POPULAR SCIENTIFIC ABSTRACT

[Martin Sundahl Laursen]
[Mining Electronic Health Records: Turning Unstructured Text into Research Data Using Natural Language Processing]

The electronic health record (EHR) contains detailed information about the patients' medical history and is therefore an important source of clinical research data. It is estimated that 80% of the information in the EHR is unstructured, which complicates locating relevant information. The medical history is also registered in a structured format as codes, e.g. for specific diseases, and in medical databases.

Clinical studies often use the coded information as research data, but it has some disadvantages: some information from the medical history is not coded; the codes can be inaccurate; medical databases only exist for some patient groups; and the codes are registered some time after the patient contact, which introduces a delay. Therefore, some clinical studies must rely on extracting research data from unstructured EHR text by manually reading through it, which is time-consuming and expensive.

Natural language processing (NLP) is a type of artificial intelligence that can analyse and process text. NLP algorithms have the potential to automatically extract clinical research data from the unstructured EHR text. It can, e.g. analyse if the text mentions that the patient has a specific disease that the study is researching. Algorithms for automatic extraction can reduce the data collection time for studies that currently rely on manual data extraction. Furthermore, NLP algorithms can facilitate that studies currently relying on coded data can supplement, improve, or replace the research data with data automatically extracted from the unstructured EHR text.

The overall purpose of this PhD dissertation is to enable Danish clinical research by automatically extracting research data from unstructured EHR text.

The work has focused on extracting events of bleeding and blood

clots originating in the veins (VTE) from the unstructured text because erroneous coding has been demonstrated for both conditions. I present two NLP algorithms to extract bleeding and VTE events and their anatomical location from unstructured EHR text.

Furthermore, I present an NLP algorithm to automatically extract an overview of the entire medical history in the form of clinical events, their attributes, and their relations from unstructured EHR text.

There are challenges to using NLP algorithms to extract research data from unstructured EHR text. One is the need for large amounts of labelled data to train the algorithm to perform the task of analysing the text, and another is that data need to be de-identified to comply with the General Data Protection Regulation. Therefore, I present an NLP algorithm for de-identification of names, streets, and locations in Danish unstructured text. The model was successfully trained with labelled data produced by a simple search function instead of data labelled by experts. This removed the need for a time-consuming and expensive annotation process. Finally, I present vector representations of Danish words tailored to the clinical domain. They can be used to search for specific topics in the text, making them easier to find during the labelling process, and save time.