Knowledge Enhanced Representations to Reduce the Semantic Gap in Clinical Decision Support

Stefano Marchesin

Department of Information Engineering, University of Padua, Padova, Italy
stefano.marchesin@unipd.it

Abstract. The semantic gap between queries and documents is a long-standing problem in Information Retrieval (IR), and it poses a critical challenge for medical IR due to the large presence in the medical language of synonymous and polysemous words, along with context-specific expressions. Two main lines of work have emerged in the past years to tackle this issue: (i) the use of external knowledge resources to enhance query and document bag-of-words representations; and (ii) the use of semantic models, based on the distributional hypothesis, which perform matching on latent representations of documents and queries. The presented research investigates the use of external knowledge resources in both lines – with a focus on knowledge-enhanced unsupervised neural latent representations and their analysis in terms of effectiveness and semantic representativeness.

Keywords: Semantic gap · Query rewriting · Content representation

1 Motivations

Medical knowledge keeps growing exponentially. Clinicians struggle to keep up with every piece of new information and this can have a significant impact on patient care. To help clinicians in patient care, Clinical Decision Support (CDS) systems have emerged. Many tools exist for searching biomedical literature, but only a few specifically target the clinical environment. Because of that, the introduction of the TREC CDS track in 2014 triggered the creation of tools and resources necessary to evaluate Information Retrieval (IR) systems designed for CDS tasks. In 2017, the TREC Precision Medicine (PM) track succeeded to the CDS track and focused on an important use case in clinical decision support: providing useful precision medicine-related information to clinicians treating cancer patients.[1]

The outcomes of the TREC CDS and PM tracks showed that, within biomedical literature and clinical trials, the large presence of synonymous and polysemous words, along with the use of context-specific expressions, significantly impairs retrieval systems [18, 2]. Such characteristics increase the semantic gap, a long-standing studied topic in IR and Natural Language Processing (NLP).

[1] http://www.trec-cds.org/

The semantic gap can be interpreted as the difference between the low-level description of document and query contents and the high-level interpretation of their meanings.

Two main lines of work have emerged in the past years to tackle the semantic gap: (i) the use of external knowledge resources (e.g., UMLS,[2] SNOMED CT[3]) to enrich bag-of-words query and document representations; and, (ii) the use of semantic models, based on the distributional hypothesis, which perform matching on latent representations of documents and queries.

The presented research investigates the use of external knowledge resources in both lines – with a focus on knowledge-enhanced unsupervised neural latent representations. Furthermore, the research is part of the H2020 ExaMode project,[4] that has the objective of providing knowledge discovery for exascale medical data. This gives us the opportunity to design, develop and evaluate complementary approaches that can increase the understanding of the semantic gap and its relatedness with retrieval effectiveness in real case CDS scenarios.

The rest of the paper is organized as follows: Section 2 presents some background and related work; Section 3 describes the proposed research methodology and presents the obtained research results, Section 4 concludes the paper with some final remarks.

2 Background and Related Work

We describe in the following the main approaches used to tackle the semantic gap in IR. We can divide these approaches in two categories, which serve as guidelines throughout the paper: approaches that enrich bag-of-words query and document representations using external knowledge resources; approaches that perform semantic matching on latent representations of documents and queries.

Methods exploiting concepts and relations from external knowledge resources demonstrate consistent improvements over classic keyword-based systems. In [8], document and query term-based representations are shifted into concept-based representations derived from SNOMED CT. In [10], documents and queries are enhanced by using medical concepts directly relating to four aspects of the *medical decision criteria*. The work is extended in [11], where queries are expanded by inferring additional conceptual relationships from domain-specific resources as well as by extracting informative concepts from the top-ranked medical records. In [19], query reformulation techniques are proposed to address literature search based on case reports.

On the other hand, latent representation models have been used for decades in IR [5, 22]. The recent advancements in neural language models [15, 9] have led the IR community to consider them for retrieval tasks. Approaches that inject the low-dimensional text representations learned by neural models within state-of-the-art IR models have emerged [4, 6], along with approaches that learn

[2] https://www.nlm.nih.gov/research/umls/
[3] http://www.snomed.org/
[4] htttp://www.examode.eu/

representations of words and documents from scratch and use them directly for retrieval [21, 20]. However, distributed representations learned by neural language models are hampered by two main limitations: (i) polysemy [7] and (ii) synonymy [16]. Few approaches have been proposed in IR to address these problems. In [12], relational semantics are used to constrain word representations which are used in a document re-ranking scenario. In [17], latent representations are built upon concepts linked to knowledge resources and injected in a text-to-text matching process – according to a query expansion technique. In [16], a tri-partite neural language model is proposed that relies on external knowledge resources to constrain word, concept and document representations jointly. The model is then used for query expansion and in document re-ranking.

3 Research Methodology and Results

Our research is driven by the following research question:

> How can external knowledge be integrated in document/query representations in such a way that, given a medical query case, we can reduce the semantic gap between query and documents and effectively return related medical knowledge?

To answer this question, we started by proposing in [13] a retrieval framework for CDS based on document-level semantic networks. The proposed framework presents a two-step methodology, where the first step addresses the automatic creation of document-level semantic networks and the second exploits such document representations to retrieve relevant documents from medical literature.

This approach inherently addresses the semantic gap. Indeed, by providing a semantic-aware representation of documents and queries, by means of semantic networks composed of concepts and relations, the framework aims at reducing specific aspects of the semantic gap like, among the others, polysemy and synonymy. However, representing documents and queries as semantic networks, composed of concepts and relations linked to a reference Knowledge Base (KB), suffers from three main aspects. First, it requires concept and relation extraction algorithms to provide concepts and relations with a high level of accuracy, as the noise injected in the document-level semantic network creation step is propagated in the retrieval step too. Second, most state-of-the-art biomedical relation extraction techniques are developed for specific relationships, like protein-protein interactions, gene-disease interactions and so on – which cover only a fraction of the biomedical domain, not wide enough for a CDS setup. Third, the complexity of concept and relation extraction algorithms makes it difficult to scale them efficiently on IR collections – which are typically orders of magnitude larger than NLP collections.

Building on the work presented in Section 2, we started investigating alternative approaches to effectively integrate concepts and relations from external knowledge sources in the retrieval process.

Knowledge enhanced bag-of-words models for CDS. We are interested in understanding how, and to what extent, concepts and relations can be integrated within query/document representations and used to enhance the effectiveness of state-of-the-art retrieval models.

In [3], we investigated how semantic relations between concepts extracted from medical documents, and linked to a reference KB, can be employed to retrieve medical literature for CDS. We leveraged two methods for extracting relations from queries and documents: a rule-based method and a learning-based method. We found that relations – when pertinent to the information need – are highly valuable, outperforming the contribution provided by only concepts. The challenge lies in how to limit those cases where relations provide no relevant results.

We participated to the TREC PM 2018 track, focusing on the Clinical Trials task [1]. The aim of our work was twofold: (i) evaluate how a recall oriented approach based on an increasing (and more aggressive) query expansion method affects precision in this context; (ii) study whether the effectiveness of the retrieval approach can be correlated to the quality of the relations contained within the KB used for the query expansion process. The analysis of the results showed that aggressive query expansion approaches are detrimental for the retrieval effectiveness.

We deepened this analysis in [2], where we considered also the Scientific Literature task. We proposed and evaluated state-of-the-art query expansion and reduction techniques to identify whether a particular approach can be helpful in both scientific literature and clinical trials retrieval. The experimental analysis showed that no clear pattern emerges for both tasks. In general, a query expansion approach using KB concepts helps the retrieval of scientific literature, while a query reduction approach improves performances on the clinical trials task. Nevertheless, we found that a particular combination performs well in both tasks – in particular the clinical trials task – and competes with the top 10 performing runs in both TREC PM 2017 and 2018.

Currently, we are exploring the use of rank fusion approaches based on multiple query expansions. The objective is to build a robust fusion model, less sensitive to the problem of *topic drift* – which occurs when the query is expanded with concepts that are not pertinent to the information need.

Knowledge enhanced semantic models for CDS. We first proposed in [14] an IR framework that combines the implicit representations – obtained through distributional learning – and the explicit representations – derived from external knowledge sources – of documents to reduce the semantic gap for CDS retrieval tasks. Combining implicit-explicit representations aims at enriching the semantic understanding of documents and reducing the semantic gap between documents and queries. Indeed, distributional representations can capture the latent semantics existing between words relying only on the document collection as knowledge source. However, they are hampered by two main limitations that knowledge-based representations can alleviate: (i) distributional learning models fail to discriminate polysemous words [7]; and (ii) distributional learning models

fail to learn close representations for synonymous words occurring in different contexts [16]. Therefore, we are currently analyzing state-of-the-art neural representation models for IR tasks. An in-depth evaluation of their effectiveness for IR tasks, along with an analysis of their ability to retrieve documents that cannot be matched using lexical models, is fundamental to understand how neural representation models can be employed and combined effectively in IR. Besides, the comparison with traditional bag-of-words models can help identifying those components that are critical for every IR model.

Thus, based on [14] and on the analysis we are conducting on neural representation models, we are developing an unsupervised neural model for learning knowledge-enhanced latent representations of words, concepts and documents. To reduce prominent aspects of the semantic gap, the model integrates relational semantics from external knowledge sources in the learning process.

4 Final Remarks

The presented research has the final objective of reducing the semantic gap and increasing retrieval effectiveness in real-case CDS scenarios. Therefore, as a subsequent step, we will design and develop knowledge enhanced multi-stage retrieval systems – which combine bag-of-words and latent representations. The idea is to leverage the complementary nature of bag-of-words and latent representations to the best, thus advancing – both methodologically and experimentally – real-case CDS applications. The results will be then validated within the context of the ExaMode project, which allows us to collaborate with hospitals and health practicioners.

Acknowledgements

The work was supported by the ExaMode project, as part of the European Union H2020 program under Grant Agreement no. 825292.

References

1. Agosti, M., Di Nunzio, G.M., Marchesin, S.: The University of Padua IMS Research Group at TREC 2018 Precision Medicine Track. In: Proc. of the 27th TREC (2018)
2. Agosti, M., Di Nunzio, G.M., Marchesin, S.: An Analysis of Query Reformulation Techniques for Precision Medicine. In: Proc. of the 42nd SIGIR (in print). ACM (2019)
3. Agosti, M., Di Nunzio, G.M., Marchesin, S., Silvello, G.: Medical Retrieval using Structured Information Extracted from Knowledge Bases. In: Proc. of the 27th SEBD (in print) (2019)
4. Ai, Q., Yang, L., Guo, J., Croft, W.B.: Improving Language Estimation with the Paragraph Vector Model for Ad-Hoc Retrieval. In: Proc. of the 39th AMC SIGIR. pp. 869–872. ACM (2016)
5. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by Latent Semantic Analysis. JASIST **41**(6), 391–407 (1990)

6. Ganguly, D., Roy, D., Mitra, M., Jones, G.J.: Word Embedding based Generalized Language Model for Information Retrieval. In: Proc. of the 38th SIGIR. pp. 795–798. ACM (2015)
7. Iacobacci, I., Pilehvar, M.T., Navigli, R.: Sensembed: Learning Sense Embeddings for Word and Relational Similarity. In: Proc. of the 53rd ACL and the 7th IJCNLP. vol. 1, pp. 95–105 (2015)
8. Koopman, B., Zuccon, G., Nguyen, A., Vickers, D., Butt, L., Bruza, P.D.: Exploiting SNOMED CT Concepts and Relationships for Clinical Information Retrieval: Australian e-Health Research Centre and Queensland University of Technology at the TREC 2012 Medical Track. In: Proc. of the 21st TREC. pp. 1–8 (2012)
9. Le, Q., Mikolov, T.: Distributed Representations of Sentences and Documents. In: ICML. pp. 1188–1196 (2014)
10. Limsopatham, N., Macdonald, C., Ounis, I.: A Task-Specific Query and Document Representation for Medical Records Search. In: Proc. of ECIR 2013. pp. 747–751. Springer (2013)
11. Limsopatham, N., Macdonald, C., Ounis, I.: Inferring Conceptual Relationships to Improve Medical Records Search. In: Proc. of the 10th OAIR. pp. 1–8 (2013)
12. Liu, X., Nie, J.Y., Sordoni, A.: Constraining Word Embeddings by Prior Knowledge–Application to Medical Information Retrieval. In: AIRS. pp. 155–167. Springer (2016)
13. Marchesin, S.: Case-Based Retrieval Using Document-Level Semantic Networks. In: Proc. of the 41st SIGIR. p. 1451. ACM (2018)
14. Marchesin, S.: Implicit-Explicit Representations for Case-Based Retrieval. In: Proc. of DESIRES 2018 (2018)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In: Proc. of NIPS 2013. pp. 3111–3119 (2013)
16. Nguyen, G.H., Tamine, L., Soulier, L., Souf, N.: A Tri-Partite Neural Document Language Model for Semantic Information Retrieval. In: Proc. of ESWC 2018. pp. 445–461. Springer (2018)
17. Nguyen, G.H., Tamine, L., Soulier, L., Souf, N.: Learning Concept-Driven Document Embeddings for Medical Information Search. In: Proc. of AIME 2017. pp. 160–170. Springer (2017)
18. Roberts, K., Simpson, M., Demner-Fushman, D., Voorhees, E., Hersh, W.: State-of-the-Art in Biomedical Literature Retrieval for Clinical Cases: a Survey of the TREC 2014 CDS track. IRJ **19**(1-2), 113–148 (2016)
19. Soldaini, L., Cohan, A., Yates, A., Goharian, N., Frieder, O.: Retrieving Medical Literature for Clinical Decision Support. In: Proc. of ECIR 2015. pp. 538–549. Springer (2015)
20. Van Gysel, C., de Rijke, M., Kanoulas, E.: Neural Vector Spaces for Unsupervised Information Retrieval. ACM TOIS **36**(4), 38 (2018)
21. Vulić, I., Moens, M.F.: Monolingual and Cross-Lingual Information Retrieval Models based on (Bilingual) Word Embeddings. In: Proc. of the 38th SIGIR. pp. 363–372. ACM (2015)
22. Wei, X., Croft, W.B.: LDA-based Document Models for Ad-Hoc Retrieval. In: Proc. of the 29th SIGIR. pp. 178–185. ACM (2006)