Master thesis proposals - IT for Health

An overview and a comparison of the Electronic Patient Record Systems available on the market

The student should do an overview of the existing Electronic Patient Record Systems in the market and compare the regarding to some method that is used for comparing information systems. Measure points can be scalability, domain, use, interaction with other systems, implementation (customizations) costs, language, presentation the individual patient records, possible to extract statistics for business intelligence purposes.

Supervisor: Hercules Dalianis, professor, hercules@dsv.su.se and
Mia Kvist, M.D., Ph.D, IT for Health

Requirements engineering for Visualisation and Presentation of Patient Records in an Electronic Patient Records System

Capture the requirements from health personnel, physicians and nurses regarding what type of functionality they would like to have in a Electronic Patient Record System specifically in the domain of visualization and presentation of the individual patient records with respect to the need of the health personnel. Construct a number of user scenarios and present them for health personnel and use an iterative process on how to improve these user case scenarios for a final system.

Supervisor: Hercules Dalianis, professor, hercules@dsv.su.se and
Mia Kvist, M.D., Ph.D, IT for Health

Requirements engineering for Information Access in an Electronic Patient Records System

Capture the requirements from health personnel, physicians and nurses regarding what type of search, browse and/or information extraction functionality the clinicians would like to have in an Electronic Patient Record. Are ordinary search engine methods sufficient, or is the information need more complex? What type of information do clinicians want to be able to search for/extract, in what way, and how should it be presented? Construct a number of user scenarios and present them for health personnel and use an iterative process on how to improve these user case scenarios for a final system.

Supervisor: Sumithra Velupillai, fil lic, sumithra@dsv.su.se, and Mia Kvist, M.D., Ph.D, IT for Health

Automatic summarization of an electronic patient records to a discharge letter

Discharge letters is a summarization of a patient's visit at a hospital. It is a time consuming job for the physician to perform. Most of the information can be found in the electronic patient record and could be automatically processed into a summary. In this task one should construct an automatic summarization system that extracts the relevant information from a patient record and constructs a relevant discharge letter. The evaluation has to be carried out by comparing the constructed discharge letter with the real one, and assessing clinical relevance.

Supervisor: Hercules Dalianis, professor, hercules@dsv.su.se or Martin Hassel, Ph.D, xmartin@dsv.su.se, and Mia Kvist, M.D., Ph.D, IT for Health

Automatic detection of a diagnosis in a clinical free text written in Swedish

Electronic patient records contain information in structured fields such as patient name, age, gender, ICD-10 codes etc, but also in free text. It would be very handy if one could extract the diagnosis expression in the free text using automatic methods. Within the Stockholm EPR corpus we have manually annotated a small subset of the corpus for diagnosis expressions. In this master thesis one could use either rule based or machine learning based methods to detect the diagnosis in the free text. An evaluation of the quality of the result must be carried out.

Supervisor: Hercules Dalianis, professor, hercules@dsv.su.se or Sumithra Velupillai, fil lic, sumithra@dsv.su.se, IT for Health

Automatic abbreviation and acronym expansion in clinical free text written in Swedish

The free text parts of clinical records contain a large amount of domain-specific abbreviations (e.g. *p5*, *LE*) and acronyms, some of them ambiguous (e.g. *vb – vid behov* or *vederbörande*). In order to be able to extract useful information from the free text parts of clinical records, expanding abbreviations and acronyms to their full form, and disambiguating ambiguous ones, would be very useful. We will shortly have access to a dictionary of Swedish medical abbreviations and acronyms. This could be used in a rule-based system for automatically expanding abbreviations and acronyms. The dictionary could also be used as a seed set in a machine learning based system. In this master thesis, either a rule-based or machine learning based system will be developed for automatically expanding abbreviations and acronyms. An evaluation of the quality of the result must be carried out.

Supervisor: Sumithra Velupillai, fil lic, sumithra@dsv.su.se, IT for Health

Automatic assignment of ICD-10 diagnose codes combining free text and measurement data

For English there are a couple of systems that attempt to assign the correct diagnose given a passage of free text from a patient's electronic health record. This has however never been attempted for Swedish clinical text. Also, most systems just look at the text, or just the measurement data (blood pressure, body temperature, age, sex etc. in combination). We have access to a large amount of electronic patient files from a large Swedish hospital, which are already assigned with ICD-10 codes. These can together with associated measurement data be used both to train and to evaluate (so there is no strict requirement to know Swedish) a newly developed or ported/modified system using machine learning algorithms such as e.g. Latent Semantic Analysis and Decision Trees.

Supervisor: Martin Hassel, Ph.D, xmartin@dsv.su.se, IT for Health

Validation of Snomed using natural language generation

Construct a basic natural language generation system using a simple grammar and template system and as input SNOMED either in Swedish or English.

Natural language generation is a research area with several developed tools for creating natural language descriptions from input data in form of formal models (as UML) or from raw data as weather data, stock exchange data.

SNOMED stands for **S**ystematized **No**menclature of **Med**icine, and is an extensive classification system for medicine covering almost all areas of medicine. It contains around 350 000 concepts. The task in this master thesis work is to make it possible to automatically create a natural language description either in Swedish and English. (SNOMED is available in Swedish, English and some other languages) from parts of the Snomed.

Supervisor: Hercules Dalianis, professor, hercules@dsv.su.se, IT for Health

Clinical text editor, precursor – automatic suggestion of SNOMED terms

A common problem when attempting reuse of free text from patients' electronic health records, either for text mining or just simple search, is that the data is extremely noisy. For example, the same medical term can in the data be referred to with a score of (mis)spellings and ad hoc abbreviations. These need to be associated with proper standardized clinical terminology, such as SNOMED, in order for the computer to be able to map the meaning over different patient records. Approaches and metrics for this include string edit distance, co-occurrence statistics and syntactic restrictions. The resulting associations and metrics can for example be used as a form of "spelling correction" within a text

editor for the production of more consistent patient records. An evaluation of the quality of the associations/suggestions must be carried out.

SNOMED stands for **S**ystematized **No**menclature of **Med**icine, and is an extensive classification system for medicine covering almost all areas of medicine. It contains around 350 000 concepts.

Supervisor: Martin Hassel, Ph.D, xmartin@dsv.su.se, and Mia Kvist, M.D., Ph.D, IT for Health

Text mining comorbidity for Chronic Obstructive Pulmonary Disease COPD, in a large electronic patient record database

The student should use available tools such as diagnose detection and negation detection tools to extract valid diagnoses from free text in the domain of Chronic Obstructive Pulmonary Disease COPD. The diagnoses should be mapped to ICD-10 codes (International Classification of Diseases). Some of the patient records are already annotated with ICD-10 codes, these codes need also to be validated. The final step of the thesis is to present a comorbidity network with COPD in focus**.**

Supervisor: Hercules Dalianis, professor, hercules@dsv.su.se, IT for Health

Synthesizing a Clinical Swedish Treebank from General Language

Clinical language, as it occurs in e.g. electronic patient records, differs in several aspects from general language. Apart from the heavy use of jargon and ad hoc abbreviations there are also syntactic differences, such as a high degree of missing main subjects (most often nouns) and predicates (most often verbs). This master thesis project proposes to test if a Swedish general language treebank can be transformed into something syntactically akin to clinical Swedish by removing constituents. The theory is tested by training a dependency parser, the MaltParser, on the synthesized treebank and evaluating it on Swedish clinical text. The results can be compared to earlier experiments with a general language parser applied to the same data. This master thesis requires knowledge in the Swedish language but does not require formal knowledge in linguistics, although some training in interpreting parsed text will be required during the thesis work.

A treebank is a text corpus in which each sentence has been parsed, i.e. annotated with syntactic structure. These parses are often represented as trees.

Supervisor: Martin Hassel, Ph.D, xmartin@dsv.su.se, or Sumithra Velupillai, fil lic, sumithra@dsv.su.se, IT for Health

Bootstrapping a Swedish General Language Treebank with Clinical Text

Clinical language, as it occurs in e.g. electronic patient records, differs in several aspects from general language. Apart from the heavy use of jargon and ad hoc

abbreviations there are also syntactic differences, such as a high degree of missing main subjects (most often nouns) and predicates (most often verbs). This master thesis project proposes to test if a Swedish general language treebank can be extended with problematic sentences from a corpus of clinical text – the Stockholm EPR Corpus. The theory is tested by running a dependency parser, the MaltParser, on parts of the Stockholm EPR Corpus, correcting the output, adding the corrected parse trees to the Treebank and retraining the parser. Results from different iterations can then be compared. This master thesis requires knowledge in the Swedish language and presupposes at least some knowledge in linguistics, even though some training in interpreting parsed text will be required during the thesis work.

A treebank is a text corpus in which each sentence has been parsed, i.e. annotated with syntactic structure. These parses are often represented as trees.

Supervisor: Martin Hassel, Ph.D, xmartin@dsv.su.se, or Sumithra Velupillai, fil lic, sumithra@dsv.su.se, IT for Health

Visualizing comorbidity graphs

In medicine, comorbidity is the presence of one or more disorders (or diseases) in addition to a primary disease or disorder. By using diagnose code co-occurrence data this can be visualized as a graph where diagnose codes are the nodes while the weight or the length of the edges between the nodes can illustrate how strong the association is. The aim of this master thesis is to develop a prototype that visualizes these graphs in different ways and to perform a user study by interviewing a small group of clinicians on the utility of such a tool. The diagnose code co-occurrence data is collected from fixed fields in the Stockholm EPR Corpus, so there is no strict requirement to know Swedish.

Supervisor: Martin Hassel, Ph.D, xmartin@dsv.su.se, IT for Health