Chapter 1
Introduction

The amount of digitised data available from healthcare systems is increasing exponentially. This data is seldom reused, either due to ignorance of its potential importance or due to the lack of available tools to process the data, but also because of the ethical policies regarding access to sensitive data. Healthcare systems and specifically health record systems contain both structured and unstructured information as text. More specifically, it is estimated that over 40% of the data in healthcare record systems contains text, so-called clinical text, sometimes also called electronic patient record text (Dalianis et al. 2009).

Clinical text contains valuable information about symptoms, diagnoses, treatments, drug use and adverse (drug) events for the patient that can be utilised to improve healthcare for other patients. The physician also writes her or his reasoning for the conclusion of the diagnosis of the patient in the patient record (Groopman 2007).

However, clinical text also contains sensitive information such as personal names, telephone numbers and addresses of the patient and relatives. This information needs to be pseudonymised before the clinical text can be utilised for secondary use (Velupillai et al. 2009).

A large amount of the information in an electronic patient record system is unstructured in the form of free text.

Clinical text is written by various professionals such as physicians, nurses, physiotherapists, psychologists etc. Often it is written under time pressure, and contain misspellings, non-standard abbreviations and jargon, as well as incomplete sentences, and is therefore difficult to process for natural language processing (NLP) tools developed for ordinary text, such as news text or text produced for a large number of readers (Allvin et al. 2011; Smith et al. 2014). Plenty of research has been carried out for clinical text processing for text written in English, but not that much for small and under resourced languages such as for example Swedish (Névéol et al. 2018). Nevertheless, a number of projects have been carried out at DSV, Stockholm University, during the years 2007–2017 in the area of clinical text mining in Swedish.

H. Dalianis, *Clinical Text Mining*, https://doi.org/10.1007/978-3-319-78503-5_1

Karolinska University Hospital has contributed over two million patient records from the years 2007–2014 to the Clinical Text Mining Group at DSV. This patient database, called HEALTH BANK—Swedish Health Record Research Bank, has obtained seven different ethical permissions for seven research projects that have been carried out or are ongoing. An initial publication by Dalianis (2014) described these efforts, and the experience from these efforts. This publication has been developed, extended and synthesised to this textbook.

This book will however, first go back in time, explaining the historical origin of the patient record, both papyrus and paper based, and the structure and content of the patient record. It will then continue with the first computer based patient record systems and examples of experimental patient record systems. The need and requirements for electronic patient record systems from clinical personnel including physicians and nurses will be described.

The process of obtaining access to electronic patient records for research will be explained, and will include applying for ethical permission, as well as practical issues regarding extracting, storing and using the patient records for research.

Natural language processing (NLP) will be explained and how information retrieval and text mining relates to NLP. Various tools for processing the patient records will be presented and the challenges in constructing these tools discussed. Challenges creating useful manually annotated data for training for so-called supervised learning will be described, in contrast to using various resources that are not annotated beforehand to train the system or artefact, so-called unsupervised learning. More specifically supervised learning simulates the behaviour of the human annotator for an artefact or system. The *annotator* is a person that manually will mark up data for training a machine learning system. *Active learning* will be used to select the most optimal data for annotation. System and artefact will be used interchangeably in this book. Algorithm is also used, but in this context as part of a system or artefact. Sometimes also the concept tool or application will be used.

The same annotated data can partly be set aside for evaluation of the artefact. There are various evaluation methods that can be used, for example k-fold cross validation, but also methods to calculate the significance of the results. These techniques are important for calculating the behaviour of the system and will be explained in this book.

First of all, to make electronic patient record text available for research, or for developing and testing applications, the records need to be de-identified since they contain information that can identify individual patients. One important application that will be described in this book is therefore the de-identification of electronic patient record text to de-identify the sensitive information, which means to remove information such as patient names, addresses and telephone numbers in the text that would otherwise reveal the identity of the patient. The identified sensitive information can be replaced with surrogates or pseudonyms, so that the text looks realistic without having any strange gaps and hence making the text coherent.

Once this first task is completed, then one can continue with the task of trying to address specific issues and attempt to acquire more knowledge.

Detecting adverse drug events (ADEs) is an example of a specific application area. An ADE may occur when a patient is treated for a disorder and can result from the use of one or more drugs and lead to the patient developing another disorder or symptom caused by the drug or drugs.

Predicting healthcare associated infections is yet another example of a specific application area that will be described in this book. A HAI is an infection that may occur while treating the patient at a hospital. It usually occurs when the patient is admitted to the hospital or just after being discharged, but according to its definition, the patient must have been admitted for at least 48 h (Ducel et al. 2002).

Other examples of specific applications that will be described are tools that may detect early symptoms of cancer, before the cancer has been detected or diagnosed. Some symptoms are very vague, but by having access to a large set of previous cancer cases with a lot of patterns, future cases may potentially be detected and predicted.

All these applications are based on analysing both the structured data of the patient record as well as the unstructured text using natural language engineering technology, or what also sometimes is called Artificial Intelligence.

Another, type of application is the handling of specific data. For example a pathology report is written by pathologists that have examined tissues from the human body to determine the disease the patient is suffering from. The result of the examination is written in free text in a pathology report that is given to the treating physician so he or she can decide on the treatment of the patient. For statistical reasons and for research the contents of the pathology reports are also entered into cancer registries, this is usually carried out manually by well-trained coders and is very time consuming work. The process of entering the content of the pathology report into a cancer registry can be automatised and in this book some examples will be provided.

Automatic diagnosis code assignment for discharge letters is another application that will be discussed: a secondary use of the same developed tool is for use for the validation of already assigned diagnosis codes. Diagnosis codes are assigned for the medical personnel as well as for administrative purposes, to calculate the cost of treatment and for the future planning of the overall healthcare provided by the whole clinical unit or hospital.

Automatic structuring of the patient record so it becomes more readable is another interesting application that will be explained. Automatic structuring will be carried out by either extracting the most important information and presenting it in a marked up format in the form of a hypertext, or by summarising the patient record.

Moreover, there is also the possibility to simplify the patient record so that a layperson can also understand it. This is because in some countries the patients have access to the patient record on the Internet and can read it, but the record is difficult to understand. However, there are ways to simplify the contents of the clinical text and explain expressions so a layperson can also understand the patient record.

In summary, we can conclude that there are many challenges in the area of clinical text mining, as well as important basic techniques that need to be explained.

Useful methods and their potential applications will be described and demonstrated, and finally summarised in this book in clinical text mining.

1.1 Early Work and Review Articles

One of the first articles written regarding clinical text mining is the article by Pratt and Pacak (1969), where the authors outline what is needed for the automatic processing of a clinical text in English to obtain an interpretation of the content.

More recently, two excellent reviews of clinical text mining are: Meystre et al. (2008), describing the state of the art of the research area, and Meystre et al. (2010), presenting the status of tools for de-identifying patient records. Yet another review article describing the detection of adverse drug events using text and data mining techniques is by Karimi et al. (2015b), while Freeman et al. (2013) review a number of tools for detecting healthcare associated infections and Spasić et al. (2014) compare different tools for detecting cancer symptoms. Regarding the closely related area biomedical text mining, see the textbook by Cohen and Demner-Fushman (2014).

The contents of this book will have as a starting point the work described by Dalianis (2014) but will extend with research work carried out by the author during 2014–2016 as well as work carried out by other contemporary researchers. This book also treats ethical and security issues, and details regarding various clinical text mining tools. Electronic patient records are written in different languages. Languages included in this book are Swedish, English and several European languages, as well as Japanese and Chinese.

Other important articles that describe the state of the art in the research field are: Nguyen et al. (2010), Skeppstedt et al. (2014) and Velupillai et al. (2014).

Regarding SNOMED CT, there are two articles describing the use of the terminology, (Lee et al. 2013, 2014), but the articles focus on academic publications and not so much on practical implementations using SNOMED CT.