Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care

Hercules Dalianis, Martin Hassel, Aron Henriksson, Maria Skeppstedt
Department of Computer and Systems Sciences (DSV), Stockholm University
Forum 100, 164 40 Kista, Sweden
{hercules, xmartin, aronhen, mariask}@dsv.su.se

Abstract

The care of patients is well documented in health records. Despite being a valuable source of information that could be mined by computers and used to improve health care, health records are not readily available for research. Moreover, the narrative parts of the records are noisy and need to be interpreted by domain experts. In this abstract we describe our experiences of gaining access to a database of electronic health records for research. We also highlight some important issues in this domain and describe a number of possible applications, including comorbidity networks, detection of hospital-acquired infections and adverse drug reactions, as well as diagnosis coding support.

1. Introduction

Health care costs increase rapidly with aging populations. However, at the same time, information technology is making health care more efficient. Information technology has made a significant contribution to improving health care, e.g. in the form of digitized radiology images, monitoring equipment for heart and blood pressure and automatic chemical analysis of body fluids. Moreover, health records have become digitized in many parts of the world, which also allows patient data to be stored in centralized databases. These patient databases contain documentation written by experienced health care personnel and record a patient's symptoms, diagnoses, treatments, prescribed drugs, etc. In countries where each patient has a unique social security number, as in Scandinavia, longitudinal studies of patients can be undertaken across clinics. In short, clinical documentation contains valuable information concerning the treatment of real patients; however, this is rarely available for research due to its sensitive nature.

2. Our Data

2.1 Obtaining Data

In order to gain access to clinical data for research, ethical permission from the local vetting board is required. The ethical application must contain a clear research question and also describe how the data will be treated and stored. Furthermore, it must be shown to the hospital management that the research will be beneficial for them and that their data will not be misused.

2.2 Storing Data Securely

In our case, the data were obtained without patient names and the social security numbers were replaced with unique serial numbers. However, the textual fields in the data may still contain sensitive identifying information, e.g. names, phone numbers and addresses. Our data are stored on encrypted servers without network access in a locked and alarm-equipped server room.

2.3 Experts in Medicine

Medical domain experts are needed to interpret the noisy text due to the professional language that pervades health records. We have had access to three senior physicians for the annotation and interpretation of our data.

2.4 Features of the Stockholm EPR Corpus

The Stockholm EPR (Electronic Patient Record) Corpus contains data from over 512 clinical units from Stockholm City Council encompassing the years 2006–2010 and over one million patients. The whole corpus contains over 500 million tokens. Certain statistics for the years 2006–2008 can be found in Dalianis et al. (2009).

3. Other Clinical Databases

The following is a brief description of five different clinical databases from three countries that have been made available for research:

- The i2b2[1]corpus is a clinical corpus consisting of approximately 1,000 notes in English.
- The CMC[2] clinical corpora consists of 2,216 records in English from children's radiology departments, which have been automatically and manually de-identified (Pestian et al., 2007).
- The De-id[3] corpus consists of 412,509 nursing notes and 1,934 discharge summaries written in English.
- A Finnish clinical corpus [4] containing 2,800 sentences (17,000 tokens) from nursing notes, which have been manually anonymized.
- THIN[5] contains the records of almost 11 million patients from general practices. The records are from the years 1986–2003 and are written in English. The database specializes in pharmacoepidemiology (see Lewis et al., 2007).

4. Applications

There is a nice overview of NLP applications using health record texts in Meystre et al. (2009). Below are some examples of what our research group has worked on using our clinical database, the Stockholm EPR Corpus.

[1] http://www.i2b2.org
[2] http://computationalmedicine.org/catalog
[3] http://www.physionet.org/physiotools/deid
[4] http://bionlp.utu.fi/clinicalcorpus.html
[5] http://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database

4.1 Comorbidity Networks

Medical records databases can be used to create comorbidity networks, i.e. a network of patients suffering from at least two diseases. These can be explored from the perspective of age, gender and specific ICD-10 diagnosis codes; an example can be seen in our demo version of Comorbidity View[6], which uses structured data from the Stockholm EPR Corpus (2006–2008). This approach can also be used to align diseases, drugs and patient groups. By extracting relevant information from the free text, this part of the health record can be used to build various co-occurrence networks. We have annotated clinical text and developed tools based on machine learning techniques that recognize symptoms, diagnoses and drugs in the free text (Skeppstedt et al., 2012). These tools are also able to recognize if the mentioned symptoms and diagnoses are in a negated or speculative context (Velupillai, 2012).

4.2 Detecting Hospital Acquired Infections

Around ten per cent of all in-patients are affected by hospital acquired infections (Humphreys & Smyths, 2006), however many of these hospital-acquired infections (HAI) are not reported correctly to the hospital management. One way to achieve automatic detection of HAIs is to have computers analyse the free text and structured information in the patient records and search for indications of HAIs, e.g. that a patient has been treated with a particular antibiotic drug after the use of a catheter (Proux et al., 2011). We are currently annotating 180 Swedish patient records, some of which contain HAIs, in order to evaluate our rule-based and machine learning tools that retrieve information from the records and determine whether they contain a possible HAI or not.

4.3. Detecting Adverse Drug Reactions

Pharmaceutical drugs are developed through laboratory tests and clinical trials; however, these methods are very expensive. To reduce costs one is also doing simulation using mathematical models but such methods are not always sufficiently robust to identify the effects and side effects of drugs. We plan to use our medical database to confirm the known effects of drugs, but also to find unknown adverse drug reactions (ADR). A initial approach is presented in Henriksson et al. (2012), where Random Indexing (a form of distributional lexical semantics) is applied to a part of the Stockholm EPR Corpus in order to extract drug-symptom pairs. Around 50 per cent of the drug-related words were conceivable ADRs, while ten per cent were known and documented ADRs.

4.4. Diagnosis Coding Support

Accurate coding of diagnoses enables statistical monitoring of symptoms, diseases and treatments. This is, however, a time-consuming and thus costly administrative task that has led to attempts to provide computer-aided coding support. We are working on a method that uses Random Index-based word spaces, containing co-occurrence information of textual units and diagnosis codes, to recommend possible codes for an uncoded clinical note (see, e.g., Henriksson et al., 2011).

[6] http://dsv.su.se/en/research/health/comorbidityview/demo/

5. Conclusions

Obtaining ethical permission and secure access to the data is probably the most difficult part of this research for the researcher. Other challenges involve the problem of interpreting and annotating the clinical data which requires the aid of medical domain experts such as physicians and nurses. The noisy data are also an obstacle for our natural language processing tools. However, given the extra resources afforded by the records, the domain presents interesting and relatively uncharted territory for language technology that can contribute to improved health care.

References

Dalianis, H., Hassel, M. and Velupillai, S. (2009). The Stockholm EPR Corpus – Characteristics and Some Initial Findings. In Proceedings of ISHIMR 2009, pp. 243–249.

Henriksson, A., Hassel, M. and Kvist, M. (2011). Diagnosis Code Assignment Support Using Random Indexing of Patient Records – A Qualitative Feasibility Study. In Proceedings of AIME, 13th Conference on Artificial Intelligence in Medicine, Springer-Verlag, pp. 348–352.

Henriksson, A., Kvist, M., Hassel, M. and Dalianis, H. (2012). Exploration of Adverse Drug Reactions in Semantic Vector Space Models of Clinical Text. In Proceedings of ICML Workshop on Machine Learning for Clinical Data Analysis.

Humphreys, H. and Smyths, E.T.M. (2006). Prevalence surveys of healthcare-associated infections: what do they tell us, if anything? *Clin Microbiol Infect* 12, pp. 2–4.

Lewis, J., Schinnar, R., Warren, B., Bilker, W., Wang, X. and Strom, B. (2007). Validation studies of the health improvement network (THIN) database for pharmaco-epidemiology research. *Pharmacoepidem Drug Safe*, pp. 393–401.

Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C. and Hurdle, J.F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform,* pp. 128–144.

Pestian, J., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K. and Duch, W. (2007). A shared task involving multi-label classification of clinical free text, BioNLP: Biological, translational, and clinical language processing, ACL, pp. 113–120.

Proux, D., Hagège, C., Gicquel, Q., Pereira, S., Darmoni, S., Segond, F. and Metzger, M-H. (2011). Architecture and Systems for Monitoring Hospital Acquired Infections inside a Hospital Information Workflow. In Proceedings of the Workshop on Biomedical Natural Language Processing, RANLP-2011, pp. 43–48.

Skeppstedt, M., Dalianis, H., Kvist, M. and Nilsson, G.H. (2012). Detecting drugs, disorders and findings in Swedish Clinical text using CRF, (forthcoming).

Velupillai, S. (2012). Shades of Certainty: Annotation and Classification of Swedish Medical Records. PhD thesis, Department of Computer and Systems Sciences, (DSV), Stockholm University.