Implicit-Explicit Representations for Case-Based Retrieval

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

ABSTRACT

We propose an IR framework to combine the implicit representations — identified using distributional representation techniques — and the explicit representations — derived from external knowledge sources — of documents to improve medical case-based retrieval. Combining implicit-explicit representations of documents aims at enriching the semantic understanding of documents and reducing the semantic gap between documents and queries.

CCS CONCEPTS

• **Information systems → Document representation**; **Query representation**; **Information extraction**;

KEYWORDS

Relation-based information retrieval; Semantic gap

1 MOTIVATIONS AND METHODOLOGY

Medical literature published every year keeps growing drastically. Clinicians have limited time to retrieve relevant information from medical literature. Standard Information Retrieval (IR) systems are not able to cope with the amount of literature and the limited time available to clinicians. Therefore, there has been a strong interest for Clinical Decision Support (CDS) systems designed to produce effective and timely knowledge, that can help clinicians in the decision making process. Such systems are known as case-based retrieval systems. Given a medical case of interest, a case-based retrieval system should retrieve highly related medical literature from a large collection of medical literature.

A key characteristic of the medical literature is the large use of synonyms and context-specific expressions. Such characteristics increase the semantic gap between documents and queries.

To tackle the problems above, both deep representation learning methods and external knowledge sources have been used. Deep representation learning effectively discovers hidden structures that relate — through latent semantic features — the different textual components, be them words, sentences or documents [3]. External knowledge resources, such as ontologies and knowledge bases, provide factual knowledge about the meaning of words and their semantic relationships.

However, by formalizing the semantic relationships between different concepts, knowledge sources represent a partial (and specific) representation of the world. Therefore, knowledge sources do not necessarily represent implicit relations that appear in documents. By learning distributional representations of words and phrases, implicit relations within documents can be considered too. Thus, distributional and knowledge-based representations of complex semantics (i.e. words, sentences and documents) identify complementary semantic aspects of the underlying documents.

We propose to integrate documents' knowledge-based representations in case-based retrieval [2], as a form of complementary refinement for distributional representations. Recent approaches exploit semantic relations to enhance the quality of learned word or concept representations [1, 4], we propose to explicitly leverage semantic relations to model document representations. Therefore, our approach extracts concepts from documents (and queries) and connect them using the semantic relations contained within a reference knowledge source — creating a knowledge graph representation for the document (or query). The intuition is that semantic relations carry high informative power that can boost precision.

The knowledge graph representation can reduce the contextual dependency of distributional representations and help discriminating more effectively semantically similar from non-semantically similar texts. Besides, since the concepts considered are only those extracted from a document or a query, the problem of *topic drift* — occurring when the query is expanded with concepts that are not pertinent to the information need — is reduced.

We propose to combine implicit and explicit representations for case-based retrieval in two different ways: (i) considering document-level knowledge graphs as additional inputs for end-to-end neural scoring models that learn the relevance of document-query pairs via semantic features; (ii) considering document-level knowledge graphs with pseudo relevance feedback to boost documents in top positions that present a more similar graph compared with the query graph. Both approaches aim at reducing the semantic gap between queries and documents.

ACKNOWLEDGMENTS

Supported by the CDC-STARS project of the University of Padua.

REFERENCES

[1] Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2014. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166* (2014).
[2] Stefano Marchesin. 2018. Case-Based Retrieval Using Document-Level Semantic Networks. In *41st ACM SIGIR (SIGIR '18)*. ACM, New York, NY, USA, 1451.
[3] Bhaskar Mitra and Nick Craswell. 2017. Neural Models for Information Retrieval. *arXiv preprint arXiv:1705.01509* (2017).
[4] Gia-Hung Nguyen, Lynda Tamine, Laure Soulier, and Nathalie Souf. 2018. A Tri-Partite Neural Document Language Model for Semantic Information Retrieval. In *European Semantic Web Conference*. Springer, 445–461.

DESIRES 2018, August 2018, Bertinoro, Italy