SAFIR: a Semantic-Aware Neural Framework for IR

Discussion Paper

Maristella Agosti, Stefano Marchesin and Gianmaria Silvello

Department of Information Engineering, University of Padua, Via Giovanni Gradenigo 6/b, 35131, Padova, Italy

Abstract

The semantic mismatch between query and document terms – i.e., the semantic gap – is a long-standing problem in Information Retrieval (IR). Two main linguistic features related to the semantic gap that can be exploited to improve retrieval are synonymy and polysemy. Recent works integrate knowledge from curated external resources into the learning process of neural language models to reduce the effect of the semantic gap. However, these knowledge-enhanced language models have been used in IR mostly for re-ranking. We propose the Semantic-Aware Neural Framework for IR (SAFIR), an unsupervised knowledge-enhanced neural framework explicitly tailored for IR. SAFIR jointly learns word, concept, and document representations from scratch. The learned representations encode both polysemy and synonymy to address the semantic gap. We investigate SAFIR application in the medical domain, where the semantic gap is prominent and there are many specialized and manually curated knowledge resources. The evaluation on shared test collections for medical retrieval shows the effectiveness of SAFIR to address the semantic gap.

Keywords

Knowledge-enhanced retrieval, representation learning, semantic gap, medical literature

1. Introduction

This paper addresses the semantic gap, a long-standing problem in Information Retrieval (IR). The semantic gap refers to the difference between the machine-level description of document/query contents and the human-level interpretation of their meanings [2]. It can be described also as the mismatch between users' queries and the way retrieval models answer to such queries. We focus on two linguistic features related to the semantic gap: synonymy and polysemy. Synonymy occurs when different words convey the same meaning, whereas polysemy occurs when the same word has different meanings depending on the context.

In the past years, two main lines of work have emerged to bridge the semantic gap between queries and documents: (i) the use of external knowledge resources to enhance query and document bag-of-words representations, and (ii) the use of semantic models to perform matching between the latent representations of queries and documents. Semantic models, which are based on the Distributional Hypothesis, have been revived by the advent of neural language models [3]. Neural language models learn distributed representations of words, also known as word embeddings, based on the context surrounding words. However, their learning process

The full paper has been originally published in ACM Transactions on Information Systems [1]

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

stefano.marchesin@unipd.it (S. Marchesin)

relies exclusively on text corpora and does not consider any external resources, which encode factual knowledge that can help to reduce the semantic gap.

To this end, recent works integrate external knowledge into the learning process of neural language models to reduce the effect of the semantic gap between queries and documents [4, 5]. However, even though knowledge-enhanced language models have been proven effective in many Natural Language Processing (NLP) tasks, their effectiveness is limited in IR [5]. We identify two reasons causing this performance gap. First, knowledge-enhanced language models have been used in IR mostly for re-ranking [4, 5]. Secondly, IR tasks are different from NLP tasks. IR requires to match a given query to a set of relevant documents, whereas NLP mostly deals with the discovery of semantic and linguistic regularities. Therefore, (knowledge-enhanced) neural language models do not encode relevance signals or discriminative aspects between queries and documents – which are fundamental to effectively address IR tasks.

In this work, we investigate which feature between synonymy and polysemy can be exploited to reduce the semantic gap and improve retrieval, and how external knowledge resources can help to bridge the semantic gap between queries and documents. To this end, we propose the Semantic-Aware Neural Framework for IR (SAFIR), an unsupervised knowledge-enhanced neural framework for IR. SAFIR jointly learns word, concept, and document representations from scratch. The learned representations are optimized for IR and encode both polysemy and synonymy to address the semantic gap between queries and documents.

We conduct an experimental evaluation to compare SAFIR with other (knowledge-enhanced) neural models on a specific task of Clinical Decision Support (CDS): medical literature retrieval.

The rest of the paper is organized as follows: Section 2 presents SAFIR, Section 3 describes the experimental evaluation, and Section 4 concludes the paper.

2. The Semantic-Aware Neural Framework for IR

SAFIR consists of three main components: semantic indexing, representation learning, and semantic matching. Below, we give an overview of the framework and its main components. For each component, we outline the required inputs and the provided outputs and we describe its high-level functioning.

The **semantic indexing** component takes as input a corpus and a knowledge resource and applies Named Entity Recognition (NER) and Entity Linking (EL) techniques to produce a knowledge-enhanced corpus. For each word, NER detects a list of candidate concepts, if any, and ranks them from the most to the least likely. Then, EL disambiguates candidate concepts against the knowledge resource relying on the context of the concept mentions (e.g., the document). The disambiguated mention-concept pair forms the atomic constituent of each knowledge-enhanced document.

The **representation learning** component consists of a shallow neural network that relies on the outputs of the the semantic indexing component to learn word, concept, and document representations. The network models polysemy and synonymy while optimizing representations for document retrieval via multi-task learning. For polysemy, word and concept representations are combined to form contextual representations for word-concept pairs, thus conveying a unique meaning in the vector space. For synonymy, the distance between the representations

Table 1
Retrieval performances of considered models. **Bold** values represent the highest scores.

		P@10			Recall@1000		
		CDS14	CDS15	CDS16	CDS14	CDS15	CDS16
BoW	BM25	0.1667	0.2600	**0.2167**	0.2503	0.1826	0.2286
	BM25/RM3	0.1833	0.2433	0.2067	0.3151	0.2884	**0.3059**
CD	word2vec	0.1133	0.1900	0.1167	0.2200	0.2194	0.1515
	NVSM	0.2033	0.2333	0.1600	0.3833	0.3093	0.2617
KE	rword2vec	0.1267	0.1967	0.1133	0.2221	0.2151	0.1414
SAFIR	SAFIR_s	0.1967	0.2267	0.1733	0.3607	**0.3134**	0.2545
	SAFIR_p	**0.2333**	**0.2633**	0.1700	**0.3846**	0.3098	0.2782
	SAFIR_{sp}	0.2200	0.2467	0.1633	0.3733	0.3110	0.2747

of words presenting a synonymy relation within the knowledge resource is minimized in the vector space. Regarding retrieval, contextual representations are learned to be close to the representations of documents that contain the corresponding word-concept pairs. This entails a matching relation specific to IR between word, concept, and document representations.

The **semantic matching** component uses the learned representations to perform semantic matching between knowledge-enhanced query and documents. Documents are ranked in decreasing order of the similarity score computed between query and document representations.

3. Experimental Evaluation

Experimental Setup. As test collections, we consider TREC Clinical Decision Support 2014 (CDS14), 2015 (CDS15), and 2016 (CDS16).[1] As knowledge resource, we adopt the 2018AA release of the UMLS Metathesaurus[2].

We use P@10 and Recall@1000 to evaluate systems and we consider three categories of retrieval models: bag-of-words models, corpus-driven models, and knowledge-enhanced models. As Bag-of-Words (BoW) models, we consider BM25 and BM25/RM3. As Corpus-Driven (CD) models, we consider word2vec [3, 6] and the Neural Vector Space Model (NVSM) [7]. As Knowledge-Enhanced (KE) models, we consider retrofitted word2vec (rword2vec) [4] and three variants of the Semantic-Aware Neural Framework for IR (SAFIR): SAFIR_{sp}, which integrates both synonymy and polysemy; SAFIR_{s} which integrates synonymy but not polysemy; and SAFIR_{p} which integrates polysemy but not synonymy.

Experimental Results We present the experimental results below. Table 1 shows model performances for medical literature retrieval on the considered collections.

The experimental results for document retrieval show that all SAFIR variants provide effective results in the considered collections. This indicates that SAFIR effectively encodes the text

[1] http://www.trec-cds.org/
[2] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/

matching signals required to perform retrieval regardless of the linguistic feature(s) modeled. Among the three variants, SAFIR$_\mathrm{p}$ provides the best results in most cases. Regarding SAFIR$_\mathrm{s}$ and SAFIR$_\mathrm{sp}$, they exhibit performances close to or slightly lower than those of NVSM and SAFIR$_\mathrm{p}$. This suggests that the impact of synonymy in CDS collections might be limited or even detrimental. In particular, we expect that modeling polysemy helps to order relevant documents in top positions of the ranking list, while modeling synonymy helps to retrieve a higher number of relevant documents which contain synonyms of the query terms. While the results confirm this trend for polysemy, they do not for synonymy. The negative results of rword2vec – which models only synonymy – compared to those of word2vec further support this intuition. Therefore, the results suggest that polysemy impacts more than synonymy on retrieval performances for CDS collections.

4. Conclusions and Future Work

We introduced the Semantic-Aware Neural Framework for IR (SAFIR), an unsupervised knowledge-enhanced neural framework for IR. The evaluation showed that the integration of knowledge resources into the learning process of neural IR models is effective and helps to bridge the semantic gap between queries and documents. The learned representations encode text matching signals, necessary for IR tasks, and linguistic features to retrieve relevant documents that are most affected by the semantic gap. In particular, the results showed that modeling polysemy is effective, whereas, on average performance, the impact of synonymy is marginal.

As future work, we plan to integrate deeper neural architectures into SAFIR representation learning component to better model linguistic features and their interactions with IR-oriented objective functions. Two other directions are the extension of SAFIR to phrase-concept associations and the sensitivity of the learned representations to the NER and EL components.

Acknowledgments

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

References

[1] M. Agosti, S. Marchesin, G. Silvello, Learning Unsupervised Knowledge-Enhanced Representations to Reduce the Semantic Gap in Information Retrieval, ACM Trans. Inf. Syst. 38 (2020) 38:1–38:48.

[2] B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, M. Lawley, Information retrieval as semantic inference: a Graph Inference model applied to medical search, Inf. Retr. Journal 19 (2016) 6–37.

[3] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, CoRR abs/1301.3781 (2013). arXiv:1301.3781.

[4] X. Liu, J. Y. Nie, A. Sordoni, Constraining Word Embeddings by Prior Knowledge - Application to Medical Information Retrieval, in: Proc. of AIRS 2016, Springer, 2016, pp. 155–167.

[5] L. Tamine, L. Soulier, G. H. Nguyen, N. Souf, Offline Versus Online Representation Learning of Documents Using External Knowledge, ACM Trans. Inf. Syst. 37 (2019) 42:1–42:34.

[6] I. Vulić, M. F. Moens, Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings, in: Proc. of SIGIR 2015, ACM, 2015, pp. 363–372.

[7] C. Van Gysel, M. de Rijke, E. Kanoulas, Neural Vector Spaces for Unsupervised Information Retrieval, ACM Trans. Inf. Syst. 36 (2018) 38:1–38:25.