The Stockholm EPR Corpus – Characteristics and Some Initial Findings

Hercules Dalianis, Martin Hassel and Sumithra Velupillai

Department of Computer and Systems Sciences, DSV, KTH/Stockholm University,
Forum 100, 164 40 Kista, Sweden, {hercules,xmartin,sumithra}@dsv.su.se

This paper describes the characteristics of the Stockholm Electronic Patient Record Corpus (the SEPR Corpus), an important resource for performing research on clinical data. The whole SEPR corpus contains over one million patient records from over 2 000 clinics. We compare parts of the SEPR corpus with the Swedish PAROLE Corpus and describe the differences and similarities. We also describe a set of experiments we have initiated on the SEPR corpus, experiments whose outcome we believe will, in the long run, contribute to the medical research as well as the daily life of the clinician. Moreover, this corpus contains characteristics that are very interesting from a linguistic point of view, such as domain specific compounds and abbreviations, and various narratives.

Keywords

automatic ICD coding, clinical free text in Swedish, hypothesis generation, text mining, uncertainty in diagnoses.

1. Introduction

In recent years the interest for performing research on biomedical and clinical data within language technology has increased immensely. There are many reasons for this. For instance, such domain specific data contains vocabularies and language use that is very interesting and not previously studied from a linguistic point of view. Also, such data contains a potentially large amount of information that could be useful for other research areas such as Medical Informatics, Epidemiology and Biomedicine, to mention only a few. However, research on clinical data is still very limited, since many privacy issues need to be solved and many ethical aspects need to be taken into account when working with Electronic Patient Records (EPRs). In order to obtain access to clinical data, issues concerning integrity and privacy need to be properly secured. Moreover, the data needs to be fully de-identified.

This paper describes some general characteristics of a large corpus of EPRs written in Swedish, which our research group plans to use for further research. We believe such research will be very interesting within the area of for instance Information Access, Text Mining and Medical Informatics. The corpus has been granted access by the hospital management from which the corpus is derived after approval from the Regional Vetting Board.

Natural Language Processing (NLP) research within the biomedical domain is currently very vivid. In this paper we focus on research performed on clinical data, which may be viewed as a sub-domain of the biomedical domain.

2. Related work

Clinical data is, per se, sensitive, since it contains personal information about health and social status of individuals, which puts the individual at risk of being identified. Access to clinical data for research purposes is therefore in many cases very difficult. However, in the UK for instance, there is a medical research database created within the Health Improvement Network called THIN, [1]. This database contains anonymised patient records, though there is not much information regarding the free text contents of this data.

At the Mayo Clinic, a set of information extraction tools has been developed specifically designed for clinical data, [2]. These tools have primarily been used on clinical data from the clinic itself. This work is also part of the Open Health Natural Language Processing OHNLP Consortium [3] (which is an initiative to establish an open source consortium for research on clinical and medical NLP research. Moreover, efforts on evaluating information extraction research on clinical data have been achieved through shared tasks, such as the i2b2 smoking status identification challenge in 2006, [4], and the Medical NLP challenge in 2007, [5]. The corpus used in the i2b2-challenge is described in more detail in [6]. A thorough review of information extraction research performed on clinical data is presented in [7].

However, most research on clinical data has been performed on EPRs written in English. For Swedish, there is still a lot of research needed, both regarding the creation of EPR corpora, and regarding the creation of NLP tools that could be useful within this domain. Some research has been performed on smaller Swedish clinical corpora. For instance, in [8], research carried out on discharge letters written in Swedish, with promising results, is described.

There are, to our knowledge, not many studies that compare the contents contained in different electronic medical record systems from a medical point-of-view. Naturally, different systems have different solutions when it comes to how necessary information is recorded, and how the relationship between structured and unstructured, free text entries are solved. It seems to be common, at least in Sweden, to have free text entries linked to keywords covering the central parts in the health care process, i.e. *Anamnes* (conversation with the patient), *Status*, *Bedömning* (assessment, analysis) and *Åtgärd* (planned action), [9]. Such entries are often added through a controlled vocabulary, which may differ between different hospitals and clinics. A study from 1992 [9] showed that the contents of these keywords were very similar between different institutions.

3. The Stockholm EPR Corpus - Characteristics

In Sweden a unique social security number for each citizen is used from birth to death. This fact makes it easy to register each encounter a citizen has with the health care system. This also makes it effective to build large centralized database systems with each patient represented in each clinic. We have gained access to a large body of electronic patient records, the Stockholm EPR corpus, from one of the largest Electronic Medical Record systems in Sweden, encompassing the years 2006, 2007 and the first half of 2008 and covering clinics and their patients from one of the largest county councils in Sweden.

Table 1 Statistics from the first five months of 2008 of the corpus.

2008 5 months	Total	SEPR Corpus	Percent
Men	408 144	188 238	46%
Women		219 906	54%
Average no of tokens per record		269	
Free text categories	6 164	2 631	43%
ICD-10 codes	35 185	16 211	46%
Missing ICD coding		138 890	34%
No of clinics		888	

This data consists of both structured information such as gender and age of the patient, as well as unstructured information in the form of free text. A calculation of a subset of the Stockholm EPR corpus showed that around 40 percent of the data entries are unstructured, the rest being structured. The unstructured entries contain more data on the other hand and constitute a larger total amount.

Moreover, there is a lot of duplicate information, as there are many authors to each record and patients may visit different clinics.

The free text is in itself semi-structured, as free text entries are put in connection to a set of free text categories that can be used in the system (and that may vary from clinic to clinic), such as *Bedömning* (Assessment), *Aktuell status* (Current status), *Social Bakgrund* (Social Background), etc. We have analyzed one fifth of the corpus, i.e. the first five months of 2008.

We can see in Table 1 that not even half of the free text categories are used and also that not even half of the ICD-10 codes [10], are used to describe symptoms and diseases. Notable is the fact that 34 percent of the records seem to lack ICD-10 coding (these might have ICD-codes from previous years, though). We have observed from our studies that the average number of tokens used in each record is not evenly distributed, some clinics and some records contain more free text than others, a fact that also holds within clinics. We have compared the SEPR corpus with the Swedish standard corpus PAROLE [11], see Table 2, in order to get a picture over the differences and similarities in distributions and vocabulary. All results are taken from raw data, i.e. no pre-processing has been performed on either corpus.

Table 2 Some comparisons between the SEPR corpus and the Parole corpus

	SEPR Corpus		**PAROLE Corpus**	
No of tokens	109 663 052		18 765 888	
No of types	853 341		550 766	
Frequencies tokens				
hapax legomena = 1	467 706	55%	292 217	53%
dis legomenon =2	107 636	13%	75 752	14%
tris legomenon=3	51 161	6%	36 376	7%
< 10	732 150	86%	481 380	87%
> 100	34 245	4%	12 881	2%
Average token length				
hapax legomena = 1	11.76		12.70	
freq > 100	8.919		7.553	
freq > 100 000	**3.909**		2.864	
All tokens	**5.478**		5.440	
Vocabulary comparison: SEPR and PAROLE Corpus				
Matching tokens		97 738 798		89.1%
Non-matching tokens		11 924 254		10.9%
Matching types		121 020		14.2%
Non-match types		732 321		85.8%

What we can see in Table 2 is that the distributions between the SEPR corpus and the PAROLE Corpus are very similar, but with slightly longer tokens in the frequencies above 100 and above 100 000 for the SEPR corpus.

Swedish morphology is more complex than English and contains a large amount of inflections. Swedish also produces compounds in a very productive, and often creative, way. Therefore there is a clear need for both stemmers or lemmatizers as well as decompounders in order to improve information retrieval and to create more representative language models. How such tools would work on the SEPR corpus needs to be investigated.

The SEPR corpus is interesting since it contains various writing styles. The records contain writings that are handovers for different shifts (often written in some form of dialogue), descriptions of the social situations with family, social care, and home conditions, descriptions and discussions of symptoms and diagnoses where other clinical specialists are consulted, careful descriptions of medical treatments, etc. Generally speaking the text is very uneven in the "writing" quality and also contains many ad-hoc abbreviations or non-standard and very domain dependant abbreviations such as *p5.*

If we take a look on some spelling errors in the SEPR corpus*:*

slemninnor, (mucous jembrane),

tarapeut (tharapist),

behasndlingsbeslut (treastment decision),

pllacera (pllace),

branmorska (mdiwife),

We can observe that most spelling errors are so called Damerau type errors, which often are keyboard slipping related errors due to fast typing. In a study carried out on patient records written in French, they found that there are up to 10 percent spelling errors in patient records (compared to 1-2 percent spelling errors in ordinary typed text), [12].

Our findings are also supported by the observations in [13], where they found that a tagger trained on a clinical corpus performed better than a tagger trained on general English when used on clinical free text.

If we take a look at some compounds in the SEPR corpus, we can observe that they are very productive:

antiepileptikadoserna, (antiepileptics dosage),

korallstensformation (formation of coral stones),

strålbehandlingsplaneringsdatortomografi, (radiation-treatment-planning-computer-tomography)

leverkirurgkonferens (conference for liver surgeons)

We realize easily that a decompounder would in these cases improve information retrieval and analysis.

The free text entries may vary in length, style and content. Some entries are very short and express uncertainty, e.g:

Viros som genes till feber? (Viros as source for fever?)

Others are longer (this example also displays a content where the medical assessment is fairly certain*):*

> *Igår och idag helt stabil i sternum vid palpation och helt oretat sår i övrigt. Odlingar hittills intesat något annat än förekomst av jästsvamp i urin, dock ett observandum då pat är immunosupprimerad efter tidigare NTx. (Yesterday and today quite stable in the sternum at palpation and completely not irritated wounds in general. Trials so far has not proved anything other than the presence of yeast in the urine, however, an observandum that pat are immunosuppressed after previous NTx)*

Furthermore, some entries are in a dialogue-style, containing a lot of language errors:

> *Beklagar nissförstånd rek ayt provar mindre smaker som innehåller mindre Kolhydrater 8vilket pat benämner som smaken sött som diasip, komplett näring naturell samt provide x-tra tomat. Ut tar upp dessa till avd för utprovning .Vi ska se vad vi kna göra med de näringsdrycker som finns i hemmet då pat är åter hemma... (Sorry for the nisunderstanding rec tto try less flavours that contain less Carbohydrates 8which pat name as taste sweet like diasip, complete nutrition natural as well as provide x-tra tomato. Ut takes these to clin for try out. Let's see what we cna do with the nutrition drinks in the house when pat is back home...).*

4. Planned experiments

In this section we describe a set of experiments that we plan to perform and some initial findings on the SEPR corpus. These experiments are chosen to give an outcome that we believe will, in the long run, contribute to the medical research as well as the daily life of the clinician. To facilitate this we will primarily tailor existing pre-processing tools for Swedish, in order for them to handle this domain-specific vocabulary, such as lemmatizers, decompounders, PoS-taggers and syntactic analyzers.

4.1 Hypothesis generation

The data contained in the EPRs contains a potentially large amount of information that is previously unknown and largely unchartered. Such research falls within the area of Text Mining. We have prepared a set of experiments where the idea is to generate new hypotheses regarding medical conditions through document clustering techniques, by exploiting both the structured and unstructured information in the EPRs.

In [14] this method is introduced and applied on a large set of epidemiological questionnaire data where the hypothesis that *farmers smoke less than the average* was generated from revealing relations between a closed answer regarding smoking habits and an unstructured answer where the respondents described their occupation in free text. The hypothesis was confirmed through a literature study.

We have applied this method on a subset of the Stockholm EPR Corpus containing records of patients from geriatric clinics. We have experimented on free text heavy entries such as *Bedömning* (Assessment) and the structured entry *Gender*. With no prior medical knowledge on common geriatric diagnoses we have generated the hypothesis that *women suffer from brittleness of the bones more often than men*. This hypothesis needs to be tested and confirmed properly, but preliminary literature studies support our finding.

4.2 Uncertainty and certainty detection

The free text parts in the EPRs that describe situations where diagnoses are stated or reasoned about may contain many expressions of uncertainty and speculation. In the Stockholm EPR Corpus, the free text entry *Bedömning* (Assessment) contains a lot of reasoning about the patient's status and planned actions.

Such language use is very interesting from an Information Extraction (IE) perspective – and related Information Access perspectives – and is not captured well by standard language models used in IE systems. We believe that it would be of great interest to identify such parts in the data in order to distinguish the degrees of certainty or uncertainty in these texts, especially across clinics and even diagnoses and diagnose codes.

We have initiated work on identification of speculative language in the Stockholm EPR Corpus by annotating a subset divided into randomly picked sentences from the total amount of free text written under the entry *Bedömning (Assessment)*. In order to make the annotated set comparable to similar research, we have based the annotation work on the ideas and guidelines for the BioScope corpus [15], in which clinical data is included.

From a preliminary analysis we have found that from 7 900 sentences, 8 447 instances of uncertain, certain or undefined expressions were identified. Undefined expressions are expressions that could not be classified as either certain or uncertain. In some cases, sentences have been divided into sub-parts, where some parts are annotated as uncertain and others as certain, which explains the larger total amount of annotations compared to the number of sentences. Within each sentence, the number of words differed from only a few to several. In total, 16 percent (1 344 instances) of the instances were annotated as uncertain expressions. Within these, 1 699 words were annotated as speculative words. 12 percent (1 016 instances) of the original 8 447 instances were also annotated as negations. These results correlate well with the findings in [15]. 7 percent were annotated as undefined, leaving 77 percent annotated as certain expressions. Once we have analysed the annotated subset in more detail, we will be able to develop a set which could be used for developing tools that identify such instances automatically.

4.3 Diagnose code suggestion

A common problem for clinicians is to choose the right ICD-10 code, or even to navigate among the 35 188 codes currently active in Sweden, for the correct classification of the symptoms and diagnosis of the patient. A solution to this problem could be to let a computer program propose a number of ICD-10 codes based on the entered textual description of the symptoms or diagnosis – codes that the clinicians then could choose from. We have utilised a word space model, [16], built on the entire scope of free text fields as well as the ICD-10 codes entered in conjunction with these fields. This word space can then be used to look up what symptoms or medical terms that in actuality relate to a certain diagnose code, as well as looking up what codes relate best to a certain set of words found in an unclassified record.

Here follows one example on some initial runs. We have entered hosta (cough) to the vector space and obtained ten examples on ICD-10 codes, where the first has the highest rank.

Hosta (cough)

J18.9 - Pneumoni, ospecificerad (Pneumonia, unspecified)
J15.9 - Bakteriell pneumoni, ospecificerad (Bacterial pneumonia, unspecified)
H66.9 - Mellanöreinflammation, ej specificerad som varig / icke varig
(Otitis media, unspecified)
J20.9 - Akut bronkit, ospecificerad, (Acute bronchitis, unspecified)
B34.9 - Virusinfektion, ospecificerad, (Viral infection, unspecified)
G96.9 - Sjukdom i centrala nervsystemet, ospecificerad
(Disorder of central nervous system, unspecified)
I50.9 - Hjärtinsufficiens, ospecificerad (Heart failure, unspecified)
F48.9 - Neurotiskt syndrom, ospecificerat (Neurotic disorder, unspecified)
C34.9 - Icke specificerad lokalisation av malign tumör i bronk & lunga
(Bronchus or lung, unspecified)
L64.9 - Androgen alopeci, ospecificerad (Androgenic alopecia, unspecified)

This result is very encouraging and we will continue to work on this approach. We will also in the near future present the results with an evaluation of the quality of the ICD-10 code suggestions.

4.4 Synonym generation

We have used a similar word space approach [16] to dynamically generate lists of closely related words. Such lists are of great interest for terminologists, both in the task of creating as well as maintaining guidelines for consistent use of terminology – a key point in any larger information management system. By capturing the paradigmatic context of a word (i.e. words that are used in similar contexts in *different* documents), it is possible to generate associative words, which can be interpreted as synonyms. From some initial experiments we have, for instance, generated the following words:

rosslig (wheezy)

andning (breathing)
slemmig (phlegm)
låter (sounds)
hostar (coughs)

Such word lists could also be useful for language generation on a grander scale. For instance, we find the application of generating patient-friendly versions of a patient's file where the, to the patient, medical gobbledygook is identified and generalised to fit the patient's level of domain competence is a very important future research direction.

5. Conclusions

In this paper we have described the Stockholm Electronic Patient Corpus and its characteristics. We have found that the corpus contains slightly longer words than a Swedish standard corpus and also, as expected, that the vocabulary differs a lot from general Swedish. Also described is a set of initiated experiments we have carried out on the corpus. We have provided some preliminary results, which will be further analysed. What we can see is that EPRs contain from several perspectives interesting data, both in the form of vocabulary and in the form of narrative – data that potentially contains a large amount of information that could be used for both further NLP research as well as further research in for instance Medical Informatics and Epidemiology.

The Stockholm EPR corpus is partly de-identified and is therefore currently not available for a broader group of researchers. We strive to make a subset of the Stockholm EPR corpus available but to make this possible we need to de-identify the corpus fully for patient confidence reasons. We plan to further analyse the domain-specific properties of the Stockholm EPR corpus in order to identify which language technology tools need adaptation and which could be used directly.

References

[1] Bourke A, Dattani H and Robinson M. Feasibility study and methodology to create a quality-evaluated database of primary care data. In: Inform Prim Care. 2004;12(3):171-7

[2] Savova G, Kipper-Schuler K, Buntrock J, and Chute C. UIMA-based clinical information extraction system. In: Proceedings of LREC 2008 -- The 6th International Conference on Language Resources and Evaluation: Towards enhanced interoperability for large HLT systems: UIMA for NLP. Marrakech, Morocco.

[3] OHNLP. Open Health Natural Language Processing, https://cabig-kc.nci.nih.gov/Vocab/KC/ index.php/OHNLP. Accessed April 23, 2009.

[4] i2b2. Informatics for Integrating Biology and the Bedside. Available at: https://www.i2b2.org. Accessed April 23, 2009.

[5] Pestian JP, Brew C, Matykiewicz PM, Hovermale DJ, Johnson N, Cohen KB, Duch W. A shared task involving multi-label classification of clinical free text. In: Proc. of ACL BioNLP; 2007 Jun; Prague.

[6] Uzuner Ö T C, Sibandam Y, Luo Y, and Szolovits P. A De-identifier for Medical Discharge Summaries. In: Journal of Artificial Intelligence in Medicine, Jan;42(1):13-35.

[7] Meystre S M, Savova G K, Kipper-Schuler K C, and Hurdle J E. Extracting information from textual documents in the electronic health record: a review of recent research. In: IMIA Year-book of Medical Informatics 2008. Methods Inf Med 2008; 47 Suppl 1:138-154.

[8] Kokkinakis D, and Thurin A. Identification of Entity References in Hospital Discharge Letters. In: Proc. 16th Nordic Conf. of Computational Linguistics NODALIDA-2007. University of Tartu, Tartu, 2007.

[9] Peterson G, and Rydmark M. 1996. Medicinsk Informatik. Almqvist & Wiksell Medicin, Liber Ut-bildning, (In Swedish)

[10] ICD-10. International Classification of Diseases (ICD), http://www.who.int/classifications/icd/en/. Accessed April 23, 2009.

[11] Gellerstam M, Cederholm Y, and Rasmark T. The bank of Swedish. In: Proceedings of LREC 2000 -- The 2nd International Conference on Language Resources and Evaluation, pages 329–333, Athens, Greece.

[12] Ruch P, Baud R, and Geissbühler A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. In: Artificial intelligence in medicine 2003;29(1-2):169-84.

[13] Coden A, Pakhomov S, Ando R, Duffy P and Chute C G. Domain-specific language models and lexicons for tagging. Journal of Biomedical Informatics, 2005: 38 (2), 422-430.

[14] Rosell M, and Velupillai S. Revealing Relations between Open and Closed Answers in Questionnaires through Text Clustering Evaluation. In Proceedings of LREC 2008 -- 6th International Language Resources and Evaluation, Marrakech, Morocco, May 28--30 2008.

[15] Vincze V, Szarvas G, Farkas R, Móra G, and Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes BMC Bioinformatics. 2008; 9 (Suppl 11): S9. Published online 2008 November 19. doi: 10.1186/1471-2105-9-S11-S9.

[16] Hassel M. Resource Lean and Portable Automatic Text Summarization. PhD thesis, School of Computer Science and Communication, Royal Institute of Technology, Stockholm, Sweden, June 2007. pages 19-29.