Undefined 1 (2019) 1–5
IOS Press

Deep Fusion of Multiple Term-Similarity Measures For Biomedical Passage Retrieval

Andrés Rosso-Mateus [a], Manuel Montes-y-Gómez [b], Paolo Rosso [c] and Fabio A. González [a,*]
[a] *Universidad Nacional de Colombia, Bogotá, Colombia*
E-mail: {aerossom,fagonzalezo}@unal.edu.co
[b] *Laboratorio de Tecnologías del Lenguaje INAOE, Puebla, Mexico*
E-mail: mmontesg@inaoep.mx
[c] *Universitat Politècnica de València, Valencia, Spain*
E-mail: prosso@dsic.upv.es

Abstract. Passage retrieval is an important stage of question answering systems. Closed domain passage retrieval, e.g. biomedical passage retrieval presents additional challenges such as specialized terminology, more complex and elaborated queries, scarcity in the amount of available data, among others. However, closed domains also offer some advantages such as the availability of specialized structured information sources, e.g. ontologies and thesauri, that could be used to improve retrieval performance. This paper presents a novel approach for biomedical passage retrieval which is able to combine different information sources using a similarity matrix fusion strategy based on a convolutional neural network architecture. The method was evaluated over the standard BioASQ dataset, a dataset specialized on biomedical question answering. The results show that the method is an effective strategy for biomedical passage retrieval able to outperform other state-of-the-art methods in this domain.

Keywords: Biomedical passage retrieval, Neural networks, question answering, Deep Learning

1. Introduction

Biomedical question answering (QA) systems are an important asset to support clinical decision processes and personal health information needs [12]. The manual effort to find useful information in large biomedical scientific repositories is a challenging task that can be alleviated with accurate automatic QA systems.

Passage retrieval methods analyze a narrow set of documents with the aim of identifying snippets that help to answer a specific question. Increased performance in passage retrieval task tends to produce a significant gain on the overall QA task [25].

Passage retrieval over biomedical domains has received less attention comparably with open domains. This fact reflects on the few QA datasets available for the biomedical domain, as Wasim et al. [27] showed. The biggest collection of question-answer passages for the biomedical domain is the dataset released by BioASQ Question Answering Challenge [25] with 2,747 questions-answer pairs. On the other hand, open domain QA has larger resources with more training data, such as SQuAD dataset with more than 100,000 questions [18], or WikiQA with 3,047 questions [29]. More recently, Google released a new open domain dataset with 307,373 questions. Biomedical passage retrieval remains a real challenge where the presence of technical terminology, compound concepts, complex entities, and elaborated queries make the task harder. In this domain, the specialized information sources have not been fully exploited and most of the approaches take advantage exclusively of textual data as a unique source of information.

In this paper, we present a novel method for biomedical passage retrieval. The model has the ability to combine different information sources. Specifically, the model fuses different term similarity mea-

*Corresponding author. E-mail: fagonzalezo@unal.edu.co.

sures which capture different aspects of the question-passage relationship. The similarities are optimally combined by a convolutional neural network. In the reported experiments three different similarity measures were used: biomedical word2vec embedding cosine similarity, term co-occurrence, biomedical concept co-occurrence. The main assumption is that the three similarities are complementary to each other and offer different views of the semantic relations between question and candidate answers. To validate the model performance and compare it against state-of-the-art models, we carried out a systematic evaluation with the BioASQ biomedical question answering challenge dataset [25]. The results show that the proposed model is an effective strategy for passage retrieval in the biomedical domain with results that equal and, in most of the cases, exceed state-of-the-art methods' performance.

The paper is organized as follows: Section 2 discusses the background and the related work; Section 3 shows the details of the proposed method; Section 4 presents a systematic evaluation of the method; finally, Section 5 exposes some conclusions and discusses our future work.

2. Background and related work

2.1. Biomedical Passage Retrieval

Biomedical queries are usually more specific than open domain search queries. This makes difficult the use of general-purpose retrieval models [25]. In the case of biomedical passage retrieval, most methods tend to only use textual information, a representative one is proposed by Brokos et al. [2]. They used an attention-based convolutional neural network to model question-answer pairs (ABCNN) [30]. This approach obtained the best results in the BioASQ 2018 challenge [5]. Other approaches use convolutional neural networks (CNN) and long-short-term-memory networks (LSTM) to represent the question and passage information to obtain a similarity score to rank the candidate answers [15].

2.2. Structured information sources

A typical biomedical query could be e.g. "How could we infer functional associations from gene fusion events?" This query involves specific domain knowledge related to genetics. Terminology databases and ontologies such as the Unified Medical Language System (UMLS) [8] encode, in a structured way, a good portion of this domain knowledge. Although specific domain knowledge resources are relevant in the sense that they can be used to alleviate problems as polysemy or synonymy disambiguation, they are not used frequently in passage retrieval tasks [11]. Despite this, there are some passage retrieval methods for open domain that exploit structured representations using different approaches: categorical grammars [31,6], synchronous context-free grammars [28], reinforcement learning [32], dependency-based trees for compositional semantics [20,24], and tree transducers [13].

In the open domain field, there are some works that try to incorporate structured based information. Das et al. [3] have demonstrated that a mixed data representation for question answering is better than using either a structured source or text source alone. In this approach the authors produce a joint representation with knowledge base facts and textual information (Universal Schema). A memory network [23] makes use of the produced join representation as an attention mechanism which is combined with the proposed query to select a related entity that answers the query. HAWK is another method developed by Usbeck et al. [26] where textual information is combined with linked data using an 8-step pipeline, which comprises: POS-tagging, NER, dependency parsing and linguistic pruning heuristics among others in order to discard no connected resulting graphs.

2.3. Similarity measures and passage retrieval

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) [1], 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 simi-

larity can be calculated following the minimal path between two nodes [22]. Ramage et al. have proposed a random walk algorithm [19] 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 [14], the semantic relation strength between a pair of terms is given by the occurrence in context windows. This parameter election 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 [9].

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.

3. Method

3.1. Overall architecture

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

3.1.1. 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 [17]. 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 [21].

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 [16]. 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 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 that in a binary vector. The resulting matrix will have a dimension $Ntimes M$, 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.

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 1.

$$Tc\_norm = (XX^T)(1/diag(XX^T) \quad (1)$$

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

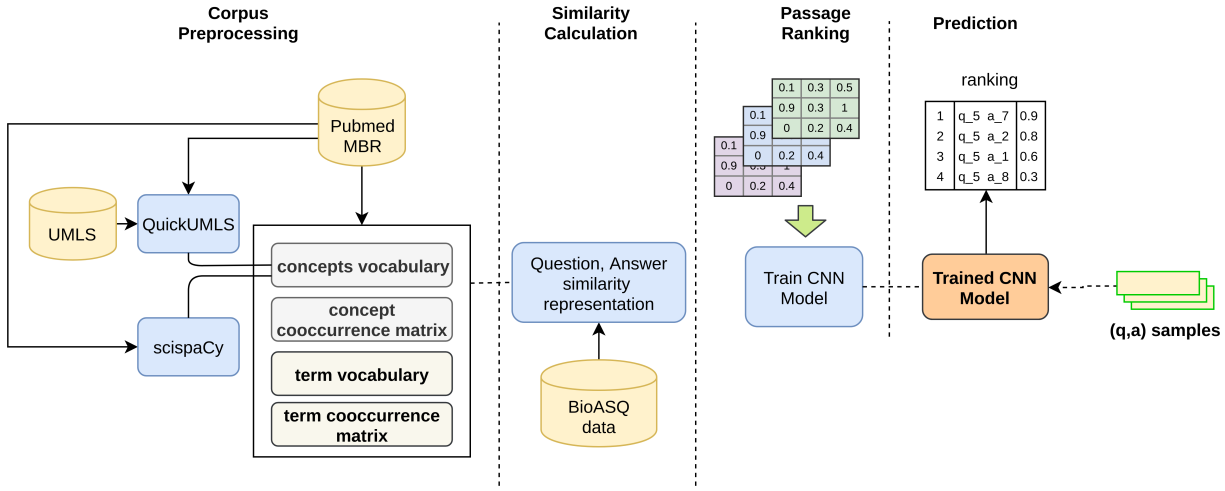


Fig. 1. Passage retrieval overall architecture

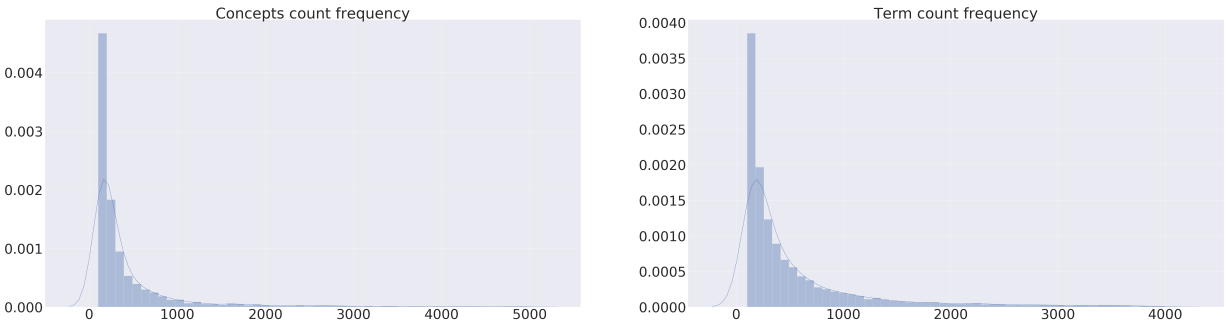


Fig. 2. Term and concept count frequency

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).

3.1.2. Co-occurrence similarity

In order to transform each question and answer term pair (q, a), we have to retrieve the correspondent co-occurrence for the related pair. This process is also done for concepts if the word is identified as such. When a word is not in the vocabulary (term or concept) we fill the related cell with 0, this allows us to align the similarity matrices representation in all the three tensor dimensions.

3.1.3. 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 trans-

formed 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 the embedding.
- **Step 3: Matrix weighting** M_t: as not all terms are equally informative for measuring text similarities [10,4], 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. 2 shows.

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

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

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

The value of $imp(x)$ function is based on the POS-tagging label. We consider verbs, nouns, and adjectives to be "important" [10,4]. 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.

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 3.

It can be observed that the cosine similarity matrix does not have a high value for "Bertolotti's" 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.

Another example of how similarity measures contribute to a improved representation is presented in the following example:

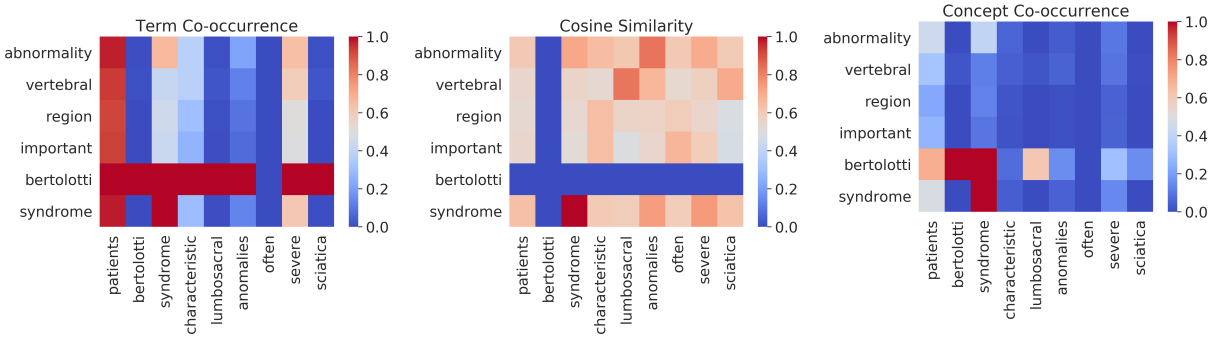
Q: Are defects in recombination repair involved in carcinogenesis?

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 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.

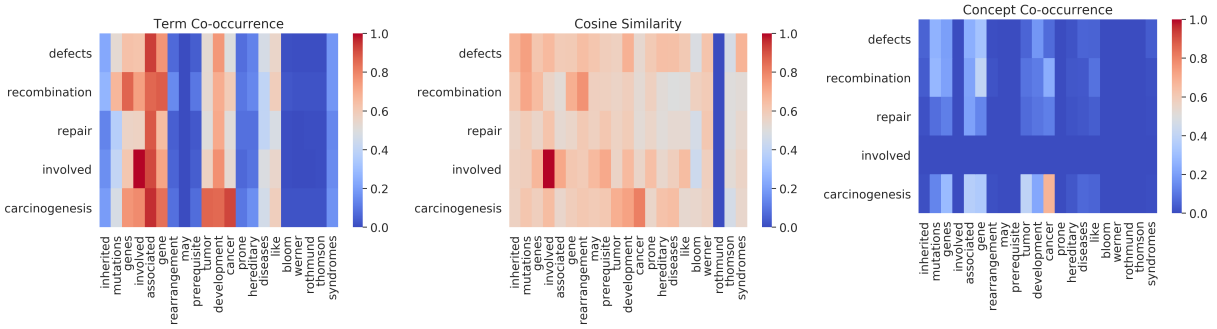
The objective with the incorporation of additional and complementary information is to feed the neural model with meaningful features that allows the model to identify when a question and answer pair are highly correlated. During the training phase, the CNN model has to determine those similarities patterns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

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



(a) Term co-occurrence similarity matrix (b) W2V cosine similarity (c) Concept co-occurrence similarity matrix

Fig. 3. Example 1. Similarity matrices



(a) Term co-occurrence similarity matrix (b) W2V cosine similarity (c) Concept co-occurrence similarity matrix

Fig. 4. Example 2. similarity matrices

3.1.4. 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. 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 sub-sampled 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.

3.2. 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.

4. 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.

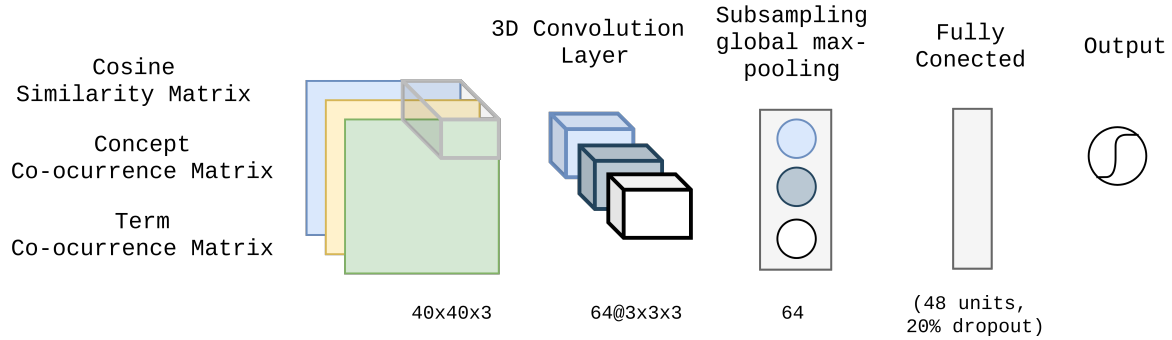


Fig. 5. Multiple channel convolutional neural network

4.1. Data set

The training was done with the question and answer pairs from 2016, 2017 and 2018 BioASQ Task B training datasets. The obtained dataset was very unbalanced, only 18% 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.

4.2. 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 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.

4.2.1. 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 [7].

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.

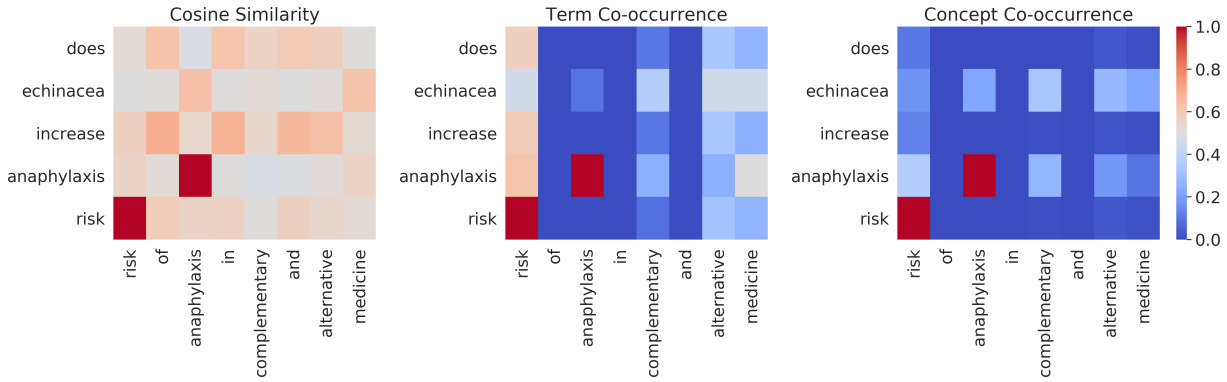
4.3. Results and Discussion

The results for different model configurations are reported in Table 1.

Method	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 1
Snippet retrieval results combining similarity matrices

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



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

Fig. 6. Similarity matrices example where concept co-occurrence have a better performance over the others

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?

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 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.

4.3.1. 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 [2]) 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 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 2 for aueb-nlp-5 [2] 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.

Method	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
concept + term + w2v	**0.2322**	**0.3838**	**0.3571**	**0.2409**	**0.1163**

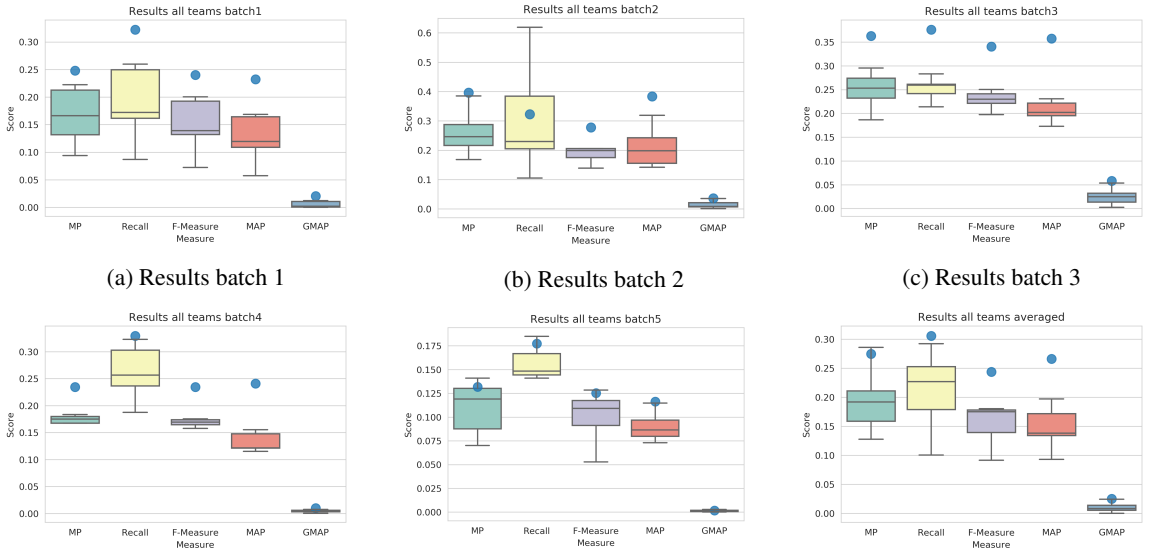
Table 2

Snippet retrieval results using the documents provided by AUEB [2]

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 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 [2].

In batch1, batch3 and batch4 we reached the best results as ilustrated in the boxplot. In batch2, the only



(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

Fig. 7. 15 best systems results for task 6b, blue points correspond to the proposed model

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.

5. CONCLUSION

In this paper, 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 the Scispacy model. The multi-similarity representation is exploited by a convolutional neural network which extracts similarity patterns and produces a semantic relatedness score, which is further used to rank the answer 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. Acknowledgement

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.

References

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

[2] Georgios-Ioannis Brokos, Polyvios Liosis, Ryan McDonald, Dimitris Pappas, and Ion Androutsopoulos. Aueb at bioasq 6: Document and snippet retrieval. *arXiv preprint arXiv:1809.06366*, 2018.

[3] Rajarshi Das, Manzil Zaheer, Siva Reddy, and Andrew McCallum. Question answering on knowledge bases and text using universal schema and memory networks. *arXiv preprint arXiv:1704.08384*, 2017.

[4] Li Dong, Furu Wei, Ming Zhou, and Ke Xu. 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, 2015.

[5] Ioannis A Kakadiaris, George Paliouras, and Anastasia Krithara. Proceedings of the 6th bioasq workshop a challenge on large-scale biomedical semantic indexing and question answering. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, 2018.

[6] Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. Lexical generalization in ccg grammar induction for semantic parsing. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1512–1523. Association for Computational Linguistics, 2011.

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

[8] Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. The unified medical language system. *Methods of information in medicine*, 32(04):281–291, 1993.

[9] Pierre Lison and Andrey Kutuzov. Redefining context windows for word embedding models: An experimental study. *arXiv preprint arXiv:1704.05781*, 2017.

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

[11] Jihen Majdoubi, Mohamed Tmar, and Faiez Gargouri. 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, 2009.

[12] Prodromos Malakasiotis, Ion Androutsopoulos, Agiatis Bernadou, Nephelie Chatzidiakou, Eliza Papaki, Panos Constantopoulos, Ioannis Pavlopoulos, Anastasia Krithara, Yannis Almyrantis, Dimitris Polychronopoulos, et al. Challenge evaluation report 2 and roadmap. *BioASQ deliverable D*, 5, 2014.

[13] Pascual Martínez-Gómez and Yusuke Miyao. Rule extraction for tree-to-tree transducers by cost minimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 12–22, 2016.

[14] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12, 2013.

[15] Diego Molla. Macquarie university at bioasq 6b: Deep learning and deep reinforcement learning for query-based summarisation. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, pages 22–29, 2018.

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

[17] National Institutes of Health. Pubmed baseline repository.

[18] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

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

[20] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4974–4983, 2017.

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

[22] Josef Steinberger and Karel Jezek. Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, 4:93–100, 2004.

[23] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.

[24] Raymond Hendy Susanto and Wei Lu. Semantic parsing with neural hybrid trees. In *AAAI*, pages 3309–3315, 2017.

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

[26] Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Lorenz Bühmann, and Christina Unger. Hawk–hybrid question answering using linked data. In *European Semantic Web Conference*, pages 353–368. Springer, 2015.

[27] Muhammad Wasim, Waqar Mahmood, and Usman Ghani Khan. A survey of datasets for biomedical question answering systems. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, 8(7):484–488, 2017.

[28] Yuk Wah Wong and Raymond Mooney. Learning synchronous grammars for semantic parsing with lambda calculus. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 960–967, 2007.

[29] Yi Yang, Wen-tau Yih, and Christopher Meek. Wikiqa: A challenge dataset for open-domain question answering. In *EMNLP*, volume 1, pages 2013–2018, 2015.

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

[31] Luke S Zettlemoyer and Michael Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *arXiv preprint arXiv:1207.1420*,

2012.

[32] Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*, 2017.