

Deep Metric Learning for Effective Passage Retrieval in the BioASQ Challenge*

Andrés Rosso-Mateus^{1,*†}, León A. Muñoz-Serna^{2,†}, Manuel Montes-y-Gómez^{3,†} and Fabio A. González^{4,†}

¹Keo World, Miami, US

²Keo World, Bogotá, Colombia

³LabTL, INAOE, Puebla, Mexico

⁴MindLab, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

This paper describes our participation in BioASQ 2023 Challenge for task 11b phase A, document retrieval and snippet retrieval. For document retrieval we have used BM25 scoring function and semantic-similarity as a re-ranking strategy, for passage retrieval our approach makes use of a metric learning method adapted for NLP. 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 by fusing different similarity measures through a siamese convolutional network. We also present a sampling strategy that selects challenging training samples which leads to an increase in the accuracy of the 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. Our approach reached the second position for snippet retrieval task.

Keywords

Deep Metric Learning, BioASQ, Passage Retrieval, Question Answering

1. Introduction

In the biomedical domain, the continuous growth of published documents poses challenges for researchers seeking relevant information. To address this issue, Question Answering (QA) systems have gained attention as they provide concise and natural retrieval of information, offering precise answers and supporting passages. The interest in developing QA systems for the biomedical domain has been increasing [1], as evidenced by the growing research and recognition of their potential in improving closed domain information access, representing the next evolution in information retrieval systems. In this paper we are going to describe the methods used for BioASQ 11 challenge phase A (document and passage retrieval) [2], but giving

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece


*Corresponding author.


†These authors contributed equally.

✉ andresr@keoworld.com (A. Rosso-Mateus); leonm@keoworld.com (L. A. Muñoz-Serna); mmontesg@inaoep.mx (M. Montes-y-Gómez); fagonzalezo@unal.edu.co (F. A. González)

🌐 <https://andresrosso.github.io> (A. Rosso-Mateus)

🆔 0000-0001-6015-4771 (A. Rosso-Mateus)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

focus in the passage retrieval task where a Deep Metric Learning method has been used to solve the related task.

Metric learning has been broadly used in face identification and other image processing tasks [3, 4]. 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 [5], 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.

Our deep metric learning method fuses different similarity measures through a siamese convolutional architecture. The proposed approach learns a metric between questions and passages, bringing semantically related pairs closer together. A sampling strategy is also presented to select both easy and hard negative samples during training, improving model performance.

The architecture combines aspects of triplet networks and siamese architectures but incorporates multiple question-passage internal similarity measures to capture important semantic features. This provides a complementary view of the relatedness between questions and passages, including structured information that is often available in domain-specific problems.

The proposed model reached the second position in passage retrieval in all batches even though the document retrieval approach was not very competitive.

The training dataset used along the whole process is the one suggested for the challenge committee [6].

2. Overall System Description

Our approach consists of two main components: the document retrieval and the passage retrieval modules, as shown in Figure 1.

The model implementation is publicly available in Github ¹.

2.1. Document Retrieval

The first module focuses on retrieving a set of documents that may contain the answer to a given question. The HayStack ² framework is employed for document retrieval, using the PubMed papers indexed in the 2023 PubMed Baseline Repository (MBR). The retrieval process involves collecting the 100 most relevant documents based on the BM25 ranking function. The query used for retrieval is the original question, and it is compared with the concatenation of each document's title and abstract.

Once the candidate documents are obtained, they are filtered down to a subset of no more than 10 documents using a re-ranking approach as follows.

¹DMLPR source code <https://github.com/andresrosso/col-un-bioasq11>

²HayStack NLP framework <https://haystack.deepset.ai/overview/intro>

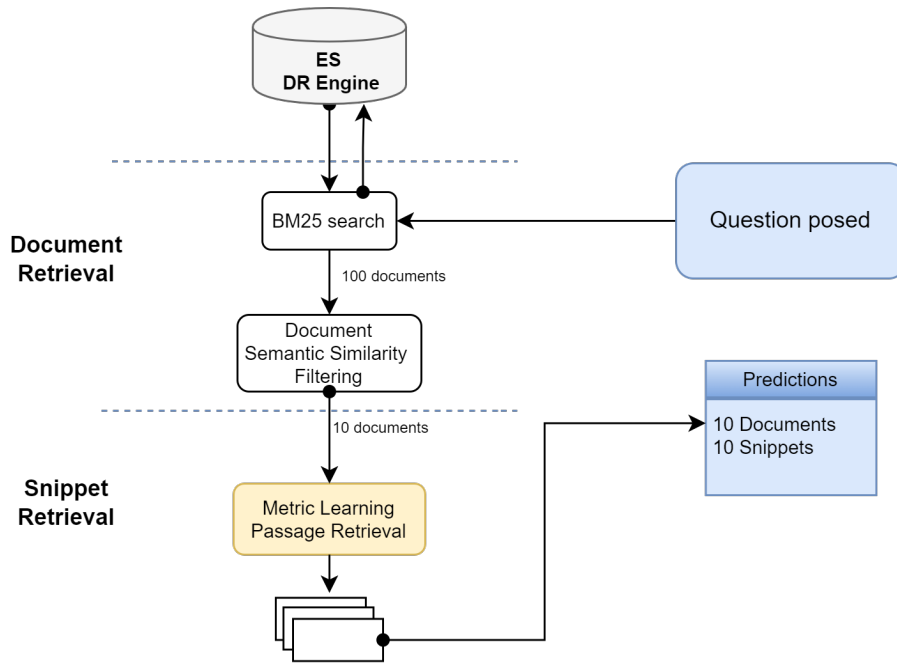


Figure 1: BioASQ Model Diagram

2.1.1. Document re-ranking

To address the issue of imprecise results from ES (Elasticsearch), a re-ranking strategy is implemented in a second stage. This re-ranker utilizes a pre-trained cross-encoder [7] to score the relevancy of all document candidates for a specific query. The re-ranking process selects the top-n documents that exhibit stronger semantic relevance to the query, leveraging the cosine distance in the embedding space. This process is illustrated in Figure 2.

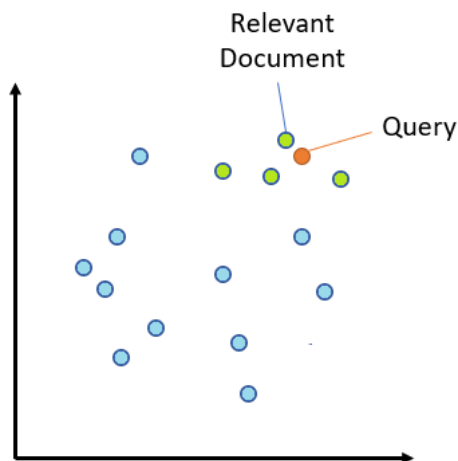


Figure 2: Semantic search for document re-ranking

To represent question and document we have used the pre-trained large language model (SBERT) [7] that has been fine-tuned using a siamese network structure to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity.

The selected subset of related documents is further analyzed to identify snippets of text that contain the answers to the question.

2.2. Passage Retrieval: A Deep Metric Learning Approach

A deep metric learning model is trained for this task, creating a metric space where question and passage pairs that are highly related are located close together. This allows for the ranking and exclusion of question-answer passages based on their similarity in the metric space.

Unlike traditional metric learning methods [3, 4], which aim to learn embedding spaces for individual samples, our approach focuses on learning a joint question-passage embedding that captures the relationship between pairs. A detailed description of the proposed architecture will be provided in the subsequent section.

2.2.1. Deep Metric Learning Model Architecture

The model architecture, depicted in Figure 3, works over three text sequences: the question, a positive passage that answers the question, and a negative passage that does not contain a valid answer.

The first step involves calculating the relatedness between the question and passages using various term-level question-passage similarity measures. These measures are represented as matrices for the positive (q, p_+) and negative (q, p_-) pairs.

These matrices are then fed into a siamese convolutional model, which identifies internal patterns in the question-passage interactions. These patterns are utilized to compute a measure of semantic relatedness, denoted as $dis_{(q,p_+)}$ and $dis_{(q,p_-)}$ for the positive and negative pairs, respectively.

The model is trained by minimizing the loss function defined in Equation 1, where the distances for positive pairs are encouraged to be close to zero, while negative pairs should have a distance greater than a margin α . The batch size is denoted as N .

$$\frac{1}{N} \sum_i^N [dis(q, p_+) - dis(q, p_-) + \alpha] \quad (1)$$

The two main blocks of this model, the input layer and convolutional layer, are described in the following subsections.

2.2.2. Input layer: Similarity Measures Calculation

The input training samples consist of a question and two passages, one positive and one negative. To represent a question-passage pair, internal semantic interactions are analyzed using three different similarity measures: 1) word embedding with cosine similarity, 2) term co-occurrence, and 3) concept co-occurrence. These measures were introduced in a previous work [8], where

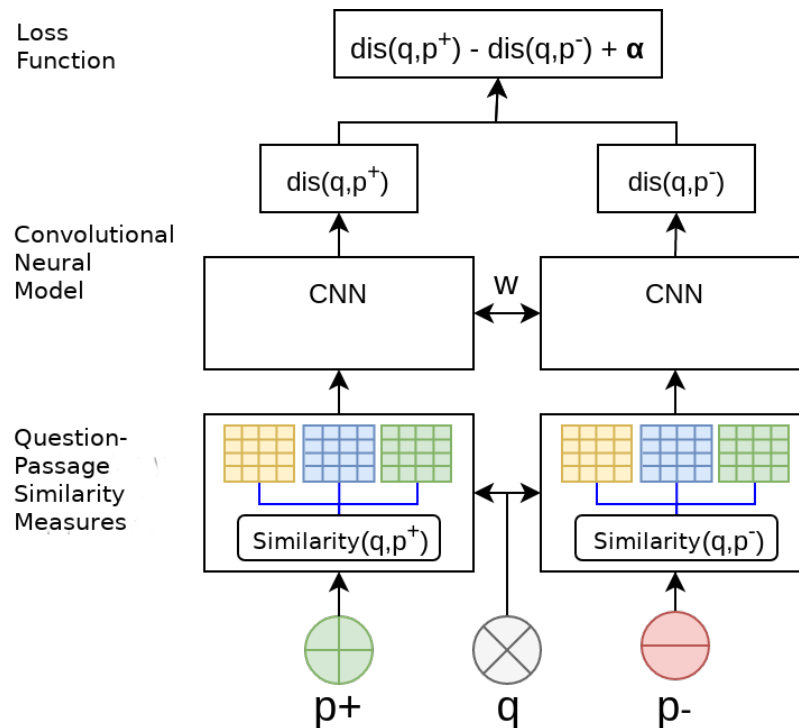


Figure 3: Overall model architecture; the input is composed of a question and a positive and negative passage, it includes a convolutional layer and a loss function that compares the distances between the positive and negative pairs. W means that the CNN sub-model weights are shared.

the interactions are captured through three similarity matrices. These matrices compare each term in the question (q_i) with each term in the candidate passage (p_j). Below is a brief description of those matrices.

Cosine similarity: it captures the relatedness of terms using the BioNLP pre-trained word embeddings and measuring their cosine similarity.

Term and concept co-occurrence measures: in order to capture statistical term by term and biomedical concepts coincidences at sentence level we pre-calculated co-occurrences matrices using the abstracts from BioASQ training data sentences. Our conceptual database is built using UMLS Meta-thesaurus³, QuickUMLS tool [9] as well as Scispacy tool [10].

To visualize the information captured with the three similarity matrices and to emphasize their complementariness, Figure 4 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.

³UMLS Meta-thesaurus <http://umlsks.nlm.nih.gov>

In the provided example, the concept similarity matrix demonstrates higher semantic similarity values for the question term 'echinacea' and its related answer passages, such as 'complementary', 'alternative', 'medicine', and 'anaphylaxis', indicating significant relationships. The cosine similarity also yields higher values for the question term 'increase' and its corresponding row. Term co-occurrence exhibits similar behavior to concept co-occurrence, but the latter places more emphasis on important terms. In this specific example, concept co-occurrence proves to be the more informative modality, highlighting a significant relationship between 'echinacea' and the terms 'anaphylaxis', 'alternative', and 'medicine'. This relationship suggests that echinacea is associated with adverse anaphylaxis allergic reactions, as documented in the medical literature.

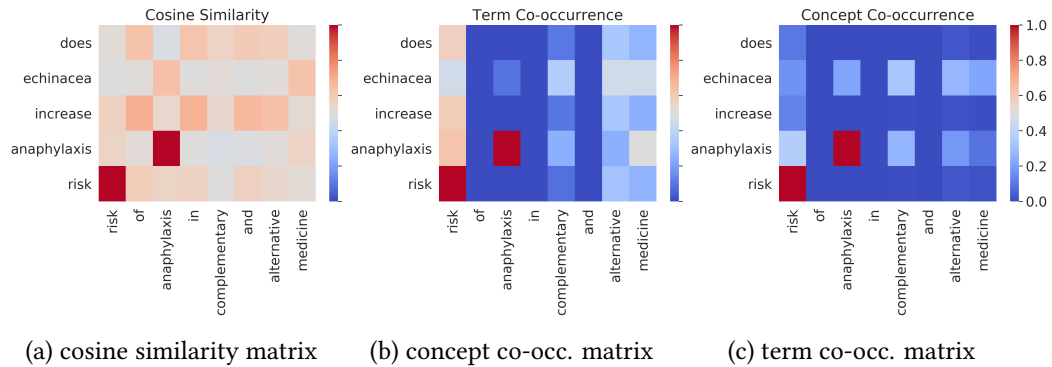


Figure 4: An example of the similarity matrices for a given question (rows) and passage (columns), aiming to visualize the sequence's internal interactions.

2.2.3. Convolutional Neural Layer

The result of the question-passage similarity calculation is a tensor with three similarity channels. This tensor is similar to the multi-channel representation used in images. The model being proposed has a siamese architecture with shared weights, where each subnet processes a pair of negative or positive input samples. The output of each subnet provides an estimation of the distance for the corresponding input pair, as is depicted in Figure 5.

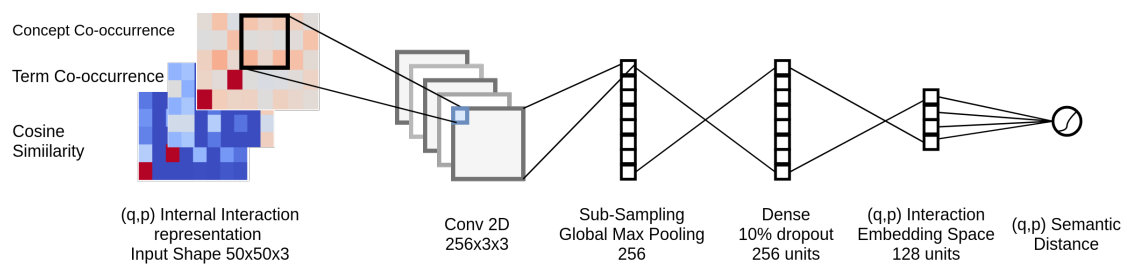


Figure 5: Convolutional model used in siamese architecture, each sub-net employ this architecture

The first layer of each subnet is composed of 256 3x3 convolutional with a Relu activation function serving as a feature extraction component. The extracted patterns are subsequently

condensed by a global max-pooling layer, which connects to a fully connected layer comprising 128 units with Relu activation. Finally, the estimated distance measure is produced by a sigmoid unit.

2.2.4. Informative Negative Passage Identification

In deep metric learning, selecting informative training samples is crucial. Previous works have emphasized this importance [11, 12]. Our approach focuses on discriminating hard negative samples based on the semantic relatedness of question and passage pairs, utilizing cosine similarity over BiosentVec sentence embeddings [13]. . During training, we first provide the model with easy negative samples and then introduce more challenging hard negative samples. The process of filtering these samples involves representing them in an embedded space using BioSentVec embeddings [13], calculating the similarity between question and passage pairs, estimating the densities for positive and negative samples; refer to Figure 6, and filtering the hard negative samples based on the likelihood of being positive, we determined whether it is hard or easy by comparing $p(x \in \text{positive})$ against $p(x \in \text{negative})$.

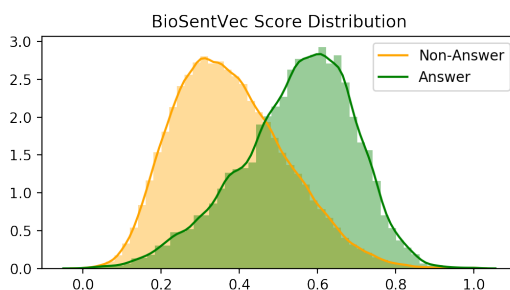


Figure 6: Cosine similarity density distribution for BioASQ negative and positive sample pairs

3. Results

To validate the performance for document retrieval and passage retrieval approaches, we have used the test dataset of BioASQ 10 version. We report results averaging official metrics over the 5 batches of the related dataset.

3.1. Document Retrieval Results and Discussion

The averaged results over the five 10b batches are presented in Table 1. We present the results for BM25 method and the comparison with the same method but using the described semantic-similarity re-ranking approach which has a slightly better performance over the BM25 initial document set.

Table 1

Document Retrieval results for BioASQ 10 (summarized)

Model	Mean precision	Recall	F-Measure	MAP	GMAP
BM25	0.2074	0.4724	0.2174	0.1319	0.0257
BM25_v2_semantic_similarity	0.2084	0.4706	0.2186	0.1332	0.0228

3.2. Passage Retrieval Results

The results of the passage retrieval task largely depend on the performance obtained in the document retrieval stage. To validate the performance of the proposed Deep Metric Learning for passage retrieval (DMLPR), we used in all experiments the same set of documents and then compare the results against the following baseline methods.

We have used three different baseline models for comparison in this paper. The first model is the **(Bert fine-tuned model)**, pre-trained on biomedical texts, and fine-tuned for question-answering tasks using BioASQ dataset [14]. The second model is a **(Siamese model)** that uses BioNLP word embeddings⁴ to represent the question and passage [15]. The third model is a conventional **(Triplet loss w2v-rep)** that also uses BioNLP word embeddings to represent the input sequences [3]. The fourth model combines a conventional triplet network with the multi-similarity representation **(Triplet loss sim-rep)**, here we use tensors to represent similarities between question-answer pairs.

These models were employed to explore and compare their performance against the proposed deep metric learning approach (DMLPR).

Table 2 presents the obtained results. The proposed method outperformed all baseline methods according to the averaged MAP score. With respect to the **(Triplet loss sim-rep)** an improvement of 10% was observed. It is important to mention that using multiple similarities as input yielded significantly better results compared to using non-interacting sequences, surpassing the **(Siamese model)** and **(Triplet loss w2v-rep)** by approximately 65%. The Bert model has moderate performance scores, and the margin concerning the proposed model is wide.

Table 2

Passage retrieval results for the proposed baselines and the best models in BioASQ challenge 10b task [16]

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
DMLPR	0.243	0.358	0.231	0.294	0.030

⁴BioNLP word vector representation, trained with biomedical and general-domain texts <http://bio.nlp.lab.org>

4. Discussion

The proposed method demonstrated a significant improvement over baseline methods. The success of the proposed model can be attributed to several factors. Firstly, the representation based on three similarity matrices proved to be much more effective in capturing the semantic relatedness between question and answer sequences compared to using independent representations. This differs from most current works that solely rely on learned text representations, incorporating domain knowledge in the form of important concepts and calculating a complementary similarity enhances the question-passage representation. Additionally, the combination of a metric learning approach with a CNN applied to text-similarity matrices was a distinctive feature of this work. The results demonstrated that this approach successfully captured the interactions between questions and passages. This strategy is not commonly employed in passage retrieval methods, and the study showcased its highly positive impact.

5. Conclusion

The paper makes use of a deep-metric learning approach for biomedical passage retrieval that outperforms baseline methods over BioASQ dataset. The model incorporates a multi-similarity representation, a convolutional neural network (CNN), and a siamese design. The training strategy involves identifying hard and easy negative samples to enhance the model's performance. The results indicate promising outcomes, leading to future research exploring alternative methods to integrate structured knowledge sources and different forms of metric learning approaches.

6. Acknowledgments

KEO World LLC 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) which also provided technical support for this work.

References

- [1] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artiéres, A.-C. N. Ngomo, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, G. Paliouras, An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition, *BMC Bioinformatics* 16 (2015) 138. URL: <http://www.biomedcentral.com/1471-2105/16/138>. doi:10.1186/s12859-015-0564-6.
- [2] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, L. Gasco, M. Krallinger, G. Paliouras, Overview of BioASQ 2023: The eleventh BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: *Experimental*

IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), ????

- [3] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.
- [4] J. Lu, J. Hu, J. Zhou, Deep metric learning for visual understanding: An overview of recent advances, *IEEE Signal Processing Magazine* 34 (2017) 76–84.
- [5] D. Bonadiman, A. Kumar, A. Mittal, 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, pp. 68–75.
- [6] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, BioASQ-QA: A manually curated corpus for Biomedical Question Answering, *Scientific Data* 10 (2023) 170.
- [7] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.
- [8] A. Rosso-Mateus, M. Montes-y Gómez, P. Rosso, F. A. González, Deep fusion of multiple term-similarity measures for biomedical passage retrieval, *Journal of Intelligent & Fuzzy Systems* (2020) 2239–2248.
- [9] L. Soldaini, N. Goharian, Quickumls: a fast, unsupervised approach for medical concept extraction, in: MedIR workshop, sigir, 2016.
- [10] M. Neumann, D. King, I. Beltagy, W. Ammar, Scispacy: Fast and robust models for biomedical natural language processing, *arXiv preprint arXiv:1902.07669* (2019).
- [11] M. Bucher, S. Herbin, F. Jurie, Hard negative mining for metric learning based zero-shot classification, in: European Conference on Computer Vision, Springer, 2016, pp. 524–531.
- [12] M. Kaya, H. Ş. Bilge, Deep metric learning: a survey, *Symmetry* 11 (2019) 1066.
- [13] Q. Chen, Y. Peng, Z. Lu, Biosentvec: creating sentence embeddings for biomedical texts, in: 2019 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, 2019, pp. 1–5.
- [14] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: pre-trained biomedical language representation model for biomedical text mining, *arXiv preprint arXiv:1901.08746* (2019).
- [15] M. Feng, B. Xiang, M. R. Glass, L. Wang, B. Zhou, Applying deep learning to answer selection: A study and an open task, in: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, 2015, pp. 813–820.
- [16] A. Nentidis, A. Krithara, G. Paliouras, L. Gasco, M. Krallinger, Bioasq at clef2022: The tenth edition of the large-scale biomedical semantic indexing and question answering challenge, in: Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II, Springer, 2022, pp. 429–435.