

Neural IR for Domain-Specific Tasks

Discussion Paper

Óscar Espitia, Gabriella Pasi

University of Milano-Bicocca, Milan, Italy
Department of Informatics, Systems, and Communication (DISCo)
Information and Knowledge Representation, Retrieval, and Reasoning (IKR3) Lab
<https://ikr3.disco.unimib.it>

Abstract

Several specific features such as the volume of data, document size, structure of the documents, jargon, the way information needs are defined, among others, are features that justify that retrieval models should not handle all information equally when it comes to domain-specific retrieval tasks (e.g., in law and healthcare). Neural IR models can deal with such features, especially those related to contextual elements (e.g., expert knowledge). The ongoing project will consider domain-specific embeddings to contextualize within a neural ranking setup document retrieval in domain-specific tasks.

Keywords

Domain-specific search, Neural IR, Contextual IR, Embeddings

1. Introduction

Traditional information retrieval (IR) models are designed as generic tools applicable in different retrieval tasks. However, several specific features justify that retrieval models should not handle all information equally when it comes to domain-specific (DS) retrieval tasks (e.g., in law and healthcare) [1]. The volume of data, document size, structure of the documents, jargon, the way information needs are defined, among others, are features that define gaps between generic search tools and DS search tasks.

Neural IR is a growing field in IR; neural networks (NN) represent a tool for learning text representations, defining ranking models capable of handling different kinds of documents, and even introducing additional elements in the retrieval process, such as DS features.

Some of those features can be considered as contextual elements, i.e., as factors influencing how an IR system is used and how its performance should be accordingly evaluated. The concept of context in IR has been extensively studied, giving rise to the so called contextual IR, which aims at optimizing the retrieval performance by defining the search context and by taking it into account in the information selection process and in assessing the search outcome [2].

In last years, there have been developments in representing DS corpora; e.g., Legal-BERT [3] and BIO-BERT [4]; which perform well in representing texts that come from the legal and health domains, respectively. Even though these attempts are not strictly developed within


IIR 2021 – 11th Italian Information Retrieval Workshop, September 13–15, 2021, Bari, Italy

✉ o.espitiamendoza@campus.unimib.it (Ó. Espitia); gabriella.pasi@unimib.it (G. Pasi)

🆔 0000-0003-2725-2972 (Ó. Espitia); 0000-0002-6080-8170 (G. Pasi)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

contextual IR, using such representation models in retrieval frameworks fits into the ways of tailoring the retrieval process with contextual factors; in this case the context is represented by specialized corpora, which help to disambiguate terms and include DS jargon.

The main aim of the ongoing project is to define DS embeddings that can serve the purpose of contextualizing search within a neural ranking setup in DS tasks. More specifically, our aim is to evaluate the effectiveness of pre-trained embeddings with respect to traditional lexical matching-based models, and to boost the performance of neural models by fine tuning the DS embeddings in a semantic-based retrieval scenario.

2. Neural IR

A generalized Learning to Rank problem is focused on finding the optimal ranking function, according to existing neural IR models [5]; such a function could be abstracted by the formulation $f(q, d) = F(\Phi(q), \Phi(d))$, where q is a query and d is a document from a collection. Depending on whether the model is focused on defining F or Φ or both, one can get different anatomies of models. First, models focused on learning representations of the input texts. This kind of model considers as its inputs some basic representations such as one-hot encoding at character, term, n-graph level, with the aim of learning more complex and dense text representations or embeddings by using NN; in this case F is usually a similarity measure. In contrast, another kind of models assume the representation function Φ as the input layer of the model, so it can be either a simple or a complex representation to find interactions between the inputs, and from those interactions the model learns relevance patterns for ranking; thus, $F = I \circ G$ is composed both of the interaction function (I) and deep models for ranking (G).

3. DS tasks examples

This section shortly introduces two examples of DS search tasks on which we are working.

Patient- clinical trials matching: clinical trials are experiments conducted in the development of new medical treatments, drugs, or devices. Recruiting candidates for a trial motivates the task of matching eligible patients (q) to clinical trials (d) [6].

Legal case retrieval: this task involves finding precedent cases (d) that are relevant, i.e., could support the decision concerning a given case (q) in the set of candidate cases [7].

Compared to traditional ad-hoc text retrieval, the above tasks are relatively more challenging since the query is much longer and more complex than common keyword-based queries. Besides that, the definition of relevance of a document to a query is beyond general topical relevance, and as such its assessment requires expert knowledge.

4. Building blocks for designing DS neural search models

This section presents some strategies that have the potential to address DS search tasks and the building blocks for model design.

Several works have tried to learn text representations (Φ) that incorporate contextual elements into the model by learning embeddings from text corpora that gather information potentially

useful to better represent the information need in a given context [8, 9]. On the other hand, it is also known that there are several DS embeddings, such as legal-BERT, Bio-BERT, that were created with the motivation of having good representations of texts that are DS.

Given the representations of both a query and a document, one can match different parts of the query with different parts of the document and then aggregate values as partial evidence of relevance. Interactions-based approaches model this matching using an interaction matrix formed by sweeping both the query and the document representations with a sliding window that instantiates a function (I) for aggregating. Each instance of the window over the query interacts with each instance of the window over the document. This process could capture local interactions based on different matching levels (e.g., word-level and passage-level). Defining the input representations, the aggregation function, and the matching level affects the model's performance; then, all those parameters must be defined according to the specific task and its features.

Finally, there are different kinds of neural networks that are especially designed for some purposes that could match some of the issues described above and could be used for building a ranking model (G):

Pooling layers sweep a kernel (with no weights) across the entire input, similar to the convolutional layer. The kernel applies an aggregation function, which could either select the element with the maximum value to send to the output array (Max-pooling), or compute the average or the sum (Average-pooling and sum-pooling, respectively) within the area covered by the kernel while it moves across the input, leading to different architectures. Intuitively, pooling layers conduct dimensionality reduction, reducing the number of parameters in the input. This operation is useful to reduce complexity, improve efficiency, and limit the risk of overfitting.

Encoders process the input and compress the information into a context vector (also known as sentence embedding or vector) with fixed length. This representation is expected to be a good summary of the meaning of the whole input. Encoders are recurrent neural networks, i.e., LSTM or GRU units, which can model sequential data.

Attention mechanisms have become an integral part of sequence models in several tasks, allowing the model to learn dependencies without regard to their distance in the input or output sequences. Attention mechanisms are usually used in conjunction with recurrent networks but also as the core of the transformer architectures [10]. An attention function outputs a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of a given context vector with the values in the input. In this way, the attention mechanism is used to infer the importance of each position of the input.

5. Aim of the project

The ongoing project hypothesizes that the building blocks described above can be exploited to deal with the challenges of DS search. Then, we can design a neural ranker for DS tasks that leverages contextual information. The contextual information can be included by analysing DS features in search tasks and text representations that pay attention to the expert knowledge contained within DS corpora. Two retrieval scenarios are considered following a re-ranking technique: first, a lexical-based retrieval phase is followed by a re-ranking phase based on

semantic similarity; second, both the first phase and the re-ranking phase will be based on semantic search. These scenarios will allow us to explore different features of the models, for example, the value of exact matching and the effect of the vocabulary mismatch problem. On the other hand, different training strategies will be considered for text representation and text classification to build a more suitable semantic space for the specific task and to learn interactions between the inputs for re-ranking. This will allow us to evaluate if the tasks can benefit from fine-tuning text representations and if the learned representations are suitable for the first-stage retrieval and re-ranking.

Acknowledgments

This work is supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval (H2020-EU.1.3.1., ID: 860721).

References

- [1] A. Hanbury, M. Lupu, Toward a Model of Domain-Specific Search, OAIR '13: Open research Areas in Information Retrieval (2013).
- [2] Z. A. Merrouni, B. Frikh, B. Ouhbi, Toward Contextual Information Retrieval: A Review and Trends, *Procedia Computer Science* 148 (2019) 191–200.
- [3] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The Muppets straight out of Law School, *Findings of Empirical Methods in Natural Language Processing* (2020).
- [4] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* (2019).
- [5] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, X. Cheng, A Deep Look into neural ranking models for information retrieval, *Information Processing and Management* 57 (2020) 102067.
- [6] B. Koopman, G. Zuccon, A test collection for matching patients to clinical trials, *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2016).
- [7] J. Rabelo, M. Y. Kim, R. Goebel, M. Yoshioka, Y. Kano, K. Satoh, A Summary of the COLIEE 2019 Competition. In: Sakamoto M., Okazaki N., Mineshima K., Satoh K. (eds) *New Frontiers in Artificial Intelligence. JSAI-isAI 2019, Lecture Notes in Computer Science* 12331 LNAI (2020).
- [8] J. Yao, Z. Dou, J. R. Wen, Employing Personal Word Embeddings for Personalized Search, *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020).
- [9] Y. Zhou, Z. Dou, J. R. Wen, Encoding History with Context-aware Representation Learning for Personalized Search, *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020).
- [10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need *Ashish* (2017).