Learning Concept-Driven Document Embeddings for Medical Information Search

Gia-Hung Nguyen, Lynda Tamine, Laure Soulier, Nathalie Souf

Learning Concept-Driven Document Embeddings for Medical Information Search

Gia-Hung Nguyen[1], Lynda Tamine[1], Laure Soulier[2], and Nathalie Souf[1]

[1] Université de Toulouse, UPS-IRIT, 118 route de Narbonne, 31062 Toulouse, France
[2] Sorbonne Universités-UPMC, Univ Paris 06, LIP6 UMR 7606, 75005 Paris, France

Abstract. Many medical tasks such as self-diagnosis, health-care assessment, and clinical trial patient recruitment involve the usage of information access tools. A key underlying step to achieve such tasks is the document-to-document matching which mostly fails to bridge the gap identified between raw level representations of information in documents and high-level human interpretation. In this paper, we study how to optimize the document representation by leveraging neural-based approaches to capture latent representations built upon both validated medical concepts specified in an external resource as well as the used words. We experimentally show the effectiveness of our proposed model used as a support of two different medical search tasks, namely health search and clinical search for cohorts.

Keywords: Medical information search, representation learning, knowledge resource, medical concepts

1 Introduction

The importance of medical information access through a diversity of targeting tasks has attracted attention of many researchers from a variety of disciplines including health sciences, social psychology, and information retrieval (IR) [17]. More specifically, those tasks include evidence-based diagnosis, health-related search, and clinical trial patient recruitment; and beyond, diverse secondary tasks, such as population health management and translational research. Practically, from the IR area perspective, information access implies searching in large corpora of documents (e.g., electronic health records (EHRs), medical scientific reviews) for relevant information. The latter is retrieved by 1) matching user's queries, formulated through sets of keywords, with documents (e.g., search for diagnosis according to symptom description as a query input) and 2) matching documents with each other (e.g., identifying potential eligible patients for a clinical trial by matching their EHRs data to the textual description of the clinical trial requirements). However, numerous research studies have shown that such matching is complex, leading to system failure, mainly because of the gap between low-level document features and high-level meaning. This is referred to as the semantic gap [6]. In the medical domain, the semantic gap is prevalent and could be implied by three core issues: [6, 11]: 1) vocabulary mismatch: if the compared texts expressing similar word senses do not have overlapping keywords (e.g., *Melanome* vs. *skin cancer*); 2) granularity mismatch: if the compared texts contain instances of general entities (e.g., *Anti-inflammatory drug* vs. *Neodex*), and 3) logical implication: if the compared texts contain evidence allowing to infer implications that could not be automatically assessed (e.g., *anorexia* and *depression*). This problem has been faced so far by adopting two main approaches. The first one deals with the use of external domain knowledge resources, mainly to enhance text representations through document or query expansion [24, 5]. However, previous research has shown that even if using concepts from controlled vocabularies (such as UMLS) leading to meaningful representations of texts, using only the latter is significantly less effective than keyword-based retrieval, or a combination of both [24]. This may be explained by errors in the concept extraction or the limitations of the hand-labelled concept vocabulary expressed in knowledge resources. The second approach relies on dimensionality reduction techniques that attempt to reduce the representation size of the document vocabulary using the hypothesis of the distributional semantic [9]. Recent research trends show that one effective and efficient way for dimensionality reduction is based on neural language models. The latter projects words in a latent semantic space, called embedding [18], by learning their semantic relationships from their context. However, it is well-known today that such representations do not allow to capture the different meanings of words [10].

In this paper, we address the issue of the medical document representation which is a critical step in the matching process. To cope with the vocabulary and granularity mismatch issues mentioned above, we advocate for the use of a neural-based approach to capture latent representations of documents built upon manually validated medical concepts specified in an external resource. To overcome the limitations of the concept vocabulary and to capture additional distributional semantic extracted from corpora that would face the issue of logical implication assessment, we attempt to achieve the optimal document representation through a refinement using both concept-based and keyword-based raw representations as inputs. The key contributions of this work are two-fold: 1) we develop a model for learning and refining neural based representations of documents using semantics from a medical knowledge resource; 2) we assess the effectiveness of the proposed model by conducting an extensive evaluation using two different medical tasks within a major evaluation benchmark (TREC[3]): a) health-related search using a corpus of scientific reviews and b) clinical search for cohorts using a corpus of patient discharge summaries.

The paper is structured as follows. Related work is discussed in Section 2. The model for learning the concept-driven document embeddings is detailed in Section 3. Section 4 presents the experimental evaluation based on two medical search tasks. Section 5 concludes the paper and suggests avenues for future work.

[3] Text Retrieval Conference (http://trec.nist.gov/)

2 On the semantic gap problem in medical search

In the medical domain, the semantic gap problem is even more challenging for several reasons [24, 11]: high variability of language and spelling, frequent use of acronyms and abbreviations, and inherent ambiguity for automated processes to interpret concepts according to contexts. The semantic gap is one of the critical factors that likely leads to dramatically decrease the IR effectiveness. We detail below the two lines of works that cope with this problem.

Knowledge-based enhancement of documents and retrieval. Early and intensive work has been undertaken to use resources, such as MeSH, UMLS, and SNOMED, to enhance the semantics of texts, either documents or queries, by performing smoothing techniques including query expansion [1, 15, 21] and/or document expansion techniques [13, 8]. For instance, Lu et al. [15] investigate query expansion using MeSH to expand terms that are automatically mapped to the user query via the Pubmed's Automatic Term Mapping (ATM) service. In [8], authors combine both query expansion and document expansion using the MeSH thesaurus to retrieve medical records in the ImageCLEF 2008 collection. More concretely, for each MeSH concept, its synonyms and description are indexed as a single document in an index structure. The query is matched to the index to identify the best-ranked MeSH concepts. Finally, identified terms denoting MeSH concepts are used to expand both the document and the query. While knowledge-based document representations perform well in major benchmark evaluation campaigns [25], it is worth mentioning that they generally require a combination with keyword-based approaches [24]. The main underlying explanation is related to the limited expressiveness of concepts and/or the inaccuracy of the concept extraction method. More recently, an emerging line of work consists in enhancing the text-to-text matching model using evidence from external resources [16, 11, 26]. All of these contributions share the same goal: making inferences about the associations between raw data and the concept layer in the resource by building a relevance model. For instance, in [11], the relevance model is built upon a graphical probabilistic approach in which nodes are information units from both raw data and the knowledge resource. Another graph-based approach [27] aims at representing concept using a spectral decomposition within a electric resistance network. The extended query is obtained according to a resistance distance-based metric. To the best of our knowledge, this work is the first one to integrate lower dimensional representations for query expansion.

Representation learning of documents and concepts from a knowledge resource. Major approaches for unsupervised learning of the word representation from unlabelled data are the Skip-gram and CBOW models [18]. Basically, Skip-gram tries to predict the context of a given word, namely its collocated words, while jointly learning word representations. With the same objective of representation learning, CBOW rather relies on a prediction model of a word given its context. These models have been extended in several ways to represent documents [12], concepts of knowledge resources [20] as well as knowledge resources through concept-relation-concept triplets [2]. Beyond, several work focuses on the use of knowledge resources by leveraging concepts and their relations to updating (retrofitting) the latent representations of words [7, 28]. For instance, Faruqui et al. [7] propose a "retrofitting" technique consisting in a leveraging lexicon-derived relational information, namely adjacent words of concepts, to refine their associated word embeddings. The underlying intuition is that adjacent concepts in the knowledge resource should have similar embeddings while maintaining most of the semantic information in their pre-learned distributed word representations. In the medical domain, an increasing number of work attempts to learn concept representations [4, 19, 3, 14], with some of them [4, 19, 14] exploiting those representations within an information access task. Authors in [14] first extend the retrofitting of concepts proposed by [29] by weighting each word-to-word relation using its frequency in the corpus. Second, they 1) build the document representation by summing up the related word embeddings and then, 2) linearly combine relevance scores of documents obtained by matching the query with the bag-of-word representation and word embedding-based representation of documents. In [3], the authors propose the Med2vec model which leverages the sequence of patient visits to learn a latent representation of terminological concepts and visits using the Skip-gram model.

Unlike most of previous work [4, 19, 29, 14, 3] that learns contextless concept representations, our aim here is to jointly learn document representations by leveraging both the distributional semantic within text corpora and the concept word senses expressed in knowledge resources. In contrast to the closest previous work [3], we do not infer temporal dependencies between documents (namely visits) and concepts from a user point of view but rather address an information access task since our model leverages corpus-based semantics to cope with the problem of logical implication inference between words.

3 Model

3.1 Problem formulation

The literature review highlights that: 1) using knowledge resources allow to enhance raw text representations while using them solely gives rise to both vocabulary limitation and inaccuracy; 2) neural language models explicitly model semantic relations between words but they are unable to highlight diverse word meanings. In this paper, we address these shortcomings by conjecturing that: 1) incorporating concepts in the learning process of high-quality document representations, rather than word representations, should build knowledge-based semantic document representations that cope with the limitations of both the resource vocabulary and the concept extraction process; 2) for a targeted document, the optimal representation expected to be achieved in the low dimensional space requires the closeness of the two distinct embeddings built upon the knowledge-based and corpus-based representations.

With this in mind and in order to achieve the goal of enhancing medical document representations with the perspective of performing effective matching, we propose a method to incorporate a semantic medical knowledge into the learning of document embeddings while leveraging from raw text representations.

Given an optimal representation of documents highlighting activated concepts, a query expansion-based matching, for instance, could then be performed for retrieving relevant documents within an information search task.

Formally, document d is a set of two elements $d = \{\mathcal{W}_d, \mathcal{C}_d\}$, where $\mathcal{W}_d$ and $\mathcal{C}_d$ express respectively sets of ordered words w_i and ordered concepts c_j in document d, namely $\{w_1, ..., w_i, ..., w_n\}$ and $\{c_1, ..., c_j, ..., c_m\}$. Using the *Distributed Version of the Paragraph Vector* model (PV-DM) [12], document d is modeled through a word-based embedding vector $\hat{d}^{(PV-DM)}$. Our first objective is to build a concept-based embedding vector $\hat{d}_i^{(cd2vec)}$ that captures the explicit semantics expressed in knowledge resources. To do so, we propose the *cd2vec (conceptualDoc2vec)* model (Section 3.2). Assuming that the concept-based embeddings might suffer from limitations related to the vocabulary and the concept extraction process, our second objective is to find the optimal real-valued representation $\hat{d}$ of document d such that the knowledge-based embedding $\hat{d}_i^{(cd2vec)}$ and the corpus-based embedding $\hat{d}^{(PV-DM)}$ are nearby in the latent space. This problem could be formulated as the minimization of this objective function:

$$\Psi(D) = \sum_{d \in D} \psi(d) = \sum_{d \in D} \left[(1-\beta) \times \|d - \hat{d}^{(cd2vec)}\|^2 + \beta \times \|d - \hat{d}^{(PV-DM)}\|^2 \right] \tag{1}$$

where D is the document collection, $||x - y||$ the euclidean distance between x and y vector representations and β is a weighting factor experimentally tuned.

3.2 Learning the concept-based representation of documents

Inspired by the Distributed Memory model (PV-DM)[12] which learns short text representations using the set of ordered words within each text, we propose the *conceptualDoc2vec* model that rather focuses on a set of ordered concepts. The PV-DM model originally stands for paragraphs and attempts to consider them as context meanings that are jointly learned with the word vectors. Considering the problem addressed in this paper, our intuition is that document vectors can play the role of context meaning and contribute to a prediction task about the next concept given many concept-based contexts sampled from documents in the corpus. The *conceptualDoc2vec* architecture is illustrated in Figure 1. More particularly, document vectors $\hat{d}^{(cd2vec)}$ are learned so they allow predicting concepts in their context. More specifically, the goal of the *conceptualDoc2vec* is to maximize the following log-likelihood:

$$\varphi = \sum_{c_j \in \mathcal{C}_d} logP(c_j \mid c_{j-W} : c_{j+W}, d) \tag{2}$$

where c_j is j^{th} concept of ordered set $\mathcal{C}_d$. W is the length of the context window. $c_{(j-W)} : c_{(j+W)}$ represents the set of concepts ranged between the $(j-W)^{th}$ and the $(j+W)^{th}$ positions in document d, without concept c_j.

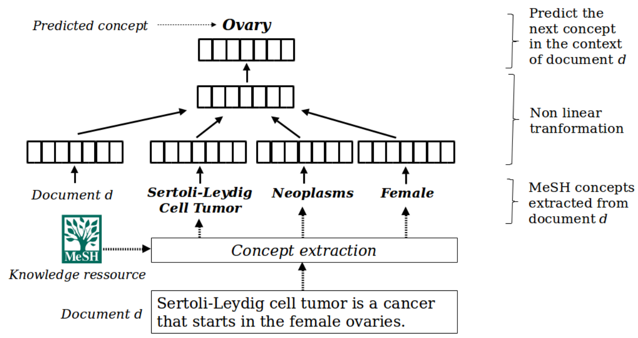

Fig. 1: Architecture of the *conceptualDoc2vec* model

The probability $P(c_j \mid c_{j-W} : c_{j+W}, d)$ is defined using a soft-max function:

$$P(c_j \mid c_{j-W} : c_{j+W}, d) = \frac{exp((\overline{v}_j^W)^\intercal . v_{c_j})}{\sum_{c_k \in \mathcal{C}_d} exp((\overline{v}_k^W)^\intercal . v_{c_k})} \tag{3}$$

where v_{c_j} is the representation of concept c_j, and $\overline{v}_j^W$ is the averaged representation of concepts in window $[j - W; j + W]$, including document d.

3.3 Solving the optimization problem

Our objective is to solve the optimization problem (Equation 1) that infers the optimal latent representation of document d to be semantically close to the knowledge-based and keyword-based latent document representations in the low dimensional space. To do so, we learn the optimal document embeddings using a stochastic gradient descent. More particularly, this method updates, for each document d its representation using the first derivative $\varDelta = \frac{\partial \psi(d)}{\partial d}$ of function ψ with respect to d with a step size of α, as illustrated in Algorithm 1.

1 Learning document representation using stochastic gradient descent

Input: $\hat{d}_i^{(PV-DM)}, \hat{d}_i^{(cd2v)}$
Output: d
$d = randomVector()$
$\psi(d) = (1 - \beta)\|d - \hat{d}^{(cd2v)}\|^2 + \beta\|d - \hat{d}^{(PV-DM)}\|^2$
while $\psi(d) > \epsilon$ **do**
$\varDelta = 2 \times (1 - \beta) \times (d - \hat{d}^{(c2vec)}) + 2 \times \beta \times (d - \hat{d}^{(PV-DM)})$
$d = d - \alpha \times \varDelta$
end while
return d

4 Experiments

4.1 Experimental Setup

Tasks. We evaluate our document representations on two medical search tasks:

- **Task1: health-related search**. This task refers to the situation where a physician seeks for relevant scientific articles providing with a fruitful assistance to achieve an accurate diagnosis/prognostic and/or to suggest a treatment considering the medical case. We use the standard OHSUMED collection consisting of a set of 348,566 references from MEDLINE and 63 queries. This dataset is known as a large-scale standard collection for ad-hoc medical IR [23]. An example of query is "*adult respiratory distress syndrome*".

• **Task2: clinical search for cohorts.** The task consists in identifying cohorts in clinical studies for comparative effectiveness research. We use the standard TREC Med collection in which queries specify particular disease/condition sets and particular treatments or interventions, expressed by physicians in a natural language form; this document collection includes over $17,000$ de-identified medical visit reports and 35 queries. An example of query is "*find patients with gastroesophageal reflux disease who had an upper endoscopy*".

Query expansion-based retrieval. We inject the optimized document embeddings d learned as detailed above in a text-to-text matching process according to a query expansion technique (noted $\mathbf{Exp_d}$). The latter enhances the input text with the top activated items (words or concepts) of the top-ranked documents within the closest embedding space (respectively, words or concepts depending on the β value). To do so, we extract a relevance score for each pairwise item/top-ranked document expressing the probability of item to express the semantics of document within its embedding space. Then, we use a CombSum merging technique applied on those scores to identify the top activated items.

Baselines. We use the state-of-the-art query expansion models:
• **Rocchio**, a query expansion model based on pseudo-relevance feedback [22].
• **LM-QE**, a language model applying a concept-based query expansion technique [21] in which candidate terms are ranked based on their similarity with descriptions in the knowledge resource. Default parameters are used.
To show the value of our enhanced document representation over the conceptualDoc2vec $\hat{d}^{cd2vec}$ and over the text-based document embedding $\hat{d}^{PV-DM}$, we run the query expansion method on these two embeddings:
• $\mathbf{Exp_{\hat{d}^{cd2vec}}}$ the concept-based embedding estimated without the optimization of the conceptualDoc2vec embeddings (see Section 3.2).
• $\mathbf{Exp_{\hat{d}^{PV-DM}}}$ the text-based embedding obtained through PV-DM [12].

Evaluation metrics. In order to measure the IR effectiveness, we use standard evaluation metrics, namely 1) *P@20* and *Recall@20* representing respectively the mean precision and recall values for the top 20 retrieved documents, and 2) *MAP (Mean Average Precision)* calculated over all queries. The average precision of a query is computed by averaging the precision values computed for each relevant retrieved document at rank $k \in (1..N)$, where $N = 1000$.

Table 1: Effectiveness of our Exp_d model on two medical search tasks.

	Health-related search			Clinical search for cohorts		
Model	MAP	P_20	Recall_20	MAP	P_20	Recall_20
LM-QE	0.0265	0.0686	0.0288	0.0793	0.1091	0.0519
Rocchio	0.0925	0.2262	0.0917	0.2096	0.2603	0.1701
$Exp_{\hat{d}^{PV-DM}}$	0.1017	0.2556	0.1086	0.3254	0.3971	0.2278
$Exp_{\hat{d}^{cd2vec}}$	0.0956	0.2365	0.0980	0.2255	0.2676	0.1319
Exp_d	0.1020	0.2556	0.1086	0.2996	0.3426	0.1989

Implementation details. We use the MeSH terminology which is mostly used in the biomedical domain [23] and the Cxtractor[4] tool to extract concepts. When learning the document representation,we set the minimum loss $\epsilon = 10^{-7}$ and the learning rate $\alpha = 0.01$. We set the size of word-based and concept-based embedding to 200. To tune the β parameter (Equation 1), we perform a cross-validation between both datasets and obtain optimal values in the training phase to 0.9 and 0.6 for respectively TREC Med and OHSUMED datasets. These values highlight that combining both words and concepts for representing documents is useful, with a higher prevalence of words for the TREC Med dataset. This could be explained by the fact that queries in this collection are more verbose.

4.2 Results

We present the performance of our model on two tasks: healt-related search and clinical search for cohort. Table 1 shows the retrieval effectiveness in terms of *MAP, P@20* and *Recall@20* for our embedding model Exp_d and the baselines as well. In general, we can observe that embedding-based expansion models ($Exp_{\hat{d}^{PV-DM}}$, $Exp_{\hat{d}^{cd2vec}}$, and Exp_d) allow to achieve better performance in both tasks than the two classic baselines, namely $LM-QE$ and *Rocchio* making use of raw-level representations of concepts and words respectively. For instance, the text-based embeddings-based expansion $\hat{d}^{PV-DM}$ achieves better results (MAP=0.2996) than the Rocchio-based expansion which obtains a MAP value equal to 0.2096. Especially, our embedding expansion model Exp_d reports significant better results in both tasks over all the three metrics compared with the $LM-QE$. These observations highlight the fact that the embedding models can improve the query expansion with help of learned latent semantics of words and/or concepts. Interestingly, by comparing the type of document embeddings used in our query expansion framework, we could outline that our optimized vector allows to slightly increase the IR effectiveness with respect to text-based embeddings $\hat{d}^{PV-DM}$ or the concept-based ones $\hat{d}^{cd2v}$. This result shows that our document representation allows to overpass, on one hand, the raw level ambiguity challenge within texts, and on the other hand, the limitation underlying the vocabulary and/or the concept extraction. Table 2 shows an illustration (query 131 of the TREC Med) in which our model Exp_d leveraging both evidences is able to find more relevant items for query expansion than

[4] https://sourceforge.net/projects/cxtractor/

Table 2: Example of terms/concepts expanded for query 131 in Trec Med

Query text	patients underwent minimally invasive abdominal surgery
Extracted Concepts	Patients; General Surgery;
Added by $Exp_{\hat{d}^{PV-DM}}$	myofascia; ultrasonix; overtube
Added by $Exp_{\hat{d}^{cd2vec}}$	Mesna; Esophageal Sphincter, Upper; Ganglioglioma
Added by Exp_{d}	*umbilical*; *ventral*; *biliary-dilatation*

other scenarios $Exp_{\hat{d}^{PV-DM}}$ and $Exp_{\hat{d}^{cd2vec}}$. More precisely, even if the high-level meaning of “abdominal surgery” is captured only by non-fined grained concept from MeSH (“General Surgery”), our model Exp_d is able to identify relevant candidate words for query expansion (“ventral”, “biliary dilatation”). Unlikely, the $Exp_{\hat{d}^{PV-DM}}$ model identifies less meaningful candidate words such as “myofascia”. This observation reinforces our intuition about the usefulness of combining latent representations of texts and concepts for achieving IR tasks. This result is consistent with previous works [24].

5 Conclusion and Future Work

We propose to tackle the semantic gap issue underlying medical information access. Our contribution investigates how to leverage semantics from raw text and knowledge resources to achieve high-level representations of documents. We propose an optimization function that achieves the optimal document representation based on both text embedding and concept-based embedding which relies on an extension of the PV-DM model. The overall obtained results reinforce the rationale of our proposed model. This work has some limitations. For instance, to identify the activated words and concepts reinjected in a query expansion method, we assume that the latent space built using our model is close to the initial embedding spaces. This could be addressed in the future by building a unified framework for learning the latent representations of documents. The latter would offer research opportunities to perform more accurate query expansion techniques fitting with other types of text mismatch faced in the medical domain.

References

1. S. Abdou and J. Savoy. Searching in MEDLINE: Query expansion and manual indexing evaluation. *Information Processing Management*, 44(2):781–789, 2008.
2. A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, 2013.
3. E. Choi, M. T. Bahadori, E. Searles, C. Coffey, and J. Sun. Multi-layer representation learning for medical concepts. *KDD*, pages 1495–1504, 2016.
4. L. De Vine, G. Zuccon, B. Koopman, L. Sitbon, and P. Bruza. Medical semantic similarity with a neural language model. In *CIKM*, pages 1819–1822, 2014.
5. D. Dinh and L. Tamine. Combining Global and Local Semantic Contexts for Improving Biomedical Information Retrieval. In *ECIR*, pages 375–386, 2011.
6. T. Edinger, A. Cohen, S.Bedrick, K. A. K, and W. H. W. Barriers to retrieving patient information from electronic health record data: Failure analysis from the trec medical records track. In *AMIA Annual Symposium*, pages 180–188, 2012.

7. M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith. Retrofitting word vectors to semantic lexicons. In *NAACL*, 2015.
8. J. Gobeill, P. Ruch, and X. Zhou. Query and document expansion with Medical Subject Headings terms at medical Imageclef 2008. In *CLEF*, pages 736–743, 2009.
9. T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.
10. I. Iacobacci, M. T. Pilehvar, and R. Navigli. Sensembed: Learning sense embeddings for word and relational similarity. In *ACL*, pages 95–105, 2015.
11. B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley. Information retrieval as semantic inference: A graph inference model applied to medical search. *Information Retrieval*, 19(1-2):6–37, Apr. 2016.
12. Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196, 2014.
13. T.-D. Le, J.-P. Chevallet, and T. B. T. Dong. Thesaurus-based query and document expansion in conceptual indexing with UMLS. In *RIVF'07*, pages 242–246, 2007.
14. X. Liu, J.-Y. Nie, and A. Sordoni. Constraining word embeddings by prior knowledge – application to medical information retrieval. In *AIRS*, 2016.
15. Z. Lu, W. Kim, and W. J. Wilbur. Evaluation of query expansion using MeSH in PubMed. *Information Retrieval*, 12(1):69–80, 2009.
16. J. Mao, K. Lu, X. Mu, and G. Li. Mining document, concept, and term associations for effective biomedical retrieval: Introducing mesh-enhanced retrieval models. *Information Retrieval*, 18(5):413–444, Oct. 2015.
17. C. Marton and C. W. Choo. A review of theroretical models on health information seeking on the web. *Journal of Documentation*, 68(3):330–352, 2012.
18. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
19. J. Minarro-Gimenez, O. Marin-Alonso, and M. Samwald. Exploring the application of deep learning techniques on medical text corpora. *Stud Health Technology Information*, 205:584–8, 2014.
20. Y. Ni, Q. K. Xu, F. Cao, Y. Mass, D. Sheinwald, H. J. Zhu, and S. S. Cao. Semantic documents relatedness using concept graph representation. In *WSDM*, 2016.
21. D. Pal, M. Mitra, and K. Datta. Improving query expansion using wordnet. *JASIST*, 65(12):2469–2478, 2014.
22. J. J. Rocchio. Relevance feedback in information retrieval. In *The SMART retrieval system*, pages 313–323, 1971.
23. N. Stokes, Y. Cavedon, and J. Zobel. Exploring criteria for succesful query expansion in the genomic domain. *Information retrieval*, 12:17–50, 2009.
24. D. Trieschnigg. *Proof of Concept: Concept-based Biomedical Information Retrieval*. PhD thesis, University of Twente, 2010.
25. E. Voorhees and W. Hersh. Overview of the trec medical records track. In *TREC*, 2012.
26. C. Wang and R. Akella. Concept-based relevance models for medical and semantic information retrieval. In *CIKM*, pages 173–182, 2015.
27. S. Wang and M. Hauskrecht. Effective query expansion with the resistance distance based term similarity metric. In *SIGIR*, pages 715–716, 2010.
28. C. Xu, Y. Bai, J. Bian, B. Gao, G. Wang, X. Liu, and T.-Y. Liu. Rc-net: A general framework for incorporating knowledge into word representations. In *CIKM*, 2014.
29. M. Yu and M. Dredze. Improving lexical embeddings with semantic knowledge. In *ACL*, pages 545–550, 2014.