Medical Retrieval using Structured Information Extracted from Knowledge Bases
(Discussion Paper)

Maristella Agosti, Giorgio Maria Di Nunzio, Stefano Marchesin, and Gianmaria Silvello

Department of Information Engineering, University of Padua, Italy
maristella.agosti, giorgiomaria.dinunzio, stefano.marchesin, gianmaria.silvello@unipd.it

Abstract. We investigate how semantic relations between concepts extracted from medical documents, and linked to a reference knowledge base, can be employed to improve the retrieval of medical literature. Semantic relations explicitly represent relatedness between concepts and carry high informative power that can be leveraged to improve the effectiveness of the retrieval. We present preliminary results and show how relations are able to provide a sizable increase of the precision for several topics, albeit having no impact on others. We then discuss some future directions to minimize the impact of negative results while maximizing the impact of good results.

Keywords: Information extraction · Knowledge bases · Medical information retrieval.

1 Motivations

The volume of medical literature published every year keeps growing at a very fast pace. The time required by clinicians to retrieve relevant information from such an amount of literature using standard systems is often prohibitive. Therefore, there has been a strong interest in Clinical Decision Support (CDS) systems [4] designed to produce effective and timely information that can help clinicians in the decision making process for patient care. Within this context, we focus on medical case-based retrieval – i.e., given a medical case of interest, the CDS system should retrieve highly related medical literature from a large collection of medical publications. Due to severe time constraints, clinicians must take fast decisions without having the possibility to thoroughly read the literature; for this reason, medical case-based retrieval favors precision over recall [5].

A key characteristic of the medical literature is the large use of synonyms and context-specific expressions. To address this term heterogeneity, Knowledge Bases (KBs) have often been exploited by Information Retrieval (IR) systems. The current availability of medical KBs offers us the opportunity to develop techniques that better capture the semantics of medical documents, leading to the following research question:

> How can we employ the rich semantic information within medical case reports and related literature to boost retrieval performances and ease the clinical decision process?

Semantic relations are a key aspect within the semantics of a document. They have been mainly used to find relevant concepts to expand a user query, but not as semantic elements to be indexed and retrieved. We hypothesize semantic relations can provide a higher semantic representation of medical cases and literature.

In this work, we present an initial study on the effectiveness of the use of semantic relations for the retrieval of medical literature. We define an approach comprising two methods: a *rule-based* method and a *learning* method. In the rule-based method, we assign a relation to a pair of concepts – contained within the same sentence – when it holds within a reference KB. In the learning method, we train a sentence-level relation extractor that is able to infer relation between a pair of concepts given the sentence context.

We evaluated our approach by using the publicly shared OHSUMED collection [8]. OHSUMED provides rather short queries which represent a hard task for our approach, since limited information – e.g. concepts and relations – can be extracted from them. Testing with OHSUMED allows us to assess the potentials and limitations of the approach. The remainder of the paper is organized as follows: Section 2 presents the background and related work, Section 3 describes the proposed approach, Section 4 presents experiments and results and Section 5 draws some conclusions and outlines future work.

2 Related Work

Concept-based IR aims at making use of external sources (like thesauri and ontologies) to provide additional knowledge and context that may not be explicit in a document collection and users' queries. Concept-based methods can be categorized in two types: (i) methods that use concepts in both indexing and retrieval stages [6], and (ii) methods that apply concept analysis in one specific stage, such as concept-based query expansion [7]. The approach we propose extends the use of concepts to relations and uses them both at the indexing and at the retrieval stages. An approach like the one we adopt is more challenging, but it allows for a finer semantic representation of documents and queries.

In the biomedical domain – where there are authoritative and curated ontologies – concept-based approaches demonstrate consistent improvements over classic keyword-based systems. In [10], 'is-a' relationships between concepts are

used to weight documents containing concepts subsumed by the query concepts. [12] proposes a method to represent medical records and queries by focusing only on medical concepts essential for the information need of a medical search task. In [11], 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.

The field of Biomedical Information Extraction (BioIE) is highly relevant for CDS. [13] reviews the recent advances in learning-based approaches for BioIE tasks. BioIE tasks comprise entity linking [23], event identification [2] and relation extraction [18, 20]. Being targeted to CDS – i.e. voted to the extraction of key relations that can facilitate clinical decision making – our problem setup is fundamentally different from the conventional biomedical setups. Most of state-of-the-art biomedical relation extraction techniques are developed for specific relations, like protein-protein interactions, gene-disease interactions and so on — which cover only a fraction of the biomedical domain.

Regarding relations in IR, [19] studies the problem of finding human readable descriptions of a given relationship in a knowledge graph. [17] applies supervised relation extraction to documents that are relevant for an information need Q and studies how many of the extracted relations are indeed relevant for Q. [9] explores current state of the art in unsupervised relation extraction (OpenIE) for the task of finding support passages to complement an entity ranking with human-readable explanations of how those retrieved entities are connected to the information need. Conversely, our approach applies supervised relation extraction to extract semantic relations that are used in both the indexing and the retrieval stages. Hence, relations play a pivotal role in the actual retrieval of documents.

3 Methodology

We present a new approach that uses semantic relations for medical case-based retrieval. The methodology is composed of the information extraction step, that is applied both at the indexing and the retrieval stages (Subsection 3.1), and the specific information retrieval stage (Subsection 3.2).

3.1 Information Extraction

The information extraction step is divided into an entity linking component and a relation extraction component.

The entity linking component extracts entity mentions within the text and links them to a reference KB; this reduces the high number of synonyms, abbreviations and context specific expressions that are present in the medical literature. For entity linking we adopt MetaMap,[1] an authoritative tool to detect medical entity mentions in free-text. MetaMap analyses biomedical free-text and identifies concepts belonging to the Unified Medical Language System (UMLS),[2]

[1] https://metamap.nlm.nih.gov/
[2] https://www.nlm.nih.gov/research/umls/

associating each mention with a number of concepts from the UMLS Metathesaurus[3] — which comprises more than 3 million distinct concepts. Within UMLS, a substantial understanding of the medical domain is included, comprising medical concepts, relations, definitions and so on.

The relation extraction component detects semantic relations between pairs of concepts within a sentence. To be consistent with concepts extracted with MetaMap, we consider semantic relations from UMLS Metathesaurus as well. Furthermore, since our task requires a high coverage of the medical domain, considering UMLS Metathesaurus relations – which are coarse-grained relationships that span to a high number of concepts – allows us to increase the recall of extracted relations.

We define two methods for the extraction of relations from documents and queries: a rule-based method and a learning method.

Rule-based: a relation is assigned to a pair of concepts if it relates them within UMLS. We assume that a UMLS relation between two concepts always occurs, even when it is not explicitly mentioned in the sentence containing the two concepts.

Learning: we train a distantly supervised [14] sentence-level Bidirectional Long Short-Term Memory (BiLSTM) neural network to detect if a relation exists between two concepts based on the context of the sentence. The network architecture is composed of an input (word embedding) layer of concatenated word features and positional features. Words are first converted into pre-trained word embeddings trained on 26 million abstracts and citations in PubMed — released by [15]. Then these word features are concatenated with two sets of positional features — to explicitly account for the pairs of words to which we expect to assign relations [21]. We apply a max-pooling layer right after the bidirectional recurrent layer and before the output layer — in order to combine segment-level features that, although not very strong in representing the entire sentence, represent local patterns well [22]. In this way, we try to overcome the tendency of recurrent connections to forget long-term information too quickly, leading the supervision at the end of the sentence to be hardly propagated to early steps in model training (due to gradient vanishing [3]).

3.2 Information Retrieval

Before to be in the condition to retrieve documents, it is necessary to index the documents, so the documents are indexed by considering all terms as in the Bag-of-Words (BoW) representation, but we also extend the BoW representation to both concepts (BoC) and relations (BoR) by considering for the indexing all the extracted concepts and relations respectively. Afterwards the ranking is obtained using Okapi BM25 ranking function [16].

Since relations are extracted at sentence level, we also index passages – i.e. groups of consecutive sentences – by considering all the relations occurring within each group of sentences (passage-level BoR). Relevant passages should contain

[3] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/

a higher number of relations related to the information need when compared to non relevant passages — being more similar in their semantic contents to the query. Therefore, documents that contain more relevant passages can be considered more relevant to the query.

We define a weighting scheme such that a document score is computed as the weighted sum of its passages scores, where scores are computed using BM25 as above. The passage-level weighting scheme is as follows:

$$score(q, d) = \sum_{p \in d} \frac{|R_p \cap R_q|}{|R_q|} BM25(p, q) \quad (1)$$

where d is the document, q is the query, p is a passage belonging to document d, R_q is the set of relations extracted from query q and R_p is the set of relations extracted from passage p.

4 Experiments and Results

We employed the OHSUMED test collection which contains 348,566 references from the on-line medical information database MEDLINE,[4] consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987-1991). The available fields are: title, abstract, MeSH[5] indexing terms, author, source, and publication type. There are 106 queries in the collection. Each query is composed of two sentences: title + description. Title is the brief summary of the medical case at hand, description is the information need required to answer a specific question for the case.

Experimental Setup:
We performed two experiments:
i) One using the rule-based method to extract relations out of documents and queries.
ii) The other using the learning method to extract relations out of documents and queries. We compared the results obtained applying BM25 to the three representations (i.e. BoW, BoC and BoR) and we evaluated the results using the nDCG measure.

Results:
i) The rule-based method was able to extract relations from a subset of 44 queries. Therefore, to investigate the effectiveness of relations, we restrict the experiments to this subset only — since the remaining queries lead to no results when considering relations. Of these 44 queries, only 39 have relations matching with some documents. Regarding the relations, we obtained the best results with the passage-level approach. We set the passage length to 2, in order to be compliant with query length. Document score was computed using the formula shown above (1). The nDCG results on these 39 queries are variable – ranging from 0 (18 cases) to 1 (5 cases), as can be seen in Figure 1. Such a variance

[4] https://www.nlm.nih.gov/bsd/medline.html
[5] https://meshb.nlm.nih.gov/search

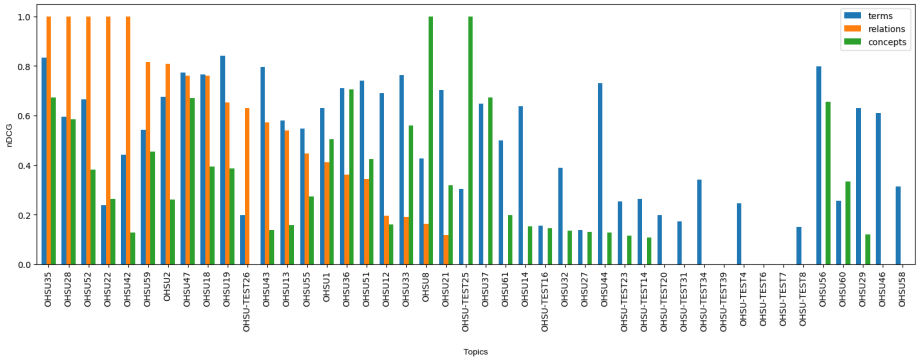

Fig. 1. nDCG values for topics containing relations. OHSU{56,60,29,46,58} returned NA values for the BoR representation. Queries are sorted in descending order first by BoR (relations) nDCG values, then by BoC (concepts) nDCG values.

gives us some hints about the informative power of relations. When properly extracted, relations can be highly effective, indeed; we compared the average nDCG values of concepts and relations on only those topics where relations give a result different than 0 and we found a statistically significant average improvement of 20%. A *t-test* was performed to validate the improvement. Regarding the comparison between relations and terms, the behavior of relations is similar to the one of terms (baseline approach), and there is no statistically significant difference between the two.

ii) The learning method was able to extract relations from a subset of 25 queries. Of these 25 queries, only 12 have relations matching with some documents. The results on these 12 queries are comparable to those presented for the rule-base method, with nDCG values ranging from 0 (in 7 cases) to 1 (in 1 case). The reason for this is two-fold: (a) the shortness of queries that limits the relations that can be extracted; and, (b) the highly different syntactic structure of queries if compared to the sentences within the medical abstracts leading to a mismatch between the query-relations and abstract-relations.

5 Conclusion

In this work, we proposed and evaluated the effectiveness of semantic relations as basic constituents for a CDS system. We defined two methods for extracting relations from queries and documents: a rule-based method and a learning method. We found that relations – when pertinent to the initial information need – are highly valuable, outperforming concepts. The challenge lies in how to limit those cases where relations provide no relevant results for the information need. To this end, considering collections where queries present a long and

narrative structure (e.g. TREC CDS tracks[6]) might be a possible direction to balance such issue.

Furthermore, defining more IR-oriented relation extraction approaches that are capable of overcoming the high precision-low recall nature of state-of-the-art methods is a direction that can be investigated.

Finally, we could compare the relation extraction approaches in terms of quality of the results to verify if the extracted relations are semantically correct. This can further clarify whether relations' limited effectiveness in IR tasks lies in current state-of-the-art relation extraction approaches or in the poor representativeness of relations themselves for IR tasks.

An initial version of this paper has been presented at the ACM 12th International Workshop on Data and Text Mining in Biomedical Informatics (DTMBio), held in conjunction with ACM 27th Conference on Information and Knowledge Management (CIKM) [1].

Acknowledgements

The work was partially supported by the CDC-STARS project of the University of Padua, Italy,[7] and by the ExaMode project,[8] as part of the European Union H2020 research and innovation program under grant agreement no. 825292.

References

1. Agosti, M., Di Nunzio, G.M., Marchesin, S., Silvello, G.: A relation extraction approach for clinical decision support. arXiv preprint arXiv:1905.01257 (2019)
2. Ananiadou, S., Pyysalo, S., Tsujii, J., Kell, D.B.: Event extraction for systems biology by text mining the literature. Trends in Biotechnology **28**(7), 381–390 (2010)
3. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks **5**(2), 157–166 (1994)
4. Berner, E.S.: Clinical decision support systems, vol. 233. Springer (2007)
5. Burke, D.T., DeVito, M.C., Schneider, J.C., Julien, S., Judelson, A.L.: Reading habits of physical medicine and rehabilitation resident physicians. American Journal of Physical Medicine & Rehabilitation **83**(7), 551–559 (2004)
6. Egozi, O., Markovitch, S., Gabrilovich, E.: Concept-based information retrieval using explicit semantic analysis. ACM Trans. Inf. Syst. **29**(2), 8:1–8:34 (Apr 2011)
7. Grootjen, F.A., Van Der Weide, T.P.: Conceptual query expansion. Data & Knowledge Engineering **56**(2), 174–193 (2006)
8. Hersh, W., Buckley, C., Leone, T., Hickam, D.: Ohsumed: an interactive retrieval evaluation and new large test collection for research. In: SIGIR'94. pp. 192–201. Springer (1994)

[6] http://www.trec-cds.org/
[7] http://datacitation.dei.unipd.it/
[8] https://www.examode.eu/

9. Kadry, A., Dietz, L.: Open relation extraction for support passage retrieval: Merit and open issues. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1149–1152. ACM (2017)
10. 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: The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)[NIST Special Publication: SP 500-298]. pp. 1–8 (2012)
11. Limsopatham, N., Macdonald, C., Ounis, I.: Inferring conceptual relationships to improve medical records search. In: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval. pp. 1–8 (2013)
12. Limsopatham, N., Macdonald, C., Ounis, I.: A task-specific query and document representation for medical records search. In: European Conference on Information Retrieval. pp. 747–751. Springer (2013)
13. Liu, F., Chen, J., Jagannatha, A., Yu, H.: Learning for biomedical information extraction: methodological review of recent advances. arXiv preprint arXiv:1606.07993 (2016)
14. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. pp. 1003–1011. Association for Computational Linguistics (2009)
15. Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., Ananiadou, S.: Distributional semantics resources for biomedical text processing. In: Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan. pp. 39–43 (2013)
16. Robertson, S.E., Walker, S.: Okapi/keenbow at trec-8. In: TREC. vol. 8, pp. 151–162. Citeseer (1999)
17. Schuhmacher, M., Roth, B., Ponzetto, S.P., Dietz, L.: Finding relevant relations in relevant documents. In: European Conference on Information Retrieval. pp. 654–660. Springer (2016)
18. Uzuner, Ö., South, B., Shen, S., DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association **18**(5), 552–556 (2011)
19. Voskarides, N., Meij, E., de Rijke, M.: Generating descriptions of entity relationships. In: European Conference on Information Retrieval. pp. 317–330. Springer (2017)
20. Wang, C., Fan, J.: Medical relation extraction with manifold models. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 828–838 (2014)
21. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 2335–2344 (2014)
22. Zhang, D., Wang, D.: Relation classification via recurrent neural network. arXiv preprint arXiv:1508.01006 (2015)
23. Zheng, J., Howsmon, D., Zhang, B., Hahn, J., McGuinness, D., Hendler, J., Ji, H.: Entity linking for biomedical literature. BMC Medical Informatics and Decision Making **15**(1), S4 (2015)