Master thesis proposal:
Detecting clinical entities using machine learning

We at the Clinical Text Mining Group, http://dsv.su.se/en/research/research-areas/health/clintextgroup, have a set of annotated clinical corpora in Swedish that each can be used for training and evaluation of different machine learning approaches as CRF, Random forest, etc, or evaluation of rule based approached to detect these instances.
The annotated corpora are:
1) Stockholm EPR PHI Corpus contains a set of annotated PHI instances.
2) Stockholm EPR Abbreviation Corpus contains a set of annotated abbreviations
3) Stockholm EPR Sentence Uncertainty Corpus (sommarmängden), contains annotated negated and speculative word and their corresponding expressions.
4) Stockholm EPR Diagnosis Factuality Corpus contains six different annotated factuality levels for diagnosis expressions.
5) Stockholm EPR Clinical Entity Corpus contains annotated findings, diagnoses, body parts and drugs.
Master students can chose to use the above corpora as resources for their thesis, or take part of any of the on-going projects in Health Informatics.
The Clinical Textmining Group presently has projects in the fields of Text simplification of clinical text, Detection of health care associated infections, and High-Performance Data Mining for Drug Effect Detection.

Prerequisites
Knowledge in data- and textmining. Knowledge in a programming language or a scripting language. Knowledge in Swedish or English language.

Contact:
Professor Hercules Dalianis, Stockholm University hercules@dsv.su.se, ph 08-674 75 47
Clinical Text Mining Group
http://dsv.su.se/en/research/research-areas/health/clintextgroup

Master thesis proposal:
Detection of new side effects of Drugs using a database of patient records

This master thesis proposal is part of a research project called High-Performance Data Mining for Drug Effect Detection, http://dsv.su.se/en/research/research-areas/datatextmining/drug-effect funded by Swedish Foundation for Strategic Research. We have access to the Stockholm EPR corpus with de-identified electronic patient records from Karolinska University hospital and the TakeCare patient record system from the years 2009-2010 encompassing 600 000 patient from 500 clinical units. These patient records contain structured information such as patient age, gender, ICD-10 codes, ATC drug codes, blood and lab values, admission and discharge dates, as well as unstructured textual information, containing assessment notes, discharge notes, nursing notes, etc.
Our research group can provide automatic methods that can help to answer your research questions regarding side effects of drugs as well as drug interactions. We need a toxicologist or equivalent that can formulate hypotheses that we can test on our clinical material. The master student may also provide us with annotations of drug and side effects so we can train our automatic tools explore the whole data base.

Prerequisites for the master student: Toxicology, basic level, interest in drug interaction, side effect of drugs.

Contact:
Professor Hercules Dalianis, Stockholm University hercules@dsv.su.se, ph 08-674 75 47
Clinical Text Mining Group
http://dsv.su.se/en/research/research-areas/health/clintextgroup

Master thesis proposal:
Automatic mapping of ICD-10 to SNOMED-CT

ICD-10 diagnosis codes, is a large set of codes used to diagnoses diseases, it contains around 16 000 concepts, that are available both in Swedish and English and several other languages. SNOMED-CT, is a very large hierarchical medical terminology, it contains around 300 000 concepts, that are available both in Swedish and English and several other languages. In English synonyms are available but not in Swedish. There have been manual approaches on mapping ICD-10 to SNOMED. In this master thesis proposal we are going to map a subset of the ICD-10 and SNMOED with each other, preferable in one domain. The mapping of the ICD-10 to SNOMED will be carried out in an automatic way, by matching each ICD-10 concepts name and textual description with each SNOMED concept name and textual description. We will also use the hierarchical structure of SNOMED for the mapping. After the mapping has been carried out we will do an evaluation of how well the mapping performed.

Prerequisites
Knowledge in a programming language or a scripting language. Knowledge in Swedish or English language.

Contact:
Professor Hercules Dalianis, Stockholm University hercules@dsv.su.se, ph 08-674 75 47
Clinical Text Mining Group
http://dsv.su.se/en/research/research-areas/health/clintextgroup

Master thesis proposal:
Automatic ICD-10 diagnosis code assignment

An ICD-10 diagnosis code is assigned to the diagnosis of a patient by the physician and written in the patient record. This work is not trivial since there are over 16 000 diagnosis code to choose among. We would like to construct a tool that assigns codes automatically. We have a previously annotated clinical corpora - the Stockholm EPR Corpus that contains over 600 000 in inpatients and their corresponding 5 million ICD-10 codes assigned by physicians during 2006-2010. Stockholm EPR corpora encompasses over 500 clinical units from the Karolinska University Hospital. The corpora contains both unstructured doctor notes and discharge letters (in Swedish). and structured information such as the ICD-10 diagnosis codes, patient age, gender, ATC drug codes, blood and lab values, admission and discharge dates.
The student should use some machine learning based technique or rule based method to assign the correct ICD-10 code to the discharge letter and then evaluate the constructed tool.

Prerequisites
Knowledge in data- and textmining. Knowledge in SQL, Knowledge in a programming language or a scripting language. Knowledge in Swedish or English language.

Contact:
Professor Hercules Dalianis, Stockholm University hercules@dsv.su.se, ph 08-674 75 47
Clinical Text Mining Group
http://dsv.su.se/en/research/research-areas/health/clintextgroup

Master thesis proposal:
Drug view – presentation of drug use of Stockholm EPR corpus

We would like to construct a navigator tool for drug use in a population of inpatients at Karolinska University Hospital. We have previously constructed a demonstrator that presents comorbidities called Comorbidityview, http://www2.dsv.su.se/comorbidityview-demo/, of the same population.
The population origins from the Stockholm Electronic Patient Corpus that contains 600 000 inpatients from 500 clinical units at Karolinska University Hospital, their drug prescriptions expressed in ATC-codes, time points for drug prescriptions etc, from the years 2009-2010.
The user of Drug View can be a toxicologist or a clinical researcher, but also persons in the hospital management that would like to collect and present statistics.
We have previously used Cytoscape as our presentation that can be used for this master thesis proposal. The research question can be which drug pairs are the mostly used?

Prerequisites
Knowledge in data- and textmining. Knowledge in SQL, Knowledge in a programming language or a scripting language. Knowledge in Swedish or English language.

Contact:
Professor Hercules Dalianis, Stockholm University hercules@dsv.su.se, ph 08-674 75 47
Clinical Text Mining Group
http://dsv.su.se/en/research/research-areas/health/clintextgroup

Master thesis proposal:
Cascaded Diagnosis Code Classification of Clinical Notes using Distributional Semantics

Diagnosis codes (ICD-10) need to be assigned to health record entries for administrative and epidemiological purposes. This is a time-consuming and non-intuitive task that could be supported by information technology. One approach to achieve this is to exploit the availability of large amounts of clinical notes with manually assigned diagnosis codes to create a statistical model that is able to predict possible diagnosis codes for new clinical notes that have not yet been assigned any diagnosis codes. A method that tackles this problem is currently being developed in the Clinical Text Mining Group at DSV by using Random Indexing (a model of distributional semantics) to learn semantic similarity estimates between the words used in clinical notes and diagnosis codes. These (semantic space) models are then used to generate a list of possible diagnosis codes for a given note. It would, however, be desirable to reduce the complexity of this challenging multi-class, multi-label classification problem (there are over 12,000 diagnosis codes). This thesis would propose and investigate a potential solution that exploits the hierarchical structure of ICD-10 in an attempt to improve the current performance of the method.

Contact:
Dr Martin Duneld, Stockholm University, xmartin@dsv.su.se ph 08- 674 74 14
Clinical Text Mining Group
http://dsv.su.se/en/research/research-areas/health/clintextgroup

Master thesis proposal:
Medical Synonym Extraction with a Semantic Space Created from Multiple Text Sources

Models of distributional semantics have been applied to a large corpus (text collection) in order to extract potential synonyms of a given term automatically. These models exploit word co-occurrence patterns in large corpora to capture the semantic similarity of terms (two terms that repeatedly appear in similar contexts are assumed to share semantic properties). In the medical domain, semantic spaces have been constructed from a corpus of clinical text (free-text in health records) in order extract synonyms of medical terms. Clinical data is, however, sometimes scarce and difficult to get ahold of (and models of distributional semantics rely on large amounts of data). It would therefore be interesting if such data could be supplemented with documents from other, more readily available sources, such as Wikipedia, blog posts, forums and Twitter. This thesis would investigate whether a semantic space constructed from multiple text sources would perform better on the synonym extraction task.

Contact:
Dr Martin Duneld, Stockholm University, xmartin@dsv.su.se ph 08- 674 74 14
Clinical Text Mining Group
http://dsv.su.se/en/research/research-areas/health/clintextgroup