A Mixed Information Source Approach for Biomedical Question Answering: MindLab at BioASQ 7B

Mónica Pineda-Vargas[1], Andrés Rosso-Mateus[1], Fabio A. González[1], and Manuel Montes-y-Gómez[2]

[1] Universidad Nacional de Colombia, Bogotá, Colombia
{mppinedav,aerossom,fagonzalezo}@unal.edu.co
[2] Laboratorio de Tecnologías del Lenguaje, INAOE, Puebla, México
{mmontesg}@inaoep.mx

Abstract. This paper describes the participation of the MindLab research group in the BioASQ 2019 Challenge for task 7b, document retrieval and snippet retrieval. For document retrieval, Elastic Search was used for the initial document retrieval step with BM25 as a scoring function. In the second stage, the top 100 retrieved documents were re-ranked with several strategies to exploit embedding semantic similarity. For the snippets retrieval subtask, the proposed approach was based on textual and conceptual information similarity patterns that were combined into a feature matrix that was subsequently processed by a convolutional neural network architecture. Our approach reached the third and second positions for the document retrieval and snippet retrieval task respectively.

Keywords: bioasq · snippet retrieval · biomedical document retrieval.

1 Introduction

In the biomedical domain, experts constantly search in previous works to support their research hypothesis, investigate causes, diseases symptoms, etc. The number of published documents is growing continuously, more than 3000 articles are indexed every day in biomedical journals [17], making it harder to find and access valuable information.

Question Answering (QA) systems can help to retrieve concise information naturally, given the precise answer and supporting passages for any information need. The interest in QA systems in the biomedical domain has been growing [1] [17] and is playing an important role in the closed domain information access and is considered to be the next step in information retrieval systems [19].

BioASQ is a closed domain information retrieval challenge over biomedical articles [17], this challenge has helped to advance the research in the biomedical information retrieval field. Mindlab team has participated in the last two editions. Here we will describe our second participation for the seventh edition in task B.

The goal of the target task is: given a question the system must return relevant concepts, relevant documents (from 2018 PubMed articles baseline [11]), relevant snippets (extracted from articles), and relevant Resource Description Framework (RDF) triples from designated ontologies [17]. In this year, our focus was document and snippet retrieval. Our method was based on a convolutional neural network model that takes as input a question-snippet similarity matrix. It combines different embeddings of words and medical concepts with the purpose of building a more meaningful representation.

The structure of this paper is as follows: Section 2 describes the system architecture, the strategies used for document retrieval are presented in Section 3 using Okapi-BM25 and Elastic Search as the first filter for efficiency and then representing documents and questions as word embeddings [10] for Doc Centroid Rerank and Word Mover's Distance as re-ranking functions. In Section 4, we present the passage retrieval module with the proposed method. Some performance analysis experiments were performed with the BioASQ6 data to evaluate and compare the proposed strategies for document and snippet retrieval, these results are shown in Section 5. The results of the current challenge are presented in section 6 and finally, conclusions and future works in Section 7.

2 The MindLab System At A Glance

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

The first module has the goal of producing a set of documents where the answer for a posed questions can reside. The Elastic Search (ES) information retrieval platform [6] was configured to index and query the PubMed Baseline Repository (MBR) document set [11]. This year the BioASQ challenge used the 2018 MBR. Those documents are bio-medical papers with title and abstract sections, also they contain meta-information such as MESH Terms, year of publication and keywords.

Based on a posed question, a query string is submitted to the ES search engine, the engine returns a set of 100 documents that are relevant to the query. The next step is a fine-grained document filtering using the query and document terms, the goal is to reduce the number of documents to 10. Our approach for document filtering is based on Word Mover's Distance (WMD) and Document Centroid semantic match.

The selected relevant documents are analyzed in depth. Snippets of the 10 most relevant documents are extracted and ranked with our Convolutional Neural Network (CNN) model that exploits semantic similarity patterns.

In the end, the top 10 scored documents and snippets are submitted in descending order to the BioASQ server.

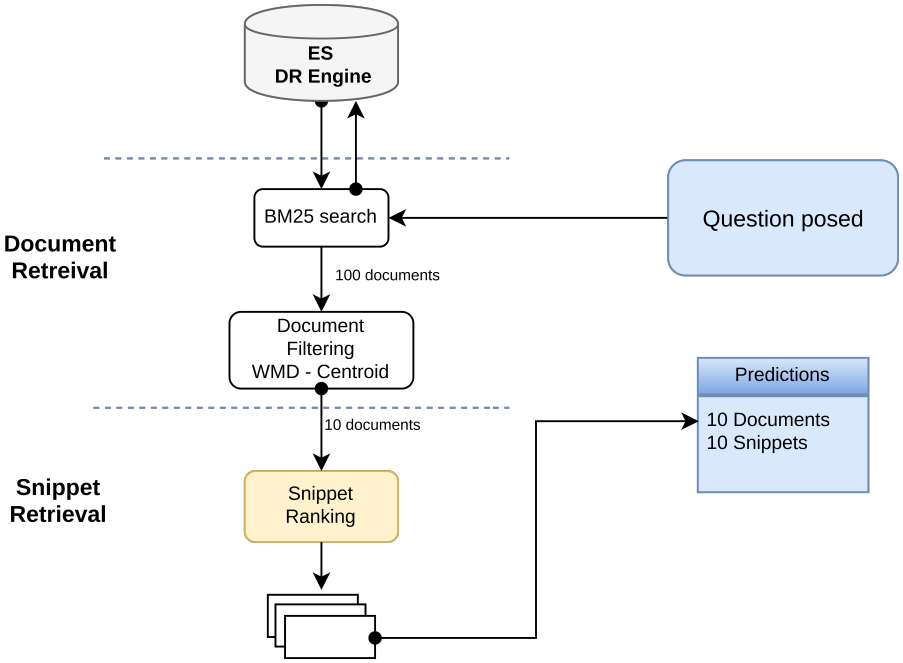

Fig. 1. BioASQ Model Diagram

2.1 Document re-ranking

The results produced by ES suffer from lack of precision. To alleviate this, we perform a re-ranking of the top-n documents using a more precise, but costly, semantic matching method based on semantic similarity. In our experiments, we evaluated two embedding similarity measures: Word Mover's Distance and Document Centroid.

Word Mover's Distance: The first was Word Mover's Distance (WMD) [8], a particular case of the Earth Mover's Distance [13]. The query and each document are represented as a weighted point cloud of embedded words as shown in Figure 2. The distance between them is the minimum cumulative distance that words from the query need to travel to match exactly the point cloud of the document.

Fig. 2. Word Mover's Distance between two documents.

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

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

Document centroid similarity: For a given query q and each document d we compute the centroid of the corresponding word vectors. The similarity of a query and a document corresponds to the cosine similarity between their corresponding centroids.

2.2 Snippet retrieval

Traditional passage retrieval methods use only textual information to identify the semantic match between the question and the answer. However, in domains such as biomedicine, there is a good number of structured knowledge resources in the form of ontologies, thesaurus, and taxonomies. The use of structured sources could provide advantages such as unambiguous knowledge representation, the possibility of applying automated reasoning methods [18], and the facility of linking different information facts, among others. The proposed approach for passage retrieval takes advantage of the huge amount of textual data in the biomedical domain in conjunction with structured-knowledge data sources.

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

The proposed method has two stages as Figure 3 shows. The first one (training phase) has the objective to learn the similarity patterns for question-answer

pairs. In the second stage (testing), the trained similarity model is used to obtain the ranking scores of a set of candidate answers (snippets) for a particular question. The method uses two representations schemes, textual and conceptual, for both answers and questions. Both representations are used as input independently. The representations are combined using a different fusion strategy.

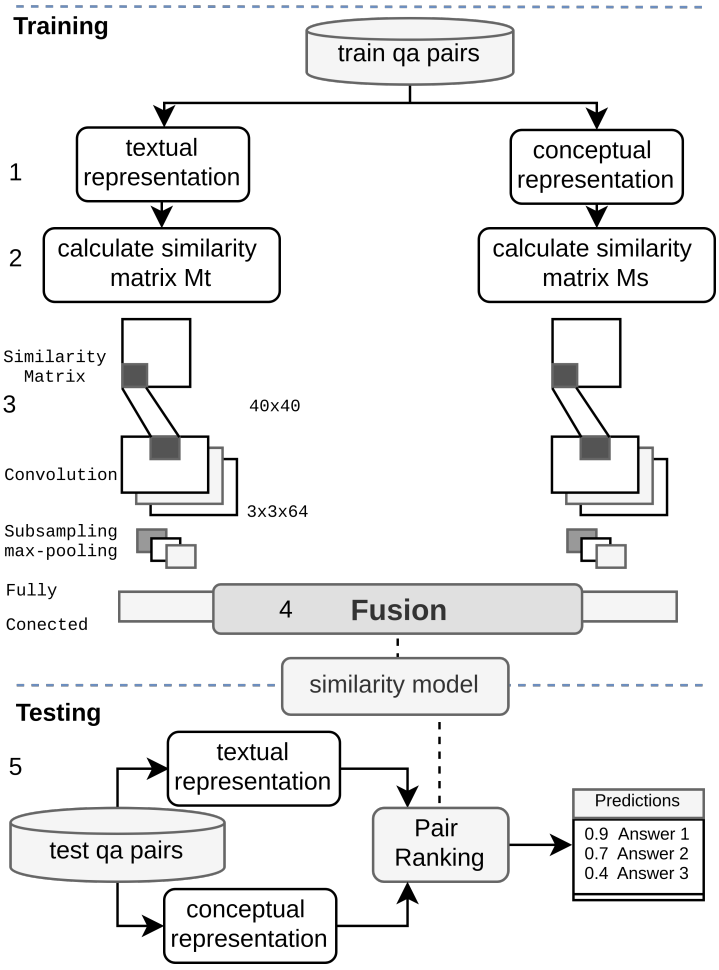

Fig. 3. Passage Retrieval Process

The details of the steps depicted in Figure 3 are discussed next.

– **Step 1 - Extract the representation**: The question and answer pairs are transformed to feed the neural network, the process is different for each modality.

- **Textual Representation**: First the text is cleaned and tokenized, a grammatical tagging is carried out with NLTK POS-tagger to extract syntactical information that will be used in salience weighting; each term is transformed later in a vector embedding using a pre-trained word2vec model provided by NLPLab, which is trained on Wikipedia and PubMed documents. [3]
- **Conceptual representation**: To identify medical concepts we use QuickUMLS [15] which is an unsupervised biomedical concept extraction tool. Those identified concepts are then transformed into a continuous vector representation using a cui2vec embedding. This embedding maps medical concepts instead of words. Concepts are referred by their concept unique identifier (CUI) from the Unified Medical Language System (UMLS) thesaurus [2]. In contrast with textual representation, there are fewer words, 4 concepts in average per question, in the text fragments that can be embedded in the conceptual representation as it is shown in Figure 4. This has to do with the reduced size of the cui2vec vocabulary. To overcome this restriction we applied expansion to question CUI embeddings following the centroid method proposed by Kuzi et al. [9].

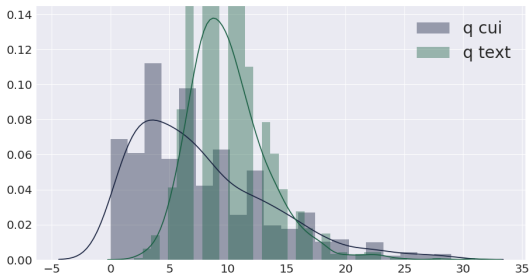

Fig. 4. Question Terms and Cuis distribution

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

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

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

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

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

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

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

- **Concept similarity matrix** M_s: In the case of conceptual information we calculate just the cosine similarity between cui2vec concept vectors.

– **Step 3. Convolutional model**: The architecture of the convolutional model is shown in Figure 3 step 3. A convolutional layer is fed with both similarity matrices M_t and M_s, CNN layer will identify element-similarity patterns to rank the relevance of a question-answer pair using both knowledge representations. Patterns identified by each CNN filter are sub-sampled by a pooling layer. The pooling layer for all the filters is merged with two fully connected layers, first one with 64 units and the second one with 16 units, a regularized dropout of 10% is used between them. The total number of parameters for the fusion model is 6,173.
– **Step 4. Multimodal fusion**: The dense outputs of the modalities are merged in a unique dense layer, which feeds another dense layer. Finally, the output score of the model is generated by a sigmoid unit on top of the last dense layer.
– **Step 5. Pair Ranking**: Candidate answers $(a_1, a_2, ..., a_k)$ are ranked against the query q using the trained similarity model. The model produces the final similarity score taking into account information from both modalities.

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

– Approach 1: Only textual representation. Questions and candidate answers are represented using only the textual embedding.

- Approach 2: Mixed data representation intermediate method –MIF. In this model, the fusion of textual and conceptual representations is carried out in an intermediate dense layer after the textual and conceptual patterns are identified by CNN's layers 3. The merged layer is then connected to the sigmoidal output unit with dropout as regularization strategy.
- Approach 3: Mixed Data Representation Late Fusion –MLF. In this approach each model (textual and conceptual) independently calculates a score for each question-answer pair, *score_t* for textual representation, and *score_s* for conceptual representation. Lastly, a linear combination produces the final score *f_score*, as shown in Equation 5. The *alpha* value was found using cross-validation with the validation partition; it was set to 0.73.

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

3 Model Performance Tuning

To evaluate the performance of the proposed methods, we used the BioASQ 6 dataset. The experimentation process is divided into two phases, the first one focused on the document retrieval process and the second one for snippets retrieval.

3.1 Document Retrieval

We indexed the full data of the 2018 MBR in ES version 6.2.2 with the default configuration, this is our baseline. For b and $k1$ BM25 parameters we evaluated the default values and the values proposed by CLEF eHealth evaluation lab 2016 [4].

The number of processed files was 928 and the total number of medical articles was 26,759,399. For each article, we extracted the title, MESH concepts and abstract to be indexed. The indexing time was around 18 hours in an Intel Xeon processor Intel(R) at 2.60GHz with 82 GB RAM and GeForce GTX TITAN X.

We evaluated four experiments: retrieve the 10 most relevant documents with *BM25 default configuration* (**BM25_v1**), the second one is to retrieve documents with *BM25 clef tuned parameters* (**BM25_v2**), the third is using *BM25 clef tuned parameters* and re-rank using Word Mover's Distance (**BM25_v2_WMD**). The last one is to retrieve documents with *BM25 clef tuned parameters* and re-rank using Doc Centroid Rerank (**BM25_v2_centroid**).

The averaged results over the five 6b batches are presented in Table 1. Adjusting BM25 parameters (b) for document length scaling and ($k1$) for document term scaling have a positive impact on the overall ES document retrieval performance. The centroid re-ranking strategy was not successful, but the use of Word Mover's Distance has a slightly better performance over the BM25 initial document set.

Table 1. Document Retrieval results for BioASQ 6 (summarized)

Model	Mean precision	Recall	F-Measure	MAP	GMAP
BM25_v1	0.2074	0.4724	0.2174	0.1319	0.0257
BM25_v2	0.2147	0.4810	0.2241	0.1354	0.0279
BM25_v2_WMD	0.2256	0.4910	0.2384	0.1483	0.0286
BM25_v2_centroid	0.2084	0.4706	0.2186	0.1332	0.0228

3.2 Snippet retrieval

The training was done with the question and answer pairs from 2016, 2017 and 2018 BioASQ Task B training datasets. The total number of used question-answer pairs were 124,144. The obtained dataset was very unbalanced, only 18% of the total number of pairs were labeled as a relevant answer. To balance the dataset, the sample extraction in the training phase was done with the same number of positive and negative samples, this strategy is also applied in the validation phase.

The model training was done using RMSprop optimization algorithm with 32 samples per mini-batch and the loss function was binary cross entropy. The number of maximum epochs was set to 50. In each epoch, we evaluated MAP and MRR, and after 5 epochs without any improvement in MAP metric, we applied early stopping to avoid over-fitting.

We have conducted our experimentation with the released batches for BioASQ 6, as it was done for the document retrieval stage. The scores obtained from BioASQ results submission page [17] are presented in Table 2. We have included the results from the winning team at BioASQ 6 challenge snippet retrieval subtask [3] for performance comparison.

Table 2. Snippet retrieval results averaged at BioASQ 6b

Method	Mean Precision	Mean Recall	F-measure	MAP	GMAP
aueb-nlp-5	0.3807	0.3655	0.3452	0.3320	0.0536
mindlab	0.2074	0.2437	0.2021	0.2102	0.0076
Only Textual	0.2074	0.2437	0.2021	0.2102	0.0161
MIF	**0.2181**	**0.2517**	**0.2161**	**0.2201**	**0.0098**
MLF	0.2014	0.2217	0.2100	0.2095	0.0086

The experimental evaluation shows that the incorporation of conceptual information can improve the performance in passage retrieval. Moreover, multimodal intermediate fusion outperforms the use of each modality individually and late fusion approaches.

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

- **Convolution parameters:** The number of convolutional filters used are 64, width 3 and length 3, the stride used is 1 without padding.
- **Convolution activation function:** After a convolutional layer, it is useful to apply a nonlinear layer [5]. We tested different activation functions and RELU gave us the best performance.
- **Pooling layers:** For the pooling layer, we used max pooling.
- **Dropout layer:** We add a dropout layer as a regularization strategy [16], setting the parameter in 10%.

4 Results at BioASQ 2019 and Discussion

4.1 Document retrieval

The results are shown in Table 3. It shows that our document retrieval implementation could still improve. In most batches, the top competitor gets approximately the double of our score.

The best result was obtained in batch 3 (Recall = 0.5213 and GMAP = 0.0070), with the tuned parameters for BM25, *Index-v2*. This improves our results obtained in the past competition [12], but the team leader in this batch reached 0.6128 in Recall and 0.0164 in GMAP, an important difference.

Document retrieval is relevant for snippet retrieval task, because it is the first information filter. The snippet retrieval method works on the top 10 documents retrieved in this phase, if it has low precision, the performance of the snippet retrieval method is affected.

Table 3. BioASQ 7 Document retrieval results

Batch	Model	Mean precision	Recall	F-Measure	MAP	GMAP
7b1	Our model	0.1120	0.5087	0.1660	0.0742	0.0039
	Top Competitor	0.1190	0.5216	0.1746	0.0809	0.0047
7b2	Our model	0.0950	0.4733	0.1444	0.0579	0.0021
	Top Competitor	0.1260	0.5967	0.1905	0.0771	0.0075
7b3	Our model	0.1280	0.5213	0.1887	0.0803	0.0070
	Top Competitor	0.3599	0.6128	0.4034	0.1102	0.0164
7b4	Our model	0.1040	0.5103	0.1573	0.0726	0.0033
	Top Competitor	0.3332	0.6141	0.3783	0.1015	0.0116
7b5	Our model	0.0550	0.3476	0.0888	0.0326	0.0005
	Top Competitor	0.0710	0.3937	0.1120	0.0425	0.0010

4.2 Snippet retrieval

As table 4 shows, the snippet retrieval approach obtained a good performance despite the low precision and recall in document retrieval. The results were enough to reach the second position in the last 4 batches and the first position in the first one.

It is important to highlight that the number of parameters to learn in our model is not large (5,192) compared to other QA Deep Learning approaches which are in order of millions and hundreds of thousands [14],[7].

Part of the error in the snippet retrieval phase is propagated from document retrieval. If we improve the results in document retrieval, we could have better results in this phase because we only retrieve the snippets present in the top 10 documents from the document retrieval phase. Even though, the incorporation of conceptual information can improve the performance in passage retrieval because the medical terms are referred by their concept unique identifier, which is unique for different words associated to the same concept.

Table 4. BioASQ 7 Snippet retrieval results

Batch	Model	Mean precision	Recall	F-Measure	MAP	GMAP
7b1	Our model	0.0951	0.2447	0.1253	0.0808	0.0008
	Top Competitor	-	-	-	-	-
7b2	Our model	0.0900	0.2243	0.1212	0.0893	0.0004
	Top Competitor	0.1447	0.3722	0.1855	0.1438	0.0019
7b3	Our model	0.1371	0.2519	0.1617	0.1404	0.0009
	Top Competitor	0.2159	0.3634	0.2472	0.2206	0.0081
7b4	Our model	0.1587	0.2760	0.1723	0.1527	0.0013
	Top Competitor	0.2060	0.4039	0.2365	0.2114	0.0075
7b5	Our model	0.0440	0.1823	0.0656	0.0499	0.0001
	Top Competitor	0.0542	0.2411	0.0818	0.0631	0.0003

5 Conclusion

In this paper we have presented the approaches and results obtained in our second participation for the seventh BioASQ challenge version. The proposed method for document retrieval was based on the Elastic Search platform with BM25 as scoring function, then a second document filtering was carried out using Word Mover's Distance (WMD) and Document Centroid semantic match.

For the snippet retrieval sub-task, the selected approach was based on a convolutional neural network that extracts similarity patterns over mixed data input representation. We have tested different fusion approaches to combine information coming from conceptual and textual sources.

The results obtained are promising for snippets retrieval, where we reach the second position despite the not very good performance of our approach for

document retrieval. This motivates our future work which will focus on improving the document retireval phase.

References

1. Michael A Bauer and Daniel Berleant. Usability survey of biomedical question answering systems. *Human Genomics*, 6(1):17, sep 2012.
2. Andrew L Beam, Benjamin Kompa, Inbar Fried, Nathan P Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. Clinical concept embeddings learned from massive sources of medical data. *arXiv preprint arXiv:1804.01486*, 2018.
3. 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.
4. Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Leif Hanlen, Aurélie Névéol, Cyril Grouin, João Palotti, and Guido Zuccon. Overview of the clef ehealth evaluation lab 2015. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 429–443. Springer, 2015.
5. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.
6. Clinton Gormley and Zachary Tong. *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*. " O'Reilly Media, Inc.", 2015.
7. Hua He and Jimmy J Lin. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *HLT-NAACL*, volume 1, pages 937–948, 2016.
8. Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.
9. Saar Kuzi, Anna Shtok, and Oren Kurland. Query expansion using word embeddings. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 1929–1932. ACM, 2016.
10. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. word2vec, 2014.
11. National Institutes of Health. Pubmed baseline repository.
12. Andrés Rosso-Mateus, Fabio A González, and Manuel Montes-y Gómez. Mindlab neural network approach at bioasq 6b. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, pages 40–46, 2018.
13. Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.
14. Aliaksei Severyn and Alessandro Moschitti. Learning to rank short text pairs with convolutional deep neural networks. In *38th ACM SIGIR*, 2015.
15. Luca Soldaini and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, 2016.
16. Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
17. 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.

18. William Tunstall-Pedoe. True knowledge: Open-domain question answering using structured knowledge and inference. *AI Magazine*, 31(3):80–92, 2010.
19. Lotfi A Zadeh. From search engines to question answering systems—the problems of world knowledge, relevance, deduction and precisiation. *Capturing Intelligence*, 1:163–210, 2006.