A Deep Metric Learning Method for Biomedical Passage Retrieval

Andrés Rosso-Mateus
MindLab
Universidad Nacional de Colombia
Bogotá, Colombia
aerossom@unal.edu.co

Manuel Montes-y-Gómez
LabTL
INAOE
Puebla, Mexico
mmontesg@inaoep.mx

Fabio A. González
MindLab
Universidad Nacional de Colombia
Bogotá, Colombia
fagonzalezo@unal.edu.co

Abstract

Passage retrieval is the task of identifying text snippets that are valid answers for a natural language posed question. One way to address this problem is to look at it as a metric learning problem, where we want to induce a metric between questions and passages that assign smaller distances to more relevant passages. In this work, we present a novel method for passage retrieval that learns a metric for questions and passages based on their internal semantic interactions. The method uses a similar approach to that of triplet networks, where the training samples are composed of one anchor (the question) and two positive and negative samples (passages). However, and in contrast with triplet networks, the proposed method uses a novel deep architecture that better exploits the particularities of text and takes into consideration complementary relatedness measures. Besides, the paper presents a sampling strategy that selects both easy and hard negative samples which improves the accuracy of the trained model. The method is particularly well suited for domain-specific passage retrieval where it is very important to take into account different sources of information. The proposed approach was evaluated in a biomedical passage retrieval task, the BioASQ challenge, outperforming standard triplet loss substantially by 10%, and state-of-the-art performance by 26%.

1 Introduction

Scientific documents are a valuable source of information for both physicians and researchers in medical sciences. Relevant information is continuously produced, more than 3,000 articles are published every day (Tsatsaronis et al., 2012). Manually finding relevant information in this among of data is an enormous challenge (Sarrouti and El Alaoui, 2017). Passage retrieval methods can alleviate the manual scanning of documents by automatically finding a subset of relevant passages that speed-up and improve the search process. Their goal is to return the highest correlated passages that are a valid answer for a given question.

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 (Bonadiman et al., 2019), 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 paper 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 paper presents a sampling strategy that chooses easy and then hard negative samples in the training phase, improving the overall model performance. The

Proceedings of the 28th International Conference on Computational Linguistics, pages 6229–6239
Barcelona, Spain (Online), December 8-13, 2020

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 (Tsatsaronis et al., 2012), and comparing with 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 of this work 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 (Brokos et al., 2018), which have in the order of millions while ours in the order of thousands.

The paper is organized as follows: Section 2 discusses the related work in Biomedical passage retrieval; Section 3 shows the details of the proposed metric learning method; Section 4 present the sampling strategy; Section 5 presents a systematic evaluation of the method; Section 6 discusses the results against the state of the art models; finally, Section 7 exposes some conclusions and discusses our future work ideas.

2 Related Work

Passage retrieval methods analyze the content of documents to identify snippets that likely answer a given specific question. This is a sub-task of the more general problem of question answering (QA). It has been extensively studied in open domain scenarios (Sarrouti and El Alaoui, 2017), but has received less attention in the biomedical domain. The launch of the biomedical QA track at BioASQ has boosted the research in this specific domain (Tsatsaronis et al., 2015). BioASQ challenge is the widest challenge for Biomedical indexing and information retrieval; for five consecutive years, they have shared a set of questions and related documents with annotated answer passages that contribute to research in this area. Numerous passage retrieval approaches have been proposed at BioASQ. For example, (Galkó and Eickhoff, 2018) 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 (Brokos et al., 2018). This approach is based on an ABCNN architecture (Yin et al., 2016), 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 USTB team approach combine different strategies to enrich query terms, such as sequential dependence models, pseudorelevance, fielded sequential dependence models and divergence from randomness models (Jin et al., 2017).

Finally, (Telukuntla et al., 2019) used Bert contextual word embeddings (Lee et al., 2020) to represent question and passage pairs, and fine-tuned the model to produce a ranking score. Most recent works 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 structured knowledge as biomedical terminology databases and ontologies, most of the approaches have not made use of these resources (Majdoubi et al., 2009). An exception is the work presented by (Bhogal et al., 2007), where ontologies are used to expand query terms. Structured resources offer information that is complementary to textual information and that can be used to alleviate problems as polysemy or synonymy disambiguation.

This work proposes a deep learning approach that takes as input different similarity representations that came from text and structured information. The proposed representations offer a diverse and complete view of question and passage interactions, which are then transformed into patterns by convolutional filters, to be finally projected into a metric learning space which locates question and related passages close to each other. The non-related passages are distant from the question with a minimum margin constrain. Metric learning has been used with success in medical diagnosis, image classification, voice recognition, among others (Kaya and Bilge, 2019). The use of metric learning in Natural Language Processing tasks has gained attention in the last years with approaches like the one presented by (Bonadiman et al., 2019), where a smoothed deep metric loss function is considered to identify repeated questions for open-domain community QA portals. This work is one of the few approaches that employ metric learning in the passage retrieval task. Another approach was presented by (Feng et al., 2015) where a metric learning approach based on a convolutional neural siamese architecture is applied to question and passage sequences. An important difference of this work with ours is that the convolutional neural network represent individually questions and answers following the architecture of a vanilla siamese network. In the model presented in this paper, the convolutional neural network learns to represent the interactions between questions and answers in an architecture that combines ideas from siamese and triplet networks.

3 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 (Schroff et al., 2015; Lu et al., 2017). 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 described in detail in the following sections.

3.1 Model Architecture

Our model architecture is presented in Figure 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 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.

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

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

[1] DMLPR source code `https://github.com/andresrosso/dmlpr_coling2020`

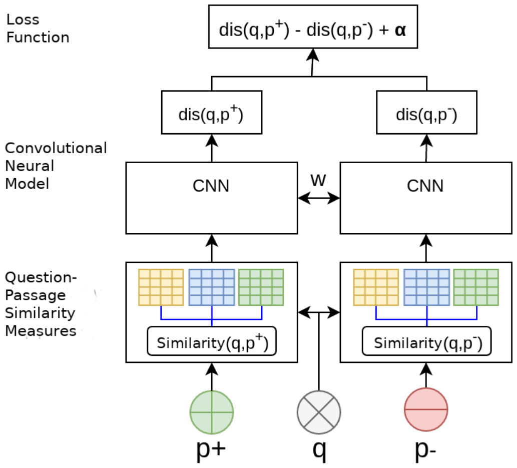

Loss Function
dis(q,p+) - dis(q,p-) + α
dis(q,p+)
dis(q,p-)
Convolutional Neural Model
CNN
W
CNN
Question-Passage Similarity Measures
Similarity(q,p+)
Similarity(q,p-)
p+
q
p-

Figure 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. W means that the CNN sub-model weights are shared.

3.2 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 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 (Rosso-Mateus et al., 2020), where the internal interactions are defined by three similarity matrices comparing each term in the question q_i with 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 (Liu et al., 2009; Dong et al., 2015).

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 (Soldaini and Goharian, 2016). To increase the concept identification coverage, a second check was done with the Scispacy tool (Neumann et al., 2019).

To visualize the information captured with the three similarity matrices and to emphasize their complementariness, Figure 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.

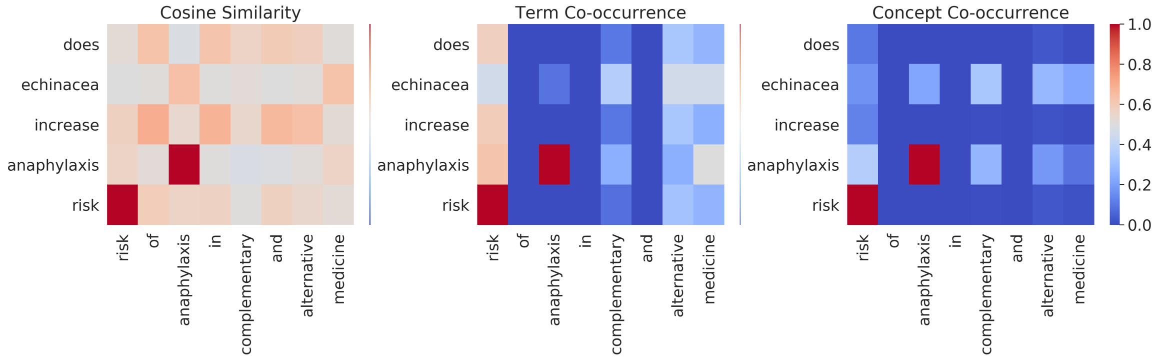
In the presented example, the concept similarity matrix offers higher semantic similarity values for question row term ’echinacea’ and the related answer passages ’complementary’, ’alternative’, ’medicine’, and ’anaphylaxis’ highlighting important relationships. Cosine similarity gives higher values to ’increase’ question term and its related row. Term co-occurrence has a similar behaviour to concept

[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

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

3.3 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 3.

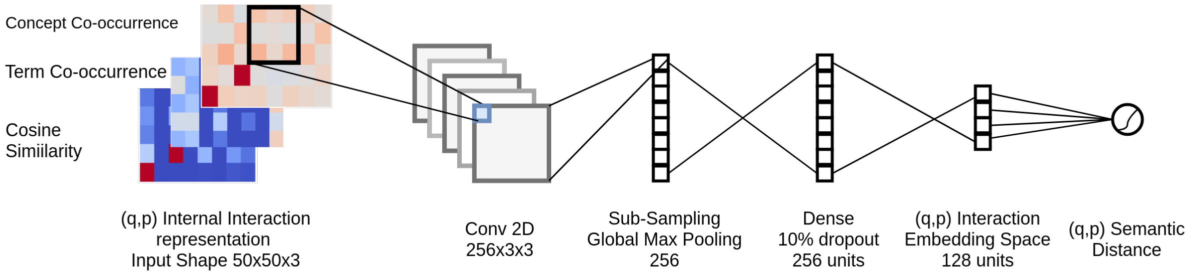

Concept Co-occurrence
Term Co-occurrence
Cosine Simiilarity
(q,p) Internal Interaction representation Input Shape 50x50x3
Conv 2D 256x3x3
Sub-Sampling Global Max Pooling 256
Dense 10% dropout 256 units
(q,p) Interaction Embedding Space 128 units
(q,p) Semantic Distance

Figure 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.

4 Informative Negative Passage Identification

Selecting informative training samples is very important in deep metric learning, as it is described in previous works (Bucher et al., 2016; Kaya and Bilge, 2019). Our approach discriminate hard negative samples based on the semantic relatedness of question and passage pairs using the cosine similarity over BiosentVec sentence embeddings (Chen et al., 2019). 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 (Chen et al., 2019); 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 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'.

5 Experimental Evaluation

5.1 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.

5.1.1 BioASQ Challenge Dataset

The BioASQ challenge provides a dataset for biomedical passage retrieval consisting of questions and related text fragments taken from PubMed abstracts (Tsatsaronis et al., 2015). The original dataset only provide 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:

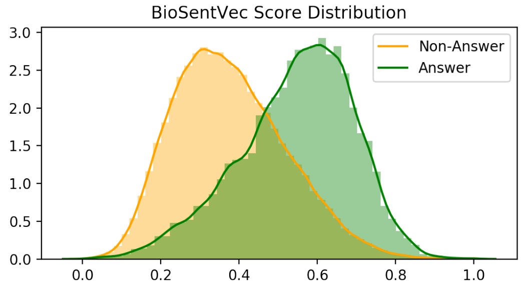


Figure 4: Cosine similarity density distribution for BioASQ negative and positive sample pairs

1. **Removal of repeated positive passages**: As there are a significant number of repeated passages, duplicated passages were removed based on the Levenshtein Distance (Yujian and Bo, 2007), as implemented in the FuzzyWuzzy tool[5].
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 (Leys et al., 2013).
3. **Selection of homogeneous negative passages**: Positive and negative passages should have similar lengths. We identified that 95% of the positive passages are between 13 and 55 terms long, therefore, we selected the negative passages that allowed a similar distribution to the positive ones.

Table 1 presents the statistics of the BioASQ training dataset after filtering out positive and adding negatives examples using the strategy discussed in Section 4 [6].

For testing, we used the test data set that was provided in the 2018 challenge version. The dataset is comprised of 5 batches, each containing 100 questions and variable candidate response passages. [7]

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

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

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

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

Table 1: BioASQ dataset with negative samples

5.1.2 Baselines

- **Bert fine-tuned model**: We used Bert model pretrained on biomedical texts (BioBert, (Lee et al., 2019)) 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 (Feng et al., 2015). Both text sequences were represented with BioNLP word embeddings [8].

- **Triplet network w2v-rep**: This is a conventional triplet network (Schroff et al., 2015) 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 by the DMLPR could by matched by a conventional triplet network using the same representation.

5.1.3 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 (Brokos et al., 2018).

5.2 Experimental Results

5.2.1 Ablation Study

The following results are intended to evaluate and compare the different configurations of the models, 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 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 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.

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

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

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

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 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.

5.2.2 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 (Brokos et al., 2018). 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.

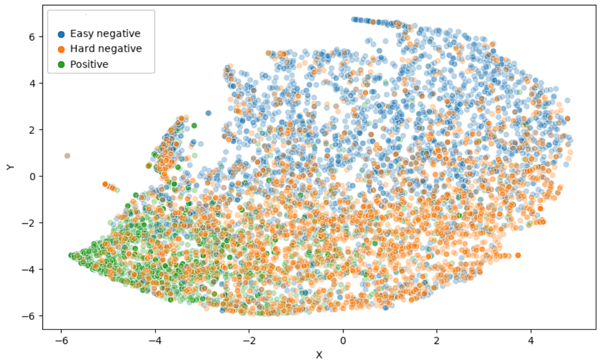


Figure 5: Visualization for the generated metric space using 2D tSNE dimensional reduction, points are BioASQ positive, hard-negative and easy-negative test-partition samples.

Table 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.

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	0.175	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 4: Passage retrieval results for the proposed baselines and the best models in BioASQ challenge 6b task (Nentidis et al., 2018)

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 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.

5.3 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.

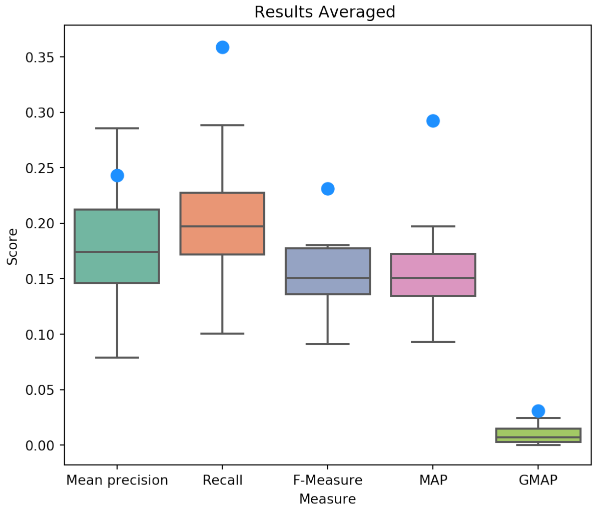


Figure 6: 15 best systems results for BioASQ task 6b, blue points correspond to the DMLPR model.

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.

6 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.

Motivated by the results obtained, future work will focus on exploring other alternatives to fuse information coming from structured knowledge sources. It will also be important to experiment with other forms of metric learning approaches.

Acknowledgements

COLCIENCIAS, REF. Agreement #727, 2016 provided financial as well as logistical and planning support. Mindlab research group (Universidad Nacional de Colombia sede Bogotá) with the cooperation of INAOE (Instituto Nacional de Astrofísica, Óptica y Electrónica) and Universitat Politècnica de València wich also provided technical support for this work.

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

References

Jagdev Bhogal, Andrew MacFarlane, and Peter Smith. 2007. A review of ontology based query expansion. *Information processing & management*, 43(4):866–886.

Daniele Bonadiman, Anjishnu Kumar, and Arpit Mittal. 2019. Large scale question paraphrase retrieval with smoothed deep metric learning. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 68–75.

George Brokos, Polyvios Liosis, Ryan McDonald, Dimitris Pappas, and Ion Androutsopoulos. 2018. Aueb at bioasq 6: Document and snippet retrieval. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, pages 30–39.

Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. 2016. Hard negative mining for metric learning based zero-shot classification. In *European Conference on Computer Vision*, pages 524–531. Springer.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE.

Li Dong, Furu Wei, Ming Zhou, and Ke Xu. 2015. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 260–269.

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

Ferenc Galkó and Carsten Eickhoff. 2018. Biomedical question answering via weighted neural network passage retrieval. In *European Conference on Information Retrieval*, pages 523–528. Springer.

Zan-Xia Jin, Bo-Wen Zhang, Fan Fang, Le-Le Zhang, and Xu-Cheng Yin. 2017. A multi-strategy query processing approach for biomedical question answering: Ustb_prir at bioasq 2017 task 5b. In *BioNLP 2017*, pages 373–380.

Mahmut Kaya and Hasan Şakir Bilge. 2019. Deep metric learning: a survey. *Symmetry*, 11(9):1066.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

Feifan Liu, Deana Pennell, Fei Liu, and Yang Liu. 2009. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of human language technologies.*, volume 1, pages 620–628. ACL.

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

Jihen Majdoubi, Mohamed Tmar, and Faiez Gargouri. 2009. 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*, pages 277–284. Springer.

Anastasios Nentidis, Anastasia Krithara, Konstantinos Bougiatiotis, Georgios Paliouras, and Ioannis Kakadiaris. 2018. Results of the sixth edition of the BioASQ challenge. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, pages 1–10, Brussels, Belgium, November. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Andrés Rosso-Mateus, Manuel Montes-y Gómez, Paolo Rosso, and Fabio A González. 2020. Deep fusion of multiple term-similarity measures for biomedical passage retrieval. *Journal of Intelligent & Fuzzy Systems*, (39):2239–2248.

Mourad Sarrouti and Said Ouatik El Alaoui. 2017. A passage retrieval method based on probabilistic information retrieval model and umls concepts in biomedical question answering. *Journal of biomedical informatics*, 68:96–103.

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

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

Sai Krishna Telukuntla, Aditya Kapri, and Wlodek Zadrozny. 2019. Uncc biomedical semantic question answering systems. bioasq: Task-7b, phase-b. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 695–710. Springer.

George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. 2012. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *2012 AAAI Fall Symposium Series*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4:259–272.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.