Terminology Extraction in Electronic Health Records. The ExaMode Project*

Stefano Marchesin[1], Giorgio Maria Di Nunzio[1] and Gianmaria Silvello[1]

[1]Department of Information Engineering, University of Padua, Italy

Abstract

Medical free-text records store a lot of useful information that can be exploited in developing computer-supported medicine. Nevertheless, extracting terminological knowledge from unstructured text is difficult because the volume of medical texts created every year keeps growing at a very fast pace and it is highly dependent on the language under examination. In this work, we present an initial study of a Natural Language Processing pipeline in order to extract terminological information from pathology reports and link this information to medical ontologies.

Keywords

Terminology Extraction, Electronic Health Records, Named Entity Recognition, Entity Linking

1. Introduction

Modern medical specialties rely on clinical context for an accurate interpretation of medical data. The literature of medical imaging analysis shows an important trend where both the Electronic Health Records (EHR) and medical images are leveraged in an approach called 'fusion paradigm' for solving complex tasks that cannot be tackled by a single modality [1]. In fact, medical free-text records store a lot of useful information that can be exploited in developing computer-supported medicine [2]. These medical free-text records can also be produced by patients in the so-called patient-reported diagnosis [3]. This type of documents can reveal if a patient left a medical encounter knowing the diagnosis explained to them and can ultimately inform on whether there are language differences between training health care professionals and patients without medical training.

However, extracting terminological knowledge from unstructured text is difficult for at least two reasons: firstly, the volume of medical texts created 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, 5] designed to produce effective and timely

1st International Conference on "Multilingual Digital Terminology Today. Design, representation formats, and management systems", June 16–17, 2022, Padua, Italy

stefano.marchesin@unipd.it (S. Marchesin); giorgiomaria.dinunzio@unipd.it (G. Di Nunzio); gianmaria.silvello@unipd.it (G. Silvello)

https://www.dei.unipd.it/~marches1 (S. Marchesin); https://www.dei.unipd.it/~dinunzio (G. Di Nunzio); https://www.dei.unipd.it/~silvello (G. Silvello)

0000-0003-0362-5893 (S. Marchesin); 0000-0001-9709-6392 (G. Di Nunzio); 0000-0003-4970-4554 (G. Silvello)

information that can help clinicians in the decision making process for patient care.[1] Secondly, the extraction of relevant terminological knowledge highly depends on the language (and not only on the specialized language). For example, in [6], the authors explore terminology related to the semantic field of terms that indicate or suggest the presence of implants in electronic medical records EHRs written in Swedish with techniques that are highly optimized (in a justified way) for that language, making that software only partially reusable for other languages.

In this work, we focus on digital pathology, a specialized field that studies digital histopathology images to diagnose cancer cases and related diseases. In particular, we propose a Natural Language Processing (NLP) pipeline in order to extract terminological information from pathology reports and link this information to medical ontologies. We evaluate the effectiveness of this approach on entity linking and text classification tasks, considering different use-cases concerning different types of cancer. Moreover, an unsupervised multilingual hybrid knowledge extraction system that combines rule-based techniques with pre-trained deep neural models to extract knowledge from pathology reports will be assessed against the manual extraction of terms from the same medical reports.

2. Proposal

This work has been carried out under the umbrella of the ExaMode project,[2] an interdisciplinary Horizon 2020 project the goal of which is to design efficient methodologies to manage the vast amount of heterogeneous medical data produced every day in different forms, and to ensure easier, faster discovery and consultation with little human supervision. The ultimate goal of the ExaMode project is to have these resources adopted not only by specialists, but also by non-experts, so that the latter can achieve a better understanding of medical information.

The dataset for this study is a selection of real medical reports produced in the Hospital of Catania, one of the project partners. These reports are divided into four pathologies, namely three types of cancer (cervix, colon and lung) and celiac disease for a total of 200 medical reports.

The proposed method is based on a NLP pipeline that adopts a combination of pre-trained Named Entity Recognition (NER) models [7] and unsupervised Entity Linking (EL) methods [8] to extract key terms from the medical reports and to link them to the reference ontology. NER is the task of identifying and categorizing key information, or "entities", within a text. An entity can be any multi-word term that consistently refers to the same concept. Each entity is therefore classified, i.e. "linked", into a predefined category, such as *disease* or *protein*.[3] Entity Linking is therefore the task of assigning unique meanings to entities mentioned in a text. Our approach proposes a combination of ad-hoc and similarity matching techniques to connect the extracted entities to unique concepts. Finally, the approach uses a set of rules to merge entities into multi-word terms. For example, the terms "colon" and "transverse" may be considered as separate entities in a text, while the "transverse colon" is the correct entity to be linked to the ontology.

[1] https://www.trec-cds.org
[2] https://www.examode.eu
[3] We use *italics* to indicate categories.

Acknowledgment

This work was partially supported by the ExaMode Project, as a part of the European Union Horizon 2020 Program under Grant 825292.

References

[1] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M. P. Lungren, Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, npj Digital Medicine 3 (2020) 136. URL: https://doi.org/10.1038/s41746-020-00341-z. doi:10.1038/s41746-020-00341-z.

[2] A. G. Dobrakowski, A. Mykowiecka, M. Marciniak, W. Jaworski, P. Biecek, Interpretable segmentation of medical free-text records based on word embeddings, Journal of Intelligent Information Systems 57 (2021) 447–465. URL: https://doi.org/10.1007/s10844-021-00659-4. doi:10.1007/s10844-021-00659-4.

[3] K. Gleason, M. R. Dahm, How patients describe their diagnosis compared to clinical documentation, Diagnosis 9 (2022) 250–254. URL: https://doi.org/10.1515/dx-2021-0070. doi:doi:10.1515/dx-2021-0070.

[4] E. S. Berner, Clinical decision support systems : theory and practice, 3rd ed., Springer, 2016.

[5] M. Agosti, G. Di Nunzio, S. Marchesin, G. Silvello, Medical retrieval using structured information extracted from knowledge bases, in: SEBD, volume 2400 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019.

[6] O. Jerdhaf, M. Santini, P. Lundberg, A. Karlsson, A. Jönsson, Focused terminology extraction for cpss the case of "implant terms" in electronic medical records, in: 2021 IEEE International Conference on Communications Workshops (ICC Workshops), 2021, pp. 1–6. doi:10.1109/ICCWorkshops50388.2021.9473700.

[7] A. Goyal, V. Gupta, M. Kumar, Recent named entity recognition and classification techniques: A systematic review, Computer Science Review 29 (2018) 21–43. URL: https://www.sciencedirect.com/science/article/pii/S1574013717302782. doi:https://doi.org/10.1016/j.cosrev.2018.06.001.

[8] X. Liao, Z. Zhao, Unsupervised approaches for textual semantic annotation, a survey, ACM Comput. Surv. 52 (2019). URL: https://doi.org/10.1145/3324473. doi:10.1145/3324473.