bonadiman, Daniele ; Moschitti, Alessandro: Convolutional neural networks vs. convolution kernels: Feature engineering for answer sentence reranking. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, S. 1268–1278

[126] Unger, Christina ; Bühmann, Lorenz ; Lehmann, Jens ; Ngonga Ngomo, Axel-Cyrille ; Gerber, Daniel ; Cimiano, Philipp: Template-based question answering over RDF data. In: *Proceedings of the 21st international conference on World Wide Web* ACM, 2012, S. 639–648

[127] Van-Tu, Nguyen ; Anh-Cuong, Le: Improving question classification by feature extraction and selection. In: *Indian Journal of Science and Technology* 9 (2016), Nr. 17

[128] Vaswani, Ashish ; Shazeer, Noam ; Parmar, Niki ; Uszkoreit, Jakob ; Jones, Llion ; Gomez, Aidan N. ; Kaiser, Łukasz ; Polosukhin, Illia: Attention is All you Need. In: *Advances in Neural Information Processing Systems*, 2017, S. 5998–6008

[129] Veyseh, Amir Pouran B.: Cross-lingual question answering using common semantic space. In: *Proceedings of TextGraphs-10: the workshop on graph-based methods for natural language processing*, 2016, S. 15–19

[130] Voorhees, Ellen M. [u. a.]: The TREC-8 question answering track report. In: *Trec* Bd. 99, 1999, S. 77–82

[131] Wang, Di ; Nyberg, Eric: A long short-term memory model for answer sentence selection in question answering. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, S. 707–712

[132] Wang, Mengqiu ; Smith, Noah A. ; Mitamura, Teruko: What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA. 7 (2007), S. 22–32

[133] Weston, Jason ; Bordes, Antoine ; Chopra, Sumit ; Rush, Alexander M. ; van Merriënboer, Bart ; Joulin, Armand ; Mikolov, Tomas: Towards ai-complete question answering: A set of prerequisite toy tasks. In: *arXiv preprint arXiv:1502.05698* (2015)

[134] Wu, Zhibiao ; Palmer, Martha: Verbs semantics and lexical selection. In: *32nd Proceedings ACL* Bd. 1 Association for Computational Linguistics, 1994, S. 133–138

[135] Xu, Kun ; Reddy, Siva ; Feng, Yansong ; Huang, Songfang ; Zhao, Dongyan: Question answering on freebase via relation extraction and textual evidence. In: *arXiv preprint arXiv:1603.00957* (2016)

[136] Yahya, Mohamed ; Berberich, Klaus ; Elbassuoni, Shady ; Ramanath, Maya ; Tresp, Volker ; Weikum, Gerhard: Natural language questions for the web of data. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* Association for Computational Linguistics, 2012, S. 379–390

[137] Yang, Hui ; Chua, Tat-Seng ; Wang, Shuguang ; Koh, Chun-Keat: Structured use of external knowledge for event-based open domain question answering. In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* ACM, 2003, S. 33–40

[138] Yang, Liu ; Ai, Qingyao ; Guo, Jiafeng ; Croft, W B.: anmm: Ranking short answer texts with attention-based neural matching model. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* ACM, 2016, S. 287–296

[139] Yang, Runqi ; Zhang, Jianhai ; Gao, Xing ; Ji, Feng ; Chen, Haiqing: Simple and Effective Text Matching with Richer Alignment Features. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy : Association for Computational Linguistics, Juli 2019, S. 4699–4709

[140] Yang, Wei ; Xie, Yuqing ; Lin, Aileen ; Li, Xingyu ; Tan, Luchen ; Xiong, Kun ; Li, Ming ; Lin, Jimmy: End-to-End Open-Domain Question Answering with BERTserini. In: *NAACL-HLT (Demonstrations)*, 2019

[141] Yang, Yi ; Yih, Wen-tau ; Meek, Christopher: WikiQA: A Challenge Dataset for Open-Domain Question Answering. In: *EMNLP* Citeseer, 2015, S. 2013–2018

[142] Yih, Scott Wen-tau ; Chang, Ming-Wei ; He, Xiaodong ; Gao, Jianfeng: Semantic parsing via staged query graph generation: Question answering with knowledge base. (2015)

[143] Yin, Wenpeng ; Schütze, Hinrich ; Xiang, Bing ; Zhou, Bowen: Abcnn: Attention-based convolutional neural network for modeling sentence pairs. In: *Transactions of the Association for Computational Linguistics* 4 (2016), S. 259–272

[144] Yoon, Seunghyun ; Dernoncourt, Franck ; Kim, Doo S. ; Bui, Trung ; Jung, Kyomin: A compare-aggregate model with latent clustering for answer selection. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, S. 2093–2096

[145] Yu, Lei ; Hermann, Karl M. ; Blunsom, Phil ; Pulman, Stephen: Deep learning for answer sentence selection. In: *arXiv preprint arXiv:1412.1632* (2014)

[146] Yu, Lei ; Hermann, Karl M. ; Blunsom, Phil ; Pulman, Stephen: Deep learning for answer sentence selection. In: *NIPS deep learning workshop* (2014)

[147] Yujian, Li ; Bo, Liu: A normalized Levenshtein distance metric. In: *IEEE transactions on pattern analysis and machine intelligence* 29 (2007), Nr. 6, S. 1091–1095

[148] Zadeh, Lotfi A.: From search engines to question answering systems—The problems of world knowledge, relevance, deduction and precisiation. In: *Capturing Intelligence* 1 (2006), S. 163–210

[149] Zhai, Chengxiang ; Lafferty, John. *A study of smoothing methods for language models applied to Ad Hoc information retrieval.* 2001

[150] Zhang, Dell ; Lee, Wee S.: Question classification using support vector machines. In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* ACM, 2003, S. 26–32

[151] Zheng, Guoqing ; Callan, Jamie: Learning to reweight terms with distributed representations. In: *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* ACM, 2015, S. 575–584