Swedish Health Data – Information Access and Representation

Swedish Health Data – Information Access and Representation

Sumithra Velupillai

Stockholm
University

REPORT SERIES NO. 09-007
ISSN 1101-8526
ISNR SU-KTH/DSV/R–09/7–SE

Printed in Sweden by US-AB, Stockholm 2009

Distributor: Department of Computer and Systems Sciences, Stockholm University

Abstract

Health related research is an interdisciplinary, broad and growing research area. With the growth of digitalised systems that simplify and make work processes more efficient in many companies and organisations, the amount of available data is now immense. The information contained in health related digital data sets could be used for further research and also, in the long run, for improving health care, health care processes, and public health.

A large amount of the information contained in these data sets is often in unstructured, free text. Health related texts can comprise various types of text, such as scientific articles, questionnaire answers, (electronic) health records, information on web sites, and e-mail. What these texts all have in common is above all the use of a domain-specific vocabulary.

Information access methods applied to textual data require a language model. Many human language technology tools have been developed in order to improve and simplify representation models, primarily for English, and predominantly for general language use. For Swedish, several human language technology tools have been developed. How these tools work on domain-specific data such as health data, is still a relatively unchartered research area. We have investigated what properties the language use in Swedish electronic health records have compared to a large, general-purpose Swedish corpus, in order to identify if and where adaptation is necessary. We have also created a representation model based on phrases instead of words for Swedish scientific medical text.

Health related texts also contain a potentially large amount of previously unknown information, which could be valuable to exploit in further research. We have developed an iterative and interactive method for exploring large text sets, based on document clustering, where both structured and unstructured information is used for generating hypotheses from epidemiological questionnaire data and electronic health records.

One of the most important factors that influence the possibilities of performing research on health related data sets is availability. Although digital information is easy to store and obtain automatically, this type of data often contains sensitive and private information that makes it impossible to distribute for further research, unless identifiable information is deleted or replaced. We have initiated work on automatic de-identification for Swedish and created a manually annotated gold standard, which could be used both for evaluating de-identification systems as well as for training new systems.

Sammanfattning

Hälsorelaterad forskning är ett tvärvetenskapligt, brett och växande forskningsområde. Med framväxten av digitala system som förenklar och effektiviserar arbetsprocesser i många verksamhetsområden har också mängden tillgänglig data ökat explosionsartat. Informationen som finns i hälsorelaterade digitala datamängder skulle kunna användas för vidare forskning och utnyttjas för att i det långa loppet förbättra vård, vårdprocesser och den allmänna folkhälsan.

En stor del av informationen som finns i dessa datamängder består ofta av fri text. Hälsorelaterade texter kan innefatta en rad olika typer av text, såsom vetenskapliga artiklar, enkätsvar, patientjournaltexter, information på webbsajter och e-post. Det som är gemensamt för dessa är framför allt att de har ett specifikt vokabulär som skiljer sig från "vanlig" text.

Metoder för informationsåtkomst ur textmängder kräver en språkmodell. Många språkteknologiska verktyg har utvecklats för att förbättra och förenkla språkliga representationsmodeller, främst för engelska, och främst för generellt språkbruk. För svenska finns det en hel del språkteknologiska verktyg, men hur väl lämpade de är för den här typen av domänspecifik text är hittills ett område det forskats lite på. Vi har undersökt vilka egenskaper språket i patientjournaltexter skrivna på svenska har jämfört med en stor, allmänspråklig svensk textsamling, för att identifiera vad som eventuellt skulle behöva anpassas. Vi har även byggt en språklig representation baserad på fraser istället för ord för svensk medicinsk vetenskaplig text.

Hälsorelaterade texter innehåller också en potentiellt stor mängd tidigare okänd information som skulle vara värdefull att utnyttja för vidare forskning. Vi har tagit fram en iterativ och interaktiv utforskningsmetod som bygger på dokumentklustring och utnyttjar både strukturerad och ostrukturerad information för att generera hypoteser från epidemiologiska enkäter och patientjournaltexter.

En av de viktigaste faktorerna som påverkar möjligheten till att forska på hälsorelaterade datamängder är tillgängligheten. Trots att digital information är enkel att lagra och anskaffa, innehåller den här typen av datamängder ofta känslig och privat information som gör det omöjligt att skapa fritt tillgängliga resurser utan att först ersätta identifierbar information. Vi har påbörjat arbete med att skapa automatisk avidentifiering för svenska samt skapat en manuellt annoterad guldstandard som skulle kunna användas både för att utvärdera avidentifieringssystem och för att träna nya system.

List of Papers

This thesis is based on the following papers:

I Magnus Rosell and Sumithra Velupillai. 2005. *The Impact of Phrases in Document Clustering for Swedish.* In Proceedings of NODALIDA'05 – 15th Nordic Conference on Computational Linguistics, Joensuu, Finland, May 20–21 2005.

II Magnus Rosell and Sumithra Velupillai. 2008. *Revealing Relations between Open and Closed Answers in Questionnaires through Text Clustering Evaluation.* In Proceedings of LREC '08 – 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, May 28–30 2008.

III Hercules Dalianis, Martin Hassel and Sumithra Velupillai. 2009. *The Stockholm EPR Corpus – Characteristics and Some Initial Findings.* To be presented at the 14th International Symposium for Health Information Management Research (ISHIMR 2009), Kalmar, Sweden, October 14-16, 2009.

IV Sumithra Velupillai, Hercules Dalianis, Martin Hassel and Gunnar H. Nilsson. 2009. *Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial.* In International Journal of Medical Informatics (2009). In Press. doi:10.1016/j.ijmedinf.2009.04.005

Acknowledgements

I am indebted to many colleagues and friends for invaluable support, inspiration, encouragement and comments, in particular:

- My supervisors Hercules Dalianis and Martin Hassel
- My co-authors Magnus Rosell and Gunnar Nilsson
- All participants in the KEA project, especially:
    - Ulf Jacobsson and everyone from the Stockholm City Council
    - Bo Wikström, Stefan Engqvist and everyone from Karolinska University Hospital
    - Vinnova, for funding
- Viggo Kann and all members of the Language Technology Group at CSC, KTH
- All participants in the Infomat project
- All supervisors and fellow PhD students at the Swedish National Graduate School of Language Technology (GSLT), especially Oscar Täckström and Yvonne Samuelsson
- Jussi Karlgren and all the other Language Technologists at SICS
- All my colleagues at the Department of Computer and Systems Sciences:
    - Paul Johannesson and all members of SYSLAB
    - Everyone at the Administration Department
    - Everyone at the IT System Group (DMC)
- The Helge Ax:son Johnson Foundation for generous scholarships
- Friends and family, especially my sisters, my friends Camilla and Mona, and Johan, for everything.

Contents

1. Health Data – Language Modelling and Information Access

1.1 Introduction

Today, numerous articles, books, theses and other types of scientific as well as popular science writings address the fact that the amount of information contained in digital form is vast and ever growing. This fact is no longer something ahead of us, but something that we are all in the midst of. Such huge amounts of data is hard, if not impossible, to organize, structure and overview manually. Automatic methods for dealing with such information are therefore needed.

The area of health research is broad and covers many scientific disciplines, such as epidemiology and medicine. Also, such research relies heavily on the access of empirical data, from which further research can be performed. A large amount of data is needed. As such data becomes digitalised, the possibilities of minimising the time-consuming manual efforts of gathering data grow, as well as the possibilities of automating if not all, at least parts of the process of accessing and analysing the data.

However, the data can be stored in many different ways, for example in the form of structured entries, or in the form of unstructured, free text. The latter poses specific problems and possibilities when it comes to accessing and structuring information. Here, language models and Human Language Technology (HLT)[1] tools might prove very useful.

1.2 Research Issues

This licentiate thesis presents research that has been performed on health related textual data in Swedish. The overall purpose is to facilitate the access of information from such domain-specific textual data, and to address language-specific issues that come with this type of work.

In order to develop Information Access (IA) methods on textual data, language models are needed. In order to build language models, one needs to analyse the language. Language analysis requires language resources, which

[1] The research area covering studies on natural (human) language and computer science is, in the literature, called Computational Linguistics (CL), Natural Language Processing (NLP) and Human Language Technology (HLT) interchangeably. Here, the term HLT is used.

are not always easy to obtain. Health related data may, for instance, in many cases be protected due to confidentiality reasons. One important aspect of the research presented here, is the work of gathering, analysing and modelling health related language resources in Swedish.

With language resources and models, it is possible to develop IA methods and tools for textual data. IA methods applied on textual data have matured over the years, but most methods are developed based on general language models. We believe that many aspects of IA methods can be improved and that powerful tools can be developed by refining both language representation models and approaches to IA in general.

For instance, visualisation and user interaction techniques are still not commonly implemented in IA tools, and we believe there is much potential in such methods. Moreover, with a thorough language analysis, we believe that alternative language models and representations may be very useful in IA settings. In particular, we believe that a tailoring of such models for domain- and language-specific issues might improve IA tools. Health related language is domain-specific, and studies on the properties of this language use in Swedish are needed in order to improve IA methods.

In a long term perspective, the hope is that such research will aid research concerning public health, disease prevention, decision support, and the like. This type of research has gained a lot of interest recently – but there is still a lot of knowledge lacking, especially when it comes to language- and domain-specific issues.

2. Health Data and Information Access

In information sciences, it is common to discriminate the terms *data*, *information* and *knowledge*. Simplified, the differences and relations could be stated as follows. *Data* consists of facts. *Information* is related to novelty and order, and is obtained by the application of knowledge to data. Relationships between data are defined by *knowledge* (Coiera, 2003). This division is of course a simplification, and not self-evident (and subject to much philosophical debate). However, it is important to discriminate between the terms, and the distinction is important for the work presented here. A more thorough discussion on these concepts can be found in, for instance, Coiera (2003).

Health related research is, as stated above, a very broad concept which covers many diverse scientific disciplines. In particular, health related data can come in many different forms. In general, what is meant by health data in this thesis, is data that concerns health issues in any way. For instance, such data can be in the form of scientific articles and writings, questionnaire answers from, e.g., epidemiological research projects, electronic health records (EHRs)[1], health related web content, population data and biomedical data.

These types of data sets have several properties in common. One is the fact that they communicate and document aspects related to research that, in a long-term prespective, has the goal of analysing, improving and developing public health. Another is the empirical basis – such research relies on the access of large amounts of data in order to generate and test workable hypotheses. Moreover, for processing such data automatically, it needs to be represented in some sort of structured form.

However, there are also many situations where findings, irregularities and the like are better expressed in the form of unstructured, free text. Hence, many large health related data sets contain both structured data entries such as measure values, closed answers in questionnaires, and controlled keywords, as well as unstructured entries in the form of free text, such as open answers in questionnaires, articles, and clinical notes.

Numerous statistical methods can be used for the automatic analysis of structured entries, such as closed answers in questionnaires or diagnose codes in EHRs, in order to extract and obtain new information and knowledge. The

[1] The (electronic) document covering patient information from hospitals is, in the literature, called both electronic medical record (EMR), electronic patient record (EPR), and electronic health record (EHR). Here, the term EHR is used.

unstructured entries, on the other hand, pose many problems, as they are written in natural language and need to be transformed into a more structured representation. These problems are discussed in more detail in Section 3.

As more and more organisations, companies and research groups have the possibility of digitalising their data, the amount of electronic data increases. With this, the need for methods or systems that aid the process of accessing relevant information in these data sets escalates, and the possibilities of utilising these data sets for further research becomes apparent.

2.1 Health Data Information – Examples of Types and Content

Communicating observations and research results is central to all research. As the scientific community grows, the amount of published data grows with it. For this, scientific conferences, journals, and the like are indispensable. In the 1750's, around 10 scientific journals existed. In 1899, the number of biomedical journals had already increased to around 1 000 (Petersson and Rydmark, 1996). Today, around 1 000 journals are published every week, and the medical literature is growing with a new article every 26 seconds or less (Coiera, 2003). Health related research nowadays have special searchable databases, forums, etc. for communications (PubMed[2], for instance). It is important that users obtain the information they need. As the contents of such articles are domain-specific, alternative ways of representing them might improve search engines and other Information Access (IA) tools compared to other domains.

In epidemiological research, large sets of demographic and lifestyle data are needed for many types of studies. Such data can be collected through distributing questionnaires to individuals or organisations. Digitalising the work of collecting and analysing such data is both efficient and effective (Ekman and Litton, 2007). Questionnaires predominantly contain closed answer alternatives for the respondents, as they are easy to analyse statistically. However, providing questions where open answers may be given makes it possible for the respondents to express themselves in more detail. Moreover, closed answers may not always provide sufficient alternatives and respondents may provide answers that are only partially correct, which is difficult to capture during automatic analysis.

In hospital settings, one moves more and more towards digital solutions for documenting the patient status and health progress. Medical documentation has a long history; in Sweden, the first systematic medical documentation process in a hospital environment was initiated in 1752 (Nilsson, 2007). Here, all documentation regarding the patient status and health progress is called a *health record*. At first, the health record was mainly a convenient way of com-

[2] http://www.ncbi.nlm.nih.gov/pubmed/

munication between physicians, as well as a way of keeping mental notes, where data about the patient such as age, name, admittance time, diagnosis, etc. was recorded. Medical practice and science has since developed, and the (patient) health record has changed in many aspects, but it still remains a central document in the health care process. The EHR of today is more complex, contains a much larger amount of data, and has to conform both to legal issues as well as to patient demands. For a thorough account on the history of medical documentation in Sweden, see Nilsson (2007).

The first attempts of introducing EHR systems were made in the U.S. in the late 1950s and the early 1960s. With the demand for rationalising the medical care systems, and the technical advances in society with more effective and less costly PCs, EHR systems were acknowledged and put into use both in the United States and in Sweden in the early 1990s (Petersson and Rydmark, 1996). Dick *et al.* (1997) present a thorough review of the status of the use of computer-based patient records in 1997.

In the United States, the Mayo Clinic in Minnesota has been using EHRs since 1994 and have more than 16 million EHRs (Pakhomov *et al.*, 2005c). In Sweden, there are currently at least three large EHR systems; TakeCare[3], Melior [4] and Cambio[5], that are used in hospitals throughout the country. For accessing information in such huge amounts of data, automatic methods are necessary.

Many other health related data sets exist of course, both in digital form and in archives of other types. For instance, there are many portals on the Internet for health related information, such as Vårdguiden[6], Netdoktor[7] and Web4Health[8], as well as health related online communities and expert systems, such as Viktop[9].

In the research presented in this thesis, three types of health related data resources have been used (all containing written free text in some form and all in Swedish):

- Scientific Medical Text[10] (used in Paper I). This set will be called SweSciMed.
- Questionnaire data from the Swedish Twin Registry[11] (used in Paper II). This set will be called SweTwin.
- EHR data[12](used in Paper III and Paper IV). This set is called the Stockholm EPR Corpus.

[3] http://www.profdoc.se/sjukhus/produkter/takecare/

[4] http://www.medical.siemens.com/

[5] http://www.cambio.se/

[6] http://www.vardguiden.se/

[7] http://www.netdoktor.se/

[8] http://web4health.info/sv/answers/project-this-site-info.htm

[9] http://viktop.se/viktop/

[10] Gathered, with permission, from Läkartidningen: http://www.lakartidningen.se/

[11] http://ki.se/ki/jsp/polopoly.jsp?d=9610&l=sv

[12] Gathered from TakeCare (developed by Profdoc: http://www.profdoc.se/)

These types of data sets naturally have different properties. However, what they all have in common is the health related content, which can be used for the purpose of improving health related research in the long run. SweSciMed is the only set that solely contains written free text, with the exception of manual keyword assignments of MeSH-terms (MeSH, 2008) (see Section 2.3). Those keywords can be used and treated as relevant structured entries.

The other two sets, SweTwin and the Stockholm EPR Corpus, on the other hand, have interesting properties that may be of great importance when exploiting them for IA research. Structured information such as gender, age, measure values, and diagnoses can be, and often is, included in both questionnaires and EHR data. Analysing patterns and relations between such entries is a straightforward process that can be used for health related research. However, the free text parts of such data sets potentially contain even more detailed and subtle information that may alter or give a more thorough account of the contents in the data sets, even more so if they are linked with the structured information.

2.2 Privacy Issues and Automatic De-identification

Health related research often deals with information gathered from individuals. This means that there are many ethical issues that need to be considered both when gathering the data, when working with the data and when communicating results obtained from the data. The integrity of the individual should never be at risk. However, the importance of having access to large data sets in order to be able to perform reliable, comparable, and useful research is of high priority. Although identifiable information is easy to delete or replace from structured entries (e.g. social security numbers), the risk of finding indentifiable information in the free text entries is potentially high. In EHR data, for instance, hospital staff may write phone numbers, names of relatives, or even names or nicknames of the individual patients. Suominen *et al.* (2007) present a thorough discussion on ethical issues regarding research on sensitive health data.

In the literature, the terms *anonymisation* and *de-identification* are often distinguished, where *anonymisation* refers to the process of *removing* identifiable information, while *de-identification* refers to *masking* or *replacing* identifiable information (see Kokkinakis and Thurin (2007) for a discussion on this). Whichever approach one chooses, the identifiable entities are the same.

When it comes to health information, the Health Insurance Portability and Accountability Act, HIPAA (HIPAA, 2003), enforced in the United States, defines 18 Protected Health Information (PHI) instances that should be replaced from EHR data or similar research data sets in order for them to be considered de-identified. These instances are names, locations, phone numbers, etc. (see e.g. HIPAA (2003) for a full list).

In Sweden, there are no similar explicit regulations. However, access to research data that involves individuals and that may require sensitive information is granted by the regional Vetting Boards, where the planned research must be clearly described. Also, there is a law concerning patient data (Patientdatalagen, SFS 2008:355[13]) that states what type of information EHRs must and must not contain. For instance, in § 2, it is regulated that patient integrity must be respected.

De-identifying large health data sets manually is both time-consuming and costly. Instead, automatic methods are needed. The process of automatically de-identifying data resembles Named Entity Recognition (NER) methods, that have been successfully developed for other domains. NER methods automatically extract entities such as personal names, places, time expressions and organisations. Research on NER techniques is extensive with, for instance, an international forum for evaluation of NER systems through the The Message Understanding Conferences, MUC (Grishman and Sundheim, 1996).

Automatic de-identification methods of health related data sets often use NER techniques. Dictionaries and gazetteers are frequently used as external resources for the identification of named entities. In Paper IV, we describe the porting of an existing de-identification software (developed for English) to Swedish. This system, called De-id, is rule-based and relies heavily on external resources (Neamatullah *et al.*, 2008). In EHR data, the vocabulary may often be ambiguous and noisy, which makes the use of external resources difficult. In such cases, other approaches where context and other language properties are used may yield better results (see e.g. Uzuner *et al.* (2007) for a more thorough discussion on approaches for automatic de-identification).

Current automatic de-identification systems show very promising results. In Uzuner *et al.* (2007), the results and evaluation of a shared task initiated at the i2b2 Center (i2b2, 2009) is described. The best system (Szarvas *et al.*, 2007) reported very good results. Uzuner *et al.* (2008) describe the evaluation of a system, Stat De-id, that uses local context and Support Vector Machines (SVMs), showing very promising results for handling noisy texts. In this work, several approaches are compared, including rule-based methods and a Conditional Random Fields De-identifier (CRFD), and Stat De-id outperforms them all. For Swedish, Kokkinakis and Thurin (2007) have developed a system on discharge letters which was successful. Our work on automatic de-identification of the Stockholm EPR Corpus (described in Paper IV) did not result in a workable system due to difficulties in porting the existing software. However, we developed a manually annotated gold standard that can be used for evaluating and training such systems. The need for and purpose of having manually annotated resources is further discussed in Section 3.2.

[13] The law text can be found in its entirety here: http://www.notisum.se/rnp/sls/sfs/20080355.PDF (Accessed on August 8, 2009)

2.3 Information Access Methods and Health Data

Information Access (IA) is also a very broad concept, which includes research on for instance Information Retrieval (IR), Information Extraction (IE), and Data and Text Mining. Such research is interdisciplinary as it incorporates computer science, linguistics (when it comes to working with free text), statistics, mathematics, and informatics, to mention only a few.

IR methods tackle the problem of searching for and retrieving relevant information from documents, document collections, databases, etc. For instance, search engines are examples of the successful application of IR techniques.

IE can be seen as a type of IR, where the difference lies in the fact that IE methods try to extract structured, well-defined information parts, such as named entities or keywords from a controlled vocabulary, or passages in documents. For further definitions and discussions on these research areas, see for instance Baeza-Yates and Ribeiro-Neto (1999).

Data and Text Mining methods on the other hand try to extract hidden, previously unknown information from large data sets (Hearst, 1999). All methods are, however, related in that they deal with gathering, structuring and handling large data sets.

Applying IA methods on health data is important for many reasons. As stated above, the amount of data is enormous and impossible to organise and structure manually. Hence, automatic methods for finding relevant information for different purposes is of great importance. Moreover, the contents of health data can, in the long run, be used for improving knowledge, treatments and processes in public health.

In the biomedical domain, research on improving IA on biomedical scientific articles, especially for finding biomedical entities such as gene and protein names, has been ongoing for a while. For instance, the BioCreAtIvE (Critical Assessment of Information Extraction in Biology) challenge has provided a set of evaluation tasks for research on text mining for biological data (Hirschman *et al.*, 2005). Another example is the TREC Genomics track (for an overview, see Hersh and Voorhees (2009)).

For medical data, and especially EHR data, IA methods combined with HLT tools have shown promising results. MedLEE, for instance, is a system used at the Columbia-Presbyterian Medical Center (CPMC) in New York. This is a modular system designed to extract clinical information from EHRs (more details can be found in e.g. Friedman *et al.* (1995), Friedman (1997), Friedman *et al.* (2004) and Mendonca *et al.* (2005)).

At the Mayo Clinic in Minnesota, similar research has been performed on their EHR data (Pakhomov *et al.*, 2005a). The work at the Mayo Clinic is also part of the Open Health Natural Language Processing Consortium (OHNLP, 2009), an initiative to establish an open source consortium for research on clinical and medical NLP research. An extensive overview of IA-related research on EHR data can be found in Meystre *et al.* (2008).

The use of controlled vocabularies, thesauri, and the like is extensive in the medical domain. Many such resources are exploited in health related IA systems. For instance, the articles in SweSciMed are all manually indexed with keywords from the MeSH vocabulary (MeSH, 2008). MeSH is translated into several languages. MeSH terms are mainly used for indexing scientific publications for more efficient retrieval (with respect to both users and time) in search engines such as PubMed. The Unified Medical Language System (UMLS, 2009) is a resource where many health related knowledge resources are collected, such as the Metathesaurus (a vocabulary database with concepts, names, and relationships) and the Semantic Network (a set of semantic types and relationships related to the contents of the Metathesaurus), which all cover different biomedical controlled vocabularies that are developed to facilitate the development of IA methods for health related data. However, most such resources currently exist only for English.

Text mining methods for extracting previously unknown information from health related data in, for instance, medical literature have also been developed (see e.g. Swanson and Smalheiser (1997)). In Paper II we present a method for generating hypotheses from health related questionnaire data. This method could be applied to any data set that contains both structured and unstructured information, from which interesting, new relations can be revealed. For instance, we have applied the method on EHR data in an initial experiment presented in Paper III, with promising results.

When it comes to the free text parts of health related data resources, many domain-specific properties make “traditional” IA approaches insufficient. A more thorough discussion on these issues is presented in Section 3.4.4.

3. Processing and Representing Unstructured Health Data

The main focus for the research presented in this thesis lies in the gathering, (pre-)processing and representation of the unstructured, free text parts of health related data sets written in Swedish, whether or not the data sets also consist of structured entries. These free text parts potentially contain information that is valuable for further research, but they are currently seldom used for anything but storage, as they need to be converted to a representation that can be used for further processing. Most importantly, in order to be able to create, develop and analyse such representations, resources in the form of corpora that capture health related language are needed.

3.1 Working with (Health) Corpora

Language resources are very important for linguistic studies and the development of Human Language Technology (HLT) tools. However, it is not trivial to model language and language use without empirical data. Nowadays, the Internet is an invaluable source for gathering large corpora that can be used for creating language models of different types (see e.g. Hassel (2001)). Moreover, many types of texts can easily be digitalised. However, many obstacles may occur; copyright issues, privacy issues (see Section 2.2 for a discussion on this) and other factors may make distributions of corpora very difficult. It is important for the research community to be able to share such corpora, in order to be able to develop and evaluate methods and results.

There exist a large amount of language resources in the form of both corpora and HLT tools for English. For small languages such as Swedish on the other hand, resources are more scarce. Table 3.1 shows the number of documents and tokens[1] in the three health related text sets used in the research presented in this thesis. We see that they differ both in token size and document size. Moreover, for the Stockholm EPR corpus, there is no natural way of breaking up the data into documents. Here, the number of patients is presented. However, one could also divide the set into “documents” consisting of clinics, patients per clinic, etc.

The SweSciMed corpus contains the largest amount of tokens per document. This is not surprising, as the documents are scientific articles. On aver-

[1] The individual occurences of a word.

Text set	docs	tokens
SweSciMed	2 422	4 383 169
SweTwin	43 341	453 105
Stockholm EPR corpus	408 144*	109 663 052

Table 3.1: *Number of documents and tokens in the health text sets used in the presented research. *For the Stockholm EPR corpus, the notion of document is not straightforward. The presented number is the total number of patients (each patient may belong to several clinics), from one fifth of the Stockholm EPR Corpus (the first five months of 2008).*

age, each document contains around 1 800 tokens. The SweTwin and Stockholm EPR corpus, on the other hand, contain less than 30 tokens per document on average (10 and 27 respectively). The Stockholm EPR corpus is the most complex data set, as it also contains a large amount of structured data that is related to the unstructured free text in different ways. A study of a subset of the Stockholm EPR corpus showed that at least around 40 percent of the data entries are unstructured, the rest being structured (see Paper III). The unstructured entries contain more data on the other hand and constitute a larger total amount of tokens. Moreover, there is a lot of duplicate information, as there are many authors to each record and patients may visit different clinics. Examples of different types of structured information contained in these sets are given in Section 2.1.

3.2 Annotated Language Resources

Within the HLT research community, there exist many large language resources in the form of corpora that have been annotated for various types of linguistic aspects, such as part-of-speech tags and syntactic representations. One example is the Penn Treebank (Marcus *et al.*, 1994). There are also resources available for discourse analysis, anaphoric resolutions, word senses, predicate-argument analysis, etc. See, for instance, PropBank (Palmer *et al.*, 2005), and GNOME (Poesio, 2004). Many resources annotated with biomedical information such as gene and protein names have also been created (e.g. the GENIA project (Collier *et al.*, 1999)).

For Swedish, there exist some annotated resources as well. The Swedish Parole (Gellerstam *et al.*, 2000), for instance, is a large corpus annotated with part-of-speech information. Another smaller resource is the SUC corpus (Ejerhed *et al.*, 2006), which is manually annotated with part-of-speech information, and also named entities, resolved pronouns, etc.

Such resources are valuable for many different reasons. First, such resources make it possible to make empirical claims about language and language use from different perspectives. Second, such resources can be used to train computational models that try to mimick, predict, or create similar language properties. Third, they may be used as gold standards, or references, so that similar research can be evaluated against the same standardised and agreed-upon material.

Although it is laborious and time-consuming to create annotated language resources manually, they are very useful for research. In Paper IV, we have initiated the work of manually annotating a subset of the Stockholm EPR Corpus for Protected Health Information (PHI) instances. We hope to be able to use this set as a gold standard for evaluating and training de-identification software, as well as for empirical analysis of the types and properties of PHI instances in this type of data.

3.3 Evaluation of Annotated Language Resources

When working with annotated resources, one needs to know how reliable the annotations actually are. This is important since the way reliability is calculated and interpreted affects many aspects that may be problematic when using such resources. For instance, the use of so called *expert* coders or so called *naive* coders may influence the annotation results a great deal. The difference between an *expert* coder and a *naive* coder lies mainly in the amount of knowledge the coder has about the task, and how complex the task is. A coder with more prior knowledge and experience in the task at hand (*expert*) is expected to perform differently compared to a coder with less knowledge (*naive*). Furthermore, the way categories are distributed, the (dis)agreement distributions among annotators, and the chosen calculation metrics may affect the reliability figures in unexpected ways.

The κ statistic, which is used for measuring reliability by taking into account expected agreement by chance, was introduced by Carletta (1996), and has become a widely used measure for reliability. Many resources created prior to the introduction of the κ statistic (e.g. the Penn Treebank (Marcus *et al.*, 1994)) mainly reported raw accuracy and overlap percentages, which are difficult to analyse.

Artstein and Poesio (2008) give an exhaustive account for different reliability and agreement measures that have been employed in different annotation tasks. They also discuss different criteria that have been formulated in order to ensure a high level of reliability and reproducibility in an annotated corpus. In the field of content analysis, recommendations for the analysis and creation of annotated resources have been developed. In particular, Klaus Krippendorff has stated the following criteria for an annotated corpus to be considered reproducible, and thus, scientifically valid (quoted from Artstein and Poesio (2008)):

- It must employ an exhaustively formulated, clear, and usable coding scheme together with step-by-step instructions on how to use it;
- It must use clearly specified criteria concerning the choice of coders (so that others may use such criteria to reproduce the data);
- It must ensure that the coders that generate the data used to measure reproducibility work independently of each other.

Once an annotated corpus is shown to be reliable, two important properties have been established (Craggs and McGee Wood, 2005):

1. The categories onto which the units are mapped are not inordinately dependent on the idiosyncratic judgments of any individual coder.
2. There is a shared understanding of the meaning of the categories and how data are mapped onto them.

There is, however, no way of giving standard confidence thresholds that can be used to claim that a coding scheme or an annotated resource is reliable. Moreover, if the purpose of creating the annotated corpus is to train a computational model, high agreement levels are very important. Otherwise, the model could imitate the inconsistent behavior of human annotators. If the purpose is to make empirical claims about the annotated phenomena, lower agreement scores may not affect the results as drastically (Craggs and McGee Wood, 2005).

To summarise, it is important to define what the purpose of using an annotated resource is clearly, and to know what the annotated resource actually contains. The PHI annotations we created in Paper IV showed promising reliability results given the chosen approach, but will be analysed in more detail in the future for the reasons mentioned above. In this work, we did not use the κ statistic for measuring the inter-annotator agreement, as we believe chance agreement would have little overall effect, given the large amount of annotation classes. However, in order to identify where the results are problematic and where they are stable, calculating κ statistics for individual annotation classes is probably useful for an in-depth analysis.

3.4 Representation Models and Information Access

There are many possible ways of representing language use. The most widely used representation model in Information Access (IA)-settings (especially for IR applications such as search engines) is the so-called *vector space model* (Salton *et al.*, 1975). In this model, each document in a document data set is represented by a vector.

The vectors are members of a high-dimensional space, where the number of dimensions is given by the number of chosen representation units. The most commonly used representation units are single words. With a large set of documents, the number of single words is usually very high, resulting in a vector space with as many dimensions as the number of single words in the whole document collection.

Each word is assigned a value in the vector based on its occurence in the text. The values are often calculated according to some weighting scheme, where variants of augmented term frequency measures such as the *tf-idf* (term frequency-inverse document frequency) scheme are most common. These measures model the importance and discriminative properties of the words in different ways, taking into account combinations of frequency of occurrence and occurrence specificity. More details on weighting schemes can be found in, for instance, Manning *et al.* (2008).

Vectors that are close to each other in the created space are judged to be similar in content. Closeness and similarity can be calculated in different ways – this is often done by calculating the *cosine* of the vector angles between vectors. For further details on the theoretical foundations of this representation model, calculation variants, etc., see, e.g., Van Rijsbergen (1979), Manning and Schütze (1999), Baeza-Yates and Ribeiro-Neto (1999) or Manning *et al.* (2008) to name but a few.

In most IA settings, it is easy to define what a *document* is (articles for instance), but when it comes to electronic health records (EHRs), such distinctions may be harder. For instance, it is possible that different situations require different definitions of a document. In some cases, a document could be defined as *all* EHRs for a patient, in others only as a patient's EHR from a specific clinic, or as specific individual entries within a patient's EHR, etc. Furthermore, defining representation units is not always straightforward (what is a word, phrase, etc. and what is best suited) and may depend on the domain and language. This issue is discussed further in Section 3.4.3 and 3.4.4.

3.4.1 Document Clustering

When working with large text sets it is often necessary to organise, structure or group the set into smaller subsets, in order to get a better overview of the content of the set. In such cases, one wants the texts in the subsets to be similar in content, and ideally also the subsets to be as dissimilar as possible. Document clustering techniques can be used to automatically group the texts in a text set according to how similar they are. These techniques are unsupervised, i.e. no training data is needed. Instead, the representation model and some similarity measure are used to automatically create clusters or subsets.

There exist many different clustering algorithms. For instance, *hierarchical* algorithms create hierarchical cluster structures, while *partitioning* algorithms create flat cluster structures. K-means is one of the most widely used partitioning clustering algorithms and is very efficient. This algorithm is described in, for instance, Jain *et al.* (1999), and consists of the following basic steps:

1. Initialize a partition with *k* groups (clusters).
2. Calculate the cluster centroids (normally the mean vector for each cluster).
3. Place each object in the cluster with the most similar centroid.
4. Repeat step 2 and 3 until some stopping criterion is met.

Bisecting K-means is a top-down, divisive hierarchical clustering algorithm in which K-means is used for splitting the worst clusters in two (Steinbach *et al.*, 2000). Both algorithms require that the number of desired clusters is chosen in advance.

Naturally, different clustering algorithms are better suited for different situations. Choosing one instead of the other depends on the task at hand. Using clustering techniques for health related data can be beneficial for many situations.

In Paper I we use the Bisecting K-means algorithm in our experiment, where our goal is to evaluate whether a representation model based on phrases may improve results for clustering scientific medical texts written in Swedish, compared to using a representation model based on single words. In Paper II we use the K-means algorithm for revealing previously unknown relations in questionnaire data. In this experiment, we cluster the text set several times, choosing different numbers of clusters each time in order to find those clusters that seem stable. As the proposed method is interactive, fast computations are required, and K-means is thus an adequate algorithm. For further details on document clustering techniques, especially with application to Swedish text data, see, e.g., Rosell (2009).

3.4.2 Evaluation of Information Access Methods

IA evaluation is inherently difficult. There are many factors that need to be taken into account when judging whether a method or system produces good results. An information access need is highly dependent on the situation, the individual, the context, and the purpose, which makes it difficult to measure reliably.

Distinguishing between *intrinsic* and *extrinsic* evaluation is important. *Intrinsic* evaluation deals with measuring results of a method or system with respect to a standardised measure or a gold standard. *Extrinsic* evaluation, on the other hand, deals with measuring whether a method or system performs well given a specific task. In such cases, human judgment is necessary.

In the work presented in this thesis, the methods have primarily been evaluated by *intrinsic* measures. For *extrinsic* evaluations, user cases or similar settings would be required, which is both time-consuming and costly. For an overview on evaluation issues for HLT research, see Spärck-Jones and Galliers (1995).

3.4.3 Language-specific Issues

When using the vector space model one has to decide which representation units one wants to use in the vectors. As stated above, single words are most commonly used. Using a representation model where all words are used in their raw forms creates extremely large and sparse vectors.

In order to reduce the size of the vectors, as well as to create a representation model which better captures the contents in the set of documents, different measures may be taken. For instance, deleting units that do not contribute to the core contents of the texts by using so called stop word lists (lists where non-content bearing words such as function words and/or high-frequent words are included) or deleting words that occur infrequently reduces the number of dimensions and makes computations more efficient.

More importantly, normalizing the representation units (in this case, words) to more general forms through morphological analysis may be very useful. Two common methods for doing this are *stemming* and *lemmatization*. In *stemming*, words are stripped from their affixes, producing word *stems*. The words *caring* and *cars* for instance, can be conflated to the common stem *car*, which might be problematic. *Stems* are not necessarily valid linguistic units, but are instead mainly used for efficient internal representations. In *lemmatization* on the other hand, words are conflated to the same linguistic base form. The examples above would be conflated to the two base forms *care* and *car*. Swedish is a highly inflectional language, and the use of such normalizing techniques has proved very successful in IA settings (see e.g. Carlberger *et al.* (2001), and Rosell (2003)). For other less inflectional languages such as English, the effects are not as dramatic.

Furthermore, compounding is very common in the Swedish language. It is possible to create very complex solid compounds, forming single word units that are very long. The word *strålbehandlingsplaneringsdatortomografi* (radiation-treatment-planning-computer-tomography) for instance, consists of five individual words in different morphological forms, together with the Swedish affix *s* which is commonly used for compound creations. Also, splitting compounds to their individual parts have improved results in IA settings (see, e.g., Dalianis (2005)).

How these factors affect language representations of health related text data in Swedish is experimented with in Paper I. In Paper III we initiate a study on the vocabulary contained in Swedish EHRs, where we plan to analyse the effects of domain-specific morphological and linguistic analysis further.

3.4.4 Domain-specific Issues

Domain-specific language, alternatively called sublanguage, is a specialized (natural) language, used within a specific domain or subject (Grishman and Kittredge, 1986). Grishman (2001, p. 1) states that "A sublanguage is characterized by a specialized vocabulary, semantic relationships, and in many cases specialized syntax."

Creating language models for domain-specific, health related language use in the same way as for general language use may not produce good results. Many studies have shown that the vocabulary in for instance EHR data differs a great deal from general vocabulary (see e.g. Coden *et al.* (2005) for a discussion on this). Here, general vocabulary is defined as such vocabulary that can be found in (large) corpora that cover several different textual, general sources. These should represent a word collection that can be useful in many different situations, such as for building a general lexicon, and hence lack domain-specific vocabulary. In Paper III we show that the Stockholm EPR corpus have word distributions similar to a general language corpus, but that the vocabulary differs a great deal.

EHR data is very noisy and contains a large amount of misspellings, ad-hoc abbreviations, and domain-specific vocabulary and language use. In Ruch *et al.* (2003), they find that French EHR data contains up to 10 percent spelling errors (compared to 1-2 percent in general language texts). Misspelled words may make the representation models skewed and produce worse results in an IA system. Ruch (2002) show that the application of spelling correction (using automatically induced spelling errors) on noisy data improves results in an IR system.

Abbreviations and acronyms are also common in EHR data. In particular, they are often ad-hoc constructions and in some cases also very internal. The abbreviations *vc* and *våc* are both used for *vårdcentral* (health centre), for instance. Another example is the abbreviation *vb*, which can be used either for *vid behov* (when needed) or *vederbörande* (the (person) concerned). Pakhomov *et al.* (2005c) present methods for disambiguating acronyms and abbreviations in EHR data.

Part-of-speech (PoS) information may be very useful for the creation of representation models and for IA systems. In Named Entity Recognition (NER) techniques, for instance, PoS information may be important for disambiguating entities. Applying PoS-taggers developed for general language on EHR data has been shown to produce worse results than using a PoS-tagger trained on English EHR discourse (Pakhomov *et al.*, 2005b). However, for German, Wermter and Hahn (2004) and Hahn and Wermter (2004) show that the use of PoS-taggers trained on general language works well on German EHR data. This may be due to the fact that German is highly inflectional, whereas English is not, and PoS-taggers often exploit suffix information when tagging unknown words. Swedish is also highly inflectional, which might indicate that the studies on German might hold for Swedish as well. Thus far, no such study has been performed to show this.

Language use from a syntactic point-of-view also differs in medical language. Campbell and Johnson (2001) presents a thorough analysis of differences and similarities for English. In general, the mentioned studies all conclude that both part-of-speech and syntactic patterns are less complex in EHR data.

Domain-specific language contains specific terminology. Justeson and Katz (1995) for instance, show that technical terminology in English consists mostly of multi-word terms being noun phrases with nouns, adjectives or the preposition *of*. Hence, building representation models based on single words may be insufficient.

In Paper I we experiment with building a representation model based on noun phrases instead of words, based on the idea that medical terminology contains phrases to a larger extent than general language. Phrases are, however, used differently in Swedish, where compounds form single word units. Hulth (2001) found that 75 percent of the entries in a keyword thesaurus used for indexing bills at the Swedish parliament consisted of one noun. We used a Swedish general language spell checking tool (Kann *et al.*, 2001) to split compounds in the experiments presented in Paper I. Tailoring the compound analysis to domain-specific vocabulary might improve the results.

4. Overview of Included Papers

In this section the papers included in this thesis are summarised and commented. Moreover, the work distribution among the co-authors is briefly described.

4.1 Paper I

The Impact of Phrases in Document Clustering for Swedish (Rosell and Velupillai, 2005)

Representing texts through a distributional model of words has, as stated in Section 3.4, shown promising results in many Information Access (IA) applications. However, the idea that a different representation model might improve results for document clustering was put forward within the Infomat project[1]. In particular, the hypothesis that phrases might be a better representation unit especially for medical texts, where many domain-specific expressions were assumed to be in the form of phrases, was the focus for this work. Unfortunately, the results were not very encouraging. However, a tendency that a phrase representation for a corpus of medical scientific texts written in Swedish gave better results compared to a corpus of Swedish newspaper articles is shown, and opens the possibility of experimenting further with domain-specific representation models. Furthermore, phrases are, in Swedish, often produced as solid compounds. A more domain-specific compound analysis may be more suitable for this task.

I was responsible for collecting a corpus of medical scientific text written in Swedish (SweSciMed), categorizing it and extracting the phrases. The representations and the clustering experiments were performed by Magnus Rosell. The analysis of the results was made jointly. The paper was written by Magnus Rosell with my assistance.

[1] http://www.csc.kth.se/tcs/projects/infomat/

4.2 Paper II

Revealing Relations between Open and Closed Answers in Questionnaires through Text Clustering Evaluation (Rosell and Velupillai, 2008)

Working with and analysing large text sets is a difficult task, and the information contained in written free text is potentially very important for further research. Developing methods for the aid in exploring and extracting previously unknown information from such text sets is highly valuable for the research community. Data sets that contain both structured and unstructured information open the possibility of extracting relations between the two which might be useful for further studies.

In this paper, we present a method where we used document clustering techniques for hypothesis generation from an epidemiological questionnaire (SweTwin), exploiting the relations between the open and closed answers. By treating the closed answers as categorizations, we used these for evaluating the clusters created from the open answers (consisting of answers where the respondents described their occupation) through an external evaluation measure. This method is interactive and requires human exploration and judgments in crucial steps, where the most interesting clusters are presented to the user through sorting the clusters according to their quality. We were able to generate the hypothesis that “farmers smoke less than the average”, a hypothesis that was verified by literature studies.

The Infomat tool (Rosell, 2009) was developed by Magnus Rosell. The original idea on generating hypothesis by exploiting structured information in conjunction with unstructured information and document clustering techniques was developed by Magnus, but the method was developed jointly while working on the experiments, as was the writing of the article.

4.3 Paper III

The Stockholm EPR Corpus – Characteristics and Some Initial Findings (Dalianis *et al.*, 2009)

The Stockholm EPR Corpus (obtained within the KEA-project[2]) contains a large amount of information that is currently never reused for further research. In particular, the unstructured, or free text, parts are currently only used as a support in the documentation process. In order to start working with the information contained in the unstructured parts, an overview of the information and language use contained in the data set as a whole is useful. Such an overview may be used to identify what kind of research could be possible to perform on such data. Specifically, getting an overview of the characteristics of the un-

[2] http://researchprojects.kth.se/index.php/kb_7795/io_9851/io.html

structured part is of interest in order to develop representation models that are suitable for IA methods. Also, comparing the frequency distributions in the Stockholm EPR Corpus with a standard Swedish corpus (the Swedish Parole (Gellerstam *et al.*, 2000)) gives a picture of what kind of data and language we are dealing with.

In this paper we also describe planned experiments on the Stockholm EPR Corpus. Some of the described experiments have been initiated and preliminary results are presented. These include work on hypothesis generation on Electronic Health Record (EHR) data, automatic diagnose coding, uncertainty and certainty detection and synonym generation.

I was responsible for extracting the frequency information from the Stockholm EPR Corpus as well as the Parole corpus. All authors were active in the analysis, evaluation and writing process.

4.4 Paper IV

Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial (Velupillai *et al.*, 2009)

The importance of developing methods for automatic de-identification of EHRs can not be understated. In order to be able to make a data set (or, at least, parts of it) such as the Stockholm EPR Corpus available for further research and to other research groups, the integrity of the individuals needs to be secured. For this, automatic de-identification methods may prove invaluable.

As the KEA-project gained access to the Stockholm EPR Corpus, Hercules Dalianis initiated the process of developing an automatic de-identifier for EPRs written in Swedish. He also worked on the porting of an existing de-identification software for English into Swedish. Martin Hassel also worked on this part. Moreover, a gold standard was needed in order to evaluate the performance of the automatic de-identifier, as well as having a possibility of empirically analysing the existence of identifiable information in the EHRs. I was responsible for developing the annotation guidelines and preparing the annotation work as well as extracting a subset for the annotation work and analysing the results on the gold standard. Both Hercules Dalianis, Gunnar Nilsson and I was involved in the annotation work. The article was written jointly.

5. Main Contributions

The research presented in this licentiate thesis describes work on Information Access (IA) from Swedish health related data. Research on health related data is interdisciplinary and covers many different research questions. Health related data often contains a large amount of unstructured, free text. In order to access information from free text, the language needs to be modelled appropriately and suitable representations need to be created.

As the research area covered in this thesis is very broad, many different research questions have been addressed. Overall, the intention has been to create a comprehensive overview of the possibilities, challenges and different perspectives that need to be considered when trying to access information from health related (textual) data in Swedish. We have focused on some of the main questions within this still relatively unchartered research area, attempting to systematically unravel different parts and aspects that need to be scrutinized and that have potential for improving IA systems specifically designed for health related data.

Resources – Gathering, Distributing, De-identifying

When developing methods for IA systems, resources that represent the real-life data one needs to be able to handle are essential. Such resources are often hard to access, especially when it comes to health related data. Privacy issues, copyrights and the like restrain the possibilities of performing research that can be properly evaluated and shared. Also, as Swedish is a very small language, the amount of available resources is even more scarce than for English, for instance. In our research, we have gained permission to use different resources: scientific medical text (Paper I), epidemiological questionnaire data (Paper II) and electronic health records (EHRs) (Papers III and IV), all in Swedish.

In Paper III, we present the Stockholm EPR Corpus, which is probably one of the largest corpora covering EHR data in Swedish. This resource is invaluable for research, but restricted due to its sensitive contents. In Paper IV we describe the creation of a de-identified subset of this corpus, which can be used for training and evaluating automatic de-identification systems. Our aim is to make (at least) this subset available to other research groups, in order to enhance the possibilities of performing reliable and comparable research in this area.

Also, in the same paper, we try to introduce important questions that arise with such work, i.e. how should one define identifiable information? What is an appropriate level of de-identification in order to be able to distribute data for further research? Resources for working with unstructured health data are needed, especially for small languages such as Swedish.

Chartering Health Related Language in Swedish

What specific properties does health related language in Swedish have? Does health related text differ from other types of text from a linguistic point-of-view? We have initiated the linguistic analysis of this domain-specific language type, particularly for EHR language in Swedish. In Paper III we present some preliminary findings, which correspond well to previous research done within this area for other languages. Health related text sets, especially EHR text sets, have properties that differ from general text. For instance, the vocabulary is more specific, and contains a lot of domain-specific abbreviations and spellings that need to be handled in order to build useful IA systems. No previous studies have, to our knowledge, been published discussing the properties of Swedish EHR langugage. There are many issues left unexplored, which we plan to pursue in future experiments.

Representation Modelling

What kind of representation is optimal for modelling health related language, for instance for improving IA-systems? Modelling health related language in Swedish successfully for IA purposes is still a relatively open research question. In Paper I we present an attempt to model such data in alternative ways, using a phrase representation for Swedish scientific medical text. The hypothesis was that medical language contains more phrases than general text, and that this representation would improve clustering results. Surprisingly, the results were not very supportive, however many parameters might have affected these findings. Such an approach has not previously been used for Swedish text. Despite the discouraging results, we believe that alternative representation models for health related language, and also for other types of domain-specific language, are needed.

Automatic Generation of New Hypotheses

Is it possible to extract previously unknown information from health related text written in Swedish through clustering techniques? Health related data, especially free text parts in such data sets, are currently not utilised for automatic analysis and for extracting previously unknown information. In Paper II we present a method for accessing and revealing new information and relations from epidemiological questionnaire data, thus initiating the possibilities of exploiting and analysing open answers in questionnaires automatically. With this method, an analysis can be performed in minutes instead of several weeks of manual labour. Specifically, the method is interactive and iterative, involving the user for the interpretation of the information contained in the data. This approach is, to our knowledge, completely novel and we believe it has great potential. The method could of course be applied to similar data sets, such as EHR data. In Paper III we present findings of a preliminary study on applying this method on EHR data with promising results. We believe that visualisation- as well as user interaction techniques are invaluable when developing IA tools, especially for large data sets from which hypotheses may be generated.

Future Directions

With the knowledge gained from the research presented here, more focused work on modelling Swedish health related language is possible. In particular, methods for normalizing the domain-specific contents of EHR data in order to create more coherent and representative language models will be developed. For instance, analysing whether domain-specific compound splitting produces considerably better IA results and representation models is an interesting and important research question.

Also, abbreviations are very common in EHR data. A thorough investigation of their use, ambiguity level and number of different constructions for the same expansions may provide valuable information on how much a representation would change if this was taken into consideration. Spelling errors are also interesting to scrutinize, as they might contribute to skewed representations.

EHR data contains a potentially large amount of hidden information that may be of great importance for health research. Developing IA methods that aid researchers in exploring and analysing the contents is very important. We believe that user interaction is a key ingredient for using such information in the generation of new hypotheses.

A particularly interesting property of EHR data is the existence of speculative language. Physicians express their findings, thoughts and planned actions in free text. Whenever there is any uncertainty involved in a decision, this is difficult to capture with traditional IA methods. Such instances are, however, important to discriminate from other expressions, as they alter the outcomes drastically. Identifying such speculative language is an interesting research area that will be studied in future experiments. The hope is that this type of research will, in the long run, aid other research areas such as epidemiology, medical informatics and the like, as well as the crucial work performed on a daily basis at hospitals.

6. Bibliography

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596. ISSN 0891-2017.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley. ISBN 0-201-39829-X.

David Campbell and Stephen B. Johnson. 2001. Comparing Syntactic Complexity in Medical and non-Medical Corpora. In *Proceedings of the AMIA Annual Symposium*, pages 90–95.

Johan Carlberger, Hercules Dalianis, Martin Hassel, and Ola Knutsson. 2001. Improving precision in information retrieval for Swedish using stemming. In *Proceedings of the 13th Nordic Conference on Computational Linguistics (NODALIDA) '01*, Uppsala, Sweden.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254. ISSN 0891-2017.

Anni R. Coden, Sergey Pakhomov, Rie K. Ando, Patrick H. Duffy, and Christopher G. Chute. 2005. Domain-specific language models and lexicons for tagging. *Journal of Biomedical Informatics*, 38(6):422–430. ISSN 1532-0464.

Enrico Coiera. 2003. *Guide to health informatics*, 2nd edition. Arnold Publication, London. ISBN 0-340-76425-2.

Nigel Collier, Hyun S. Park, Norihiro Ogata, Yuka Tateishi, Chikashi Nobata, Tomoko Ohta, Tateshi Sekimizu, Hisao Imai, Katsutoshi Ibushi, and Jun ichi Tsujii. 1999. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL-99)*, pages 271–272.

Richard Craggs and Mary McGee Wood. 2005. Evaluating Discourse and Dialogue Coding Schemes. *Computational Linguistics*, 31(3):289–296. ISSN 0891-2017.

Hercules Dalianis. 2005. Improving search engine retrieval using a compound splitter for Swedish. In *Proceedings of the 15th Nordic Conference on Computational Linguistics (NODALIDA) '05*, Joensuu, Finland.

Hercules Dalianis, Martin Hassel, and Sumithra Velupillai. 2009. The Stockholm EPR Corpus – Characteristics and Some Initial Findings. In *Proceedings of The 14th International Symposium for Health Information Management Research (ISHIMR-09)*, Kalmar, Sweden, October 14-16 2009.

Richard S. Dick, Elaine B. Steen, and Don E. Detmer, editors. 1997. *Computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition*. National Academies Press.

Eva Ejerhed, Gunnel Källgren, and Benny Brodda. 2006. Stockholm Umeå Corpus Version 2.0, SUC 2.0. ISBN 91-631-5876-0.

Alexandra Ekman and Jan-Eric Litton. 2007. New times, new needs; e-epidemiology. *European Journal of Epidemiology*, 22(8):285–292.

Carol Friedman. 1997. Towards a Comprehensive Medical Language Processing System: Methods and Issues. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Fall Symposium*, pages 595–599.

Carol Friedman, George Hripcsak, William DuMouchel, Stephen B. Johnson, and Paul D. Clayton. 1995. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1(1): 83–108.

Carol Friedman, Lyudmila Shagina, Yves Lussier, and George Hripcsak. 2004. Automated Encoding of Clinical Documents Based on Natural Language Processing. *Journal of the American Medical Informatics Association*, 11(5):392–402.

Martin Gellerstam, Yvonne Cederholm, and Torgny Rasmark. 2000. The bank of Swedish. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC) 2000*, pages 329–333, Athens, Greece.

Ralph Grishman. 2001. Adaptive information extraction and sublanguage analysis. In B. Nebel, editor, *Proceedings of the Workshop on Adaptive Text Extraction and Mining at the 17th International Joint Conference on Artificial Intelligence (IJCAI'01)*, Seattle, USA.

Ralph Grishman and Richard Kittredge, editors. 1986. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Lawrence Erlbaum Association, Hillsdale, NJ.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference–6: A Brief History. In *Proceedings 16th International Conference on Computational Linguistics (COLING)*, pages 466–471, Morristown, NJ, USA. Association for Computational Linguistics.

Udo Hahn and Joachim Wermter. 2004. High-Performance Tagging on Medical Texts. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, Morristown, NJ, USA. Association for Computational Linguistics.

Martin Hassel. 2001. Internet as Corpus - Automatic Construction of a Swedish News Corpus. In *Proceedings of NODALIDA'01 - 13th Nordic Conference on Computational Linguistics*, Uppsala, Sweden, May 21-22 2001.

Marti A. Hearst. 1999. Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 3–10, Morristown, NJ, USA. Association for Computational Linguistics. ISBN 1-55860-609-3.

William Hersh and Ellen Voorhees. 2009. Trec genomics special issue overview. *Information Retrieval*, 12(1):1–15. ISSN 1386-4564.

HIPAA. 2003. Health Insurance Portability and Accountability Act (HIPAA), Privacy Rule and Public Health Guidance. From CDC and the U.S. Department of Health and Human Services. Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm. Accessed on May 26, 2009.

Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. 2005. Overview of BioCreAtIvE: critical Assessment of Information Extraction for Biology. *BMC Bioinformatics*, 6 Suppl 1. ISSN 1471-2105.

Anette Hulth. 2001. *The Gist of Written Documents: Automatic Derivation and Evaluation of Content Descriptors*. Licentiate thesis, Department of Computer and Systems Sciences, Stockholm University.

i2b2. 2009. Informatics for Integrating Biology & the Bedside. Partners Healthcare. Available at: https://www.i2b2.org/. Accessed on May 26, 2009.

A. K. Jain, M. N. Murty, and P. J. Flynn. 1999. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323. ISSN 0360-0300.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Viggo Kann, Rickard Domeij, Joachim Hollman, and Mikael Tillenius. 2001. *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Ludek Hrebicek*, volume 60, chapter Implementation aspects and applications of a spelling correction algorithm. WVT Wissenschaftlicher Verlag Trier.

Dimitrios Kokkinakis and Anders Thurin. 2007. Identification of Entity References in Hospital Discharge Letters. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA) 2007*, Tartu, Estonia.

Chris D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press. ISBN 0521865719.

Chris D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA. ISBN 0-262-13360-1.

Mitchell P. Marcus, Beatrice Santorini, and Mary A. Marcinkiewicz. 1994. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Eneida A. Mendonca, Janet Haas, Lyudmila Shagina, Elaine Larson, and Carol Friedman. 2005. Extracting Information on Pneumonia in Infants Using Natural Language Processing of Radiology Reports. *Journal of Biomedical Informatics*, 38(4):314–321.

MeSH. 2008. Medical Subject Headings, National Library of Medicine. Available at: http://www.nlm.nih.gov/mesh/. Accessed on May 26, 2009.

Stéphane M. Meystre, Guergana K. Savova, Karin C. Kipper-Schuler, and John E. Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *IMIA Yearbook of Medical Informatics 2008. 47 Suppl 1:138-154*.

Ishna M. Neamatullah, Margaret Douglass, Li wei H. Lehman, Andrew Reisner, Mauricio Villarroel, William J. Long, Peter Szolovits, George B. Moody, Roger G. Mark, and Gari D. Clifford. 2008. Automated de-identification of free text medical records. *BMC Medical Informatics and Decision Making*, 32(8).

Inga Nilsson. 2007. *Medicinsk dokumentation genom tiderna*. Enheten för medicinens historia, Lunds universitet. ISBN 978-91-633-1987-7. In Swedish.

OHNLP. 2009. Open Health Natural Language Processing Consortium. Available at: https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/OHNLP. Accessed on May 26, 2009.

Serguei Pakhomov, James Buntrock, and Patrick Duffy. 2005a. High Throughput Modularized NLP System for Clinical Text. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI, June 2005.

Serguei Pakhomov, Anni Coden, and Christopher G. Chute. 2005b. Developing a corpus of clinical notes manually annotated for part-of-speech. *International Journal of Medical Informatics*, 75(6):418–429. ISSN 1386-5056.

Serguei Pakhomov, Ted Pedersen, and Christopher G. Chute. 2005c. Abbreviation and Acronym Disambiguation in Clinical Discourse. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 589–593.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106. ISSN 0891-2017.

Göran Petersson and Martin Rydmark, editors. 1996. *Medicinsk informatik*. Almqvist & Wiksell Medicin, Liber Utbildning. In Swedish.

Massimo Poesio. 2004. Discourse Annotation and Semantic Annotation in the GNOME corpus. In Bonnie Webber and Donna K. Byron, editors, *Proccedings of the Association for Computational Linguistics (ACL) 2004 Workshop on Discourse Annotation*, pages 72–79, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Magnus Rosell. 2003. Improving clustering of Swedish newspaper articles using stemming and compound splitting. In *Proceedings of the 14th Nordic Conference on Computational Linguistics (NODALIDA) '03*, Reykjavik, Iceland.

Magnus Rosell. 2009. *Text Clustering Exploration – Swedish Text Representation and Clustering Results Unraveled*. PhD thesis, School of Computer Science and Communication, Royal Institute of Technology, Stockholm, Sweden. ISBN 978-91-7415-251-7.

Magnus Rosell and Sumithra Velupillai. 2005. The Impact of Phrases in Document Clustering for Swedish. In *Proceedings of the 15th Nordic Conference on Computational Linguistics – NODALIDA '05*, Joensuu, Finland.

Magnus Rosell and Sumithra Velupillai. 2008. Revealing relations between open and closed answers in questionnaires through text clustering evaluation. In European Language Resources Association (ELRA), editor, *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, Marrakesh, Morocco, May 2008.

Patrick Ruch. 2002. Using contextual spelling correction to improve retrieval effectiveness in degraded text collections. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 345–353, Taipei, Taiwan.

Patrick Ruch, Robert Baud, and Antoine Geissbühler. 2003. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial Intelligence in Medicine*, 29(1-2): 169–184.

Gerald Salton, A. Wong, and C. S. Yang. 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM*.

Karen Spärck-Jones and Julia R. Galliers. 1995. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Number 1083 in Lecture Notes in Artificial Intelligence. Springer.

Michael Steinbach, George Karypis, and Vipin Kumar. 2000. A comparison of document clustering techniques. In *Proceedings of the Workshop on Text Mining, Sixthth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, USA.

Hanna Suominen, Tuija Lehtikunnas, Barbro Back, Helena Karsten, Tapio Salakoski, and Sanna Salantera. 2007. Applying language technology to nursing documents: Pros and cons with a focus on ethics. *International Journal of Medical Informatics*, 76 Suppl 2:293–301.

Don R. Swanson and Neil R. Smalheiser. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91(2):183–203. ISSN 0004-3702.

György Szarvas, Richard Farkas, and Robert Busa-Fekete. 2007. State-of-the-art anonymization of medical records using an iterative machine learning framework. *Journal of the American Medical Informatics Association*, 14: 574–580.

UMLS. 2009. Unified Medical Language System, National Library of Medicine. Available at: http://www.nlm.nih.gov/research/umls/. Accessed on May 26, 2009.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the State-of-the-Art in Automatic De-identification. *Journal of American Medical Informatics Association*, 14:550–563.

Özlem Uzuner, Tawanda C. Sibanda, Yuan Luo Y, and Peter Szolovits. 2008. A De-identifier for Medical Discharge Summaries. *Artificial Intelligence in Medicine*, 42(1):13–35. ISSN 0933-3657.

Keith C. J. Van Rijsbergen. 1979. *Information Retrieval, 2nd edition*. Deptartment of Computer Science, University of Glasgow.

Sumithra Velupillai, Hercules Dalianis, Martin Hassel, and Gunnar H. Nilsson. 2009. Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics (2009)*. In Press. doi:10.1016/j.ijmedinf.2009.04.005.

Joachim Wermter and Udo Hahn. 2004. Really, is medical sublanguage that different? Experimental counter-evidence from tagging medical and newspaper corpora. In *Proceedings of the 11th World Congress on Medical Informatics (MEDINFO 2004)*.

7. Included Papers

Paper I.
The Impact of Phrases in Document Clustering for Swedish (Rosell and Velupillai, 2005)

Paper II.
Revealing Relations between Open and Closed Answers in Questionnaires through Text Clustering Evaluation (Rosell and Velupillai, 2008)

Paper III.
The Stockholm EPR Corpus – Characteristics and Some Initial Findings (Dalianis *et al.*, 2009)

Paper IV.
Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial (Velupillai *et al.*, 2009)

Paper I

The Impact of Phrases in Document Clustering for Swedish

Magnus Rosell and Sumithra Velupillai. 2005. *The Impact of Phrases in Document Clustering for Swedish.* In Proceedings of the 15th Nordic Conference on Computational Linguistics (NODALIDA '05), Joensuu, Finland, May 20–21 2005.

The Impact of Phrases in Document Clustering for Swedish

Magnus Rosell and **Sumithra Velupillai**
KTH Nada
100 44 Stockholm
Sweden
{rosell, sumithra}@nada.kth.se

Abstract

We have investigated the impact of using phrases in the vector space model for clustering documents in Swedish in different ways. The investigation is carried out on two text sets from different domains: one set of newspaper articles and one set of medical papers.

The use of phrases do not improve results relative the ordinary use of words. The results differ significantly between the text types. This indicates that one could benefit from different text representations for different domains although a fundamentally different approach probably would be needed.

1 Introduction

For document clustering one normally uses the vector space model to represent texts. It is based on the distribution of single words over the texts in a set. We have investigated the impact of introducing phrases in this representation for Swedish in different ways and in different domains. Our hypothesis was that phrases would improve results and that the improvement would be greater for the medical papers than for the newspaper articles as we believe that phrases carry more significance in the medical domain.

To calculate similarity between documents with respect to their phrases we use a word trie (in one set of experiments). This approach has a lot in common with the method presented in (Hammouda and Kamel, 2004). They show improvements in clustering results on web pages using phrases combined with single words, using other algorithms than we. Another related method is the Phrase-Intersection Clustering method which has been proven efficient on web pages (Zamir and Etzioni, 1998). It is based on word-n-grams rather than phrases.

2 Text Sets

We have used a set of 2500 newspaper articles from KTH News Corpus (AB) (Hassel, 2001) and a set of 2422 medical papers from Läkartidningen[1] (Med). In Table 1 some statistics for the sets are given.

We need categorizations of the text sets for the evaluation. The newspaper articles have been categorized by the paper into five sections such as Economy and Sports etc.

The medical papers are categorized with The Medical Subject Headings (MeSH) thesaurus[2]. This thesaurus is (poly)hierarchical with a term and a unique code at each place in it. The terms are not unique and may occur at several places in the hierarchy. There are 15 *broad headings* at the top level.

Each paper has one or more terms from the thesaurus assigned to it. This categorization is very extensive, but also very hard to handle for clustering evaluation. Hence we have made four attempts to flatten and disambiguate it so that each paper belongs to only one of a set of non overlapping categories.

We have made three categorizations where we try to put each document into one of

[1] http://www.lakartidningen.se/

[2] http://www.nlm.nih.gov/mesh/meshhome.html

Text Set	Categories	Documents	Words	Unique Words
AB	5	2500	119401	5896
Med	15, 814	2422	4383169	26102

Table 1: Text Sets

15 categories corresponding to the 15 broad headings. The first, which we call General, is constructed by choosing the broad heading to which most of the MeSH-terms assigned to the paper belongs.

By choosing the broad heading under which the most specific term (the term deepest into the hierarchy) is found for each paper we have constructed the second categorization, which we call Specific.

Many of the papers have as one of the terms assigned to it one or several broad headings. In the third categorization we have chosen this (always one) as the categorization of those papers. The other papers are categorized using the same system as for our categorization Specific. We call this categorization Combined.

We have made a fourth categorization which we call Term. In this every paper is assigned the MeSH-term that has the highest frequency among the terms assigned to it. This leads to a categorization with 817 categories.

The categorizations General and Combined are those that seem most trustworthy. A paper may probably have a very specific term assigned without having its broad heading as the general focus (see Specific). Terms at different levels of the MeSH-hierarchy probably make up an unequal categorization (see Term).

3 Linguistics

We used the grammar checking program Granska[3] to extract nominal phrases from the texts and a stemmer (Carlberger et al., 2001) to stem all words. To prevent very similar but not identical phrases to be deemed unsimilar we removed stopwords within the phrases as well as from the single words.

Swedish solid compounds often correspond to phrases (or compounds) in other languages. We use the spell checking program Stava (Kann et al., 2001) to split them. An earlier study (Rosell, 2003) has proven this to improve clustering results for newspaper articles. We also try to represent the split compounds as phrases and try to split compounds within phrases (see Section 5).

4 Similarity

When calculating the similarity between two documents using phrases two natural alternatives are at hand. Either one chooses to deem phrases similar only if they are identical or one looks at the overlap of words between them. We have tried both. In the first case we have calculated the weight for each phrase in a document as the frequency of its appearance in that document multiplied by the sum of the idf-weight for the single words in it.

To find the overlaps of phrases in documents we have built a trie based on words for each document from the phrases appearing in them. Each phrase is put into the trie in its entire and with all but the first word, with all but the first two words, etc. In each node of the trie we save the number of times it has been reached. To calculate the overlap of phrases between two documents we follow all common paths in the tries and multiply relative appearances in each node weighted by the sum of the idf-weights for the words along the path.[4]

5 Representations

From the phrases and single words we built several different representations. Refer to Table 2 through this section.

Combining all the described possibilities (full phrases or overlap, using split com-

[3] http://www.nada.kth.se/theory/projects/granska/

[4] Compare with Phrase-Intersection Clustering in (Zamir and Etzioni, 1998).

Repr.	Description			
Worst	The worst possible result			
Rand	Random partiton of the set – average for ten iterations			
Best	The best possible result			
1	Only words, stemming			
2	Only words, stemming and splitting of compounds			
3	P	PM	NSP	NSC
4	P	PM	NSP	SC
5	P	PM	SP	NSC
6	P	PM	SP	SC
7	P	POM	NSP	NSC
8	P	POM	NSP	SC
9	P	POM	SP	NSC
10	P	POM	SP	SC
11	P&W	PM	NSP	NSC
12	P&W	PM	NSP	SC
13	P&W	PM	SP	NSC
14	P&W	PM	SP	SC
15	P&W	POM	NSP	NSC
16	P&W	POM	NSP	SC
17	P&W	POM	SP	NSC
18	P&W	POM	SP	SC

Abbr.	Explanation
P	Similarity only between phrases
P&W	Similarity using both phrases and words
PM	Phrase-match
POM	Phrase-overlap-match
SP	Use splitted compounds as phrases
NSP	Do not use splitted compounds as phrases
SC	Split compounds within phrases
NSC	Do not split compounds within phrases

Table 2: Representations

pounds as phrases or not, and split compounds within phrases or not) we get eight different representations based on phrases. By combining[5] these with the ordinary single word representation with split compounds we get eight more. This gives 16 representations (representations 3 through 18 in Table 2). We also made the reference representation (only words, 1) and the representation where solid compounds have been split (2), giving in total 18 different representations.

Finally, for comparison we also try a random "clustering" (Rand) and in the evaluation we present the theoretical worst (Worst) and best (Best) possible results (see Sections 7 and 8).

6 Clustering Algorithm

The clusterings have been made using the divisive algorithm Bisecting K-Means (Steinbach et al., 2000) which splits the worst cluster (i.e. largest) in two, using K-Means, until the desired number of clusters are reached. We have let the K-Means algorithm iterate ten times and for each split we ran it five times and picked the best split (evaluated using the similarity measure). Average results are calculated over ten runs to ten clusters for each representation.

7 Evaluation

As we compare different representations we use extrinsic evaluation measures that requires a categorization of the the same text set to compare with. Among the extrinsic evaluation measures that have been used for text clustering are *the purity* and *the entropy.* These measures are well suited for evaluation of single clusters, but for evaluation of whole clusterings *the mutual information* is better. (Strehl et al., 2000)

Consider a text set N with n texts. Let C be a clustering with γ clusters, c_1 through c_γ. By n_i we mean the number of texts in cluster c_i ($\sum_{i=1}^{\gamma} n_i = n$). Similarly, let K be a categorization with κ categories, $k^{(1)}$ through $k^{(\kappa)}$ and let $n^{(j)}$ denote the number of texts in category $k^{(j)}$.

The γ by κ matrix M describes the distribution of the texts over both C and K; that is $m_i^{(j)}$ is the number of texts that belong to c_i and $k^{(j)}$.

The mutual information of clustering C and

[5] We use equal weight on the two different representations. In (Hammouda and Kamel, 2004) they try different weightings.

categorization K is:

$$MI(C,K) = \sum_{i=1}^{\gamma}\sum_{j=1}^{\kappa} \frac{m_i^{(j)}}{n} \log(\frac{m_i^{(j)} n}{n_i n^{(j)}}) \quad (1)$$

A theoretical tight upper bound is $MI_{max}(C,K) = \log(\kappa\gamma)/2$, the mean of the theoretical maximal entropy of the clustering and the categorization. By dividing the mutual information by this we get a normalized measure. (Strehl, 2002)

This normalization is theoretical and particular for each clustering-categorization-setting. We want to compare results on different such settings, with different text sets, having varying clustering complexity/difficulty. Therefore we need to normalize with regard to something else.

Since we want to know how much introducing phrases improve results we use the result from a clustering using only words as a reference. By comparing the results with this reference we take the complexity of the different text sets into account.

There are two simple and reasonable ways of normalizing the result using the word clustering result, $MI(C_{word}, K)$. We can divide the result by it:

$$MI_{word}(C,K) = \frac{MI(C,K)}{MI(C_{word},K)}, \quad (2)$$

or we can divide the improvement by the maximum possible improvement from the word clustering result:

$$MI_{imp}(C,K) = \frac{MI(C,K) - MI(C_{word},K)}{MI_{max}(C,K) - MI(C_{word},K)} \quad (3)$$

The first normalization is suitable when we have a decrease in performance. It puts the decrease in relation to the greatest possible decrease. The second normalization is suitable when we have an increase in performance.

8 Results

We present the results of our investigation in Tables 3 and 4. All values are average results over ten clusterings with standard deviation within parenthesis.

The first row of each part of these tables gives the results for the newspaper articles and the following the results on the medical papers compared to the different categorizations. In Table 4 we only give results for representations Term and General as the results for Combined, General and Specific are very similar.

The columns represent the different representations which were described in Section 2 and summarized in Table 2. In Table 3 we present the result for a random “clustering” (the average of 10 random partitions of the text set) and the theoretical worst and best possible results.

9 Discussion

When, in the following discussion, we refer to the results on the medical papers we consider the results on the categorization General (which is very similar to results on Combined and Specific). The results with respect to the categorization Term of the medical papers are a bit different than for the others. As we believe the other categorizations to be better we do not discuss this further.

To split *compounds* in the representation based only on words (representation 2 compared to 1) improve results when clustering the newspaper articles but not when clustering the medical papers. This may be because compounds in the medical papers would need a different analysis. We have also used a stoplist for certain words that should not be split based on other newspaper articles as described in (Rosell, 2003). An optimization for medical compounds here would perhaps improve results.

All variations of clustering using *phrases* performs worse than clustering using only words. Clustering performs worse when using only phrases (representations 3-10) than when using the combination of words and phrases (representations 11-18). Since clustering using words is superior the impact of phrases is diminished in the combined representations (11-18).

Looking at the *representations based only on phrases* (3-10) we see that results on news-

	Measures	Worst	Rand		Best	1		2	
AB	MI	0.000	0.009	(0.003)	2.822	0.947	(0.043)	1.093	(0.084)
	MI_{word}	−100.0%	−99.0%	(0.3%)	198.0%	0.0%	(4.6%)	15.4%	(8.9%)
	MI_{imp}	−50.5%	−50.0%	(0.2%)	100.0%	0.0%	(2.3%)	7.8%	(4.5%)
Combined	MI	0.000	0.038	(0.006)	3.614	0.407	(0.016)	0.415	(0.010)
	MI_{word}	−100.0%	−90.6%	(1.4%)	787.9%	0.0%	(4.0%)	2.0%	(2.4%)
	MI_{imp}	−12.7%	−11.5%	(0.2%)	100.0%	0.0%	(0.5%)	0.3%	(0.3%)
General	MI	0.000	0.041	(0.005)	3.614	0.478	(0.013)	0.486	(0.016)
	MI_{word}	−100.0%	−91.5%	(1.1%)	656.0%	0.0%	(2.7%)	1.7%	(3.4%)
	MI_{imp}	−15.2%	−13.9%	(0.2%)	100.0%	0.0%	(0.4%)	0.3%	(0.5%)
Specific	MI	0.000	0.038	(0.005)	3.614	0.396	(0.010)	0.397	(0.017)
	MI_{word}	−100.0%	−90.4%	(1.2%)	812.6%	0.0%	(2.6%)	0.1%	(4.2%)
	MI_{imp}	−12.3%	−11.1%	(0.1%)	100.0%	0.0%	(0.3%)	0.0%	(0.5%)
Term	MI	0.000	1.450	(0.008)	6.498	1.850	(0.023)	1.868	(0.018)
	MI_{word}	−100.0%	−21.6%	(0.5%)	251.2%	0.0%	(1.2%)	1.0%	(0.9%)
	MI_{imp}	−39.8%	−8.6%	(0.2%)	100.0%	0.0%	(0.5%)	0.4%	(0.4%)

Table 3: Text Clustering Results (stdv)

	Measures	3		4		5		6	
AB	MI	0.067	(0.020)	0.071	(0.017)	0.086	(0.024)	0.080	(0.032)
	MI_{word}	−93.0%	(2.1%)	−92.5%	(1.8%)	−91.0%	(2.6%)	−91.5%	(3.4%)
General	MI	0.112	(0.008)	0.117	(0.012)	0.028	(0.005)	0.030	(0.002)
	MI_{word}	−76.6%	(1.7%)	−75.4%	(2.5%)	−94.2%	(1.1%)	−93.7%	(0.4%)
Term	MI	1.547	(0.020)	1.547	(0.013)	0.574	(0.096)	0.585	(0.022)
	MI_{word}	−16.4%	(1.1%)	−16.4%	(0.7%)	−69.0%	(5.2%)	−68.4%	(1.2%)

	Measures	7		8		9		10	
AB	MI	0.095	(0.020)	0.150	(0.024)	0.071	(0.021)	0.058	(0.010)
	MI_{word}	−90.0%	(2.1%)	−84.1%	(2.5%)	−92.5%	(2.2%)	−93.9%	(1.0%)
General	MI	0.148	(0.011)	0.178	(0.015)	0.031	(0.005)	0.037	(0.025)
	MI_{word}	−69.0%	(2.4%)	−62.7%	(3.1%)	−93.5%	(1.0%)	−92.2%	(5.2%)
Term	MI	1.565	(0.033)	1.607	(0.027)	0.506	(0.045)	0.694	(0.269)
	MI_{word}	−15.4%	(1.8%)	−13.2%	(1.4%)	−72.6%	(2.5%)	−62.5%	(14.6%)

	Measures	11		12		13		14	
AB	MI	0.820	(0.051)	0.809	(0.057)	0.946	(0.078)	0.919	(0.100)
	MI_{word}	−13.4%	(5.4%)	−14.6%	(6.0%)	−0.1%	(8.2%)	−3.0%	(10.6%)
General	MI	0.148	(0.016)	0.168	(0.018)	0.210	(0.013)	0.216	(0.013)
	MI_{word}	−69.0%	(3.4%)	−64.8%	(3.8%)	−56.0%	(2.7%)	−54.9%	(2.8%)
Term	MI	1.562	(0.022)	1.566	(0.021)	1.314	(0.052)	1.336	(0.064)
	MI_{word}	−15.6%	(1.2%)	−15.4%	(1.1%)	−29.0%	(2.8%)	−27.8%	(3.5%)

	Measures	15		16		17		18	
AB	MI	0.746	(0.090)	0.734	(0.063)	0.954	(0.063)	0.940	(0.061)
	MI_{word}	−21.3%	(9.5%)	−22.5%	(6.7%)	0.8%	(6.7%)	−0.8%	(6.4%)
General	MI	0.226	(0.022)	0.230	(0.007)	0.217	(0.029)	0.247	(0.020)
	MI_{word}	−52.8%	(4.5%)	−52.0%	(1.5%)	−54.7%	(6.1%)	−48.3%	(4.3%)
Term	MI	1.642	(0.026)	1.649	(0.033)	1.460	(0.054)	1.486	(0.048)
	MI_{word}	−11.2%	(1.4%)	−10.9%	(1.8%)	−21.1%	(2.9%)	−19.7%	(2.6%)

Table 4: Results for Text Clustering with Phrases (stdv)

paper articles are almost as bad as random clustering for all of them. The performance on the medical papers, on the other hand, is better than random clustering as long as we do not use split compounds as phrases. It is also better here to use the word trie representation (POM) rather than the simple phrase match (PM). In all this is an indication that phrases contain more information in the medical papers than in the newspaper articles.

For the *combined representations* (11-18) the results are much harder to analyze as the word representation is so much better than the phrase representation. The results on the newspaper articles are much better than on the medical papers here. This could be since the phrase representations do not contain as much information for the newspaper articles as for the medical papers and they thereby obscure the clustering to a lesser extent. Concerning the medical papers, all what is stated for the representations using only phrases holds, except that here it is not negative to use the split compounds as phrases (SP). For the newspaper articles there is even a great increase in performance when using the split compounds as phrases. This could be explained if the phrase representations using split compounds gives no information, which the results for representations 3-10 indicates. There is no reliable difference between the use of simple phrase match and the word trie representations for the newspaper articles as the standard deviation is very high.

No cases show any change in performance when splitting compounds within phrases (SC) or not. The reason for this could be beacuse the amount of compounds within phrases is small.

It is important to bear the great *differences of the two text sets* in mind. The differences in results between them show that clustering works differently on corpora with different contents (i.e. medical text vs. newspaper text). However, this difference might as well to a great extent be explained by other things, such as the structure and size of the texts and the difference of the categorizations. The medical papers are much longer than the newspaper articles. This could in fact explain all of the differences between them regarding information found in the phrases and the compounds. The categorization of the newspaper articles is probably much better than our categorizations of the medical articles.

10 Conclusions and Further Work

Phrases do not improve clustering in Swedish. At least with the representations tried here. The impact of phrases is more obvious on the medical papers. Overlap match between phrases performs better than simple match. It seems to be bad to consider split compounds as phrases and it does not matter whether one splits compounds within phrases or not.

Splitting solid compounds for the ordinary word representation improves results for the newspaper articles and does not make results worse for the medical papers.

The results are very different for the two text types, the newspaper articles and the medical papers. Maybe one would need to develop different representations for different text types. The information found in the phrases of the medical papers could perhaps be exploited using some other representation. But the same information might also be found in the ordinary representation using only words.

Our results are different from those presented in (Hammouda and Kamel, 2004). This is presumably, at least partially, because of differences between Swedish and English. Swedish solid compounds often correspond to phrases in English.

It could be interesting to try other variations of the representations using phrases presented here, but to really use the information that phrases contain relative to mere words a fundamentally different approach is probably needed. One interesting obvious extension of the present work is, however, to look at word-n-grams instead of phrases as these has proven useful in other works.

Acknowledgements

Our thanks go to professor Viggo Kann for supervision, to VR (Vetenskapsrådet – the Swedish Research Council) for funding and to all participants in the project “Infomat” from KTH Nada (department of Computer Science and Numerical Analysis at the Royal Institute of Technology) and the department of Medical Epidemiology and Biostatistics (MEB) at Karolinska Institutet, both in Stockholm, Sweden.

References

J. Carlberger, H. Dalianis, M. Hassel, and O. Knutsson. 2001. Improving precision in information retrieval for Swedish using stemming. In *Proc. 13th Nordic Conf. on Comp. Ling. – NODALIDA ’01.*

K. M. Hammouda and M. S. Kamel. 2004. Efficient phrase-based document indexing for web document clustering. *IEEE Transactions on Knowledge and Data Engineering*, 16(10):1279–1296. Student Member-Khaled M. Hammouda and Senior Member-Mohamed S. Kamel.

M. Hassel. 2001. Automatic construction of a Swedish news corpus. In *Proc. 13th Nordic Conf. on Comp. Ling. – NODALIDA ’01.*

V. Kann, R. Domeij, J. Hollman, and M. Tillenius, 2001. *Text as a Linguistic Paradigm: Levels,* *Constituents, Constructs. Festschrift in honour of Ludek Hrebicek*, volume 60, chapter Implementation aspects and applications of a spelling correction algorithm.

M. Rosell. 2003. Improving clustering of swedish newspaper articles using stemming and compound splitting. In *Proc. 14th Nordic Conf. on Comp. Ling. – NODALIDA '03*.

M. Steinbach, G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques. In *Proc. Workshop on Text Mining, 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*.

A. Strehl, J. Ghosh, and R. Mooney. 2000. Impact of similarity measures on web-page clustering. In *Proc. AAAI Workshop on AI for Web Search (AAAI 2000), Austin*, pages 58–64. AAAI/MIT Press, July.

A. Strehl. 2002. *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. Ph.D. thesis, The University of Texas at Austin.

O. Zamir and O. Etzioni. 1998. Web document clustering: A feasibility demonstration. In *Research and Development in Information Retrieval*, pages 46–54.

Paper II

Revealing Relations between Open and Closed Answers in Questionnaires through Text Clustering Evaluation

Magnus Rosell and Sumithra Velupillai. 2008. *Revealing Relations between Open and Closed Answers in Questionnaires through Text Clustering Evaluation.* In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC ’08), Marrakech, Morocco, May 28–30 2008.

Revealing Relations between Open and Closed Answers in Questionnaires through Text Clustering Evaluation

Magnus Rosell[*], **Sumithra Velupillai**[†]

[*] KTH CSC
100 44 Stockholm
Sweden
rosell@csc.kth.se
[†] DSV, KTH - Stockholm University
Forum 100
164 40 Kista
Sweden
sumithra@dsv.su.se

Abstract

Open answers in questionnaires contain valuable information that is very time-consuming to analyze manually. We present a method for hypothesis generation from questionnaires based on text clustering. Text clustering is used interactively on the open answers, and the user can explore the cluster contents. The exploration is guided by automatic evaluation of the clusters against a closed answer regarded as a categorization. This simplifies the process of selecting interesting clusters. The user formulates a hypothesis from the relation between the cluster content and the closed answer categorization. We have applied our method on an open answer regarding occupation compared to a closed answer on smoking habits. With no prior knowledge of smoking habits in different occupation groups we have generated the hypothesis that farmers smoke less than the average. The hypothesis is supported by several separate surveys. Closed answers are easy to analyze automatically but are restricted and may miss valuable aspects. Open answers, on the other hand, fully capture the dynamics and diversity of possible outcomes. With our method the process of analyzing open answers becomes feasible.

1. Introduction

Questionnaires are an important source for new research findings in many scientific disciplines, as well as for commercial exploitation. They may contain both closed ended and open ended questions. The answers to these are called closed and open answers, respectively. Closed answers are restricted to a fixed set of replies, while open answers are not. Statistical methods can be used to study closed answers in large questionnaires. Open answers must be reviewed manually.

Open answers contain valuable and detailed information that is very time-consuming to analyze manually. Methods for assisting the process of analyzing open answers in questionnaires are needed. Natural Language Processing tools could aid such processes, by enhancing the quality of the methods and therefore also the end results.

In Text Mining methods for discovering new, previously unknown information from large text sets are studied (Hearst, 1999). One such method is text clustering, which divides a set of texts into groups (clusters) of texts with similar content. As the content of clusters usually is divers, human investigation and interpretation is needed. The investigation can be aided by the clustering method in several ways. For clustering to be really useful both textual and visual presentation of the clusters should allow the user to explore the results, and interactively focus on interesting and intricate parts.

Collecting large sets of demographic and lifestyle data systematically is central for epidemiological studies. In (Ekman et al., 2006) the feasibility of using web-based questionnaires is discussed. Moving towards e-epidemiology increases the possibilities of conducting large population-based studies immensely, both with respect to cost-efficiency and availability (Ekman and Litton, 2007).

We present a method for hypothesis generation using text clustering, involving human judgement in crucial steps. The method is applied to a large epidemiological questionnaire with promising results.

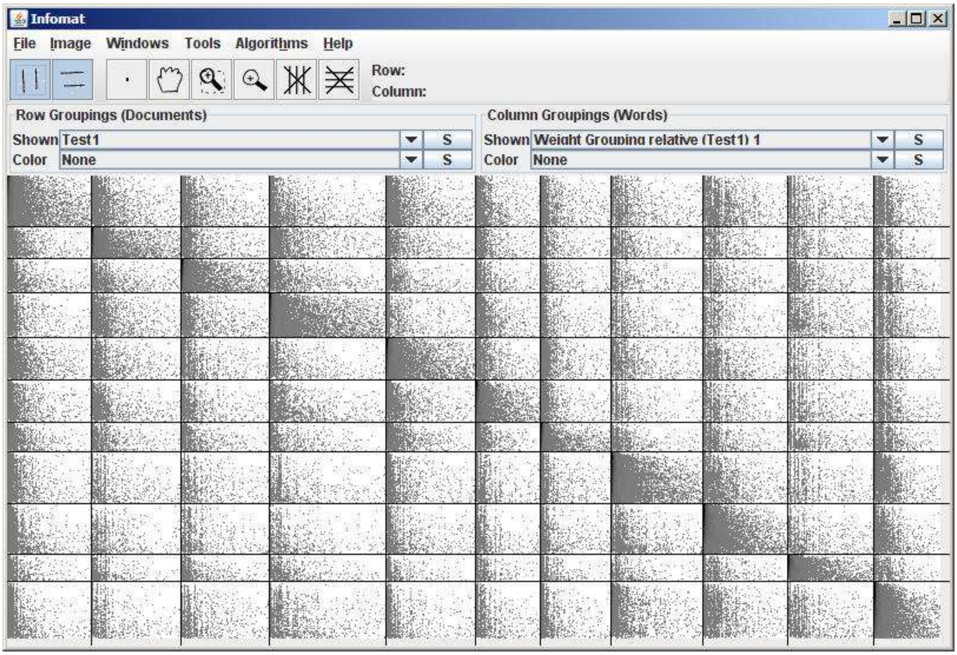
2. Related Work

Swanson and Smalheiser (1997) describe a method for hypothesis generation by linking possibly related medical literature. Their method exploits existing literature in order to discover previously unknown information and involves user interaction.

In the Scatter/Gather-system (Cutting et al., 1992) clustering is used as a tool for exploration of text sets. Clusterings are presented in a textual format and the user can interactively choose to re-cluster parts of the result, homing in on interesting themes.

To our knowledge, little research has been performed on automatically revealing new information from open answers in questionnaire data. Li and Yamanishi (2001) present a method for analyzing open answers in questionnaires using rule analysis and correspondence analysis. They describe a few other systems, but information about these is not readily found.

Central to all exploration methods is human interaction. Exploration of unstructured information requiers human interpretation.



Row Groupings (Documents)

Shown Test1

Color None

Column Groupings (Words)

Shown Weight Grouping relative (Test1) 1

Color None

Figure 1: Infomat. 41 549 texts (rows) from the questionnaire presented in Section 4. clustered to 11 clusters (K-Means), represented by 5 978 words (columns). Clusters are separated by lines. The text clusters are sorted according to smoking purity, where those with the highest amount of smokers are found at the top. The texts in each cluster are sorted in order of similarity to the cluster centroid. The words are clustered using the algorithm of Figure 3. Within each word cluster the words are sorted in order of weight in the corresponding text cluster centroid. A distinct diagonal is visible in the 11-by-11 pattern as could be expected. (The opacity of each pixel is proportional to the sum of the weights of its matrix elements.)

3. Method

We propose a method for hypothesis generation from open answers in questionnaires based on text clustering. The method could be described as follows:

1. Cluster the text set
2. Identify interesting clusters
3. Explore cluster contents
4. Formulate potential hypotheses

These steps should be repeated several times. For each repetition different settings (text representation, different clustering algorithms, etc.) could be used. Any recurring hypotheses may be further studied, through literature studies or new surveys.

The proposed method is semi-automatic and can easily be applied using the Infomat tool (see Section 3.1.). User interaction is a central part of the process. Human judgement is required to draw relevant conclusions in each step.

3.1. The Infomat Tool

Infomat[1] is a vector space visualization tool aimed at Information Retrieval (IR) and text clustering in particular (Rosell, 2007). It incorporates the ideas from the Scatter/Gather-system (Cutting et al., 1992), adding new functionality.

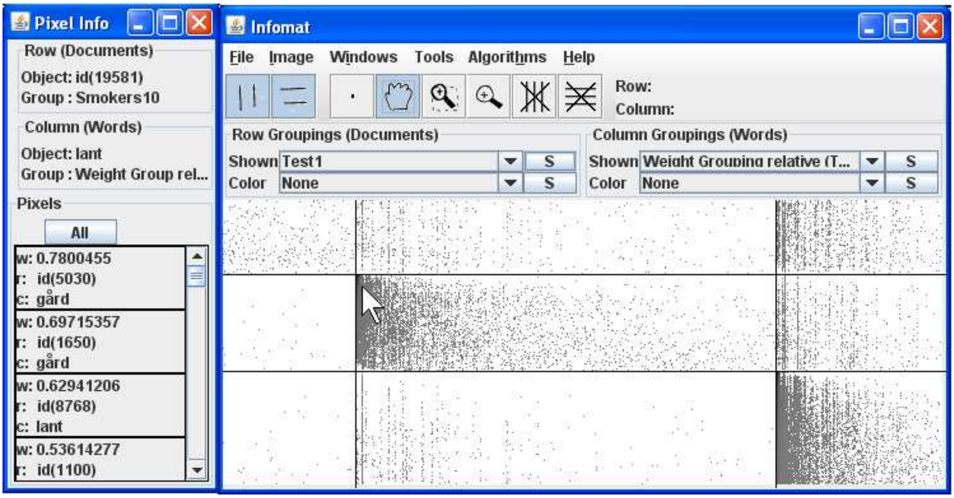
Infomat presents information stored in a matrix as a scatter plot, where the opacity of each pixel is proportional to the weight(s) of the corresponding matrix element(s). Here texts are represented in the vector space model by a text-by-word matrix, see Figure 1 for an example.

By sorting the rows (texts) and columns (words) in different ways hidden relationships between the objects may be exposed as visual patterns. Since the rows and columns represent actual objects (texts and words), the visual patterns are possible to comprehend.

Textual information about the matrix can be obtained in different ways. For instance the text(s) and word(s) of each pixel are presented when the cursor is moved over the matrix. It is also possible to zoom in and out, in order to investigate parts of the matrix in more detail, see Figure 2.

Infomat allows the user to cluster both rows and columns. The algorithm introduced in Figure 3 constructs a clustering of the words relative to a text clustering. An extensive description of the content of a text cluster is given by the combination of the visual patterns and the corresponding relative word cluster. (Naturally, reading the actual texts in the clusters can provide further insights.)

[1] http://www.csc.kth.se/tcs/projects/infomat/infomat/



Row (Documents)
Object: id(19581)
Group : Smokers10

Column (Words)
Object: lant
Group : Weight Group rel...

Pixels

w: 0.7800455
r: id(5030)
c: gård

w: 0.69715357
r: id(1650)
c: gård

w: 0.62941206
r: id(8768)
c: lant

w: 0.53614277
r: id(1100)

Row:
Column:

Row Groupings (Documents)
Shown Test1 S
Color None S

Column Groupings (Words)
Shown Weight Grouping relative (T... S
Color None S

Figure 2: Infomat zoom example. A part of the picture in Figure 1 (centered around the second row and column clusters from the bottom right corner) is shown in the Infomat main window. The *Pixel Info* window to the left gives the matrix elements (weight, text, word) that are represented by the pixel indicated by the cursor. It also shows to which groups (along rows and columns) the texts and words belong. The Swedish word *gård* means “a farm” and *lant* could be translated to “country”. There are several more words in the scroll list.

```
Input:  a text set $\mathfrak{T}$,
        a set $\mathfrak{W}$ of all words appearing in $\mathfrak{T}$,
        a clustering of the texts $\{T_i\}$.

  • For each text cluster $T_i$:
      – calculate the centroid $\overline{T_i}$
      – construct an empty corresponding word cluster $W_i$
  • For each word $w \in \mathfrak{W}$:
      – find $\overline{T_k}$ with maximal weight for $w$
      – put $w$ in $W_k$, ordered by its weight in $\overline{T_k}$

Output:  a clustering of the words $\{W_i\}$.
```

Figure 3: *The Relative Clustering Algorithm*

3.2. Identifying Interesting Clusters

A closed answer in a questionnaire may be viewed as a categorization, making it possible to measure clusterings of open answers by ordinary clustering quality measures. If the categorization distribution (measured by a quality measure) in a cluster differs significantly from the entire set the cluster is potentially interesting. Whether a categorization distribution in a cluster differs sufficiently must be judged by the user and depends on the data set, the categorization, etc. In Infomat the clusters can be sorted in order of quality measure value, identifying the clusters with extreme values as the most interesting, see Figure 1 for an example.

In the context of clustering the quality measure *precision* (p) compares each cluster i to each category j in the categorization:

$$p_{ij} = \frac{n_{ij}}{n_i}, \tag{1}$$

where n_{ij} is the number of texts from category j in cluster i, and n_i is the number of texts in cluster i. From the dominating category we get the *purity* for each cluster:

$$\rho_i = \max_j \{p_{ij}\}. \tag{2}$$

The purity is a useful measure here as it is easy to understand. This helps in formulating the hypothesis, see Section 3.4..

3.3. Exploring Cluster Contents

One of the main challenges in text clustering is to describe the contents of the clusters to a user. Other text clustering tools, Scatter/Gather (Cutting et al., 1992) for instance, usually only present a headline consisting of some of the words with the highest weights in the cluster. However, short cluster headlines only provide a partial description of the cluster content, possibly omitting important characteristics.

For each text cluster a corresponding relative word cluster created by the algorithm in Figure 3 constitutes a cluster description. It provides an extensive overview of the cluster content, which can be grasped through browsing with the Infomat tool, as described in Section 3.1..

3.4. Formulating Hypotheses

If a cluster is deemed interesting, as described in Section 3.2., a hypothesis can be formulated from the cluster content (Section 3.3.). It can be expressed as a relation between the content and the closed answer distribution in the cluster. A hypothesis that recurs over several method iterations is worth investigating further.

3.5. Filtering Hypotheses

The generated hypotheses should be seen as starting points for further analysis. Therefore the exact quality measure values (in the identification of interesting clusters) are not that important – it is the tendencies that matters. Further, the hypotheses might not be novel as they are constructed solely from the investigated questionnaire. A domain expert can make well judged decisions on which tendencies to further pursue.

If the method produces an interesting hypothesis it can be considered useful. Whether the hypotheses holds can only be determined through further studies on material separated from the questionnaire.

3.6. Method Extensions

The method could be extended in several ways. In fact, the more ways the data is processed (revealing the same relations) the better. Several clusterings of rows and columns using different clustering algorithms can provide insights when combined. An especially interesting clustering technique, which is clearly related to the relative clustering algorithm in Figure 3, is *co-clustering* (Dhillon, 2001), where text and word clusterings are constructed simultaneously.

In the identification of interesting clusters, other quality measures, for instance *entropy*, could be used. They could be interesting as an aid in a general investigation of the text set. It is, however, harder to formulate a hypothesis using more abstract and complex measures than purity.

Several closed answers could be used in the identification of interesting clusters, for instance by constructing a categorization of the combination of them. If several open answers are available, clusterings of them could be used as well, considering any one of them a categorization. Further, the Infomat tool allows the user to view a second clustering or categorization along both rows and columns by coloring matrix elements depending on which cluster/category they belong to.

As presented here, the method relies heavily on human judgement. We believe this is unavoidable (and even desirable). Still, perhaps a more automated process could aid the human further in making these judgements. For instance, a predefined scheme of clusterings (and re-clusterings of parts of clustering results) could be run. The results of these could be presented in a condensed form, by for instance only displaying clusters that have been deemed sufficiently interesting automatically. This would make the identification of recurring relations more straightforward.

4. Text Set: Questionnaire

Karolinska Institutet (Swedish Medical University) administrates The Swedish Twin Registry[2], the largest twin registry in the world, containing information about more than 140 000 twins. See (Lichtenstein et al., 2002; Lichtenstein et al., 2006) for a description of the contents and some findings that have come from it.

	Gender	Smoking
ρ	0.52 (women)	0.71 (non-smokers)

Table 1: Gender and smoking purity for the entire set

	Women	Men
ρ	0.75 (non-smokers)	0.65 (non-smokers)

Table 2: The purity of smokers by gender for the entire set

The registry is based on information from questionnaires containing both closed and open answers. The combination of these provides a large set of valuable (medical, biological, sociological, etc.) information. Manual treatment of this is slow and costly.

The work presented here does not focus on revealing twin-specific information. Instead, the text set is used as an example to show how questionnaire data can be exploited.

4.1. An Open Answer on Occupation

Between 1998 and 2002, all twins born in or before 1958 were asked, among other things, to describe their occupation in a few words or sentences (in Swedish). The described occupation is either the last or the primary occupation during the respondent's lifetime. These answers provide a large set of texts with valuable but unaccessible information.

4.2. Representation of the Open Answer

In our experiments we have used the vector space model with tf*idf-weighting to represent the texts and the cosine measure for calculating similarity between texts and clusters. After applying a stoplist, we split compounds using the spell checking program STAVA (Kann et al., 2001) and conduct lemmatization using the grammar checking program Granska[3]. In (Rosell, 2003) improvements in clustering results on Swedish news texts using such techniques are reported.

After preprocessing 41 549 texts remained, having on average 10 different words (including compound parts). There were only 5 978 different words in total and each word occurred in on average 69 texts[4].

4.3. Closed Answers: Gender and Smoking

The questionnaire has several closed answers regarding smoking habits. We have constructed a categorization where we define *smokers* as respondents that have smoked more than a year, and *non-smokers* as all other. There are 12 244 smokers, that is 71% are non-smokers. Table 1 gives the smoking and gender purity for the entire set, and in Table 2 the purity of smokers by gender is shown.

[2] http://www.meb.ki.se/twinreg/index_en.html

[3] http://www.nada.kth.se/theory/projects/granska/

[4] After removing words that only occur in one text.

Clusters	Cluster A	Cluster B	Cluster C	Cluster D
Words	boss (chef)	drive (köra)	assistant (biträde)	country (lant)
	leader (ledare)	chauffeur (chaufför)	care (vård)	forest (skog)
	personell (personal)	car (bil)	home (hem)	farm (gård)
	company (företag)	driver (förare)	food (mat)	cultivator (brukare)
	work- (arbets)	lorry- (lastbils)	old (gammal)	animal (djur)
	task (uppgift)	lorry (lastbil)	cook (laga)	agriculture (lantbruk)
	administrative (administrativ)	truck (truck)	help (hjälpa)	agriculture (jordbruk)
	lead (leda)	taxi (taxi)	service (tjänst)	cow (ko)
	project (projekt)	load (lasta)	sick (sjuk)	worker (arbetare)
	responsibility (ansvar)	road carrier (åkeri)	housing (boende)	works (bruk)
Number of texts	3747	2037	4083	2231
Number of words	3358	2483	2706	2137
ρ(non-smokers)	0.64	0.65	0.76	0.78
ρ(gender)	0.73 (men)	0.90 (men)	0.91 (women)	0.64 (men)

Table 3: Example text clusters from a clustering to 20 clusters of the occupation answers. The two top and two bottom clusters sorted in order of smoking purity. The words are the highest ranked in the corresponding word clusters and have been manually translated from Swedish. The sizes of the text and relative word clusters, as well as the smoking and gender purity are also displayed.

5. Experiment

We have applied our method on the questionnaire, described in the previous section, using the Infomat tool with the K-Means algorithm. The latter since it is fast, which makes the waiting times quite acceptable and the exploration pleasant even on an ordinary home computer.

We clustered the open answers regarding occupation several times to different numbers of clusters. Each time we also applied the relative clustering algorithm (see Figure 3) to the words. An example clustering is given in Figure 1. We also compared each clustering to the closed answer to identify interesting clusters as described in Section 3.2. The text clusters of Figure 1 are sorted in order of purity of smokers – the higher up in the picture the more smokers in the cluster.

We browsed the cluster contents as described in Section 3.1. In this particular example the cluster second from the bottom caught our eye: it has a low percentage of smokers, it is small and seemed to be coherent. In Figure 2 we have zoomed in on this cluster (and its relative cluster). After further browsing at this level we became convinced that a substantial part of the answers described occupations related to farming. Hence, we formulated a potential hypothesis, a relation between the open and closed answer: farmers smoke less than the average.

We repeated the steps of our method several times and observed the same relation in many of the iterations. Table 3 gives a textual presentation of another clustering, where *Cluster D* further supports this discovery.

After only a few hours[5] of exploration, concentrating on the most interesting clusters, we have formulated the following four hypotheses. They correspond well to the four clusters presented in textual form in Table 3.

A People working in leadership positions smoke more than the average.

B People working in the transportation industry smoke more than the average.

C Care workers smoke less than the average.

D Farmers smoke less than the average.

In the next section we try to assess hypothesis D, which was most consistent. The others may in part be explained by the gender distribution, see Tables 2 and 3, and should be studied further.

Studying the text clusters in Table 3, compared to gender regarded as a categorization, four other hypotheses could be formulated. We leave it to the reader to assess the quality of these.

6. Evaluation

With no prior knowledge of smoking habits in different occupation groups we have generated a hypothesis indicating a tendency that farmers smoke less than the average. In order to support or discard it thorough investigations and/or surveys should be performed. Lacking such possibilities, we have tried to find existing comparable surveys on smoking habits (after formulating the hypothesis).

Surveys differ in what they cover, both population sample and questionnaire formulation. The definition of a *smoker* may vary between surveys. Also, there exist many categorization systems for occupations, many of them differing in specificity and structure.

The questionnaire we have derived our hypothesis from is described in Section 4.. We have found the following comparable surveys:

- a Swedish survey by Statistics Sweden (SCB, 2006)
- two U.S.A. surveys (Lee et al., 2004; Lee et al., 2007)
- a European survey (McCurdy et al., 2003)
- an Australian survey (Smith and Leggat, 2007)

The most comparable survey is the one made by Statistics Sweden[6] (SCB), as it is conducted on the Swedish population. SCB is the central government authority for official statistics in Sweden. They provide general population statistics.

[5] Naturally, the amount of time can differ significantly depending on the questionnaire and the purpose of the investigation. The experiment demonstrates that interesting results can be obtained within a reasonable time.

[6] http://www.scb.se

The survey performed by SCB covers the years 1980 – 2005 and the ages 16 – 84. It is given almost every year and the statistics are presented from different aspects: household type, age groups, socio-economic group, etc. Here, smokers are defined as respondents who smoke daily. We focus on the years 1998 – 2003 (the time for the twin questionnaire) and the statistics for farmers as a socio-economic group.

The percentage of smokers overall in the SCB survey is smaller than in the questionnaire, as well as among farmers, see Table 4. However, the tendency that farmers smoke less than the average can also be seen here. Thus, the SCB survey supports our hypothesis.

	All workers	Farmers
SCB 1998 – 99	23.9%	8.7%
SCB 2000 – 01	24.6%	7.2%
SCB 2002 – 03	23.4%	8.9%
Questionnaire	29%	-
Cluster D	-	22%

Table 4: SCB: daily smokers in socio-economic groups in Sweden 1998 – 2003, ages 16 – 84. Questionnaire: smokers (according to definition in Section 4.3.) among twin respondents 1998 – 2002, born in or before 1958. Cluster D: one cluster from a clustering of the open answers in the questionnaire, see Table 3.

All surveys have different occupation categorization systems. The U.S.A. surveys, for instance, utilize a fine-grained categorization of farmers, and the portion of smokers differs between the subgroups. Also, the surveys cover different age groups. The European survey is focused on a younger population sample. Further, different smoker definitions are used. The Australian survey distinguishes *current*, *ex-*, and *never*-smoker groups. However, the tendency that farmers smoke less than the average is apparent in all surveys.

Considering all differences between the surveys and the twin questionnaire we can confirm our hypothesis, that farmers smoke less than the average. Thus our method is proven successful.

7. Conclusions and Future Work

We have presented a method for hypothesis generation from questionnaires through text clustering evaluation. Using the method we have generated the hypothesis that farmers smoke less than the average, which we have confirmed through literature studies. Normally, a new investigation would need to be performed.

We plan to apply the method on other questionnaires in different domains. Also, it could be applied on other types of data sets containing both textual data and data restricted to predefined values. One interesting example is electronic medical records.

Our method makes it feasible to explore and analyze open answers in large questionnaires, potentially containing hidden information. It provides a means for interactively revealing interesting parts of that information, reducing the manual work load significantly.

Acknowledgements

Our thanks go to VR (Vetenskapsrådet - the Swedish Research Council) for funding and to all participants in the project “Infomat” from KTH CSC (The School of Computer Science and Communication at The Royal Institute of Technology) and the department of Medical Epidemiology and Biostatistics (MEB) at Karolinska Institutet, both in Stockholm, Sweden.

We also thank VINNOVA (Swedish Governmental Agency for Innovation Systems) for funding and all members of the project KEA (Knowledge Extraction Agent) from DSV, KTH-Stockholm University and KTH CSC.

Furthermore, we would like to thank GSLT (The Swedish Graduate School of Language Technology).

The Swedish Twin Registry is supported by grants from VR, the Department of Higher Education, and NIH (grant AG-08724).

8. References

D. R. Cutting, J. O. Pedersen, D. Karger, and J. W. Tukey. 1992. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proc. 15th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*.

I. S. Dhillon. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD ’01: Proc. 7th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pages 269–274, New York, NY, USA. ACM.

A. Ekman and J. E. Litton. 2007. New times, new needs; e-epidemiology. *European Journal of Epidemiology*, 22:285–292(8).

A. Ekman, P. Dickman, Å. Klint, E. Weiderpass, and J. E. Litton. 2006. Feasibility of using web-based questionnaires in large population-based epidemiological studies. *European Journal of Epidemiology*, 21:103–111(9).

M. A. Hearst. 1999. Untangling text data mining. In *Proc. 37th Annual Meeting of the Association for Computational Linguistics*, pages 3–10, Morristown, NJ, USA. Association for Computational Linguistics.

V. Kann, R. Domeij, J. Hollman, and M. Tillenius, 2001. *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Ludek Hrebicek*, volume 60, chapter Implementation aspects and applications of a spelling correction algorithm.

D. Lee, W. LeBlanc, L. Fleming, O. Gómez-Marín, and T. Pitman. 2004. Trends in US smoking rates in occupational groups: the national health interview survey 1987-1994. *J. Occup. Environ Med.*, 46(6):538–48.

D. Lee, L. Fleming, K. Arheart, W. Leblanc, A. Caban, K. Chung-Bridges, S. Christ, K. McCollister, and T. Pitman. 2007. Smoking rate trends in U.S. occupational groups: The 1987 to 2004 national health interview survey. *J. Occup. Environ Med.*, 49(1):75–81.

H. Li and K. Yamanishi. 2001. Mining from open answers in questionnaire data. In *KDD ’01: Proc. 7th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pages 443–449, New York, NY, USA. ACM.

P. Lichtenstein, U. De faire, B. Floderus, M. Svartengren, P. Svedberg, and N. L. Pedersen. 2002. The swedish twin registry: a unique resource for clinical, epidemiological and genetic studies. *Journal of Internal Medicine*, 252:184–205.

P. Lichtenstein, P. F. Sullivan, S. Cnattingius, M. Gatz, S. Johansson, E. Carlstrom, C. Bjork, M. Svartengren, A. Wolk, L. Klareskog, U. de Faire M. Schalling, J. Palmgren, and N. L. Pedersen. 2006. The swedish twin registry in the third millennium: An update. *Twin Research and Human Genetics*, 9(6):875–882.

S. A. McCurdy, J. Sunyer, J. Zock, J. M. Antó, and M. Kogevinas. 2003. Smoking and occupation from the European community respiratory health survey. *J. Occup. Environ Med.*, 60(9):643–8.

M. Rosell. 2003. Improving clustering of Swedish newspaper articles using stemming and compound splitting. In *Proc. 14th Nordic Conf. on Comp. Ling. – NODALIDA '03*.

M. Rosell. 2007. Infomat – a vector space visualization tool. In M. Sahlgren and O. Knutsson, editors, *Proc. of the Workshop Semantic Content Acquisition and Representation (SCAR) 2007*. Swedish Institute of Computer Science (SICS), Stockholm, Sweden. SICS Technical Report T2007-06, ISSN 1100-3154.

Statistics Sweden SCB. 2006. Undersökningarna av levnadsförhållanden (Living condition survey). http://www.scb.se/LE0101.

D. Smith and P. Leggat. 2007. Tobacco smoking by occupation in Australia: Results from the 2004 to 2005 national health survey. *J. Occup. Environ Med.*, 49(4):437–445.

D. R. Swanson and N. R. Smalheiser. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artif. Intell.*, 91(2):183–203.

Paper III

The Stockholm EPR Corpus – Characteristics and Some Initial Findings

Hercules Dalianis, Martin Hassel and Sumithra Velupillai. 2009. *The Stockholm EPR Corpus – Characteristics and Some Initial Findings* To be presented at the 14th International Symposium for Health Information Management Research (ISHIMR '09), Kalmar, Sweden, October 14-16, 2009.

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.

Paper IV

Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial

Sumithra Velupillai, Hercules Dalianis, Martin Hassel and Gunnar H. Nilsson. 2009. *Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial.* In International Journal of Medical Informatics (2009). In Press. doi:10.1016/j.ijmedinf.2009.04.005

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and *F*-measure in a manual and computerized annotation trial

Sumithra Velupillai[a,*], *Hercules Dalianis*[a], *Martin Hassel*[a], *Gunnar H. Nilsson*[b]

[a] *Department of Computer and Systems Sciences, Stockholm University/KTH, Forum 100, 164 40 Kista, Sweden*
[b] *Department of Neurobiology, Care Sciences and Society, Center for Family Medicine, Stockholm, Sweden*

ARTICLE INFO

Article history:
Received 31 October 2008
Received in revised form
2 March 2009
Accepted 9 April 2009

Keywords:
Medical informatics applications
Natural language processing
Medical record systems
Electronic patient records in Swedish
Protected health information
Ethical issues
Annotation

ABSTRACT

Background: Electronic patient records (EPRs) contain a large amount of information written in free text. This information is considered very valuable for research but is also very sensitive since the free text parts may contain information that could reveal the identity of a patient. Therefore, methods for de-identifying EPRs are needed. The work presented here aims to perform a manual and automatic Protected Health Information (PHI)-annotation trial for EPRs written in Swedish.

Methods: This study consists of two main parts: the initial creation of a manually PHI-annotated gold standard, and the porting and evaluation of an existing de-identification software written for American English to Swedish in a preliminary automatic de-identification trial. Results are measured with precision, recall and *F*-measure.

Results: This study reports fairly high Inter-Annotator Agreement (IAA) results on the manually created gold standard, especially for specific tags such as names. The average IAA over all tags was 0.65 *F*-measure (0.84 *F*-measure highest pairwise agreement). For name tags the average IAA was 0.80 *F*-measure (0.91 *F*-measure highest pairwise agreement). Porting a de-identification software written for American English to Swedish directly was unfortunately non-trivial, yielding poor results.

Conclusion: Developing gold standard sets as well as automatic systems for de-identification tasks in Swedish is feasible. However, discussions and definitions on identifiable information is needed, as well as further developments both on the tag sets and the annotation guidelines, in order to get a reliable gold standard. A completely new de-identification software needs to be developed.

1. Introduction and background

Within hospital care there has been an explosion in the production of electronic patient records (EPRs) in digital form. A large amount of these records contain unstructured free text that is almost never reused. This information is considered very valuable for research but is also very sensitive since the records contain information that may reveal the identity of the patient. In this paper we evaluate a manual and comput-

[*] *Corresponding author.* Tel.: +46 8 16 11 74.
E-mail address: sumithra@dsv.su.se (S. Velupillai).

1386-5056/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2009.04.005

Please cite this article in press as: S. Velupillai, et al., Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial, Int. J. Med. Inform. (2009), doi:10.1016/j.ijmedinf.2009.04.005

erized annotation of identifiable information on a subset of a comprehensive hospital EPR data set, in order to compile a de-identified gold standard.

The paper is organized as follows: the rest of this section contains background information regarding previous work with protected health information in EPRs as well as work on annotated data sets. The following section, Section 2, describes the method choices made for the manual annotation trial and the automated annotation trial, as well as the evaluation metrics used. The results on the manual and automated annotation trials are discussed and described in Section 3, including some thoughts on strengths and limitations as well as implications for further research.

1.1. *Ethical and legal issues in the reuse of information in EPRs*

The health care system must take special considerations regarding ethical issues, where the Hippocratic oath is an important principle. Research on information contained in EPRs must be considered protected from an ethical point-of-view. The authorities in the US that approves research that is sensitive from an ethical point-of-view, such as for example using EPRs in research, are the local Institutional Review Boards (IRBs). In Sweden we have the corresponding regional Ethics Committees. Permission to perform research on EPRs can be approved under the condition that the clinical files are de-identified with regard to patient name and social security number. It is also required to de-identify the EPRs further, removing other types of information that may identify a patient, such as names of relatives and addresses. This can be done by using automatic methods. Once approval is given from an Ethics Committee, actual data release is authorized by hospital management.

1.2. *Identifying protected health information in EPRs*

Techniques for identifying protected health information (PHI) in EPRs have been studied mainly on EPRs written in English. Named entity recognition (NER) is a technique where categories such as names of persons, places and organizations and points in time are automatically extracted from texts. Such methods are often used in automatic de-identification systems. Sweeney [1] describes the Scrub system applied on a small subset of pediatric EPR files (275 records). The Scrub system reaches almost 99 percent precision. The De-id de-identification software described in [2], obtained a recall of 97 percent and a precision of 75 percent on EPRs written in English. This system is rule-based and relies heavily on external resources.

In Sibanda and Uzuner [3] methods for identifying PHI using local context without need for extensive amounts of external resources and hand-crafted rules show promising results. At the i2b2 Center, a fully de-identified EPR text set [4] has been developed which has been used in shared tasks. Unfortunately, there is no detailed description of the annotation process of the 889 discharge summaries that were de-identified. An evaluation of different de-identification software systems applied on the i2b2 material is described in Uzuner et al. [5] where the best system (Szarvas et al. [6]), gained an *F*-measure of 97 percent, a precision of 99 percent and a recall of 96 percent. In Uzuner et al. [7], several de-identified corpora were used to evaluate a set of de-identification tools. In particular, they develop a new de-identifier, Stat De-id, which uses local context and is based on Support Vector Machines (SVMs). The system shows promising results, especially for handling fragmented and noisy texts such as EPRs.

So far, there are not many studies on de-identifying EPRs in Swedish. Kokkinakis and Thurin [8] have worked with de-identifying 200 hospital discharge letters in Swedish achieving 97 percent precision and 89 percent recall.

As the systems described above are mainly based on different training and test data sets, it is difficult to compare the performance values. Moreover, details on the characteristics of the corpora (pre-processing choices and possible problems) as well as algorithm choices may affect the results and make comparisons even more problematic. They do, however, serve as good examples of research carried out in this area, and as pointers as to what kind of procedures, approaches and results may be considered state-of-the-art.

1.3. *Annotated data sets and Inter-Annotator Agreement*

Manually annotated data sets are often used for developing and evaluating automatic systems, as well as for supporting empirical claims, especially for different natural language processing tasks. Developing well-defined guidelines and tag sets for such annotations is important and crucial for the performance of the automatic systems. Moreover, such data sets need to be both representative and reliable. A data set is considered reliable if it can be shown that the annotations have high agreement on the tags assigned for the annotation task between the annotators [9].

By measuring the Inter-Annotator Agreement (IAA) in such data sets it is possible to identify possible weaknesses and strengths in the annotation task. Inconsistencies between the annotators indicate either that some annotations are wrong, or that the annotation scheme is inappropriate for the data set [9]. There exist many IAA measures that are more or less appropriate for different annotation tasks.

In Wilbur et al. [10], the construction of an annotated biomedical text set is described with respect to both annotation guidelines, annotation work and IAA, with results reported mainly with *F*-measure. In Uzuner et al. [7] the manual annotation of a corpus containing 90 authentic discharge letters is described. The annotation was carried out using three annotators and the IAA between them measured by Kappa was 100 percent. It is, however, not described if this value was reached after several annotation iterations or directly.

1.4. *The Stockholm-EPR corpus*

In prior work our research group has gained access to several hundred thousand EPRs from the Karolinska University Hospital and Stockholm City Council. This access has been granted by the hospital management at the Karolinska University Hospital after approval from the Stockholm Ethics Committee

Please cite this article in press as: S. Velupillai, et al., Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial, Int. J. Med. Inform. (2009), doi:10.1016/j.ijmedinf.2009.04.005

(Etikprövningsnämnden i Stockholm). These records contain both structured and unstructured entries, such as measurement values and sections of free text. The records were delivered to us "de-identified" in the sense that they did not contain any patient's personal name or social security number in the structured fields of the EPRs. However, the unstructured free text parts may still contain PHI instances. Therefore, to make the EPR data set accessible to a broader group of researchers both in medicine, linguistics and computer science, these PHIs need to be removed. In our Work in Progress Proposal [11] we have proposed certain steps to annotate a subset of the Stockholm-EPR corpus for full de-identification. In this study we will present the initial development and evaluation of our de-identification approach for this EPR data set, which we call the Stockholm-EPR-Gold-Standard.

One general aim of this project is to make it possible for researchers to use the abundant digital textual information that is available in EPRs, without risking the exposure of any patient's PHI. Specific aims are: (1) to develop and evaluate a manually de-identified gold standard of EPRs written in Swedish, and (2) to port and evaluate an automatic de-identifying software developed for American English to Swedish, in a PHI-annotation trial.

2. Methods

The EPRs we are studying originate from over 2000 clinics in the Stockholm area. The work consists of two main parts: (1) the manual creation of a gold standard with all PHI instances tagged and classified and (2) the porting, adaptation and evaluation of an existing automatic de-identification system for Swedish.

2.1. *The Stockholm-EPR-Gold-Standard*

A gold standard corpus from 100 EPRs in Swedish has been constructed. As the EPRs may vary in language usage, style and other aspects between clinics, the gold corpus has been compiled from five different clinics: Neurology, Orthopaedia, Infection, Dental Surgery and Nutrition. The records are distributed evenly genderwise (five patients per gender and clinic). The records containing the most free text per clinic and gender were included in the corpus. As the records were extracted from a medical record system database, they contained a number of columns with structured data as well as columns with free text. Although the main interest for the research carried out here lies in the free text, we included all columns in the gold standard set. This makes the calculations over types and tokens different, depending on which data is included. The manual annotations were made on the data set containing all columns, where the total number of tokens was around 380 000, the total number of types was around 31 000. Counting only the free text columns, the gold standard contains around 174 000 tokens (around 20 000 types). Naturally, these amounts may differ depending on how types and tokens are defined. EPRs contain a lot of numbers (medication prescriptions for instance) and other types of entities that may be defined in different ways when it comes to types and tokens. Here, numbers are included as tokens and types.

2.2. *Creating a gold standard*

As there are no general guidelines on which information is required to be deleted from EPRs in Sweden, we have followed the U.S. Health Insurance Portability and Accountability Act [12] and created a tag set covering the 18 PHI types given in this act. This includes the following 18 items: Names, Locations, Dates, Ages > 89 years, Telephone numbers, Fax numbers, Electronic mail addresses, Social security numbers, Medical record numbers, Health plan beneficiary numbers, Account numbers, Certificate/license numbers, Vehicle identifiers, Device identifiers and serial numbers, Web Universal Resource Locators (URLs), Internet Protocol (IP) address numbers, Biometric identifiers, and Any other unique identifying number or characteristic.

We have intentionally extended the set of PHI-tags to cover ethnicity and relations (such as sister and daughter), as we believe such instances may reveal crucial identifiable information. Names are divided into full, first and last names, and nested if applicable. They are further divided into tags covering patient, relative or clinician, as these may be useful for future research on identification and classification of semantic roles. All other names are tagged with a generic name tag. Hence a name of a nurse such as "John Smith" would be tagged the following way:

< Clinician_Full_Name >< Clinician_First_Name > John

< /Clinician_First_Name >< Clinician_Last_Name > Smith

< /Clinician_Last_Name >< /Clinician_Full_Name > .

Locations are divided into street addresses, towns, countries, municipalities, organizations and health care units. Dates are tagged either as full date (an instance containing year, month and date), date part (month and/or date) or year.

The tag set was developed in two iterations, by initially annotating a small subset of the Stockholm-EPR corpus with an early version of the tag set. The results from this annotation were used for improving and developing the second version of the tag set. The second tag set includes some more fine-grained tags, in order to distinguish some aspects of different entities. We deliberately chose not to specify the guidelines in great detail, as we wanted to discover what kind of discrepancies and coverage we would obtain by having less detailed definitions in some cases, and as we intended to make this annotation task a multi-procedure. This is, of course, problematic when it comes to comparability and reproducibility of the annotation task. The approach does however have the advantage that a deeper knowledge about the characteristics of the observed PHI instances can be further scrutinized.

The second version of the tag set was used for annotating the gold standard set of 100 EPRs. Three annotators (one senior medical researcher (SM), one senior computer science researcher (SC) and one junior computer science researcher (JC)) have annotated the set. The annotators worked separately, with no discussions during the annotation process, which we believe is useful in order to find which tags might be problematic and need to be further defined. We used the plugin Knowtator [13] within the Protégé 3.3.1 Ontology Editor and Knowledge Acquisition System [14] for the annotations.

2.3. *Porting the De-id software to Swedish*

The de-identification software package De-id [2] for EPRs written in American English is one of the few de-identification softwares which is both well documented, has shown good results for American English and which is freely available. It is rule-based and relies on lexical resources. For these reasons we decided to port it to Swedish by adapting it to the Swedish language through language-specific rules and resources. We have used a very straight on approach, creating lexical resources with very little manual interference. Specifically, we have adapted the system to take care of Swedish telephone numbers, social security numbers and date formats. The new De-id software is called Deid-Swe. Our Deid-Swe software also uses external resources such as Swedish pharmaceuticals lists, taken from FASS [15]. Added to these was a list of Swedish names of diseases, taken from Wikipedia. Furthermore, lists of addresses and person names have been gathered from the web. De-id [2] uses name lists that contain both clinician names and patient names taken from hospital databases that were connected to the EPRs. Unfortunately, we did not have access to any hospital database with names. Instead, we used a large number of names (male and female), first names and last names gathered from a web site containing Swedish names [16]. We used different sized variants of these lists covering 10 000 names in each list to over 100 000 names in each list.

The list of addresses contains all street addresses from major parts of Stockholm, taken from electronic municipality maps, and the lists of personal names was taken from a web site listing all names in the civil registry. The learning module of a Swedish Named Entity Recognizer [17] was used to extract 2000 new locations and 4000 new organizations from our set of EPRs (excluding the clinics in the gold standard corpus). The list of organizations contains 2000 clinic names and was included in the hospital lists for Deid-Swe, while the 2000 new locations were added to the address list and companies were concatenated with the company list. The address list contained 10 000 addresses and the company list contained 2000 companies. Finally, mimicking the original De-id package, two lists of high frequency tokens in the EPR set was generated from the complete set. The first of these two lists encompasses the 5000 most common tokens while the second covers the 50 000 most common tokens. These are used for the system not to annotate common tokens as PHI instances. The American test EPRs contain about 26 000 types and 336 000 tokens and the Swedish Gold corpus contains 20 000 types and 174 000 tokens (counting only the free text parts). There is less than a factor 2 in difference between the two domains. The style of the Swedish EPRs and the American English EPRs is very similar when it comes to structure, with notes that describe different sequences in the health care process.

2.4. *Evaluation metrics*

The results are mainly evaluated with Inter-Annotator Agreement (IAA): precision, recall and *F*-measure as main outcomes. Similarly to Wilbur et al. [10], we have not used the commonly used Kappa statistic for measuring IAA. This is mainly motivated by the fact that there are no random agreement models that would be suitable for this annotation task, especially given the large number of annotation classes. Moreover, Kappa measures are difficult to compare across data sets. These issues are discussed in more detail in for instance Wilbur et al. [10], with further references. Precision (also called positive predictive value, PPV) and recall (also called sensitivity) are measures commonly used in Information Retrieval and Extraction and provide a means to analyze the coverage of the annotated items by each annotator. *F*-measure is the harmonic mean of precision and recall. Reporting precision, recall and *F*-measure makes it possible to directly analyze how the annotations are actually distributed.

Moreover, as this annotation task has several annotation classes (tags), average precision, recall and *F*-measure can be calculated at a micro- or macro-level. Reporting micro-averaged results means that precision, recall and *F*-measure are calculated on a global total amount of annotations, while reporting macro-averaged results means that precision, recall and *F*-measure are calculated for each annotation class and averaged.

We have measured the IAA pairwise between the three annotators, getting as the final result the average pairwise measure. The results are also calculated over spans and classes. IAA for the manually created gold standard has been measured in Knowtator. Here, spans are defined as the exact length and position of an annotation, classes are defined as the annotation tags. This distinction may be very valuable, since high discrepancies might be due to indistinct definitions of the created tags. However, for de-identification, having high agreement results over spans is preferred over high results over classes.

Table 1 – The pairwise agreement in *spans* between the three annotators measured as recall, precision and *F*-measure for *all* PHI-tags.

Annotators	Micro			Macro		
	Precision	Recall	*F*-measure	Precision	Recall	*F*-measure
JC–SC	0.55	0.49	0.52	0.37	0.29	0.31
JC–SM	0.83	0.68	0.75	0.38	0.34	0.35
SM–SC	0.45	0.48	0.46	0.34	0.3	0.29
Average	0.61	0.55	0.58	0.36	0.31	0.32

Please cite this article in press as: S. Velupillai, et al., Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and *F*-measure in a manual and computerized annotation trial, Int. J. Med. Inform. (2009), doi:10.1016/j.ijmedinf.2009.04.005

Table 2 – The agreement in *classes* between the three annotators measured as recall, precision and *F*-measure for *all* PHI-tags.

Annotators	Micro			Macro		
	Precision	Recall	F-measure	Precision	Recall	F-measure
JC–SC	0.61	0.54	0.57	0.46	0.37	0.39
JC–SM	0.94	0.76	0.84	0.47	0.39	0.42
SM–SC	0.53	0.57	0.55	0.36	0.33	0.32
Average	0.69	0.62	0.65	0.43	0.36	0.38

Table 3 – The pairwise agreement in *spans* between the three annotators measured as recall, precision and *F*-measure for all *name* tags.

Annotators	Micro			Macro		
	Precision	Recall	F-measure	Precision	Recall	F-measure
JC–SC	0.89	0.68	0.77	0.65	0.58	0.6
JC–SM	0.95	0.87	0.91	0.57	0.54	0.55
SM–SC	0.80	0.66	0.72	0.55	0.47	0.50
Average	0.88	0.74	0.80	0.59	0.53	0.55

3. Results

3.1. *Gold standard corpus*

In total, the average number of annotations was 4794. The average IAA over spans for all PHI-tags for the three annotators was 0.58 *F*-measure (micro-averaged). The pairwise agreement ranged between 0.46 and 0.75 *F*-measure, micro-averaged (Table 1). For classes, the average IAA was 0.65 *F*-measure, with a pairwise agreement ranging between 0.55 and 0.84 *F*-measure, micro-averaged (Table 2). The macro-averaged results were consistently lower, which is probably due to the fact that some annotation tags, although being similar, were used differently, but consistently, by the annotators. The agreement was consistently higher between the two annotators SM and JC compared with the agreement with SC.

The IAA over spans and classes varied among the different subgroups of the PHI-tags. The average agreement over spans for name tags, for instance, was very high, 0.80 *F*-measure (micro-averaged), with a pairwise agreement ranging from 0.72 to 0.91 *F*-measure, micro-averaged (Table 3). The average number of annotations covering all names was 1646, amounting to 34 percent of the total number of annotations. Locations (including tags such as "Health_Care_Unit" and "Street_Address"), on the other hand, had much lower agreement results over spans, with an average agreement of 0.29 *F*-measure (micro-averaged), pairwise agreement ranging from 0.17 to 0.38 *F*-measure, micro-averaged (Table 4). These results were higher when looking at the results for class: average agreement was 0.48 *F*-measure (micro-averaged), pairwise agreement ranging between 0.35 and 0.68 *F*-measure, micro-averaged (Table 5). These results might reflect a need for more specific definitions on the usage of the location tags. In particular, the discrepancies were often due to differences in the coverage of a tag. An instance such as "Avdelning 22, Karolinska Universitetssjukhuset, Solna" could be tagged with one "Health_Care_Unit"-tag, or several, and it could also include the "Municipality" or "Town"-tag for "Solna". In total, the average number of annotations covering locations were 1370 (29 percent of the total number of annotations). Phone numbers amounted to an average of 2 percent of the total number of annotations.

3.2. *Evaluation of Deid-Swe*

The results for Deid-Swe have been measured against the manually created gold standard, by using each manually annotated set as the gold standard. In general, Deid-Swe heavily overgenerated PHI instances, which resulted in very low *F*-measures ranging between 0.04 and 0.16, where precision was 0.03–0.09 and recall was 0.56–0.76. We performed a man-

Table 4 – The pairwise agreement in *spans* between the three annotators measured as recall, precision and *F*-measure for all *location* tags.

Annotators	Micro			Macro		
	Precision	Recall	F-measure	Precision	Recall	F-measure
JC–SC	0.27	0.4	0.32	0.41	0.39	0.4
JC–SM	0.48	0.31	0.38	0.54	0.41	0.47
SM–SC	0.12	0.28	0.17	0.26	0.34	0.29
Average	0.29	0.33	0.29	0.40	0.38	0.39

Please cite this article in press as: S. Velupillai, et al., Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial, Int. J. Med. Inform. (2009), doi:10.1016/j.ijmedinf.2009.04.005

Table 5 – The pairwise agreement in *classes* between the three annotators measured as recall, precision and *F*-measure for all *location* tags.

Annotators	Micro			Macro		
	Precision	Recall	F-measure	Precision	Recall	F-measure
JC–SC	0.35	0.53	0.42	0.51	0.47	0.49
JC–SM	0.88	0.56	0.68	0.65	0.47	0.55
SM–SC	0.25	0.59	0.35	0.32	0.47	0.38
Average	0.49	0.56	0.48	0.49	0.47	0.47

ual evaluation on a small subset of the gold corpus and found that Deid-Swe performed quite well on dates, but overgenerated exceedingly on most other tags. We also tried to change the size of the lists of 5000 and 50 000 most common tokens to smaller lists but unfortunately this did not improve the results. We also tried to use name lists in different sizes, which resulted in slightly better results when using smaller lists. The smaller lists contained 10 000 first names and 20 000 last names.

4. Discussion

In this study we develop and evaluate a manual gold standard annotated for de-identifying EPRs written in Swedish, and a de-identification software for the automatic annotation of PHI instances. The main findings were that IAA was fairly high in general and very high in certain classes such as names, using both manual and computerized annotation. Unfortunately, the porting of an existing computerized de-identification system (De-id) to Swedish (Deid-Swe) did not yield good results, mostly due to the fact that Deid-Swe was difficult to adapt to Swedish EPRs.We have the impression that since De-id is rule-based, the required resources and heuristics need to be tuned and are difficult to generalize to a new domain and language. Our plan is to create a completely new de-identification software for Swedish, starting by using rules and lexical resources, after which it will be augmented using some sort of semi-supervised machine learning technique such as iterative machine learning [6] or active learning [18].

4.1. *Strengths and limitations*

The main strength is that our research to our knowledge has previously not been carried out on Swedish EPRs, and that it is unique due to the kind of textual data being used. PHI-annotation and IAA in this respect has received little attention. The gold corpus has been compiled from five different clinics which is another advantage, as it covers different language use, style and other aspects within clinics. Although Uzuner et al. [5,7] describe similar work, there are not many details regarding the creation of the manually annotated resources, especially regarding IAA results of the annotation classes and whether the resources were created in a one-stage or multi-stage procedure. Our approach on identifying PHI instances both manually and computerized in a non-English setting is not previously evaluated. We have also used annotators with different backgrounds, which is very valuable for further analysis of the results.

A main limitation is our restricted adaptation of Deid-Swe when it comes to for instance compound constructions, lemmatization, language-specific characters and misspellings in a Swedish context. Moreover, there was no documentation in De-id regarding how to balance the different lexical resources, which might have been one of the main drawbacks. We tried a large number of combinations of sizes of the lexical resources to see if that would improve our results. In particular, the approach to gather lexical resources automatically with very little manual intervention proved to be problematic. The resources clearly need to be manually scrutinized and cleaned. For instance, many names are also common tokens or ambiguous in other ways, which need to be handled separately.

Furthermore, the IAA trial on the manually created Gold standard was performed in one phase where a multi-stage procedure would have increased our figures. We believe that a consensus on the tag set can be achieved by further iterations in the annotation process. In particular, a thorough analysis of the current version of the gold standard makes it possible to identify which PHI instances might be problematic for future systems by analyzing the discrepancies. Such an approach does, however, have the disadvantage of making it more difficult to reproduce the same annotation task. Given thorough documentation on the iteration stages and steps taken, we believe a reliable and reproducible gold standard can be created.

4.2. *Gold standard and Deid-Swe*

Our results on the gold standard corpus and IAA (0.65 average *F*-measure (micro-averaged), 0.84 *F*-measure highest pairwise, Table 2) can be considered high taking account our one phase procedure and the limited training. Our results are lower, but in line with Mani et al. [19], however that study was based on annotation of protein names.

Our figures on IAA over spans and classes varied widely among the different subgroups (0.80 average *F*-measure for names (spans), Table 3, 0.29 average *F*-measure for locations (spans), Table 4), can be compared to 0.85 *F*-measure for acronym tags and 0.15 *F*-measure for array-protein tags in the study from Mani et al. [19]. This indicates that further refinements and definitions of the PHI-tag set are needed.

The results we obtained on this initial manual annotation trial may also reflect the complexity of the annotation task. We have, in contrast to the work presented in Uzuner et al. [7], fine-grained some of the PHI-tags, which has resulted in some discrepancies that might not have arisen given more general PHI-tags. Clearly, more detailed annotation guidelines are needed in order to achieve higher IAA results. This was, however, expected, as we did not have prior knowledge as to how such instances were actually represented in the EPRs. For this reason, we believe further annotation iterations are needed. We plan to analyze our results according to such performance measures as are described in Chinchor and Sundheim [20] and Hirschman et al. [21] in our future developments of the gold standard corpus.

Due to overgeneration we received very low IAA between the three annotators and Deid-Swe, because of very low precision. The results can be considered an underestimate due to the limitations mentioned above. These low quality results correspond well with the results presented in [22] for De-id ported to French and used on EPRs written in French.

4.3. *Feasibility*

Each EPR takes on average 30 mins to de-identify manually, and considered workable for the 100 EPRs. The annotation software was appropriate for the task, however limited by the format of the text files. The computer-based annotation was performed in minutes. The porting of the De-id software from American English to Swedish went smoothly when it comes to practical issues such as compiling and running the system. It did, however, require extensive work on creating appropriate resources.

4.4. *Implications for health care policy*

De-identification is crucial for the possibility of performing further research on a corpus containing sensitive and private information. However, guidelines and definitions on which PHI instances in EPRs that need to be removed in order for them to be considered secured from re-identification risk need to be developed and discussed. Another aspect is considering the appropriate level of de-identified EPRs prior to distributing them for research. Removing all instances of names, phone numbers and addresses could possibly be sufficient for giving access to research, with a more rigid security at the next level. We have shown that such instances can be identified with high accuracy. We believe that the 18 PHI instances listed in HIPAA are not appropriate for a Swedish standard on which PHI instances need to be removed for an EPR to be considered fully de-identified for research purposes. De-identifying instances covering dates and health care units for instance are not as crucial as other classes such as relations and ethnicity, which we believe contain a much higher risk for possible re-identification.

4.5. *Implications for research*

In a long-term perspective, we are planning to apply different text mining and information extraction techniques for exploiting the valuable information that this type of data sets contains. We believe that such methods may benefit many diverse research areas such as medicine and epidemiology. As a first step we are planning to do research in the area of speculative language. EPRs contain a potentially large amount of speculation, uncertainty and negation together with certainty and confirmation. This property is significant for the diagnosis and documentation procedure, and is very important to extract. For many text mining and information extraction tools, such issues are seldom taken into account, which we believe is problematic. A more detailed description of the research we propose to perform after the de-identification process can be found in Velupillai et al. [11].

5. Conclusions

Our evaluation of a manual gold standard for de-identifying EPRs revealed that the IAA in general and especially in certain classes (e.g. names) was fairly high. However, several classes have a low IAA, possibly due to both limitations in our approach as well as limits in what is possible to achieve. Using computerized annotation resulted in very low figures, but these are considered possible to increase with further system development using a different approach. Transporting a rule-based de-identification system developed for one language into another language directly is problematic and non-trivial, and such issues need to be considered when performing this type of research. Moreover, for evaluation, manual annotations are needed. Such work is very time-consuming, and depending on the difficulty of the annotation task and scheme, several iterations may be needed in order to achieve a reliable corpus. As there are no general thresholds of what results should be considered good for any given annotation task, such decisions must be made based on the task at hand. For de-identification, it is crucial to ensure the reliability of the gold corpus, as the integrity of the patient must be ensured. The gold corpus created for this work will be further analyzed and developed in order to ensure its reliability.

Summary points

What was known before the study:

- Free text parts in electronic patient records (EPRs) contain a potentially large amount of valuable information. In order to make them available for research, all instances of protected health information (PHI) need to be removed.
- Automatic methods for de-identification of EPRs have shown promising results for English, with high results in precision, recall and F-measure.
- Manually created gold standard data sets are needed for evaluation, and such sets need to be both representative and reliable.

What this study adds:

- We have created a preliminary gold standard in de-identified EPRs written in Swedish, with fairly high Inter-Annotator Agreement (IAA) results. The results show very high results for some PHI-tags, and lower for others, which indicates a need for further development and definitions of PHI instances. This is one of the first results reported on IAA for PHI.

Please cite this article in press as: S. Velupillai, et al., Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and *F*-measure in a manual and computerized annotation trial, Int. J. Med. Inform. (2009), doi:10.1016/j.ijmedinf.2009.04.005

- Porting the automatic De-Id system from American English to Swedish was problematic and non-trivial, probably due to difficulties in translating its rules to Swedish and balancing the lexical resources. It is probably more time efficient to construct a completely new software for de-identification of EPRs written in Swedish.

REFERENCES

[1] L. Sweeney, Replacing personally-identifying information in medical records, the Scrub system, in: Proc. AMIA Annu. Fall Symp., 1996, pp. 333–337.

[2] I.M. Neamatullah, M. Douglass, L.H. Lehman, A. Reisner, M. Villarroel, W.J. Long, P. Szolovits, G.B. Moody, R.G. Mark, G.D. Clifford, Automated de-identification of free text medical records, BMC Medical Informatics and Decision Making 8 (2008) 32, doi:10.1186/1472-6947-8-32.

[3] T. Sibanda, O. Uzuner, Role of local context in automatic de-identification of ungrammatical, fragmented text, in: Proc. HLT-NAACL 2006, New York, 2006.

[4] i2b2, Informatics for integrating biology and the bedside, 2008. Available at: http://www.i2b2.org (accessed October 31, 2008).

[5] Ö. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, Journal of the American Medical Informatics Association 14 (5 (September)) (2007) 550–563.

[6] G. Szarvas, R. Farkas, R. Busa-Fekete, State-of-the-art anonymization of medical records using an iterative machine learning framework, Journal of the American Medical Informatics Association 14 (2007) 574–580.

[7] Ö.T.C. Uzuner, Y. Sibandam, Y. Luo, P. Szolovits, A de-identifier for medical discharge summaries, Journal of Artificial Intelligence in Medicine 42 (1 (January)) (2008) 13–35.

[8] D. Kokkinakis, A. Thurin, Identification of entity references in hospital discharge letters, in: Proc. 16th Nordic Conference on Computational Linguistics NODALIDA-2007, University of Tartu, Tartu, 2007.

[9] R. Artstein, M. Poesio, Inter-coder agreement for computational linguistics, Journal of Computational Linguistics 34 (4 (December)) (2008) 555–596.

[10] W.J. Wilbur, A. Rzhetsky, H. Shatkay, New directions in biomedical text annotation: definitions, guidelines and corpus construction, BMC Bioinformatics 7 (2006) 356.

[11] S. Velupillai, H. Dalianis, M. Hassel, Diagnosing diagnoses in Swedish Clinical Records, in: H. Karsten, B. Back, T. Salakoski, S. Salanterä, H. Suominen (Eds.), Proc. First Conference on Text and Data Mining of Clinical Documents, Turku, Louhi'08, September 3–4, 2008, pp. 110–112.

[12] HIPAA, Health Insurance Portability and Accountability (HIPAA), Privacy Rule and Public Health Guidance, 2003. From CDC and the U.S. Department of Health and Human Services, April 11, 2003. Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm (accessed October 31, 2008).

[13] P. Ogren, Knowtator: a Protégé plug-in for annotated corpus construction, in: Proc. HLT-NAACL 2006, Morristown, NJ, USA, ACL, 2006, pp. 273–275.

[14] Protégé, 2008. Available at: http://protege.stanford.edu/ (accessed October 31, 2008).

[15] FASS, 2008. Available at: http://npl.mpa.se/mpa.npl.services/home2.aspx (accessed October 31, 2008).

[16] Svenska Namn. Available at: http://www.svenskanamn.se/ (Swedish names, in Swedish) (accessed February 27, 2009).

[17] H. Dalianis, E. Åström, SweNam—a Swedish Named Entity Recognizer, Its Construction, Training and Evaluation, Technical Report, TRITA-NA-P0113, IPLab-NADA, KTH, June 2001.

[18] F. Olsson, Bootstrapping named entity annotation by means of active machine learning, A method for creating corpora, Ph.D. Thesis, University of Gothenburg, 2008, ISBN 978-91-87850-37.

[19] I. Mani, Z. Hu, S. Bae Jang, K. Samuel, M. Krause, J. Phillips, C.H. Wu, Protein name tagging guidelines: lessons learned Comparative and Functional Genomics, 1–2, John Wiley & Sons, Ltd., 2005, pp. 72–76.

[20] N. Chinchor, B. Sundheim, MUC-5 evaluation metrics, in: MUC5'93: Proc. Fifth Conference on Message Understanding, Association for Computational Linguistics, Baltimore, MD, 1993, pp. 69–78.

[21] L. Hirschman, A. Yeh, C. Blaschke, A. Valencia, Overview of BioCreAtIvE: critical assessment of information extraction for biology, BMC Bioinformatics 6 (Suppl. 1) (2005) S1.

[22] C. Grouin, A. Rosier, O. Dameron, P. Zweigenbaum, Testing tactics to localize de-identification, in: MIE 2009: Proc. 22nd Conference of the European Federation for Medical Informatics, Sarajevo, Bosnia and Herzegovina, 2009.

Please cite this article in press as: S. Velupillai, et al., Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial, Int. J. Med. Inform. (2009), doi:10.1016/j.ijmedinf.2009.04.005