First Steps Towards Patient-Friendly Presentation of Dutch Radiology Reports

Koen Dercksen[1,2] [0000−0003−2571−9102]
Arjen P. de Vries[2] [0000−0002−2888−4202]

[1] Radboud University, Nijmegen, The Netherlands
{koen,arjen}@cs.ru.nl
[2] Radboud University Medical Center, Nijmegen, The Netherlands

Abstract. Nowadays, clinical patients are often free to access their own electronic health records (EHRs) online. Medical records are however not written with the patient in mind – the medical terminology necessary to ensure unambiguous communication between medical professionals on likelihood of pathology renders the EHRs less accessible to patients. By annotating these texts with links to external knowledge bases, the patients can be provided with additional reliable information to clarify terminology. In this paper, we present preliminary work on preparing Dutch radiology reports for named entity recognition and entity linking to provide additional information to patients. Additionally, we suggest a roadmap for further research into patient-friendly presentation of radiology reports.

Keywords: information retrieval · electronic health records · natural language processing

1 Introduction

Hospitals worldwide give patients increasingly more access to their medical electronic health records (EHRs), a trend that concurs with (and facilitates) the personalisation of health-care. EHRs include a variety of information, from basic administrative information about past and future visits to the clinic, medical communication (e.g., letters requesting treatment from the general practitioner, exchange of information between medical experts), but also the radiology reports (added within 1-7 days after authorisation of the report). Patients at the Radboud University Medical Center appreciate access to their health records and use the patient portal actively; 69% of the 120K outpatients visiting the clinic used the portal at some point, over 57K patients log in every quarter, and patients who access the portal do so six times per quarter, on average. EHRs are however primarily created to facilitate the communication between caregivers and other medical staff, and not written for exploration by laymen. Medical terminology is necessary to guarantee unambiguous, reliable, safe, uniform and directional information exchange between medical professionals on likelihood of pathology,

but renders the EHR less accessible for patients. Additionally, consider the presence of medical images in the EHR, where even the physician requires help to interpret this data correctly. Patients are likely to respond to this knowledge gap regarding the terminology in their EHR by searching for medical information online, risking patient anxiety[3] when inaccurate, inadequate, or incomprehensible information is found [24]. This in turn causes longer consultation times and additional interactions with the physician to help patients understand their medical situation, resulting in undesirable and often unnecessary extra pressure on the health care system.

A (partial) solution to the problem outlined above would be to reduce the knowledge gap by augmenting the EHR with links to background information that can clarify what is written in the report [11]. However, while abundant free-text radiology reports are available within Radboud University Medical Center, we face the specific challenge of working with Dutch medical text. First, this setting reduces the number of language resources (structured vocabularies and language models) significantly. Also, due to the intended usage of our research, we should only link to reliable and curated information, complementary to public sources like Wikipedia, in order to avoid annotations with incorrect information. Examples of targets we identified include Thuisarts, Hartwijzer, and Radiology Assistant.

In this work, we detail initial findings regarding available resources and tools for automated annotation, named entity recognition, and evaluation in the medical domain, as well as plans for future research.

2 Related work

2.1 Datasets

Named entity recognition (NER) in medical texts has been a topic of active research for a long time. However, the majority of prior research addresses medical scientific literature (often Pubmed) [8, 9, 26], where annotation focuses on entities like DNA or proteins; not that relevant to our main objective. A notable exception are the research datasets released under the n2c2 moniker, formerly known as i2b2 [23]. These datasets consist of unstructured clinical notes in English, annotated for a variety of natural language processing (NLP) tasks, including NER and concept normalisation. The THYME corpus [22] is another dataset of unstructured clinical notes annotated with medical concepts, as well as temporal events and relations. In terms of publicly available Dutch radiology report datasets, we are not aware of any in existence.

2.2 Concept extraction tools

Elaborate tools for NER and entity linking (EL) have been made available for English medical text. Many of these tools use the Unified Medical Language Sys-

[3] e.g. *cyberchondria*, excessive health concerns after repeated Internet searches for medical information.

tem (UMLS) as a knowledge base [17]. UMLS is an ontology of biomedical concepts, consisting of controlled vocabularies in different languages. MetaMap [2, 3] is a tool that extracts clinical concepts by detecting noun phrases, generating variations of those phrases based on an extensive knowledge base of synonyms and relations, and then selecting the best fitting concept from UMLS. Apache cTAKES [16] is another such tool, capable of detecting things like symptoms, diseases and medication in English unstructured clinical notes. It uses a dictionary based on (subsets of) UMLS, SNOMED CT [5] and RxNORM [14]. MedLEE [6] is another NLP tool developed to structure free-text radiology reports in English, using concept and synonym mappings. Finally, QuickUMLS [20] is a fast approximate dictionary matching tool to find UMLS concepts in text. This tool generates heuristically valid variations of token sequences within a certain text window, and finds a set of matching concepts such that any overlap between concepts is minimised. As it does not rely on other knowledge than UMLS, it can be used for Dutch as well.

3 Methodology

3.1 Data

We collected approximately 10K reports of patients that underwent a thorax CT scan at Radboud University Medical Center (Radboudumc) between 2013 and 2018. As part of the collection process, the reports are anonymised using in-house software; names, dates, phone numbers and so on are replaced with generic tokens like "NAME", "DATE", and "PHONENR", respectively. We are currently in the process of annotating a subset of these reports extensively, using BRAT [21]. BRAT is a web-based tool for structured text annotation, allowing for labelling of text spans and relations between annotations. We first apply QuickUMLS [20] to tag concepts that match closely with Dutch concepts in UMLS. However, this tool misses many entities, and does not tag e.g. relations in the text; this is why complementary manual annotations and corrections are necessary. For the annotation scheme, we follow cTAKES; we want to detect findings, diseases and signs/symptoms, as well as location identifiers and keywords that indicate negation and uncertainty. We also annotate relations between e.g. findings and their corresponding location identifiers and modifiers. Where applicable, annotations will include a UMLS concept identifier. An example of these annotations is shown in Figure 1. We use the UMLS semantic types from [20] as entity types.

3.2 Automated information extraction

Once the report annotations are obtained, we can propagate them to the rest of the reports [7]. We intend to use a combination of rule-based methods as well as recent neural models like BERT [4] in order to get the best result. BERT has been shown to work well for relation extraction and semantic role labeling [13, 18], and

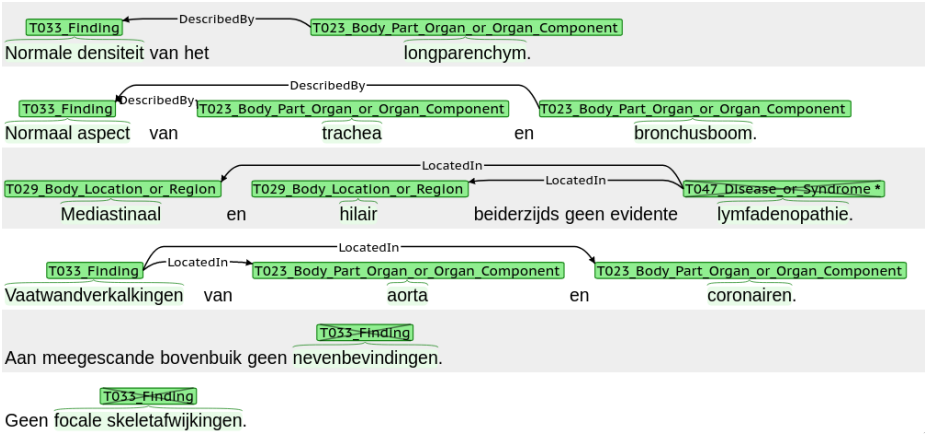

Fig. 1. Example of report annotated using BRAT. Crossed out entity types indicate a negation. An asterisk indicates uncertainty.

also for clinical NER [15, 19]. Clinical BERT embeddings are freely available [1, 12], but unfortunately none exist for clinical Dutch. We suspect that the use of contextualised word embeddings will improve automatic annotation propagation over simply propagating to identical text spans. Automatic annotations obtained using such a model can in turn be used to improve the word embeddings in a weakly supervised fashion. These embeddings can then be used to train NER/EL models.

4 Future work

While UMLS is a useful resource for NER and EL, the language used in concept definitions is not that friendly to laymen. We will need to curate data from more resources, particularly those that provide concept definitions and explanations in easily understandable language (e.g. those mentioned in Section 1).

Haridas and Kim [10] showed that clustering techniques applied to different types of clinical documents can improve automatic annotation of generic entities. We could perhaps extend this idea to different types of patients, improving NER by taking clinical context and background into account.

Radiology reports also contain information that is not necessarily relevant for a patient to read. For example, the radiologist might mention that "no pleural fluid was found". A patient could interpret this to be negative, while it is not. To retain only relevant sentences, we are planning to look into summarisation approaches as a starting point [25].

Finally, we intend to look into multi-modal enhancement of the report presentation. For example, a report might mention the location of a finding in the accompanying medical image. These mentions and images could be coupled to show patients where a specific finding is located in their body.

Acknowledgement. PhD project *MIHRacle: Multi-modal Interactive Health Records* is partially funded by the Radboud AI for Health collaboration between Radboud University and Radboudumc, and the Innovation Center for Articial Intelligence (ICAI).

References

[1] Emily Alsentzer et al. "Publicly Available Clinical BERT Embeddings". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 2019, pp. 72–78.

[2] Alan R Aronson. "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." In: *Proceedings of the AMIA Symposium*. American Medical Informatics Association. 2001, p. 17.

[3] Alan R Aronson and François-Michel Lang. "An overview of MetaMap: historical perspective and recent advances". In: *Journal of the American Medical Informatics Association* 17.3 (2010), pp. 229–236.

[4] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[5] Kevin Donnelly. "SNOMED-CT: The advanced terminology and coding system for eHealth". In: *Studies in Health Technology and Informatics* 121 (2006), p. 279.

[6] Carol Friedman et al. "A general natural-language text processor for clinical radiology". In: *Journal of the American Medical Informatics Association* 1.2 (1994), pp. 161–174.

[7] Cyril Grouin. "Controlled propagation of concept annotations in textual corpora". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. 2016, pp. 4075–4079.

[8] Maryam Habibi et al. "Deep learning with word embeddings improves biomedical named entity recognition". In: *Bioinformatics* 33.14 (2017), pp. 37–48.

[9] Kai Hakala et al. "Syntactic analyses and named entity recognition for PubMed and PubMed Central—up-to-the-minute". In: *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. 2016, pp. 102–107.

[10] Nithin Haridas and Yubin Kim. "Clustering Large-scale Diverse Electronic Medical Records to Aid Annotation for Generic Named Entity Recognition". In: *HSDM 2020 Workshop on Health Search and Data Mining*. Vol. 1. 2020.

[11] Jiyin He et al. "Generating links to background knowledge: a case study using narrative radiology reports". In: *Proceedings of the 20th ACM international conference on Information and knowledge management*. 2011, pp. 1867–1876.

[12] J Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." In: *Bioinformatics (Oxford, England)* (2019).

[13] Chen Lin et al. "A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 2019, pp. 65–71.

[14] Simon Liu et al. "RxNorm: prescription for electronic drug information exchange". In: *IT Professional* 7.5 (2005), pp. 17–23.

[15] Yifan Peng, Shankai Yan, and Zhiyong Lu. "Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets". In: *arXiv preprint arXiv:1906.05474* (2019).

[16] Guergana K Savova et al. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications". In: *Journal of the American Medical Informatics Association* 17.5 (2010), pp. 507–513.

[17] Peri L Schuyler et al. "The UMLS Metathesaurus: representing different views of biomedical concepts." In: *Bulletin of the Medical Library Association* 81.2 (1993), p. 217.

[18] Peng Shi and Jimmy Lin. "Simple BERT models for relation extraction and semantic role labeling". In: *arXiv preprint arXiv:1904.05255* (2019).

[19] Yuqi Si et al. "Enhancing clinical concept extraction with contextual embeddings". In: *Journal of the American Medical Informatics Association* 26.11 (2019), pp. 1297–1304.

[20] Luca Soldaini and Nazli Goharian. "QuickUMLS: a fast, unsupervised approach for medical concept extraction". In: *MedIR workshop, SIGIR*. 2016, pp. 1–4.

[21] Pontus Stenetorp et al. "BRAT: a web-based tool for NLP-assisted text annotation". In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 2012, pp. 102–107.

[22] William F Styler IV et al. "Temporal annotation in the clinical domain". In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 143–154.

[23] Özlem Uzuner et al. "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text". In: *Journal of the American Medical Informatics Association* 18.5 (2011), pp. 552–556.

[24] Ryen W. White and Eric Horvitz. "Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search". In: *ACM Trans. Inf. Syst.* 27.4 (Nov. 2009). ISSN: 1046-8188. DOI: 10.1145/1629096.1629101. URL: https://doi.org/10.1145/1629096.1629101.

[25] Yuhao Zhang et al. "Learning to Summarize Radiology Findings". In: *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*. 2018, pp. 204–213.

[26] Jin Guang Zheng et al. "Entity linking for biomedical literature". In: *Proceedings of the ACM 8th International Workshop on Data and Text Mining in Bioinformatics*. 2014, pp. 3–4.