Knowledge Enhanced Representations for Clinical Decision Support

Stefano Marchesin and Maristella Agosti
Department of Information Engineering, University of Padua
Padua, Italy
{stefano.marchesin,maristella.agosti}@unipd.it

ABSTRACT

The study presents a methodology that contributes to reduce the semantic gap in clinical decision support systems. The methodology integrates semantic information – provided by external knowledge resources – into unsupervised neural Information Retrieval (IR) models. The objective is to design and develop innovative methods that can be effective in real-case medical scenarios.

KEYWORDS

Semantic gap, neural IR, clinical decision support

1 INTRODUCTION AND RELATED WORK

Clinicians struggle to keep up with the pace at which medical literature is growing. This factor has given rise to Clinical Decision Support (CDS) systems. CDS systems are designed to assist clinicians in providing patient care – e.g. formulate diagnoses, decide treatments, etc. Among the different tasks that CDS systems perform, biomedical literature search is pivotal. However, very few existing tools specifically target the clinical environment. To foster their growth, the *Text REtrieval Conference (TREC)*, in 2014, introduced a CDS track. The TREC CDS track triggered the creation of tools and resources to evaluate Information Retrieval (IR) systems designed for CDS tasks. In 2017, TREC Precision Medicine (PM) became a successor to the CDS track. Focused on an important use case in clinical decision support, it provides useful precision medicine-related information to clinicians treating cancer patients.[1] TREC CDS and PM tracks highlight that the large presence of synonyms and polysemous words found in biomedical literature and clinical trials – along with the use of context-specific expressions – significantly reduces the effectiveness of retrieval systems [1, 10]. Such features of medical language increase the semantic gap, representing a long-standing problem in IR and Natural Language Processing (NLP). In IR, the semantic gap reflects the difference between low-level description of document and query contents and high-level interpretation of meanings.

To bridge the semantic gap, semantic models have long been used in IR. Recent advances in neural language models [7] have led the IR community to adopt them for retrieval tasks. Approaches that inject low-dimensional text representations learned by neural models within state-of-the-art IR models have emerged [2], along with approaches that learn representations of words and documents from scratch and use them directly for retrieval [11]. However, distributed representations learned by neural language models suffer by two main limitations: (i) polysemy [3] and (ii) synonymy [9]. Few approaches have been proposed in IR to address these problems. In [4], relational semantics are used to constrain word representations applied in a document re-ranking scenario. In [8], 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 [9], a tripartite neural language model is proposed that relies on external knowledge resources to jointly constrain word, concept and document representations. The model is then used for query expansion and document re-ranking.

The methodology proposed to address the semantic gap in CDS is shortly presented in the following and it is part of the H2020 ExaMode project,[2] whose objective is to provide knowledge discovery for exascale medical data. In Section 2, we present the structure and contents of the histopathology clinical reports used in the ExaMode project, while in Section 3 we introduce the methodology that can help reduce the semantic gap in CDS tasks.

2 CLINICAL REPORT CONTENTS

A histopathology report is a document that is written and signed by a pathologist. It contains the results of the analyses performed on specific tissues or cells to obtain a pathological-clinical diagnosis that can lead to appropriate treatment options in case of disease. To provide a general structure for pathology reports, the College of American Pathologists (CAP) has set a series of guidelines.[3] The structure and contents of pathology reports are described below.

Patient Identifier and Clinical Information: contains the patient's identifier and specific information such as name, date of birth, hospital and medical record number. Clinical information provides details such as symptoms, medical conditions or data about the target specimen. The source of the specimen sample is also provided. Additionally, the pathologist's name and signature, along with the laboratory name and address, are also specified in this section.

Macroscopic Description: describes how a specimen appears to the naked eye and details what portions of the target specimen are examined under the microscope. The description includes the size, colour, number of tissue samples and, when appropriate, the weight of the specimen. In the presence of multiple tissues or organs within the specimen, each one is described and sampled. Each sample produces a microscope slide that is listed in the pathology report.

Microscopic Description: describes how the specimen looks under the microscope compared to normal cells. It also describes

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

[2] https://www.examode.eu/

[3] https://www.cap.org/cancerprotocols/

IIR 2019, September 16–18, 2019, Padova, Italy

whether the cancer has invaded nearby tissues. All the information derived from microscopic descriptions can help provide guidelines for treatments.

Diagnosis: has to do with the final pathology diagnosis issued by the pathologist following specimen examination. Cancer diagnoses describe multiple aspects associated to the specific tumour(s). For most of these diagnoses, the grade of the tumour – determined by applying tumour-specific criteria to the microscopic features – is included.

Comments: contains additional information used by the pathologist to describe challenging cases. This section may contain additional data such as images, molecular studies, references and addendum information, useful to the care team.

The methodology we propose will be evaluated on real-case medical scenarios, where clinical reports follow the above-mentioned structure. Moreover, clinical reports provided by ExaMode consider different use-cases and are written in multiple languages. Methods capable of effectively representing the underlying medical data are fundamental to reduce the semantic gap and retrieve relevant medical data that support the pathologists in their decision making.

3 METHODOLOGY

This study explores how the semantic information contained within external knowledge resources can be integrated into unsupervised neural IR models. This offers the opportunity to design, develop and evaluate novel approaches that increase the understanding of the semantic gap and its relation with retrieval effectiveness in the real-case CDS scenarios provided by the ExaMode project.

Our research is driven by the following research question:

> How can external knowledge be integrated in document/query representations so that, given a query clinical report, the semantic gap between the query and the documents can be reduced to effectively retrieve medical knowledge?

We first addressed the research question in [5], proposing a retrieval approach for CDS based on document-level semantic networks, comprising of two steps: (i) automatic creation of document-level semantic networks, (ii) retrieval of relevant medical knowledge using document-level semantic networks. The approach provides a semantic-aware representation of documents and queries by means of semantic networks embodying semantic concepts and relations. Its aim is to reduce the semantic gap, especially considering aspects of polysemy and synonymy. Nevertheless, the representation of documents and queries as semantic networks, derived from a reference Knowledge Base (KB), has three main limitations. Firstly, it requires concept and relation extraction algorithms to achieve a high level of accuracy, since the noise in creating a document-level semantic network is likely to propagate even in the retrieval step. Secondly, most state-of-the-art techniques to extract biomedical relations are developed to detect specific relationships like protein-protein interactions, gene-disease interactions and so on. They, however, cover only a fraction of the biomedical domain which is not wide enough for CDS. Thirdly, the complexity of concept and relation extraction algorithms makes it difficult to scale them efficiently on IR collections – typically being orders of magnitude larger than NLP collections. Despite preserving the initial idea, we started investigating alternative approaches to effectively integrate concepts and relations from external knowledge sources into the retrieval process. In [6], we proposed an IR framework that combines implicit and explicit representations of documents to reduce the semantic gap for CDS. Implicit representations are obtained through distributional learning, whereas explicit representations are derived from external knowledge sources. The combination of these representations has the aim of enriching the semantic understanding of documents – which in turns reduces the semantic gap between documents and queries.

Implicit representations can thus capture the latent semantics existing between words (and documents) relying only on the document collection as knowledge source. However, such representations are hindered by two main limitations that knowledge-based representations can reduce: (i) distributional learning models fail to discriminate polysemous words [3]; and (ii) distributional learning models fail to learn close representations for synonymous words occurring in different contexts [9]. Therefore, we are currently performing an evaluation of state-of-the-art neural representation models for IR. An in-depth evaluation of their effectiveness is fundamental to understand how neural representation models can be combined effectively with external knowledge sources, so as to reduce the semantic gap and increase retrieval effectiveness. Based on [6] and on this analysis, we are developing an unsupervised neural model to learn 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.

ACKNOWLEDGMENTS

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

REFERENCES

[1] M. Agosti, G. M. Di Nunzio, and S. Marchesin. 2019. An Analysis of Query Reformulation Techniques for Precision Medicine. In *Proc. of the 42nd ACM SIGIR (in print)*. ACM.

[2] Q. Ai, L. Yang, J. Guo, and W. B. Croft. 2016. Improving Language Estimation with the Paragraph Vector Model for Ad-Hoc Retrieval. In *Proc. of the 39th ACM SIGIR*. ACM, 869–872.

[3] I. Iacobacci, M. T. Pilehvar, and R. Navigli. 2015. SENSEMBED: Learning Sense Embeddings for Word and Relational Similarity. In *Proc. of the 53rd ACL and the 7th IJCNLP*, Vol. 1. 95–105.

[4] X. Liu, J. Y. Nie, and A. Sordoni. 2016. Constraining Word Embeddings by Prior Knowledge–Application to Medical Information Retrieval. In *AIRS*. Springer, 155–167.

[5] S. Marchesin. 2018. Case-Based Retrieval Using Document-Level Semantic Networks. In *Proc. of the 41st ACM SIGIR*. ACM, 1451.

[6] S. Marchesin. 2018. Implicit-Explicit Representations for Case-Based Retrieval. In *Proc. of DESIRES 2018*.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. of NIPS 2013*. 3111–3119.

[8] G. H. Nguyen, L. Tamine, L. Soulier, and N. Souf. 2017. Learning Concept-Driven Document Embeddings for Medical Information Search. In *Proc. of AIME 2017*. Springer, 160–170.

[9] G. H. Nguyen, L. Tamine, L. Soulier, and N. Souf. 2018. A Tri-Partite Neural Document Language Model for Semantic Information Retrieval. In *Proc. of ESWC 2018*. Springer, 445–461.

[10] K. Roberts, M. Simpson, D. Demner-Fushman, E. Voorhees, and W. Hersh. 2016. State-of-the-Art in Biomedical Literature Retrieval for Clinical Cases: a Survey of the TREC 2014 CDS track. *IRJ* 19, 1-2 (2016), 113–148.

[11] C. Van Gysel, M. de Rijke, and E. Kanoulas. 2018. Neural Vector Spaces for Unsupervised Information Retrieval. *ACM TOIS* 36, 4 (2018), 38.