Chapter 12
Conclusions and Outlook

In this book research in clinical text mining from the early days in 1970 up to now (2017) has been compiled. This book provided information on paper based patient record writing as well as the basics in electronic patient records, electronic patient record systems and terminologies. This book described the basic tools for natural language processing of clinical text, including text mining and machine learning techniques and the evaluation of these tools. Lots of examples of applications of clinical text mining have been given.

The book started with the history of the earliest papyrus based patient records in form of instructions for chirurgical treatment of wounds obtained in war in the Ancient Egypt and continued to the father of medicine, Hippocrates, in the Ancient Greece, who took careful notes of the symptoms and treatment of his patients. Hippocrates also urged that these notes should be used by new physicians involved in the treatment of the patients. The Hippocratian way of documenting symptoms and treatment of the diseases was further developed by the Arabs during the Islamic Golden Age, who also introduced hospitals.

In Europe, during the Age of Enlightenment, the taxonomy system of plants and animals was invented by Carl von Linné at Uppsala University. This classification system also inspired to the first classification system of diseases also by Carl von Linné. Nils Rosén von Rosenstein, who was a colleague of Carl von Linné at Uppsala University, introduced and developed the patient record in Sweden.

The book continued with a description of the requirements of electronic patient records systems from the perspective of health personnel, and the development of the electronic patient record systems from the first systems in the 1960s. The transition process from paper based patient records to electronic patient records was also explained.

The book followed with the description of some future support tools for patient record systems, implemented as prototypes assisting the physician and healthcare personnel to obtain a quick overview of the patient record.

H. Dalianis, *Clinical Text Mining*, https://doi.org/10.1007/978-3-319-78503-5_12

The book was continued by pointing at the specific characteristics of the patient record text, such as misspellings, nonstandard abbreviations, jargon and incomplete sentence constructions, the use of Greek and Latin language in the patient records and the different influences of Greek and Latin in Swedish patient record text.

The book followed by presenting the history of the different medical terminology systems, such as ICD diagnosis codes and SNOMED CT, and many other different terminologies and classification systems used in healthcare, for example, ATC drug coding and MeSH for literature indexing.

The challenges in mapping between different terminologies and also how to use ICD-10, SNOMED CT, ATC and MeSH for understanding the meaning of concepts written in free text in patient records have been discussed.

The book described the metrics and methods for evaluating both natural language processing tools and information retrieval tools. For evaluating the tools in a quantitative manner a gold standard is needed, or what in lay language is called "the correct answers". To produce this gold standard manually annotated text was produced; therefore, the manual annotation process for textual data was described as well as the evaluation of the annotation quality.

The annotated data can also be used for the development, training and evaluation of machine learning tools which was described in this book.

The book continued by explaining the basic tools used for natural language processing, and specifically tools adapted for clinical natural language processing. These basic tools included methods for segmentation and tokenisation of string of characters or tokens, the morphological processing of words, such as lemmatisation and stemming, compound splitting, abbreviation detection and expansion. Part-of-speech tagging to reveal the function or word class of each word was explained. Since many words may be misspelled spell checking and spelling correction were also described. To understand the structure of a text syntactical analysis can be carried out, or what is called (syntactical) parsing.

To understand the meaning of a sentence or a text semantical analysis and concept extraction was performed. This was called named entity recognition and negation detection. Negation detection is important since many of the symptoms in clinical text are negated in the reasoning process undertaken by the physician to find the disorder of the patient. Part of the semantic analysis consisted of processing steps to find relations between a drug and its effect or side effect. This was called relation extraction and more research is still needed on obtaining valuable results.

Clinical text also contains many temporal expressions that need to be resolved to understand when something occurred in a time line in the discourse. Therefore, tools have been developed to process temporality. These temporal processing tools were described and the different frameworks that were developed to deal with temporality in clinical text were explained.

Computational methods such as rule-based methods and machine learning methods were described, their differences and their pros and cons. Different ready-to-use tools were demonstrated. Since manual annotation of clinical text is costly (as well as all annotation work in general), active learning has been developed. Active

learning assists in choosing the most optimal and useful data for annotation such that is reduced annotation time.

To obtain electronic patient records for research ethical permission was needed. The process of applying ethical permission was explained. Ethics and privacy for the access and use of electronic patient records were discussed as well as the safe storage of the electronic patient records.

De-identification and pseudonymisation of sensitive data in electronic patient records text was described. Sensitive data is mostly data that can identify individuals, usually this data was either removed, by removing the whole data table that contained the sensitive data, or by anonymising the data in the data table. However, the sensitive data is often present in the clinical free text in form of names, addresses and telephone numbers, which need to be automatically identified and then removed, or pseudonymised which means to change the names to fake names or surrogates, the addresses to pseudo addresses etc.

The final main chapter presented a series of useful applications using the patient record data as input to these applications. Applications ranging from the presentation of the patient records in the form of an automatic summary to presentation of the basic concepts in patient record in the form of automatically extracted key words to support for the production of patient records text in the form of spelling correction systems, to big data analytics in the form of adverse event and healthcare associated infection detection and prediction. Natural language generation to produce useful text from various data that can be used by clinicians was also presented, including generation of natural language descriptions of SNOMED CT concepts. Other useful applications are comorbidity networks to find which disease co-occur or causes other diseases and various methods for hypothesis generation from clinical texts.

Various retrieval methods to extracts cohorts of patients with certain characteristics which is also an important domain were presented in the book. Using the clinical text as input for detecting adverse drug events, healthcare associated infections, pressure sores, patient falls, device failures, nutrition problems and surgical complication are also important and was presented in this textbook. Various clinical decision systems were presented as well as tools for automatic ICD to SNOMED CT mapping. One speech application was also presented to make early detection of dementia on patients.

The book ended with a presentation of the number of research networks and performed shared tasks in the clinical text mining.

At the beginning of the book a number of research questions were posed that are answered below:

- One main research question is: *Using artificial intelligence to analyse patient records: is it possible and will it improve healthcare?* The answer will be yes, if Artificial Intelligence is considered as a smart algorithm that will support humans and not as an independent will with a *soul* or a *deterministic will*.
- Another research question, which is rather long, is: *Can one process clinical text written in Swedish with natural language processing tools developed for standard Swedish such as newspaper and web texts, to extract named entities such as*

symptoms, diagnosis, drugs and body parts from clinical text? This question can partly be answered as yes, but of course since patient record texts are very domain specific this can not be carried out with the standard NLP tools, instead tools need to be adapted to the clinical domain.

Then followed a number of research sub questions as:

- Can one decide the factuality of a diagnosis found in a clinical text? What does *Pneumonia?* or *Angina pectoris cannot be excluded* or just *No signs of pneumonia?* really mean? The answer to this question is yes, to a certain level this can be carried out.
- *Can one determine of the temporal order of clinical events? Have the symptoms occurred a week ago or two years ago?* The answer is not completely yes, since it is difficult to extract temporal relations from free text and to align relative and absolute time points.
- *Can new adverse drug effects be found by extracting relations between drug intake and adverse drug effect?* The answer is no, the surveillance systems using clinical texts have not found new adverse drug effects yet, but this is a question of time and it will soon happen.
- *How much clinical text must be annotated manually to obtain correct and useful results?* The answer is probably at least 5000 annotated entities for NER, the proposed number of annotations gives usually rather good results after training. Regarding the task of relation extraction this is difficult to say, since the results are rather poor.
- *How can patient privacy be maintained while carrying out research in clinical text mining?* This is a difficult question to answer since the more data we leave behind traces on the different public digital places and business systems the easier it will be to track citizens, and our information can then be used to break privacy.

This book will probably become the standard text book for another 10 years until is updated by another textbook in the clinical text mining area.

This book has collected a vast amount of knowledge in the area clinical text mining combined with healthcare analytics and medicine. The knowledge was presented in a pedagogical and didactic way, explaining healthcare concepts. Healthcare concepts are described in the context of natural language processing and text mining methods. This book has leveraged future research and development of useful applications for healthcare. This was carried out by enabling the unlocking of previous knowledge and experience of thousands of clinicians, such as physicians and nurses, which was documented in thousands of electronic patient record repositories throughout the world. This unlocked knowledge will improve healthcare for humanity.

12.1 Outcomes

The book will become the standard scientific book for clinical text mining. The research results will be linked to courses at DSV, the Stockholm University, along with Master's in Health informatics that is jointly administered by Karolinska Institute and Stockholm University. Karolinska University Hospital, Stockholm County Council (SLL), other county councils in Sweden and companies such as Capish Knowledge and IMS Health Sweden, now IQVIA, are other stakeholders of this book.