SHADES OF CERTAINTY

Annotation and Classification of Swedish Medical Records

Sumithra Velupillai

Doctoral Dissertation
Department of Computer and Systems Sciences
Stockholm University
April, 2012

Stockholm University
ISBN 978-91-7447-444-2
ISSN 1101-8526
DSV Report Series No. 12-002

Typeset by the author using LaTeX
Printed in Sweden by US-AB

ABSTRACT

Access to information is fundamental in health care. Today, with electronic documentation possibilities, techniques for automatic extraction of information from written documentation are used daily in many areas. However, in the clinical setting, written documentation is still unattainable for improving health care from many perspectives. For Swedish, research for improving automatic information access from medical records is still scarce. This thesis presents research on Swedish medical records with the overall goal of building intelligent information access tools that can aid health personnel, researchers and other professions in their daily work, and, ultimately, improve health care in general.

First, the issue of ethics and identifiable information in medical records is addressed by creating an annotated gold standard corpus for de-identification, and by porting an existing de-identification software to Swedish from English. The aim is to move towards making textual resources that do not risk exposure of an individual patient's information available to researchers. Results for the ported rule-based system are not encouraging, but the Inter-Annotator Agreement results for the created gold standard are fairly high.

Second, in order to be able to build accurate information extraction tools, distinguishing affirmed, uncertain and negated information is crucial. Certainty level annotation models are created and analyzed, with the aim of building automated systems. One model distinguishes *certain* and *uncertain* expressions on a sentence level, and is applied on medical documentation from several clinical departments. Differences between clinical practices are also studied. More fine-grained certainty level distinctions are presented in a second model, with two polarities along with three levels of certainty, and is applied on a diagnostic statement level from an emergency department. Overall agreement results for both models are promising, but differences are seen depending on clinical practice, the definition of the annotation task and the level of domain expertise among the annotators.

Third, using annotated resources for automatic classification of certainty levels is studied by employing machine learning techniques. Encouraging overall results using local context information are obtained. The fine-grained certainty level model is also used for building classifiers for coarser-grained, real-world e-health scenarios, showing that fine-grained annotations can be used for several e-health scenario tasks.

This thesis contributes two annotation models of certainty and one of identifiable information, applied on Swedish medical records. One of the certainty level models has been successfully applied for building automatic classifiers. Moreover, a deeper understanding of the language use linked to conveying certainty levels in Swedish medical records is gained. Three annotated corpora that can be used for further research have been created, and the implications for automated systems are presented.

SAMMANFATTNING

Tillgång till information är centralt inom hälsovården. Tekniker för automatisk extraktion av fakta ur skriftlig dokumentation används dagligen i många områden. Inom hälsovården är dock skriven dokumentation fortfarande oåtkomlig för att förbättra hälsovården utifrån flera perspektiv. Forskning på förbättrad automatisk informationsåtkomst för svenska är fortfarande knapp. Denna avhandling presenterar forskning på svenska kliniska texter med det övergripande målet att bygga intelligenta informationsåtkomstverktyg som kan assistera hälsovårdspersonal, forskare och andra yrkesutövare i deras dagliga arbete, och, i längden, förbättra hälsovården i stort.

Problemet etik och identifierbar information i elektroniska patientjournaler behandlas genom skapandet av en annoterad guldstandard för avidentifiering och översättandet av en existerande programvara för automatisk avidentifiering till svenska från engelska. Målet är att tillgängliggöra textresurser som inte riskerar exponering av en patients sekretessbelagda information för vidare forskning. Att översätta en existerande regelbaserad programvara från engelska till svenska ger inte önskvärda resultat, men den manuellt skapade guldstandarden resulterar i en tillförlitlig korpus.

För att kunna bygga intelligenta informationsextraktionstekniker behöver säker, osäker och negerad information skiljas åt. Annoteringsmodeller för osäkerhet skapas och analyseras, med målet att bygga automatiska system. En modell skiljer mellan säker och osäker information på meningsnivå, och appliceras på klinisk dokumentation från flera medicinska kliniker. Skillnader mellan olika typer av kliniker studeras också. En mer finfördelad modell presenteras i en andra modell, där osäkerhet modelleras i två polariteter tillsammans med tre nivåer av säkerhet, och annoteras på diagnosnivå från en medicinsk akutklinik. Övergripande resultat är lovande, men skillnader uppmärksammas beroende på kliniktyp, definitionen av annoteringsuppgiften och nivån av domänexpertis hos annoterarna.

Slutligen studeras användandet av annoterade textresurser för automatisk klassificering. Lovande resultat uppnås då lokal kontextinformation används. Den finfördelade osäkerhetsmodellen används också för att bygga klassificerare för e-hälsoscenarier som kräver grövre säkerhetsindelning, där det visar sig att en finfördelad annoteringsmodell framgångsrikt kan användas för flera e-hälsoscenarier.

Denna avhandling resulterar i två annoteringsmodeller för osäkerhet och en för identifierbar information, applicerat på svensk klinisk text. En av säkerhetsmodellerna används framgångsrikt för att bygga automatiska klassificerare. Dessutom uppnås en djupare kunskap om språket som används för att förmedla osäkerhet i svensk klinisk text. Tre annoterade textresurser som kan användas för vidare forskning skapas, och implikationer för utvecklandet av automatiska system presenteras.

List of Papers

This thesis is based on the following papers:

I 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 78:12 (2009) 19–26.

II Hercules Dalianis and Sumithra Velupillai. 2010. *How Certain are Clinical Assessments? Annotating Swedish Clinical Text for (Un)certainties, Speculations and Negations.* In Proceedings of LREC '10 – 7th International Conference on Language Resources and Evaluation, Valletta, Malta, May 19–21, 2010.

III Sumithra Velupillai. 2010. *Towards A Better Understanding of Uncertainties and Speculations in Swedish Clinical Text – Analysis of an Initial Annotation Trial.* In Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, Uppsala, Sweden, July 10, 2010. ACL.

IV Sumithra Velupillai, Hercules Dalianis and Maria Kvist. 2011. *Factuality Levels of Diagnoses in Swedish Clinical Text.* In Proceedings of the XXIII International Conference of the European Federation for Medical Informatics (MIE), Oslo, Norway, August 28–31 2011. IOS Press.

V Sumithra Velupillai, 2011. *Automatic Classification of Factuality Levels – A Case Study on Swedish Diagnoses and the Impact of Local Context.* In Proceedings of The Fourth International Symposium on Languages in Biology and Medicine (LBM 2011), Singapore, December 14–15, 2011.

VI Sumithra Velupillai and Maria Kvist. 2012. *Fine-grained Certainty Level Annotations Used for Coarser-grained E-health Scenarios – Certainty Classification of Diagnostic Statements in Swedish Clinical Text.* In A. Gelbukh (Ed.): CICLing 2012 Part II, LNCS 7182, pp. 450–461. Springer-Verlag Berlin Heidelberg 2012.

ACKNOWLEDGEMENTS

I have been lucky in many ways during this exciting journey. There are many people that have played a major role in the work presented here, and I am happy to have been able to be a part of such a wonderful network. First of all, I want to express my gratitude to my supervisor, Prof. Hercules Dalianis, and my co-supervisor, Dr. Martin Hassel, who have encouraged me throughout these years, and provided an exciting and intriguing work environment. Thank you for supporting me and helping me in moving forward. Hercules, thank you for always being happy and positive, and thank you both for your shared interest in the cinematic arts – I have a long list of movies to devour!

Second, I am indebted to my co-authors, Prof. Gunnar Nilsson and Dr. Maria Kvist. Without your extensive knowledge and domain expertise, this could not have happened! Thank you also, Mia, for being so curious and interested in so many things - I have really enjoyed our sessions of word puzzling and discussions, and thank you both for teaching me about how things work in the real clinical world.

When I embarked on this journey, we were a much smaller group working in this area. It is wonderful to follow, and be a part of, the development of the research group, now called *Health Care Analytics and Modeling*, and to have the opportunity of working, studying, worrying, laughing and exchanging ideas and thoughts with my fellow PhD candidates Maria Skeppstedt and Aron Henriksson.

Prof. Paul Johannesson and Thashmee Karunaratne, thank you for giving me invaluable comments on the first public draft version of this dissertation, and for the very nice discussions during the pre-doc seminar.

I also want to thank all my colleagues at the Unit of Information Systems, and everyone at the IT Systems group for helping me with technical issues whenever they have arisen. Thank you, Constantinos Giannoulis, for wanting to maintain and develop the PhD student council at DSV, and for all the fun we had doing it! Thanks also to all fellow PhD students at DSV.

A big thank you to Bo Wikström, Stockholms läns landsting (Stockholm City Council) and Stefan Engqvist, Karolinska University Hospital, for helping with obtaining access to the data used in this work. The KEA (Knowledge Extraction Agent) project was the starting point for everything, thank you VINNOVA for funding this project and my initial period as a PhD candidate.

Although Sweden is a small country, we have an impressive research environment for natural language processing. Being affiliated to the National Graduate School of Language Technology (GSLT) has been amazing, thank you for providing a wonderful platform for courses, researchers and students, and for travel funding. I also want to thank Prof. Viggo Kann and Dr. Magnus Rosell for supervising my master thesis work and for encouraging me to continue doing research in the area. Magnus, a special thank you also for our fruitful collaborations and co-authorships during your time at KTH. Thank you Oscar Täckström, Dr. Jussi Karlgren, Dr. Atelach Alemu Argaw, Dr. Anette Hulth, and all other colleagues who have helped me in so many ways.

In am honored to have been invited to NICTA, National Institute for Information and Communications Technology Australia, in 2010 and 2011, in the research discipline of machine learning, funded by the NICTA business area of eHealth. Thank you, Dr. Hanna Suominen, Dr. David Martinez, Dr. Leif Hanlen and Dr. Lawrence Cavedon for making my visits enjoyable and fruitful and for being such wonderful company! I hope our collaboration will lead to new, interesting findings. Hanna, thank you for for being such a wonderful colleague and friend.

I am also honored to have been able to initiate an international research collaboration with Dr. Wendy Chapman and her research group at the Division of Biomedical Informatics at the University of California, San Diego, through funding from the Stockholm University Academic Initiative, for the years 2011–2013. The project is called Interlock: Stockholm - San Diego - Inter-Language collaboration in clinical NLP. It was wonderful to have you here in Stockholm in June, 2011, Wendy, and I really enjoyed my one-month stay in San Diego in November 2011 – a special thank you to Danielle Mowery, our collaboration turned out wonderfully and, more importantly, thank you for being so welcoming and for our amazing excursions. I look forward to our future experiments, collaborations and adventures.

The HEXAnord network (HEalth TeXt Analysis network in the Nordic and Baltic countries), funded by Nordforsk (Nordic Council), has been a magnificent forum for meeting colleagues, learning about experiences in other Nordic countries, and

for having great fun. Thanks to all of you!

Unfortunately, the amount of carbon emissions produced during my time as a PhD student has reached unreasonable levels, but at the same time, I have been lucky to be able to attend conferences, network meetings, and give talks around the world, and through this, I have met many eminent researchers who have widened my horizon and given me new insights. This would not have been possible without the generous funding I have been awarded through the Helge Ax:son Johnson foundation and the K & A Wallenberg foundation.

All my friends and family: thank you for keeping me sane and for the constant love and support. This is dedicated to all of you.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Consider the following health care scenarios:

- A hospital administrator is working on identifying conditions or events that have happened to patients in a large hospital, that endanger their safety. Such cases, for instance hospital acquired infections, are called adverse events. All cases of adverse events need to be scrutinized, no misses can be made. To find these events, *all* relevant medical records in the hospital medical record system need to be analyzed. This means a huge amount of documentation. How is she to find these records?
- A patient is experiencing pain after operation. Pain medication is prescribed by the responsible clinician, allowing for extra dosage if necessary. Several nurses are taking care of the patient. Each gives extra pain medication and documents this, along with pain observations, in the medical record. Over time, it is obvious that the basic pain medication is insufficient and should be changed, as evidenced by the extra dosages and the pain observations. This is, however, missed by the physician, as the amount of documentation is immense, and hence the patient receives inadequate medication. How could such misses be avoided?
- A physician meets a new patient. Along with hearing the patient's description about her symptoms and problems, the clinician needs to get an

> overview of the previous medical history of the patient, which has been documented in the medical record. The amount of documentation is gigantic, and she needs to read through hundreds of pages of documentation. How could she be helped in this situation?

These scenarios reflect different aspects of the very complex reality of health care. There are different types of professions, different information needs, and, currently, different ways of supporting these scenarios. What they all have in common is, at least, the fact that the *content* of the medical record is an essential component for an automated system to ease, or support, the already heavy workload in the daily work.

Health care is complex. Clinicians, nurses and other health care professionals are faced with numerous problems and situations every day and need to make decisions based on different types of information at hand, such as the patient's description of her symptoms, the patient's previous medical history, and information from colleagues. This information comes in different forms: verbally, written, through images, etc. Subsequent decisions, reasoning and actions are documented, in order to ensure good quality of care.

Currently, there are many techniques in the natural language processing and information retrieval research communities that work well for supporting information needs in different ways. Search engines are a clear example of successful solutions for meeting certain kinds of information needs, that are used by many on a daily basis, both for professional and private use. There are also mature techniques for extracting more specific information from narratives, such as named entities, e.g., person names, times, and quantities (see Chapter 2 for further details).

However, most such techniques do not take an important issue into account: for certain information needs it is important to ensure the highest possible level of relevance to the resulting extracted information. Medical records contain a large amount of reasoning and decisions based on insecure information. It is not always clear what a patient suffers from, and this uncertainty is reflected in the medical record. Moreover, in order to ensure that *relevant* information is extracted, cases of negation and speculation should be distinguished from affirmed cases.

In the example scenarios above, for instance, the hospital administrator could get support from an automatic system that extracts *all* relevant medical records, meaning that only clearly negated cases are excluded. The clinician who misses to take action on the insufficient pain medication could get support from a system that automatically alerts her when a threshold is reached based on information written in the record. The clinician who needs to sift through extensive amounts of documentation to understand a patient's medical history could be aided by a system where an overview is given, listing all affirmed, suspected and excluded conditions separately.

1.1 RESEARCH PROBLEM, AIM AND GOAL

How can information extraction from medical records be improved for different information needs? In particular, how can such techniques be developed for Swedish medical records?

The *problem* is that, although there are information extraction techniques developed for handling expressions of uncertainty and negation, i.e. distinguish affirmations, negations and speculations, for some languages, predominantly English, none exist for Swedish. This leads to problematic information extraction results. Specifically, little is known about *how* uncertainties are expressed in Swedish medical records, knowledge that is needed for building automated tools. Moreover, it is difficult to perform research on medical records as they contain private information about patients, whose integrity needs to be ensured.

The *aims* are to 1) move towards making medical records (in Swedish) available for further research by creating a de-identified corpus of Swedish medical records, and, in particular, 2) to provide a description of how certainty levels, i.e. affirmed, speculated and negated information, are expressed in Swedish medical records, create models and corpora that capture this, and build classifiers that distinguish them, for different information needs.

The long-term *goal* is to build better information extraction systems that can aid clinicians, researchers and other professions in their daily work, and, in this way, improve health care in general.

1.1.1 RESEARCH QUESTIONS

Based on the research problem and aims, the following research questions are addressed:

Making Medical Records Available for Further Research (Paper I)

- How can a de-identified corpus of Swedish medical records be created?
- Can an existing de-identification tool built for English be ported to handle Swedish medical records?

Certainty Levels in Swedish Medical Records

- How is medical uncertainty expressed in medical records (in Swedish):
  - on a sentence level? (Papers II and III)
  - on a diagnostic statement level? (Paper IV)
- How can a corpus annotated for uncertainty on a diagnostic statement level be used for automatic classification of uncertainty levels? (Paper V)
- How can a corpus annotated for uncertainty on a diagnostic statement level be used for automatic classification of different information needs (i.e. real-world scenarios)? (Paper VI)

1.1.2 EXPECTED RESULTS

To answer the research questions, a *corpus*, i.e. a collection of representative documents, annotated for, in this case, either identifiable information or uncertainty at a sentence and a diagnostic statement level, is needed. In order to create annotated corpora, an annotation model, as well as guidelines with instructions for how the model is to be applied on the documents, is required. A number of annotators is needed, in order to evaluate and measure the reliability of the resulting corpus. This measure is ideally as high as possible, meaning that the annotators agree on the assigned annotations to the highest extent possible. A corpus with high annotator agreement can subsequently be used for a) corpus analysis, and b) building, training and evaluating an automatic classifier that is, in the ideal case, able to mimic human performance. Methods for achieving this along with success criteria are further discussed in Chapter 3.

For the research questions stated in the previous section, the expected results thus are:

- A corpus of Swedish medical records annotated for identifiable information
- A feasibility study of automatic de-identification of Swedish medical records
- A corpus annotated for sentence level uncertainty expressions
- A corpus annotated for diagnostic statement level uncertainty
- A feasibility study of automatic classification of diagnostic statement level uncertainty
- A feasibility study of automatic classification of diagnostic statement level uncertainty for different e-health scenarios

1.1.3 CONTRIBUTIONS

The main contributions along with the research process in this thesis are shown in Figure 1.1.

Three annotation models and guidelines are created iteratively, after which the corpus creation is performed by annotators who are assigned to annotate representative documents. Three annotated corpora are created; the *Stockholm EPR PHI Corpus*, the *Stockholm EPR Sentence Uncertainty Corpus* and the *Stockholm EPR Diagnosis Uncertainty Corpus*. These are, in turn, used for either corpus analysis or automatic classification, or both. These corpora can be used for further research[1].

[1] Provided that proper ethical approval is obtained.

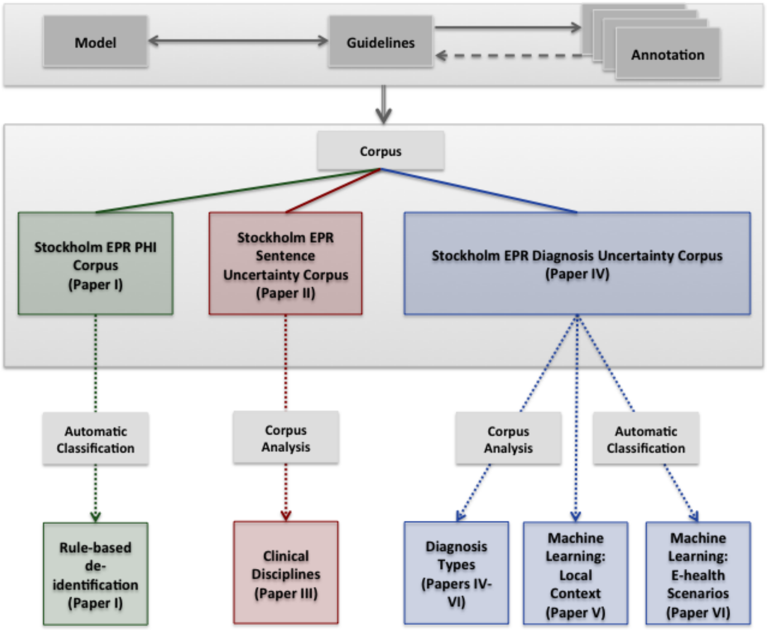

Model
Guidelines
Annotation
Corpus
Stockholm EPR PHI Corpus (Paper I)
Stockholm EPR Sentence Uncertainty Corpus (Paper II)
Stockholm EPR Diagnosis Uncertainty Corpus (Paper IV)
Automatic Classification
Corpus Analysis
Corpus Analysis
Automatic Classification
Rule-based de-identification (Paper I)
Clinical Disciplines (Paper III)
Diagnosis Types (Papers IV-VI)
Machine Learning: Local Context (Paper V)
Machine Learning: E-health Scenarios (Paper VI)

Figure 1.1: Overview: research process and contributions. By creating an annotation model along with guidelines three annotated corpora are created: the Stockholm EPR PHI Corpus, the Stockholm EPR Sentence Uncertainty Corpus, and the Stockholm EPR Diagnosis Uncertainty Corpus. These are used for corpus analysis and/or automatic classification.

A deeper understanding of the language use linked to conveying certainty levels in Swedish medical records is obtained through corpus analysis of the Stockholm EPR Sentence Uncertainty Corpus and the Stockholm EPR Diagnosis Uncertainty Corpus. Moreover, the feasibility studies of building automatic classifiers for certainty level classification using the Stockholm EPR Diagnosis Uncertainty Corpus show promising results.

1.2 RESEARCH FRAMEWORK, STRATEGY AND PROCESS

This research is interdisciplinary and draws on several research traditions: computer science, linguistics, information science and health care (all of which, in turn, are somewhat interdisciplinary in themselves). On a logical level, the underlying research theory is *inductive*. There is no predefined hypothesis to be tested and falsified, instead, we build theories based on empirical data. The type of research is both *exploratory* and *descriptive*. It is *exploratory* since we want to discover relevant features and characteristics, and *descriptive* as we want to understand and describe these problems.

The research methods are mainly quantitative. However, in order to gain a deeper understanding of the studied phenomena, qualitative methods are also used for parts of the steps taken. Hence, the much debated dichotomy between quantitative and qualitative research paradigms, and, in particular, the philosophical assumptions traditionally assigned to these, ought to be addressed. Commonly, quantitative research belongs to the *positivist* philosophical tradition, while qualitative research, on the other hand, traditionally belongs to an *'anti-positivist'* (most often an interpretive or social constructionist) tradition. In recent years, this dichotomy has been challenged (e.g. Mingers (2001) and Kaplan & Duchon (1988)), and *pluralist* approaches have been advocated. *Pluralism*, as defined and proposed in Mingers (2001), means that multi-method research is preferred, and that consideration should be given to different dimensions when designing a research project – real situations, social, material dimensions, as well as the research context.

Moreover, *pragmatism* has been suggested as a viable option, where scientific interpretations must make sense practically, and *actions* are a central unit of analysis (see, e.g. Goldkuhl (2004), Johnson & Onwuegbuzie (2004) and Hevner & Chatterjee (2010)). This is the primary influence of the conducted research, where a pluralist approach is taken. The overall goal is to build better information extraction systems that are useful in practical settings.

This stance is closely related to the *design science* framework[2]. In design science, 'designing new and innovative artifacts' (Hevner et al., 2004) is in focus.

[2] Whether design science is to be considered a research *method* or a general research paradigm is subject to some debate (see, e.g. Wayne (2010)). Here, the latter is advocated.

'It seeks to create innovations that define the ideas, practices, technical capabilities, and products through which the analysis, design, implementation, and use of information systems can be effectively and efficiently accomplished' (Hevner & Chatterjee, 2010).

In design science, three research cycles are defined: the *relevance* cycle, the *rigor* cycle and the *design* cycle (Hevner & Chatterjee, 2010). In the *design* cycle, the construction of the artifact(s), evaluation and subsequent feedback to refine the design is performed iteratively. The artifacts created in this work are three annotated corpora: the *Stockholm EPR PHI Corpus* (Paper I), the *Stockholm EPR Sentence Uncertainty Corpus* (Papers II and III), and the *Stockholm EPR Diagnosis Uncertainty Corpus* (Paper IV), see Figure 1.1. Moreover, automatic classifiers are built: for de-identification (Paper I) and for automatic classification of diagnostic statement level certainty, the latter in different variants (Papers V and VI). The annotated corpora (i.e. reference standards) are evaluated through inter-annotator agreement measures, and the classification results are evaluated against reference standards, see Chapter 3.

In the *rigor* cycle, past knowledge is provided and form the foundation to the research project to ensure that the produced results are research contributions and not 'routine designs' (Hevner & Chatterjee, 2010), and, through that, assert the research project's innovation. Here, appropriate theories and methods are to be used for constructing and evaluating the artifact(s). These are described in Chapters 2 and 3.

Finally, in the *relevance* cycle, an application context is to be defined, providing the requirements for the research as well as the acceptance criteria for evaluation of the research results. Here, the application context is positioned in the health care environment, and, more specifically, in addressing information extraction needs from Swedish medical records. However, an important part of the *relevance* cycle is also that the resulting artifacts are supposed to be returned into the application domain for utility and field testing. This latter part has not been performed in the presented work, and is left for future development.

One important aspect included in the design science framework is the dissemination of research results to different types of audiences. The research presented here has been published and presented both to a natural language processing community (Papers II, III, V and VI) and a medical informatics community (Papers I and IV).

Further details on the steps taken, motivations and limitations are presented in Chapter 3 (Method).

1.3 THESIS OUTLINE

This thesis is organized in six chapters. The Introduction chapter positions the research and states the problem, aim and goal, together with the overall research framework. The following chapters form a summary and foundation of the conducted research that is based on the six published articles included in Chapter 6.

The second chapter, Background, provides an overview over the relevant concepts that this research is based on. It also gives an account of related, relevant research on information extraction from medical records, modeling language and building classifiers, and, more specifically, approaches taken for the two main tasks addressed in this thesis: de-identification and certainty level identification of medical records.

Chapter 3, Method, details the method choices taken for the necessary steps in the conducted research along with limitations and a discussion on ethical issues, and Chapter 4, Results, gives an account of the obtained results for the different steps. In the concluding chapter, Chapter 5, contributions, conclusions, lessons learned and possible ways forward are elaborated.

CHAPTER 2

BACKGROUND

This chapter describes the larger research setting in which this research is positioned. First, the nature of medical records and their context is described, along with approaches taken for extracting information from such documents, as well as approaches taken for extracting information from other types of documents. Second, research on building corpora for language modeling is discussed. Third, general approaches for automatic classification of textual content is briefly accounted for. Finally, and more specifically, approaches for de-identification and uncertainty modeling is presented.

2.1 MEDICAL RECORDS AND INFORMATION EXTRACTION

The history of documenting encounters in hospital settings is long, in Sweden the first systematic documentation started in 1752 (Nilsson, 2007). The internal content of the medical record can be structured in different ways, e.g. 'source-oriented', 'problem-oriented' and 'time-oriented' (Tange, 1996). In the 'source-oriented' medical record, data is grouped in a hierarchy of categories originating from the source of the medical data, e.g. laboratory results, which, in turn are organized into sub-categories. The 'time-oriented' medical record is two-dimensional, enabling a presentation of both the type of data *and* time - with an emphasis on the importance of using time as the universal organizing principle. Finally, in the 'problem-oriented' medical record, the first organizing principle is the partitioning per problem, i.e. grouping observations for each problem the patient suffers from. The second principle is to organize each section according to the physician's way of thinking (Tange, 1996). Electronic health record systems were developed with the advances of technology, motivated by the need for efficiency and rationality in medical care systems, and introduced in, e.g. the U.S and Sweden in the early 1990s (Petersson & Rydmark, 1996).

Building electronic health record systems requires careful consideration of many different parameters: the users, the hospital administration, laws and regulations, consistency, interoperability and follow-up capabilities. Whether or not documented information is to be structured (i.e. belong to predefined vocabularies, terminologies and/or numerical values) or unstructured (i.e. written in free-text) is subject to much debate. The advantages of moving towards structured entries are, among others, the possibilities of ensuring consistent and measurable documentation, and the availability of statistical software that can automatically analyze this type of data. However, it has been showed that adding and enforcing structured information leads to an increased workload and errors in the health care process (Suominen, 2009). Moreover, an important aspect is lost with such a solution: the possibilities of nuanced and detailed information exchange (Lovis et al., 2000) and support for individualized care (Tange (1996) and Tange et al. (1997)).

Although there are numerous electronic health record systems on the market, both internationally and nationally in Sweden[1], none of them are designed without capabilities of documenting in free-text. It is estimated that free-text constitutes around 40% of the documented information (Dalianis et al., 2009), which means that there is a large amount of free-text information, along with structured data, that could be used for information extraction.

Techniques for information extraction from text are constantly refined and developed in the natural language processing research community. Information extraction techniques extract specific, predefined types of information from text,

[1] There are at least three different systems used throughout Sweden: TakeCare (http://www.cgmtakecare.com/, Accessed January 19, 2012), Mellior (http://www.nwe.siemens.com/sweden/internet/se/Healthcare/IT-losningar/Melior/Pages/Melior.aspx (in Swedish), Accessed January 19, 2012) and Cambio (http://www.cambio.se/ (in Swedish), Accessed January 19, 2012).

whereas, e.g. information retrieval techniques extract relevant documents, and text mining techniques extract new, previously unknown information, see for instance Jurafsky & Martin (2009), Baeza-Yates & Ribeiro-Neto (2011), and Feldman & Sanger (2007) for further details on definitions, techniques and applications. In the clinical domain, it is recognized that general language solutions are not sufficient to ensure good performance. Medical records are noisy, they contain a large amount of medical jargon, domain-specific and ad hoc abbreviations, misspellings and ill-formed syntax (Campbell & Johnson (2001), Meystre et al. (2008), Savova et al. (2010), Dalianis et al. (2009)).

In general, there are two main approaches for building automatic information extraction systems: those that rely on rules in some more or less complex form (from simple pattern matching to symbolic modeling) and those that rely on statistical methods and machine learning (Meystre et al., 2008). Rule-based systems differ from machine learning methods as they are not dependent on training data for the model creation. Machine learning is a vivid research area in itself and is applied both on textual and structured data. Two broad approaches taken in machine learning techniques are *supervised* learning, where the task is to learn a mapping from a known input to a desired output by following an automated learning algorithm, and *unsupervised* learning, where there is only input data and the aim is to find regularities in the data (see, e.g. Alpaydin (2010) for an overview of machine learning approaches, and Jurafsky & Martin (2009) for applications and approaches in natural language processing).

For English, there are several systems developed for information extraction from medical records. The MedLEE system is a rule-based system built for automated decision support and to facilitate information access at the Columbia-Presbyterian Medical Center (CPMC) (Friedman et al. (1994), Friedman et al. (1995), Friedman (1997), Friedman et al. (2004) and Mendonca et al. (2005)). The Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) is a system built at the Mayo clinic in Minnesota, USA, that has been applied to practical tasks such as ascertaining cardivascular risk factors and treatment classification (Savova et al., 2010). This system is released open-source as a part of the Open Health Natural Language Processing Consortium (OHNLP, 2012).

Research on medical records written in languages other than English is more scarce. For Finnish, research on clinical information access has been performed, including tasks such as topic segmentation from nursing narratives (Suominen, 2009). For French, techniques for improving spelling corrections (Ruch et al., 2003), creating medical lexicons (Cartoni & Zweigenbaum, 2010), and detecting risk patterns related to hospital acquired infections (Proux et al., 2009) have been proposed, for example. Within the EVTIMA project[2], research on Bulgarian patient records is performed, including tasks such as correlations in patient status data (Boytcheva et al., 2009) and structuring status descriptions (Boytcheva et al., 2010).

2.2 MODELING LANGUAGE: THE CASE OF CORPORA AND ANNOTATIONS

Research in natural language processing often requires empirical data. Collections of documents (*corpora*) marked with linguistic attributes serve as reference (or *gold*) standards. Linguistic attributes that are modeled include part-of-speech, syntax, semantic roles, etc. The marking (or annotation) can be performed either manually, semi-automatically or fully automatically. Reference standards are used to evaluate automatic systems, for training machine learning applications, and also for performing quantitative, descriptive language studies (Craggs & Wood, 2005). "'Linguistic annotation' covers any descriptive or analytic notations applied to raw language data" (quote from Bird & Liberman (2001)). Manual annotations are often needed for complicated knowledge representations.

The Penn Treebank (Marcus et al., 1994) is a large corpus containing part-of-speech tags and syntactic information. PropBank is built on top of the Penn TreeBank and contains annotations of semantic predicate-argument structures (Palmer et al., 2005). Syntactic and semantic dependencies have been annotated for many languages. For instance, the 2009 CoNLL shared task on joint parsing of syntactic and semantic dependencies included corpora in Spanish, Japanese, German, English, Czech, Chinese and Catalan (Hajič et al., 2009). The Stockholm-Umeå Corpus (Ejerhed et al., 2006) is a large collection of Swedish documents annotated for morphosyntactic information, named entities, etc. The Turku Clinical TreeBank and PropBank is a corpus of annotated Finnish intensive care nursing narratives (Haverinen et al., 2009). These examples form only a fraction of all

[2] http://lml.bas.bg/evtima/, Accessed January 20, 2012

annotated resources available for research, for different purposes. However, until now, annotated resources of Swedish medical records are very rare.

In the clinical domain, using medical records, several annotation efforts have been presented. For instance, annotated gold standards have been created for evaluation of systems in research challenges. At the i2b2 center (i2b2, 2012) several shared tasks have been conducted, such as identification of obesity (Uzuner, 2009) and medication extraction (Uzuner et al., 2010). Another example is presented in Chapman & Dowling (2006), where clinical conditions have been annotated in emergency department reports.

In the following sections, approaches for the tasks of de-identification and of uncertainty identification are discussed.

2.3 DE-IDENTIFICATION

Although medical records are supposed to maintain the highest level of confidentiality for patients, identifiable information about patients can still be found in the free-text parts of the records. From an ethical perspective, these issues are further discussed in Section 3.6. Here, approaches to building resources and classifiers for automatic de-identification from the free-text parts of medical records are presented.

For building systems that automatically extract identifiable information from free-text, the first step is to define which instances constitute identifiable information. Personal names of patients are of course typical examples of such information. However, there are other examples: phone numbers, addresses, identification numbers, etc. In the U.S., types of so called Protected Health Information are defined in the Health Insurance and Portability and Accountability Act[3] as instances to be removed or replaced from medical records in order for them to be considered fully de-identified and safe from an patient integrity point-of-view. These instances are further described in Section 3.2.

The second step is to create or use a reference standard (or *gold standard*) to which automatic systems can be evaluated. A gold standard is a collection of documents

[3] http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm, Accessed January 22, 2012

annotated with instances of the desired output information. Finally, an automatic system is to be built, and, as described above, two major approaches have been employed: rule-based or machine learning based systems, where rule-based systems do not rely on training data while machine learning based systems do.

Neamatullah et al. (2008) have built a rule-based system that relies on lexical look-up tables, regular expressions and simple heuristics to identify protected health information instances, as well as physician's names and years of dates. 2 434 nursing notes from 153 randomly selected patient records were used for creating a gold standard which was used for developing and refining the algorithm. A second test corpus of 1 836 nursing notes was used for the final evaluation of the system. Results are reported as an estimated recall of 0.943 on the test corpus, where no patient names were missed, and 0.967 recall, 0.749 precision on the development corpus (see Section 3.4 for definitions of the evaluation measures precision, recall and F-measure). A resulting corpus with surrogate protected health information is publicly available under a data use agreement, the resulting software is freely available[4].

The Scrub system is a de-identification tool that uses templates and specialized knowledge in its detection algorithms to identify proper names, address blocks, phone numbers, etc. Each entity has a detection algorithm. Final results are reported as almost 0.99 precision.

A challenge on de-identification was performed at the i2b2 Center (i2b2, 2012), where medical discharge summaries were annotated for Protected Health Information (Uzuner et al., 2007). Seven teams participated in the challenge and best performances for all categories were above 0.98 F-measure. The top performing system achieved results of 0.997 F-measure, using a machine learning based, iterative, Named Entity Recognition approach (Szarvas et al., 2007).

A machine learning based approach is presented in Uzuner et al. (2008), using local context features and a support vector machine implementation. They report a result of 0.97 F-measure.

Kokkinakis & Thurin (2007) present work on de-identification of Swedish discharge letters. They report results of 0.98 recall and 0.97 precision using a rule-based named entity recognition system, although on a different set of identifica-

[4] http://www.physionet.org/physiotools/deid/, Accessed January 21, 2012

tion classes. A review over recent approaches of de-identification is presented in Meystre et al. (2010).

2.4 EPISTEMIC MODALITY: CERTAINTY, SPECULATION AND HEDGING

Epistemic modality, as defined by Nuyts (2001, p. 21 f.), is '(the linguistic expression of) an evaluation of the chances that a certain hypothetical state of affairs under consideration (or some aspect of it) will occur, is occurring, or has occurred in a possible world which serves as the universe of interpretation for the evaluation process, and which, in the default case, is the real world [...]'. He continues: 'In other words, epistemic modality concerns an estimation of the likelihood that (some aspect of) a certain state of affairs is/has been/will be true (or false) in the context of the possible world under consideration'.

The notion of epistemic modality has been addressed from many perspectives, in particular in linguistics and logic. Although there exist different theories and definitions, the core notions are similar: an author committing to the certainty of an uttered proposition (Saurí, 2008).

Related concepts that have been addressed and studied within the natural language processing and linguistic community include for instance *subjectivity* and *hedging*. *Subjectivity*, as defined in Wiebe et al. (2001), means 'aspects of language used to express opinions and evaluations' and is here divided into two main types: *evaluation* and *speculation*, where the latter category is defined as 'including anything that removes the presupposition of events occurring or states holding, such as speculation and uncertainty'.

Hedging is a term that has been associated with linguistic uncertainty and used in many studies, in particular in the domain of scientific writing. It was introduced by Lakoff (1973), and is defined as 'words whose job is to make things fuzzier or less fuzzy'. The term can be interpreted as a linguistic means to indicate a lack of commitment to a statement (Hyland, 1998).

The levels as to which epistemic modality is best modeled is not agreed upon. Nuyts (2001) argues that the estimation of likelihood is situated on a scale, ranging from affirmed certainty to negated certainty, while others propose discrete values (see, e.g. Saurí (2008) for a discussion on this).

The examples in Figure 2.1 show different levels of certainty, ranging from affirmed to negated, with levels of uncertainty in between, expressed through linguistic markers in a fictive part of a medical record. The boundaries between (b) – (e) could differ depending on how certainty levels are interpreted, in some cases, they would be treated as one, i.e. speculation in general, without distinguishing further levels of speculation or uncertainty.

(a) Patient *has* Parkinson.

(b) Physical examination *strongly suggests* Parkinson.

(c) Patient *possibly* has Parkinson.

(d) Parkinson *cannot yet be ruled out*.

(e) *No support* for Parkinson.

(f) Parkinson *can be excluded*.

Figure 2.1: Examples of different levels of certainty, ranging from affirmed to negated, with levels of uncertainty in between, applied on hypothetical patient cases.

In this thesis, the terms uncertainty, speculation and hedging are used interchangeably, and positioned under the overall term *uncertainty*.

2.5 MEDICAL CERTAINTY

Practitioners in clinical medicine are faced with situations where many diagnostic alternatives may be present, and judgments are made difficult by uncertain, incomplete and complex data (Hewson et al., 1996). As a medical student, one is taught to manage uncertainty in different ways, based on factors such as human limitations, characteristics of the patient or disease, organizational problems, professional knowledge, etc. (Lingard et al. (2003), Lester & Tritter (2001)).

In medical practice, verbal probability terms are frequently used to convey levels of certainty in diagnostic reasoning. Studies on how these terms are interpreted by physicians reveal that such interpretations are inconsistent among professionals (e.g. Khorasani et al. (2003) and Hobby et al. (2000)).

Verbal and numerical uncertainty expressions and their role in communicating clinical information have been studied from many perspectives and for different purposes, e.g. decision-making, interpretation, impact on physicians, patients and information systems. Most often, studies have used direct and indirect scaling procedures, giving study objects a fixed number of verbal expressions to judge, and evaluating inter- and intra-subject agreement (see e.g. Clark (1990) for a review). In most cases, intra-subject agreement is found to be high, and inter-subject agreement to be low (see Chapter 3 for further details on measuring annotator agreement). Intermediate probabilities are often more difficult to agree on, while very high or low probabilities result in higher agreement (see e.g. Khorasani et al. (2003), Hobby et al. (2000), Christopher & Hotz (2004)). In many cases, the main conclusion is to recommend the use of controlled vocabularies for expressing different levels of certainty. The verbal expressions used range from one word expressions such as *definite*, *likely*, *possible*, to longer expressions such as *cannot be excluded*. The relationship between expressing probabilities verbally or numerically has also been studied (e.g. Timmermans (1994) and Renooij & Witteman (1999)), where findings suggest that verbal expressions are found to be more vague than numerical, and hence more difficult to use in decision-making.

2.6 ANNOTATED CORPORA OF CERTAINTY

Research on modeling uncertainty for natural language processing and information extraction has gained interest lately. Several annotation efforts and corpora have been created, in particular in the biomedical domain and for news documents. Some examples are given below.

The BioScope Corpus (Vincze et al., 2008) contains annotations on a sentence and keyword level applied on biomedical research articles and abstracts, as well as medical records (radiology reports). Sentences that are *speculated* or *negated* are marked. Speculative sentences are those 'that state the possible existence of a thing, i.e. neither its existence nor its non-existence'. Negated sentences are those with an 'implication of the non-existence of something'. Speculative and negation elements, or cues, are marked together with their linguistic scope.

For text mining in the biomedical domain, the need for distinguishing uncertain and negated information has gained particular focus in recent years. Wilbur et al. (2006) present a model with five qualitative dimensions: focus, polarity, certainty, evidence and directionality, applied on sentences from biomedical research articles. Certainty levels are represented as values between 0–3, where 3 is the highest level of certainty. The GENIA corpus is a large resource consisting of biomedical articles annotated for part-of-speech tags, syntactic information, terms and events. The event annotations (1 000 abstracts) also include annotations for negation and three levels of uncertainty (Collier et al., 1999). Light et al. (2004) present work on sentence level speculation identification on biomedical abstracts, where three levels of speculations are defined: *low speculative*, *high speculative* and *definite*.

FactBank is an annotated resource for event factuality applied on a news corpus (Saurí & Pustejovsky, 2009). Two polarities (positive and negative) and three levels of certainty (certain, probable, possible) are used. Rubin et al. (2006) present a study on certainty classification of news documents, with a model that also includes perspective, focus and time. They model certainty in four degrees: absolute, high, moderate or low (which can be seen as comparable to the values 0–3 in Wilbur et al. (2006)).

Subjectivity is studied in the work by Wiebe et al. (e.g. Wiebe et al. (2001) and Wiebe et al. (2005)), where the resulting MPQA Corpus has been widely used in other research studies. Here, speculation is considered a type of subjectivity. However, no degrees of speculation are modeled, as the focus lies in the differences between *subjective* and *objective*, i.e. perspective. Another example is the ACE (Automatic Content Extraction) Corpus[5], which has a *relation* annotation part where modality is included as a binary distinction: *asserted* and *other*.

[5] Version 6.0: http://www.ldc.upenn.edu/Projects/ACE/, Accessed January 22, 2012

2.7 AUTOMATIC CLASSIFICATION OF CERTAINTY IN THE BIOMEDICAL DOMAIN

Medlock & Briscoe (2007) present a weakly supervised learning model applied on a corpus of biomedical articles, where sentences are classified as either speculative or non-speculative. This corpus is also annotated for gene names and is used in Szarvas (2008) along with clinical radiology reports and used for building probabilistic learning models. The same corpora are used in Kilicoglu & Bergler (2008) for identification of speculative language, using a linguistically motivated approach with lexical resources, syntactic patterns and weighting schemes for quantifying hedging strengths.

The BioScope Corpus has been used for creating automatic uncertainty classification systems. For instance, the CoNLL 2010 Shared task included the biomedical sub-corpus for the task of detecting hedges and their scope in natural language text (Farkas et al., 2010). The top performing system obtained an overall F-measure of 0.86 for detecting uncertain sentences (Tang et al., 2010), and 0.57 for detecting in-sentence hedge cues (Morante et al., 2010). The clinical part of the BioScope Corpus is also used in Morante & Daelemans (2009) for a machine learning based classifier of uncertainty cue scopes.

In the clinical domain, using medical records, rule-based systems have been developed for distinguishing negations and uncertainties (e.g. Chapman et al. (2001) and Friedman et al. (2004)). ConText (Harkema et al., 2009), is an extension of the NegEx algorithm (Chapman et al., 2001), where three contextual features are used for identifying negated, historical, and hypothetical conditions, and conditions not experienced by the patient, in emergency department reports. RadReportMiner (Wu et al., 2009) is a context-aware search engine, taking into account negations and uncertainties, achieving improved precision results (0.81) compared to a generic search engine (0.27) using a modified version of the NegEx algorithm, including expanded sets of negation and uncertainty keywords.

Chapman et al. (2011) present an extension of the ConText algorithm for building a document-level classifier of CT pulmonary angiography reports, where certainty states of diagnoses are modeled as *uncertain*, *present* or *absent*, among other features.

The 2010 i2b2/VA challenge (Uzuner et al., 2011) included a task on assertion classification of medical problem concepts, along with two other tasks: concept extraction and relation classification. Here, a medical condition was to be classified as '*present*, *absent*, or *possible* in the patient, *conditionally present* in the patient under certain circumstances, *hypothetically* present in the patient at some future point, and mentioned in the patient report but associated with someone other than the patient'. de Bruijn et al. (2011) obtained best results for the assertion task, with an F-measure of 0.94, using different combinations of machine learning classifiers and feature representations.

A comparison between rule-based and machine learning approaches for assertion classification is presented in Uzuner et al. (2009), where NegEx is extended to cover assertions, and a machine learning based classifier (StAC) based on Support Vector Machines (SVM) is presented, applied on medical records from different domains. Here, a medical problem is assigned either presence, absence or uncertainty, or association with someone other than the patient. The machine learning based approach yields best results.

It should be noted that *negation identification* is a closely related task that has received a considerable amount of attention in research on information extraction both in general and specifically for medical records. Here, it is included in the models of uncertainty, and not treated as a separate task.

CHAPTER 3

METHOD

This chapter details the method choices made for addressing the research questions. Overall assumptions, framework and process are described in Section 1.2. First, the data used for creating the annotated gold standards is described (Section 3.1), followed by a description of the annotation models and guidelines that were used for each task (Section 3.2). Second, the approaches taken for automatic classification are given and discussed (Section 3.3). Third, evaluation methods are described: annotator agreement measures are used for evaluating the gold standard corpora, and classification performance is measured against the gold standards (Section 3.4). Limitations are elaborated in Section 3.5, followed by a discussion on ethical issues in Section 3.6.

3.1 DATA: THE STOCKHOLM EPR CORPUS

Data from the Stockholm EPR Corpus was used (Dalianis et al., 2009). This corpus is extracted from TakeCare[1], an Electronic Health Record system used in the Stockholm County Council (Stockholms läns landsting). The data covers electronic health records from this system during the years 2006, 2007 and the first half of 2008, from around 900 clinical units in the Stockholm area.

[1] http://www.cgmtakecare.com/, Accessed January 19, 2012

The health records contain both structured (e.g. age, gender, diagnosis code, measure values) and unstructured (i.e. free-text) data. Around 40% of the data entries are unstructured, constituting a majority of the total amount of data (Dalianis et al., 2009). The system allows for semi-structured free-text entries, meaning that there are predefined keywords or headings that are used for specific types of documentation, e.g. *anamnes* (patient history), *status* (patient status), *bedömning* (assessment) and *åtgärd* (planned action). These headings can be chosen freely by each clinical department (and profession) and are used in templates where any chosen number of headings and structured entries can be used. The data contains documentation from different types of health professionals such as physicians, nurses and physical therapists. General statistics from the first five months of 2008 is shown in Table 3.1 (modified version from Dalianis et al. (2009)). Three subsets were extracted from this corpus and are further described below.

Table 3.1: The Stockholm EPR Corpus: general statistics from the first five months of 2008. *Total amount of free-text headings used in the medical record system (years 2006 – 2008) = 6 164. **Total amount of ICD-10 codes in the data set = 35 185. hapax legomena = one occurrence, that is, 55% of the tokens occur only once in the corpus, 4% occur 100 times or more.

2008 (5 months)	*n*	%
Men	188 238	46%
Women	219 906	54%
Free-text headings	2 631	43%*
ICD-10 Codes	16 211	46%**
Clinics	888	
Tokens	109 663 052	
Types	853 341	
Average no of tokens per record	269	
hapax legomena = 1	467 706	55%
dis legomenon = 2	107 636	13%
tris legomenon= 3	51 161	6%
< 10	732 150	86%
> 100	34 245	4%

A gold standard for de-identification (Paper I)

Goal: extract representative documents for the de-identification task that are to be used for annotation and creating the annotated gold standard corpus.

Five different clinics were chosen: Neurology, Orthopaedia, Infection, Dental Surgery and Nutrition. For each clinic and gender, the medical records richest in free-text were included, in total five records per gender and clinic, amounting to one hundred medical records in total. A *medical record* was defined as *all* the documentation for *one patient from each clinic*. For each clinic and gender, the top five records richest in free-text were included. All available data was included, separated in columns (tab separated), i.e. structured as well as unstructured.

Different types of clinics were chosen in order to capture variations in language style but also in how potential identifiable information might differ depending on clinical discipline. Choosing the medical records richest in free-text instead of a random sampling was motivated by the fact that these might contain more instances of identifiable information. Representative documents for this task means that as many instances of identifiable information as possible were to be included in the gold standard, as opposed to a "representative" document compared to the total amount, or population, of medical records as a whole.

This choice means that there is a bias in the type of texts that were included in the corpus. For instance, there is a risk that the included records are not quite representative for a "typical" medical record from the respective clinic. However, as the task is to annotate identifiable information, a randomly extracted medical record for each clinical discipline might not result in many annotations; the most important task here is to achieve as high coverage as possible of instances that might risk exposure of individuals. Moreover, the chosen clinical disciplines might be debatable. The aim was to include as different clinical disciplines as possible, and, by consulting domain expertise, these were chosen. An alternative could have been to randomly sample the included clinics from the total amount of clinics in the Stockholm EPR Corpus, or to create a random sample of all the patients from the total amount of patients.

All types of authors were included, e.g. physicians, nurses and physical therapists. This was motivated by the fact that *all* instances of identifiable information are to be found, irrespective of profession.

These bias issues limits the usability of the resulting gold standard for other purposes. It can not, for instance, be used to infer the prevalence of identifiable information as a whole in Swedish medical records. However, it can be used as a reference standard for evaluating how well e.g. an automatic classifier is able to classify identifiable information as defined in the annotation task (see Section 3.2), compared to human annotators.

A gold standard for sentence level certainty classification (Papers II and III)

Goal: extract representative documents for the sentence level certainty classification task that are to be used for annotation and creating the annotated gold standard corpus.

In the initial uncertainty annotation task, a subset from the Stockholm EPR Corpus containing only assessment (*bedömning*) fields was chosen. Sentences[2] were randomly extracted from all assessment entries in the total data set. The assessment entry was chosen based on the knowledge that these entries are those that contain the largest amount of reasoning. The documents are written or dictated[3] by physicians.

A representative document was here defined as one assessment entry irrespective of clinic, patient or time. The aim was to understand how uncertainties are expressed in general in the parts of the medical records that contain the largest amount of reasoning.

Choosing a random sample from *all* clinical departments in the Stockholm EPR Corpus has drawbacks. The diversity between different medical disciplines may be too large, and a deeper understanding of the implications of uncertain utterances may be more fruitful to study separately for one discipline at a time. On the other

[2] Sentences are split by using a simple sentence tokenizer, based on punctuation and capitalized letter heuristics.

[3] In the case of dictation, a secretary has transcribed the dictation manually.

hand, there might also be overall similarities, and as there are no previous studies on this phenomenon in the Swedish clinical domain, a broad characterization may instead be given. The resulting annotated gold standard was used for corpus analysis, analyzing differences between different clinical disciplines (Paper III).

Using only the assessment field limits the coverage and loses context, i.e. only parts of the whole medical records were used. However, as this is where the most reasoning is documented in the medical record, it captures an essential property and serves as a good starting point for understanding how uncertainties are expressed.

A gold standard for diagnostic statement level certainty classification (Paper IV)

Goal: extract representative documents for the diagnostic statement level certainty classification task that are to be used for annotation and creating the annotated gold standard corpus.

For creating the gold standard of certainty classification on a diagnostic statement level, two steps were needed. First, defining the entities, i.e. diagnostic statements, required the compilation of a diagnostic statement lexicon.

The diagnostic statements were identified through manual analysis. Two physicians marked diagnostic statements on a subset of 150 random assessment entries from the chosen emergency department. As stated above (Section 2.1), the language in health records is noisy. Diseases can have many names, and with the time pressure involved in the daily clinical activities, they may also be misspelled and/or abbreviated in numerous ways. For instance, *Noradrenalin* (a medication) has been found in 350 different variations in Finnish health records, and 60 variations in Swedish (Allvin et al. (2011), and Suominen (2009)).

For this reason, a manually created list of diagnosis statements was preferred over using existing terminologies such as ICD-10[4] or SNOMED-CT[5], in order to cap-

[4] International Classification of Diseases, http://www.who.int/classifications/icd/en/, Accessed January 22, 2012

[5] Systematized Nomenclature of Medicine-Clinical Terms, http://www.ihtsdo.org/snomed-ct/, Accessed January 22, 2012

ture as many variations as possible. The physicians conformed to a definition of a diagnostic statement: *a medical condition with a known cause, prognosis or treatment*. All variants, including abbreviations, misspellings and inflections were marked. In total, 337 diagnostic statements were compiled in the lexicon.

There are alternative methods for compiling lexicons. For instance, the process could be (semi-)automatized by building e.g. a distributional lexical semantic model using techniques such as random indexing (e.g. Sahlgren (2006)) on a larger set of medical records, defining a number of diseases or diagnoses to look up in this model, evaluating the results manually, and compiling the lexicon from the evaluation result. This approach would, however, also have problems. For instance, deciding which diagnostic statements to look for needs to be defined. Focusing on only one type of disease or clinical department type would give a richer and more focused characterization of how knowledge certainty is expressed in such a specific context. Moreover, utilizing existing terminologies would ensure that the diagnostic statements to be judged are generally accepted by the research and health care community. A combination of these techniques would also be a viable option, using terminologies for finding related concepts in a semantic model. However, in both cases, the broad coverage would be lost, i.e. the coverage gained by letting domain experts identify diagnostic statements manually means that one ensures that all relevant variants are identified.

Second, extracting health records containing these statements was needed. For this corpus, assessment entries were also used. However, only one clinical department was chosen: a university hospital emergency ward. That is, a representative document was defined as one assessment entry from all medical records from one emergency ward. This was motivated from the fact that this is a clinical department where many different types of diseases are encountered, which makes it possible to analyze differences between different types of diagnoses – some diagnoses are clinically difficult to ascertain, while others are easier. The sampling of the assessment was randomized from the total amount of assessment entries from the chosen emergency ward.

To build the corpus for diagnostic statement level certainty annotation, a simple, automatic, string matching procedure along with a general Swedish language lemmatizer[6] was applied, with a longest string-match heuristic. All diagnos-

[6] http://www.cst.dk/online/lemmatiser/, Accessed March 21, 2012

tic statements were marked with brackets for the annotators, e.g. *Patient with <Diagnosis>diabetes mellitus</Diagnosis>*.

Matching entries from given lexicons in documents can also be performed in alternative ways. String edit distance algorithms, such as the Levenshtein distance algorithm, described in e.g. Jurafsky & Martin (2009), are powerful in their simplicity in capturing spelling variants of given entries. Other alternatives include using named entity recognizers or techniques similar to those described above, or, of course, combinations of such techniques. The simple approach chosen is time- and complexity efficient and made it possible to compile a useful corpus for the task at hand.

3.2 ANNOTATIONS AND GUIDELINES

The data sets were used as the base for creating each annotated gold standard. An annotation model and guidelines for applying the model on the data were needed for each annotation task. Annotation models consist of annotation classes that represent the information that is to be identified. Annotation guidelines contain definitions, examples, and instructions for the annotators to apply the annotation model on the data. An iterative process was employed in order to define and refine the annotation classes in each annotation model. In order to capture what the medical records actually contain, a grounded theory methodology (Strauss & Corbin, 1990) was employed for creating the annotation models, similar to the approach taken in Chapman & Dowling (2006). Moreover, through literature reviews, all annotation guidelines were based on, or inspired by, existing related resources, enabling comparison of results (to some extent).

The *goal* for each annotation task was to 1) create a model that represents the desired output, i.e. a representation of identifiable instances or a representation of uncertainty at some level in medical records, and 2) to create guidelines that are clear and understandable for the annotators, so that the annotation task can be carried out and result in high annotator agreement, and thus a reliable annotated corpus. Measuring and evaluating agreement is further discussed in Section 3.4.1.

Knowtator (Ogren, 2006), a plugin in the Protégé Ontology and Knowledge Acquisition System[7], was used for performing the annotation work.

De-identification (Paper I)

Three annotators annotated the de-identification set: one senior medical researcher (SM), one senior computer science researcher (SC) and one junior computer science researcher (JC). The senior medical researcher is a domain expert, while the other two are non-domain experts. For this task, domain knowledge is not essential, as the instances to be annotated do not require medical knowledge. No interaction between the annotators was allowed during the annotation work.

Due to the lack of specific regulations regarding which information is considered identifiable and risking patient integrity in the free-text parts of electronic health records in Swedish legislation, the U.S. Health Insurance Portability and Accountability Act (HIPAA)[8] formed the basis of defining entities to be annotated in the de-identification gold standard. HIPAA defines a number of so called Protected Health Information (PHI) types:

- Names
- Locations
- Dates
- Ages > 89 years
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social security numbers
- Medical record numbers
- Health plan beneficiary numbers

[7] http://protege.stanford.edu/, Accessed January 30, 2012

[8] http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm, Accessed January 22, 2012

- Account numbers
- Certificate/license numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) address numbers
- Biometric identifiers
- Full face photographic images and any comparable images
- Any other unique identifying number or characteristic

Through two annotation iterations, the following additions and refinements were made for the resulting annotation model:

- Names:
  - are divided into *full*, *first* or *last* names, and nested if applicable.
  - are specified into *patient*, *clinician* and *relative*. A generic name tag is used if neither is applicable.
  - Example: "John Smith" (clinician):
    < Clinician Full Name >
    < Clinician First Name >
    John
    < /Clinician First Name >
    < Clinician Last Name >
    Smith
    < /Clinician Last Name >
    < /Clinician Full Name >
- Dates:
  - Full date (year, month and date)
  - Date part (month and/or date)
  - Year
- Ethnicity
- Relations

Although these changes and additions complicate comparisons to previous approaches (for a review, see Meystre et al. (2010)), they were deemed important for the task at hand, as they represent important instances that could be used for inferring the identity of an individual in a medical record. Moreover, they can be normalized and collapsed into broader classes employed in other research efforts, to facilitate comparisons.

Sentence level certainty (Papers II and III)

The sentence level uncertainty annotation model was inspired by the BioScope corpus (Vincze et al., 2008), where biomedical articles and abstracts, as well as clinical radiology reports, are annotated on a sentence level for uncertainty and on a token level for speculation and negation cues. However, some changes were made. First, in order to capture cases where contradictory expressions were embedded in one sentence, for instance through subordinate clauses, the annotators were allowed to divide sentences into sub-expressions, by marking them separately. Second, the linguistic *scope* was not included in the token level annotations, i.e. defining the *scope* of a negation or speculation cue based on linguistic criteria, as is done in the BioScope corpus. Moreover, sentences containing question marks were annotated, which is not the case in the BioScope corpus. An annotation class for *certain* sentences was also included.

The annotation classes are: *certain_expression*, *uncertain_expression* and *undefined_expression* for sentence, or sub-sentence expressions. Certainty in this case was modeled irrespective of a sentence being in the positive or negative polarity. *Undefined_expression* was used for cases where the annotator deemed the sentence unclear with respect to certainty level. On a token level, the annotation classes are *negation*, *speculative_words* and *undefined*. The latter annotation class was used for token level keywords that the annotators were uncertain about. Token level annotations were allowed to encompass multi-word tokens. The entire assessment entry was shown to the annotators, in order to provide the surrounding context. The sentence to be annotated was marked with brackets. An example entry is shown in Figure 3.1.

Non-domain expert annotators performed the annotation task; one senior level student, one undergraduate computer scientist, and one undergraduate language con-

<s>Således tolkas som viros med övre luftvägsinfektion.</s> Får återgå till hemmet med utökad febernedsättande regim samt åter om 2-3 dagar om ej förbättrad.

<s>Hence interpreted as virosis with upper respiratory infection.</s> Can return home with increased antifebrile regime and return in 2-3 days if no improvement..

Figure 3.1: Example assessment entry. <s> = sentence. Each sentence was to be assigned a certainty class. If needed, the sentence could be broken up into sub-expressions, by marking and assigning a certainty class for each sub-expression . On a token level, keywords were to be marked either for negation or speculation. The token level classes could span over multi-word tokens, and were allowed to be nested, e.g. a negation could be nested within a speculative keyword.

sultant. They had no prior knowledge about the content of the data. The annotators worked independently while annotating, but met in even intervals to discuss the task and refine the guidelines. This was performed in order to measure the effect of problem resolving and result differences over time, as in Haverinen et al. (2009).

Choosing to base the annotation model by inspiration of an existing model facilitates comparison between results. At the time, there were not many existing resources for uncertainty annotation in biomedical texts (in particular medical records) that also provided guidelines for the annotation task. Choosing to make some changes in the model makes a comparison more difficult, but the general trends may still be comparable. The sentence level approach has some disadvantages, for instance, it does not provide information about which information that is uncertain. By allowing the annotators to mark and separate contradictory certainty levels within one sentence, a deeper understanding of this phenomenon is obtained. However, this means that the total amount of annotations may differ between annotators, which is somewhat problematic. Sentence level annotations are a more time and complexity efficient unit to analyze computationally.

Allowing the annotators to break sentences into sub-expressions also complicates the evaluation, as the measurable units might result in different total amounts. However, it was judged as important to allow for such freedom, as this was an initial attempt at characterizing the way uncertainty is expressed in Swedish medical records. The motivation for not basing the annotation schema on linguistic criteria was to focus on the clinical relevance of the information contained in the medical records.

Diagnostic statement level certainty (Papers IV, V and VI)

A more fine-grained uncertainty annotation model was chosen in order to capture a more detailed level of characterization. First, the level of analysis is on *diagnostic statements*, a more fine-grained unit than sentences. Second, the uncertainty levels are more detailed, compared to the binary sentence level annotation task described above. Inspired by the model presented in FactBank (Saurí & Pustejovsky (2009) and Saurí (2008)), a model with two polarities and three gradations was used: *Positive* and *Negative* along with the gradations *Certain*, *Probable* and *Possible*, six annotation classes in total, see Figure 3.2. The model is to be considered a rough scale, or continuum.

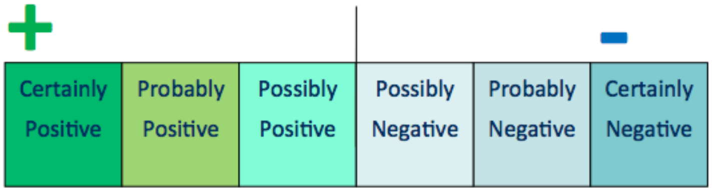


Figure 3.2: Certainty level classification of diagnostic statements: two polarities and three levels of certainty, in total six classes.

For cases where the diagnostic statement in its context meant something else (e.g. *infektion* (infection, short for clinic)) the annotation class *Not Diagnosis* was added. Moreover, *Other* was added as an annotation class for cases where the annotator could not assign any of the above-mentioned classes, or where the diagnostic statement referred to someone other than the patient.

Choosing a model with groups, and not, for instance, a linear representation of certainty levels, was motivated partly because of comparability to previous approaches, and partly due to computational complexity.

Two domain expert annotators performed the task: physicians (A1 and A2) with experience in both reading and writing medical records. The annotation guidelines were created through iterations and are publicly available[9]. An example entry is shown in Figure 3.3 (from Paper VI).

[9] http://www.dsv.su.se/hexanord/guidelines/guidelines_stockholm_epr_diagnosis_factuality_corpus.pdf

Oklart vad pats symtom kan komma av. Ingen säker <D>infektion</D>. Inga tecken till inflammatorisk sjukdom eller <D>allergi</D>. Reflux med irritation av luftrör och således hosta? Dock har pat ej haft några symtom på <D>refluxesofagit</D>. Ingen ytterligare akut utredning är befogad. Hänvisar till pats husläkare för fortsatt utredning.

Unclear what patient's (abbr.) symptoms arise from. No certain <D>infection</D>. No signs of inflammatory disease or <D>allergy</D>. Reflux with irritation of airways and therefore cough? But pat has not had any symptoms of <D>refluxoesophagitis</D>.No further urgent investigation required. Refer to pat's GP for continued investigation..

Figure 3.3: Example assessment entry. D = Diagnostic statement. Each marked diagnostic statement was judged for certainty levels. In this case, the diagnostic statements *infektion* (infection), *allergi* (allergy) and *refluxesofagit* (refluxoesophagitis) were to be assigned one of the six certainty level annotation classes.

Distinguishing fine-grained levels of uncertainty is not trivial. It has been showed that more annotation classes and detailed levels of knowledge representation lead to lower agreement results (Bayerl & Paul, 2011). However, less granular models lose expressive power. Although defining the distinctions between low levels of certainty between the polarities posed difficulties, the annotators found the resulting model functional and agreeable. Moreover, as discussed in e.g. Nuyts (2001), certainty levels are perceived as a scale by humans, which motivates the chosen model.

As discussed in Section 2.6, many different certainty level annotation models have been proposed, with different certainty level distinctions. It is difficult to compare the models, including the one proposed here, as they are applied on different corpora, and have been designed for different purposes. Models that include several levels of certainty are more costly when it comes to computational efficiency, i.e. classifying several classes (certainty levels) requires more complex computational models. Moreover, defining borders between several annotation classes in an annotation model is also intricate and requires more labor, which motivates coarser distinctions. On the other hand, coarser certainty level models do not represent the subtleties expressed through natural language, and may lose expressive power. Such issues need to be taken into consideration when designing an annotation task for modeling uncertainty. Choosing to base this model mainly on the one proposed by Saurí (2008) has drawbacks, since the *events* are not directly comparable (*diagnostic statements*, in this proposed model), and the source and temporality are not addressed. Despite these differences and disadvantages, the core model remains in focus and can be compared, at least to some extent, i.e. the assignment of levels of certainty and polarities.

3.3 AUTOMATIC CLASSIFICATION

Two different methods were used for automatic classification: rule-based for de-identification, and machine learning for diagnostic statement level uncertainty classification. The goal for both approaches was to create an automatic classifier that approaches the performance level of human annotators, i.e. to create an automatic classifier that is able to classify instances as well as humans are able to.

3.3.1 DE-IDENTIFICATION

A rule-based classifier called De-Id (Neamatullah et al., 2008), developed for English, was chosen for the de-identification task (Paper I). This software package relies on lexical resources and is well documented, is freely available and has shown good results for American English. It was adapted to Swedish by replacing the English lexical resources with Swedish equivalents with as little manual intervention as possible. A list of Swedish diseases and lists of addresses and names were extracted from the Internet from various publicly available resources (for details, see Paper I). Different sizes of the name lexicon were used, 10 000 to over 100 000 names in each lexicon. Addresses were extracted from electronic municipality maps. 2 000 new locations and 4 000 new organizations were extracted from the Stockholm EPR Corpus (excluding the medical records contained in the de-identification gold standard) using the learning module of a Swedish Named Entity Recognizer (Dalianis & Åström, 2001). The De-Id software also includes lists of the most frequent tokens from the medical record corpus for not marking common tokens as protected health information instances. For Swedish, this was generated from the Stockholm EPR corpus into two lists: one with the 5 000 most common tokens, and one with the 50 000 most common tokens.

Choosing a rule-based system instead of, e.g., a machine learning system, was motivated by the efficiency and non-reliance on training data, as the gold standard was relatively small. All external resources, i.e. lexicons and lists of words, could, of course, have been compiled in alternative ways. However, focus was put on scalability, coverage and efficiency, minimizing manual workload.

3.3.2 DIAGNOSTIC STATEMENT LEVEL CERTAINTY CLASSIFICATION

Classifying diagnostic statement certainty levels was done by using machine learning techniques. A sequence labeling machine learning approach was chosen, using Conditional Random Fields (CRF) (Lafferty et al., 2001) as implemented in the CRF++ package[10]. Token level classifiers were built: all diagnostic statements[11] belonged to one and only one certainty level class, all other tokens were assigned the class *NONE*.

All eight annotation classes from the Stockholm EPR Diagnosis Uncertainty Corpus annotation model were used for multi-class classification looking at local context features: word, lemma and part-of-speech tags (Paper V). A second classification task was also performed, where intermediate certainty levels were collapsed: *probably* and *possibly* positive and negative were grouped into *probably_possibly_[positive|negative]*, and *other* and *not_diagnosis* were grouped into one joint class, in total five classes. This was performed in order to study classifier performance on a less complex multi-class classification problem.

Following the results from Paper V, the same classifier and top performing set of features were used for classifying three e-health scenario tasks (Paper VI): *adverse event surveillance*, *decision support alerts* and *automatic summaries*. For each scenario, the fine-grained certainty level classes[12] were grouped into coarser-grained certainty classes: *existence* and *no existence*, *plausible existence* and *no plausible existence*, and *affirmed*, *speculated* and *negated*, respectively.

Sequential labeling classifiers such as Conditional Random Fields have been successful for information extraction tasks in several natural language processing experiments. There are, of course, other classification algorithms that could have been used instead. For instance, Support Vector Machines (SVM) have also produced good results for similar tasks, e.g. Uzuner et al. (2009). Results in Uzuner et al. (2009) show that local context features are most useful in a similar setting,

[10] http://crfpp.googlecode.com/svn/trunk/doc/index.html, Accessed March 21 2012.

[11] Multi-word diagnostic statements such as *heart attack* were concatenated and treated as one token: *heart_attack*.

[12] The classes *other* and *not diagnosis* were disregarded for this experiment.

which motivates the feature setting choice presented here. Moreover, the choice of learning algorithm was not central, instead, a feasibility study is in focus.

The chosen scenarios also only serve as examples for future real-world implementation settings. There are, naturally, other possible scenarios where other distinctions may be needed. However, the aim was rather to show both that there are scenarios that need these distinctions (in different ways), and that it is possible to use one fine-grained model for several coarser-grained scenarios.

3.4 EVALUATION

Statistical measures were used for evaluating both the annotated corpora (reference standards) through annotator agreement, and classification performance by comparing results against the reference standards. Annotator agreement evaluation means that one measures how well the annotators agree on the annotation task, i.e. how the application of the annotation model through applying the annotation guidelines is interpreted and agreed upon by different annotators. Ideally, annotators understand the task identically and agree on all instances, which means that the annotation model is well-defined and that the guidelines are clear and unambiguous: the resulting annotated corpus is *reliable*, indicating the *validity* of the annotation model and guidelines (Artstein & Poesio, 2008).

Evaluating classification performance means that results are measured against a reference (gold) standard, a collection of documents containing the desired output (as defined by humans). Commonly, this is also compared to a baseline, e.g. a random or majority class assignment.

For both tasks, the number of *true positives* (TP), *false positives* (FP) and *false negatives* (FN) is needed. For some measures, the number of *true negatives* (TN) is also needed. *True positives* are the correctly labeled instances,*false positives* are the instances incorrectly labeled as positives, and *false negatives* are the instances incorrectly labeled as negatives. In many text classification tasks, the number of *true negatives* is often either unknown or proportionally very large, which affects evaluation results negatively. Measures that do not require the number of *true negatives* commonly used in the natural language processing and information extraction research community are *precision*, *recall* and *F-measure*. *Precision*, or *positive predicted value, PPV*, gives the proportion of correctly classified instances from all resulting positive instances, Eq. 3.1.

$$Precision : P = \frac{TP}{TP+FP} \tag{3.1}$$

Recall, or *sensitivity*, gives the proportion of correctly classified instances from all positive instances in the data set, Eq. 3.2.

$$Recall : R = \frac{TP}{TP+FN} \tag{3.2}$$

Increasing *recall* by, e.g., assigning more instances to a class, leads to a decrease in *precision*. A combination of these two measures, such as the *F-measure*, the harmonic mean of the two with a weight (β) set for precision and recall, is often used as an indicator of the overall performance, Eq. 3.3. When equal weight is given for *precision* and *recall* ($\beta = 1$), this is also called the *balanced f-score* or F_1, which is used here.

$$F - measure : F_\beta = \frac{(\beta^2+1)P \times R}{(\beta^2 \times P)+R} \tag{3.3}$$

Qualitative error analysis is performed for all tasks. Errors are analyzed manually and categorized according to emerging types.

3.4.1 ANNOTATOR AGREEMENT

For the de-identification (Paper I) and the sentence level uncertainty (Papers II and III) gold standards, inter-annotator agreement results were calculated with precision, recall and f-measure. For these tasks, the entities to be annotated were not predefined, which means that there might be differences in span coverage and the total amount of annotations. The inbuilt agreement calculator in Knowtator (Ogren, 2006) was used for calculating agreement over classes and spans in the de-identification gold standard. For the sentence level uncertainty gold standard, an in-house built script was used for calculating exact and partial matches. Exact matches were based on a token level, while partial matches were based on a character level, i.e. for each token if all characters in a token was marked equally by two annotators, there is both an exact and a partial match. Pairwise agreement was calculated, which means that each annotator was evaluated against one other, for all annotator combination pairs. Overall average agreement is the average of the pairwise agreement results.

The annotator agreement on the diagnostic statement level uncertainty annotation task (Paper IV) was evaluated through F-measure and Cohen's κ (Cohen, 1960). Moreover, both intra- and inter-annotator agreement results were calculated. Intra-annotator agreement results were evaluated in order to measure consistency in one annotator, and was done by creating a new, randomized order for a subset of the corpus. All measures were calculated with an in-house built script.

The number of true negatives is poorly defined for the two first tasks (de-identification and sentence level uncertainty), as there may be overlaps and varying lengths. Agreement measures such as Cohen's κ are thus not possible to calculate in these cases (Hripcsak & Rothschild (2005), Wilbur et al. (2006) and Chapman & Dowling (2006)). When there is an unknown value of true negatives, the F-measure approaches κ (Hripcsak & Rothschild, 2005).

For the task of annotating diagnostic statement level uncertainty, the instances to be annotated were predefined, and there was no overlap or varying lengths of the annotations. Thus, evaluating with both κ and F-measure provided a rich picture of the agreement results.

As the certainty levels in the diagnostic statement level certainty model were considered as a scale, or continuum, weighted κ (κ_w) is a measure well-suited for analyzing the agreement (Kundel & Polansky, 2003). Through this, relative importance for disagreement in distant categories is deemed higher than those closer in the scale, or ranking, as opposed to giving equal weight to all classes, as with Cohen's κ. These agreement results have been added as an extension to the results presented in Paper IV[13].

Evaluating annotator agreement is not trivial. In particular, defining thresholds where agreement is deemed 'good' is subject to some debate, as are choices of measures (see, e.g. Artstein & Poesio (2008), Di Eugenio (2000) Di Eugenio & Glass (2004)). Landis & Koch (1977) propose threshold values for interpreting

[13] Only for the certainty level classes, not for *other* and *not diagnosis*

κ value	Strength of agreement
<0.00	Poor
0–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Table 3.2: Strength of agreement, beyond chance, measured by κ, according to Landis & Koch (1977)

the strength of agreement as measured by κ, see Table 3.2. Stricter interpretations have also been proposed, stating that a threshold of above 0.8 'ensure an annotation of reasonable quality', and that values above 0.67 allows for tentative conclusions to be drawn (Artstein & Poesio, 2008). However, as discussed in Artstein & Poesio (2008), stating specific thresholds for all purposes is not possible. Instead, reporting method choices in detail in order for readers to be able to interpret whether agreement results hide disagreements is more important. Moreover, the impact of domain expertise, the complexity of the annotation models and other factors that may have impact on annotation results is important to take into consideration (Bayerl & Paul, 2011). For the work presented here, no specific thresholds are set, although the aim is, naturally, to reach as high agreement as possible. As a minimum, a general aim is to at least achieve moderate agreement, as defined by Landis & Koch (1977), putting focus in analysis of the results.

3.4.2 AUTOMATIC CLASSIFICATION

For the automatic classification approaches, evaluation means that, given the chosen classification method and setting, the classifier is able to approach human performance as defined in the reference standard, to some lesser or greater extent.

The ported De-Id package to Swedish was evaluated against each one of the three manually annotated gold standards. Precision, recall and F-measure were used. Micro-averaged numbers are given (see below).

Results from using the Conditional Random Fields machine learning algorithm on the diagnostic statement level gold standard were evaluated by splitting the set into a training set (80%) and a test set (20%) with a stratified class distribution. The gold standard was annotated by one annotator (A1). This set extends the corpus annotated by the two annotators (A1 and A2), which was created for inter-annotator agreement evaluation. Precision, recall and F-measure were calculated using the CoNLL 2010 Shared task evaluation script conlleval.pl[14]. Micro-averaged numbers are given. 95% confidence intervals were calculated for precision and recall.

Micro-average calculations means that equal weight is given for each instance in the classification, i.e. results are calculated for each annotation instance separately. Macro-averaged calculations, on the other hand, give equal weights for each *class*, or category, in the classification, i.e. the average is calculated for the annotation class, not for the average of all instances. With skewed class distributions, meaning that some annotation classes are much more frequent than others, the latter tends to favor the majority class.

3.5 LIMITATIONS

For modeling uncertainty on both sentence and diagnostic statement levels, only notes written by physicians were used. Nurse documentation may potentially also have a large amount of reasoning which would be important to include in an overall information extraction system. However, this may be necessary to model separately, as this type of documentation differs from that written by e.g. physicians. The data is from one geographical area, and from one electronic medical record system, which limits generalizability for conditions in Sweden as a whole.

Time and resources are always limitations in research. Having more annotators, and creating larger annotated resources would, of course, be desirable. The annotators themselves are also a source of limitation: they are pooled from an educated population, from the Stockholm area, and internally from the research group.

This is not a thesis on machine learning or classification. There is a large amount of research on machine learning algorithms, feature engineering, parameter tuning and performance evaluation. The classification parts of this thesis serve as feasibility studies, pointing towards the overall goal.

[14] http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt, Accessed March 21 2012

3.6 ETHICAL ISSUES

There are of course general rules and regulations pertaining to performing research on clinical data. Access to patient data is only permitted if approved by local regional ethical boards. Applying for permission requires rigorous descriptions of the planned research study. Moreover, there is a specific law, Patientdatalagen (patient data law), SFS 2008:355[15], in which regulations about what type of information medical records must and must not contain is stated. For instance, it is stated that the purpose of the law is that personal information is to be designed so that, and treated in a way that patients and any other registered person's integrity is respected (§2). Similar to e.g. Finland (Suominen, 2009), Swedish legislation does not address natural language processing as a specific case for performing research on medical record data.

For the research presented in this thesis, permission was granted from the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission numbers 2007/1625-31/5 and 2009/1742-31/5. When applying for this permission, careful descriptions about the planned research were written, as well as detailed information about how the data itself would be stored and secured. From the hospital side, the data was de-identified in the sense that all social security numbers (personnummer) were replaced by an anonymous, random key, and the replacement key was not given to the research group, i.e. it is not possible to re-identify any social security number from the data obtained. Moreover, all names (in the structured entry) were removed.

The data is stored on an encrypted, password enforced, server in a locked and alarmed room to which only a handful of researchers have access, after signing confidentiality agreements. The data is never exposed to a network connection, ensuring that data is never unconsciously sent to a third party. Small subsets, such as those used for building the corpora described in this thesis, were extracted and stored locally on encrypted, password enforced files for annotation and development. Annotation and development was never performed while connected to any network.

Published research results contain no confidential information, and were chosen carefully to be as general as possible, ensuring that individual anonymity is kept.

[15] The law text can be found in its entirety here: http://www.notisum.se/rnp/sls/lag/20080355.htm (Accessed on January 16, 2012)

Risks

Although care has been taken in ensuring individual patient integrity by conforming to both health related research regulations, patient data regulations, and, naturally, general research ethics regulations[16], there is, of course, always risks that these precautions are not sufficient. The main risks involved in the work presented here are the possibilities of re-identifying an individual patient from the medical records by combining external information or indirect information contained in the documents. This risk is deemed minimal, as great care has been taken to use data out of its context, access to the data is severely restricted and confidentiality agreements are ensured.

[16] see, for instance, http://www.codex.vr.se/en/forskarensetik.shtml, Accessed January 22, 2012

CHAPTER 4

RESULTS

This thesis results in three annotated gold standards: one annotated for identifiable information (the Stockholm EPR PHI Corpus) and two for uncertainty: the Stockholm EPR Sentence Uncertainty Corpus, annotated on a sentence level, and the Stockholm EPR Diagnosis Uncertainty Corpus, annotated on a diagnostic statement level. The Stockholm EPR PHI Corpus was used for evaluating an automatic de-identification classifier, and the Stockholm EPR Diagnosis Uncertainty Corpus was used for training and testing a machine learning based classifier, using the certainty level model in different ways and for different purposes. Furthermore, a lexicon of Swedish diagnostic statements was produced. More detailed results are presented below, and further details are found in the respective papers. These resources are the first of their kind.

None of the resulting gold standards have, in the included work, been compiled into consensus sets, i.e. sets where disagreements have been resolved and are treated as a new 'ground truth'. Instead, they are to be considered initial steps, where agreement results reflect the success (or failure) of the annotation model and subsequent guidelines. These results are related to the *reliability* of the created corpora: the higher the agreement, the more reliable gold standards.

4.1 THE STOCKHOLM EPR PHI CORPUS

In order to make medical records available for research, it is important to ensure that patient integrity is kept. Identifiable information is found in the free-text parts of medical records, and this needs to be removed or replaced, i.e. de-identified, before releasing any data for research. A gold standard corpus of Swedish medical records annotated for identifiable information was produced: the Stockholm EPR PHI Corpus. This was created by using an annotation model containing in total 40 annotation classes along with guidelines for applying these classes on Swedish medical records.

Counting all data, the total number of tokens is 380 000 (around 31 000 types). Counting only the free-text columns, the number of tokens is 174 000 (around 20 000 types). A simple white-space tokenizer was used, including numbers as tokens and types.

The Inter-Annotator Agreement (IAA) results are 0.58 F-measure (micro-averaged), when looking at the overall average agreement between the three annotators over *spans*. The results ranged between 0.46 and 0.75 F-measure when looking at pairwise agreement, see Table 1 in Paper I. IAA for *classes* ranged between 0.55 and 0.84 pairwise F-measure, with an overall average of 0.65, Table 2 in Paper I.

34% of the total number of annotations are names. IAA for these classes was high: 0.80 F-measure overall average, pairwise agreement ranging between 0.72 and 0.91 F-measure. Locations resulted in lower agreement: 0.29 F-measure (overall average, spans) and 0.48 F-measure (overall average, classes). The location classes amount to 29% of the total number of annotations. Discrepancies were mainly due to differences in span coverage, 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*.

From the 40 defined annotation classes, some were not present in the data, e.g. *Health care beneficiary number* and *Biometric Identifiers*. A total of 28 annotation classes were identified.

The annotation model contains many fine-grained annotation classes, e.g. separate name classes for clinicians, patients, and relatives. In Table 4.1, merged annotation classes and their frequency are presented. These numbers were extracted after resolving disagreements among the annotators, and after collapsing fine-grained annotation classes into more generic annotation classes in order to lower the complexity of the task (Dalianis & Velupillai, 2010).

Annotation Class	*n*
Age	56
Date Part	710
Full Date	500
First Name	923
Last Name	929
Health Care Unit	1021
Location	148
Phone Number	136
Total	4423

Table 4.1: Frequency of annotation classes for de-identification after resolving disagreements and collapsing fine-grained classes into more generic classes.

The annotation task is complex, in the sense that the annotators themselves marked the boundaries of each annotation instance, leading to different total amounts of annotations, and in the sense that the amount of annotation classes is large. Given these complexities, the IAA results were considered reliable, in particular for annotation classes such as *names*, although caution is to be taken when interpreting results. As discussed in Section 3.4.1, defining specific thresholds for where IAA results are to be considered reliable is not trivial.

Porting the de-identification software De-Id to Swedish resulted in very poor precision results, between 0.03 and 0.09. Recall ranged between 0.56 and 0.76, yielding F-measure results between 0.04 and 0.16. The main problem was an excessive over-generation of most tags, except for dates, where the system performed quite well. Using different sizes of external lexicons did not improve results significantly, as the content of the lexicons did not reflect the content of the medical records.

4.2 THE STOCKHOLM EPR SENTENCE UNCERTAINTY CORPUS

Building information extraction tools that are able to handle and distinguish negated and uncertain information from affirmed information requires knowledge about how such information is expressed. This has not previously been studied in Swedish medical records. The first step in gathering this knowledge has resulted in the Stockholm EPR Sentence Uncertainty Corpus, where uncertain and certain expressions were annotated on a sentence level, and negation and speculation keywords were annotated on a token level.

A total of 6 740 sentences was annotated by three annotators: one senior level student (ULS), one undergraduate computer scientist (UCS), and one undergraduate language consultant (ULC). The total amount of tokens is 69 495.

Overall micro-averaged Inter-Annotator Agreement (IAA) results for the sentence level uncertainty annotation task resulting in the Stockholm EPR Sentence Uncertainty Corpus improved over time, from 0.53 to 0.79 F-measure lowest pairwise agreement, and from 0.58 to 0.80 highest pairwise agreement, exact match. Partial match agreement was higher: from 0.63 to 0.82 and from 0.67 to 0.85 pairwise agreement lowest and highest, respectively. These results are shown in Table 5 in Paper II.

The majority sentence level class, *certain_expression* resulted in high overall micro-averaged agreement: from 0.81 to 0.84 pairwise F-measure, exact match. However, *uncertain_expression* resulted in lower agreement, from 0.38 to 0.53 pairwise, overall micro-averaged F-measure, exact match. Partial match agreement was higher for all sentence level classes. These results are shown in Tables 1 and 2 in Paper II. On average, around 13 percent of all sentences are uncertain.

On the token level, *negations* resulted in high agreement: from 0.82 to 0.88 pairwise F-measure, for both exact and partial matching. *Speculative_words*, on the other hand, resulted in lower overall agreement, ranging between 0.47 to 0.58 overall pairwise F-measure, the highest agreement was seen between annotators UCS and ULC for time interval 2: 0.63 F-measure, see Tables 3 and 4 in Paper II.

From the thirteen clinical groups[1] (Table 1 in Paper III), it is showed that *geriatrics* had a low average amount of uncertain sentences and a high overall average pairwise agreement, while *neurology* had a high average amount of uncertain sentences, see Figure 4.1 (from Paper III). On average, uncertain sentences were longer than certain sentences (Figure 3 in Paper III).

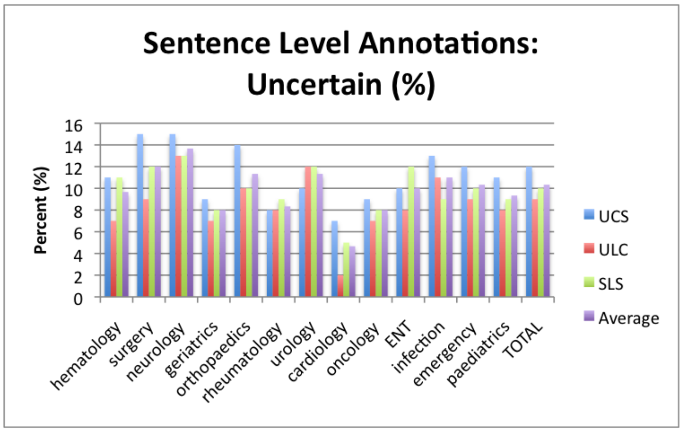


Figure 4.1: Sentence level annotation: *uncertain*, percentage per annotator and clinical practice.

Negations, in total thirteen unique tokens, were unigrams, while *speculative words* had an average token length of 1.34 and were often *n*-grams such as *kan vara* (could be) and *tyder på* (indicates that). The most common speculative words per annotator for *neurology* and *urology*, with the highest overall average of uncertain sentences, are shown in Table 3, Paper III. The longest *n*-grams, ranging between three and six tokens, were often nested with negations, such as *kan inte se några tydliga tecken* (can't see any clear signs) and *inte helt har kunnat uteslutas* (has not been able to completely exclude). Question marks were the most common tokens annotated as a speculation cue. *Sannolikt* (likely) was almost always annotated as a speculative word (over 90 percent), while *om* (if) was only annotated as a speculative word in 9 percent of all occurrences.

[1] Clinical disciplines were grouped together, and those with a total number of sentences > 100 were analyzed.

Similar to the de-identification task, stating a specific threshold at which results are considered reliable is not easy. As the annotators were allowed to define boundaries both at the sentence and token level, overlaps and different total amounts of annotations were found. However, given that the task was designed as an initial annotation study, the IAA results are considered reliable for performing corpus analysis and for refining the task further.

4.3 THE STOCKHOLM EPR DIAGNOSIS UNCERTAINTY CORPUS

Moving closer to the goal of building more efficient information extraction systems and deepening the knowledge about how uncertainties, negations and affirmations are expressed in Swedish medical records, the diagnostic statement level annotation task was created, resulting in the Stockholm EPR Diagnosis Uncertainty Corpus.

A total of 3 846 assessment entries were annotated in the corpus. 1 297 assessment entries were annotated by both annotators (A1 and A2). The remaining assessment entries were annotated by one annotator (A1), for extending the amount of annotation instances for training and evaluating an automatic classifier.

The fine-grained diagnostic statement certainty level annotations in the Stockholm EPR Diagnosis Uncertainty Corpus resulted in fairly high overall agreement: 0.7/0.58 F-measure, 0.73/0.6 Cohen's κ, Intra/Inter. Looking at only the certainty level classes, using κ_w, results were higher: 0.88/0.82 with proportional weights, and 0.95/0.92 with quadratic weights, Intra/Inter κ_w, respectively. A contingency table with the annotation class assignments for certainty level classes is shown in Table 4.2[2].

Of the total amount of diagnostic statements in the created lexicon (337), 227 were found in the data. From the total amount of assessment entries, approximately 50% contained at least one of the diagnostic statements in the created lexicon. The lexicon itself is a valuable resource for e.g. terminology analysis, and is publicly available upon request.

[2] The discrepancy in total amount of annotation instances between the two sets was caused by mismatches and missed instances.

	CP	PrP	PoP	PoN	PrN	CN	∑
CP Intra	990	78	4	0	3	4	1079
Inter	*834*	*59*	*7*	*0*	*4*	*5*	*909*
PrP Intra	20	236	55	1	1	0	313
Inter	*66*	*134*	*10*	*1*	*0*	*0*	*211*
PoP Intra	4	38	127	25	9	0	203
Inter	*11*	*149*	*180*	*41*	*45*	*1*	*427*
PoN Intra	0	0	6	14	7	1	28
Inter	*0*	*0*	*0*	*1*	*5*	*1*	*7*
PrN Intra	1	1	1	10	118	25	156
Inter	*0*	*0*	*0*	*2*	*35*	*18*	*55*
CN Intra	2	0	4	0	51	195	252
Inter	*2*	*0*	*0*	*4*	*99*	*193*	*298*
∑ Intra	1017	353	197	50	189	225	2031
Inter	*913*	*342*	*197*	*49*	*188*	*218*	*1907*

Table 4.2: Contingency table, Inter- and Intra-Annotator frequency distribution per annotation class. Columns: Annotator A1, first annotation iteration. Rows: Intra: Annotator A1, second annotation iteration (same set randomized), Inter: Annotator A2. CP = Certainly Positive, PrP = Probably Positive, PoP = Possibly Positive, PoN = Possibly Negative, PrN = Probably Negative, CN = Certainly Negative, ∑ = Total

Only approximately 50% of the diagnoses were affirmed with certainty. The lowest certainty levels in the negative polarity (*possibly negative*) was rare, and resulted in low agreement (0.35/0.03 F-measure, Intra/Inter). The majority class, *certainly positive*, resulted in high agreement, 0.9 F-measure for both intra- and inter-annotator agreement. A contingency table with results for all annotation classes for both intra- and inter-annotator agreement is shown in Table 1, Paper IV.

Patterns in certainty levels assigned to different types of diagnostic statements were observed. The fifteen most common diagnostic statements are shown in Table 4.3 and their certainty level assignments are shown in Figure 4.2.

For instance, diagnostic statements that show on the outside, e.g. eczema, urthicaria, skin infection and varicoses were dominantly *certainly positive*, as were general conditions such as overweight or asystolia, and diagnoses that are measured by an instrument, such as auricular fibrillation/ECG. Generic diagnostic statements

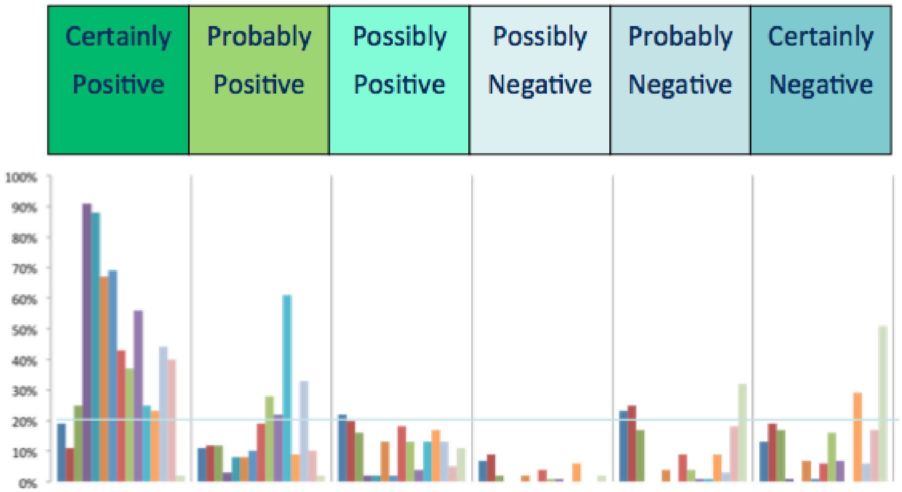


Figure 4.2: Diagnostic statement level annotation: the fifteen most common diagnostic statements and their certainty levels. *Certainly positive* is in majority. On the far right, *certainly negative*, the generic diagnostic statement *skeleton injury* is found. If a skeleton injury is confirmed, a more specific term is used. Under *probably positive*, the highest bar is *virosis*, a common condition often stated as probable when no other diseases can be confirmed.

such as *skeleton injury* were found in the negative polarity, while specific findings (e.g. fractures) were found in the positive polarity, with the specific fracture diagnosis name. Similarly, if the patient did not suffer from an ischaemic heart disease, the generic diagnostic statement *ischaemia* was often used and found in the negative polarity, while if the patient had a confirmed or probable diagnosis, *heart attack* or *angina pectoris* was used, see Figure 4.3.

When there are medical reasons for not securing certainty, for instance for common, 'fuzzy' diseases such as virosis and gasthritis, *probably positive* dominated. For some conditions, such as *hypertension*, a counterpart in the negative polarity was not found, i.e. either the patient has normal blood pressure or low (hypotension).

Although differences were seen in how certainty levels were expressed for different diseases, the markers for certainty levels were most often lexical keywords.

Diagnostic statement (Swedish)	English translation
dvt	deep venuos thrombosis, abbreviated
lungemboli	pulmonary embolism
infektion	infection
förmaksflimmer	atrial fibrilation
hypertoni	hypertension
hjärtsvikt	congestive heart failure
KOL	COPD, chronic obstructive pulmonary disease
angina	angina
pneumoni	pneumonia
allergisk reaktion	allergic reaction
viros	virosis
blödning	bleeding
uvi	urinary infection, abbreviated
hjärtinfarkt	heart attack
ischemi	ischaemia

Table 4.3: The fifteen most common diagnostic statements found in the Stockholm EPR Diagnosis Uncertainty Corpus.

Examples of some common lexical markers in the respective polarities and certainty levels are shown in Figure 4.4.

This annotation task differed from the two previous tasks: the instances to be annotated were predefined. This makes evaluation of annotation agreement results somewhat easier, there are no discrepancies in marking boundaries. Some instances were missed by the annotators, resulting in different total amounts of annotated instances, but the span coverage for all annotations were identical. When analyzing results with F-measure and Cohen's κ, agreement results can be considered *moderate*, according to the thresholds given by Landis & Koch (1977). However, given that the certainty level classes are ordered, or considered as a scale, weighted κ (κ_w)-measures are more suitable, and with these, we see that results were very encouraging, indicating reliable results.

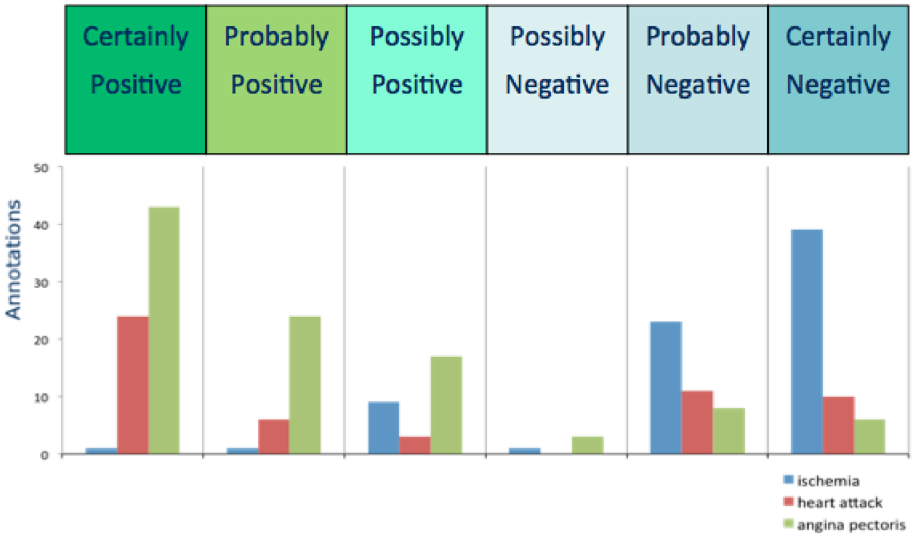
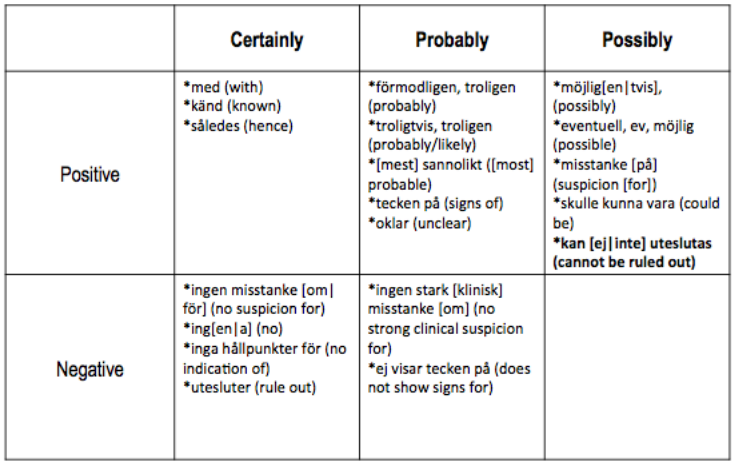


Figure 4.3: Diagnostic statement level annotation: inverted pattern, *ischaemia* in the negative polarity, *heart attack* and *angina pectoris* in the positive polarity.

4.3.1 AUTOMATIC CLASSIFICATION: LOCAL CONTEXT FEATURES

Overall micro-averaged results for a baseline where only the word itself was used as a feature was 0.56 F-measure, for all classes, and 0.60 F-measure for merged classes, see Table 4 in Paper V. A majority class baseline was 47.6%. Adding local context features step by step improved results, where the best results were obtained using words, lemmas and part-of-speech features in a window size of ±4. This setting resulted in 0.70 F-measure for all classes and 0.76 F-measure for merged classes, see Table 5 in in Paper V. *Preceding* context was more effective than *posterior* context; similar results were obtained when using the four preceding words, lemmas and part-of-speech tags (0.69/0.74 F-measure, all/merged versus 0.60/0.65). The greatest improvement was seen for *certainly negative*, where using local context features (±4) yielded 0.72 F-measure compared to 0.43. This trend was seen for all window size steps, with the greatest increase between ±2 and ±3: from 0.55 to 0.67 (all classes).



	Certainly	Probably	Possibly
Positive	*med (with) *känd (known) *således (hence)	*förmodligen, troligen (probably) *troligtvis, troligen (probably/likely) *[mest] sannolikt ([most] probable) *tecken på (signs of) *oklar (unclear)	*möjlig[en|tvis], (possibly) *eventuell, ev, möjlig (possible) *misstanke [på] (suspicion [for]) *skulle kunna vara (could be) ***kan [ej|inte] uteslutas (cannot be ruled out)**
Negative	*ingen misstanke [om|för] (no suspicion for) *ing[en|a] (no) *inga hållpunkter för (no indication of) *utesluter (rule out)	*ingen stark [klinisk] misstanke [om] (no strong clinical suspicion for) *ej visar tecken på (does not show signs for)	

Figure 4.4: Diagnostic statement level annotation: common lexical markers for the different certainty levels. Markers for certainly positive are not always explicit. For possibly negative, agreement is low, and few instances are assigned this class. The example marked with bold in *Possibly positive* could, for some diagnostic statements, be used as a marker for *Possibly negative*.

Typical errors were within the same polarity or missed instances. Local context features were not sufficient for cases where conjunctions were used, e.g. *Inga hållpunkter i lab och ekg för pågående ischemi* (no basis in lab and ecg for ongoing ischaemia), and for cases where lab results were indicators for specific certainty levels. Moreover, some sentences were very short and did not contain anything but the diagnostic statement itself, and some diagnostic statements were part of a longer discussion with many modifiers and speculations, both of which would require larger contexts and other feature models.

4.3.2 AUTOMATIC CLASSIFICATION: E-HEALTH SCENARIOS

For the three e-health scenarios presented in Paper VI, *adverse event surveillance*, *decision support alerts* and *automatic summaries*, promising results were obtained: 0.89, 0.91 and 0.8 overall micro-averaged F-measure, respectively. Each scenario is further discussed in the sections below.

Adverse event surveillance

Adverse event surveillance means that a hospital wants to avoid adverse events such as hospital acquired infections or other conditions or events that have happened to patients in a hospital and that endanger their safety. Normally, such instances are identified retrospectively by scrutinizing medical records for specific triggers that indicate the possibility of an adverse event. Only cases that are negated with the highest possible level of certainty should be excluded from an automated support system.

For this binary classification task (*existence* and *no existence*), local context features (±4) improved results considerably compared to a classifier baseline using only the word itself as a feature, from 0.66 F-measure to 0.89, but only slightly compared to a majority class baseline (88%), see Table 3 in Paper VI. For both classes, precision results were improved the most, from 0.53 to 0.93 (*existence*) and from 0.54 to 0.83 (*no existence*).

The error analysis showed that known difficulties in the distinction between the certainty level classes *probably negative* and *certainly negative* in the annotation model were reflected in the classification results. The strength of the negation was the source for most errors, i.e. there were not many errors in assigning polarity. A typical phrase that was judged differently was *inga hållpunkter för* (no indicators of), where the inconsistencies were often linked to specific diseases that are also difficult to clinically exclude, e.g. *DVT* (deep venous thrombosis). Modifiers such as *liten* (small) in phrases like *liten misstanke* (small suspicion) were ambiguous: whether emphasis was put on *liten* (*small* suspicion) or *misstanke* (small *suspicion*) yielded different interpretations in the strength of uncertainty.

Decision support alerts

This scenario reflects the situations where an alert would support a clinician in making a decision. For instance, the clinician missing the information about insufficient pain medication could receive an automated alert when the documentation about the pain observations and extra medication have reached a specific threshold. Here, the important certainty level distinction lies in separating positive (or near positive) cases from negated cases.

Overall micro-averaged F-measure results for the binary classifier (*plausible existence* and *no plausible existence*) was 0.91, an improvement over the majority class baseline (80%) and the classifier baseline (0.61 F-measure), see Table 4 in Paper VI. The minority class *no plausible existence* improved both for precision (from 0.72 to 0.92) and recall (from 0.22 to 0.79).

Sources of errors were mainly cases where clinical exclusion is difficult for a disease (e.g. *DVT*), but also cases where a test has been performed, often in order to exclude a diagnosis. It was often evident from the surrounding context that the diagnosis is unlikely, but the performing of a test is in itself an indicator of a suspicion.

Automatic summaries

An overview, or textual summary, would support clinicians in getting an overall impression of a patient's medical history and earlier conditions. For these cases, a distinction between affirmed, negated and speculated instances would ease the understanding of the patient's current situation.

The multi-class classification problem with the classes *affirmed*, *speculated* and *negated* resulted in an overall micro-average F-measure of 0.8, which was an improvement over both baselines (0.5). Precision results were improved for all three classes, from 0.79 to 0.87 (*affirmed*), 0.25 to 0.81 (*speculated*) and 0.50 to 0.81 (*negated*), see Table 5 in Paper VI.

The border between *certainly positive* and *probably positive* in the annotation model was the main source of errors. Again, assigning polarity was not the problematic issue, but rather distinguishing fine-grained levels. Diseases that are measured by e.g. an apparatus, such as *hypertoni* (hypertension), showed higher agreement, while diseases that are measured subjectively, such as *hyperventilering* (hyperventilation) and *panikångest* (panic disorder), were more often disagreed upon. Markers such as *misstänkt* (suspected) and *kliniska tecken på* (clinical signs of) were not judged consistently. Chronic diseases caused problems in some cases, where the example *troligen stressutlöst astma* (probably stress triggered asthma) could be assigned *certainly positive* (the patient has asthma) or *probably positive* (this particular event of an asthma attack is probably triggered by stress).

CHAPTER 5

CONCLUSIONS, CONTRIBUTIONS AND POSSIBLE WAYS FORWARD

The hospital administrator needing support in finding relevant medical records for identifying adverse events, the clinician missing important pain medication information in the medical record documentation, and the physician needing to sift through hundreds of pages of documentation to get an overview over a new patient, have yet to see a system that supports them automatically in these tasks. However, this research is an important step towards this goal, as it fills a knowledge gap in the essential steps that need to be taken for building such systems.

5.1 CONCLUSIONS

A number of research questions were stated in Chapter 1. These are addressed below.

5.1.1 MOVING TOWARDS MAKING MEDICAL RECORDS AVAILABLE FOR RESEARCH

One of the aims of this thesis was to move towards making medical records available for research. To achieve this, an annotation model with annotation classes covering instances of identifiable information was created and applied on Swedish medical records from five different clinics.

How can a de-identified corpus of Swedish medical records be created?

By defining an annotation model encompassing annotation classes for identifiable information, and creating guidelines for applying this model on Swedish medical records, an annotated resource was created: the Stockholm EPR PHI Corpus.

From the lack of definitions of what constitutes identifiable information in medical records in Swedish legislation, the annotation model was based on the protected health information instances defined in US regulations. These were further refined into a total of 40 annotation classes.

The gold standard in its original annotated form, i.e. annotated by three annotators, is deemed reliable when evaluating with pairwise and overall average Inter-annotator agreement measures, although it results in some problematic issues. Inter-annotator results were high for some classes, e.g. names, and the fine-grained model captures details in different types of identifiable information. However, span coverage and the application of fine-grained classes such as *Municipality* and *Town* differed between the annotators. Results are in line with similar research, e.g. Mani et al. (2005). The choice and motivation of which Protected Health Information instances to cover in a de-identification corpus and/or system has differed in previous research efforts (see, e.g., Meystre et al. (2010) for a review). With a fine-grained approach such as the one included in the Stockholm EPR PHI Corpus, it is possible to collapse classes into coarser-grained classes, thus enabling different perspectives.

The corpus is valuable for several reasons: it contains medical records from five different types of clinics and documentation from several types of clinical professions. De-identified corpora available for research most often contain medical records from one clinical department or one type of author only (e.g. Finnish intensive care nursing narratives (Haverinen et al., 2009) and American nursing notes (Neamatullah et al., 2008)). There is currently ongoing work on replacing all annotations with pseudonyms, further ensuring a minimal risk of patient integrity exposure, for creating a corpus to be released for research.

Can an existing de-identification tool built for English be ported to handle Swedish medical records?

Porting an existing rule-based de-identification software was not trivial and required extensive tailoring which might be very time-consuming. The obtained results are in line with those obtained when attempting to port the same de-identification software to French (Grouin et al., 2009). Instead, machine learning methods might be better suited for this task. In a follow-up study, the Stockholm EPR PHI Corpus has been refined into two variants and used for training and evaluating a Conditional Random Fields classifier, yielding promising results (0.8 F-measure). Moreover, in this study, through an error analysis, 49 new instances were identified, that were missed by the annotators, showing that machine learning algorithms might complement misses made by human annotators (Dalianis & Velupillai, 2010).

5.1.2 CERTAINTY LEVELS IN SWEDISH MEDICAL RECORDS

Another aim of this thesis was to provide a description of how certainty levels, i.e. affirmed, speculated and negated information, are expressed in medical records, create models and corpora that capture this, and build classifiers that distinguish them, for different information needs. This was achieved through two annotation tasks: one on a sentence level, using laymen as annotators, and one on a diagnostic statement level, using domain-experts as annotators. Moreover, feasibility studies on automatic classification of diagnostic statement level uncertainty were performed.

How is medical uncertainty expressed in medical records (in Swedish) on a sentence level?

Medical uncertainty is, to a large extent, expressed through lexical markers such as *misstänkt* (suspected), *sannolikt* (probably) and *troligtvis* (likely). Although not very common, sentences with conflicting certainty level information are found.

The Stockholm EPR Sentence Uncertainty Corpus is the first resource of Swedish medical records annotated for uncertainty information on a sentence level. Through this, differences between clinical disciplines have been identified, where *neurology* contained more uncertain sentences on average, while *geriatrics* contained fewer. In neurology, clinicians are faced with diseases that are more difficult to ascertain. The majority of all sentences were affirmed, or certain. However, an average of 13.5% of all sentences were judged as uncertain, which is similar to the clinical part of the BioScope Corpus (Vincze et al., 2008), and also to similar research on scientific articles (e.g. Light et al. (2004)). This is a considerable amount, having implications for building intelligent information extraction systems.

Most sentences did not contain conflicting certainty levels, but when they do, they need to be separated. Hence, a sentence level model is sufficient for most cases, but not for all. Speculative keywords were often longer than one token, and negations play an important role for strengthening uncertainty, e.g. *ingen typisk urinvägsinfektion* (not a typical urinary tract infection). This is an important feature also addressed in Kilicoglu & Bergler (2008), where terms indicating strong certainty ('unhedgers'), such as *typical* or *clear*, suggest uncertainty when found within the scope of a negation.

For non-domain experts, this task was difficult, which was reflected in the agreement results. The domain specific jargon and high level of clinical reasoning is not easily accessible for people without clinical expertise. However, important insights have been obtained. Negations are, in their core forms, unambiguous and easy to identify. Some speculation cues are also unambiguous, e.g. *sannolikt* (likely) while others are not, e.g. *om* (if).

How is medical uncertainty expressed in medical records (in Swedish) on a diagnostic statement level?

A broad picture of how medical uncertainty is expressed on a diagnostic statement level in Swedish clinical assessment entries from an emergency ward has been produced through the creation of the Stockholm EPR Diagnosis Uncertainty Corpus. This is the first of its kind. Domain expertise was essential for this task. Despite a difficult task, overall agreement was high, especially when weighting discrepancies closer to each other in the spectrum less than discrepancies further away on the scale.

Certainty levels for different types of diseases are expressed in different ways, an important finding for future implementations. Most certainty levels are expressed through lexical markers, but not all. Test results of different kinds are important cues and indicate different levels of certainty, depending on both the disease type and the test itself. Some diseases are very severe, and are crucial to identify even if they are very rare. As in the sentence level model, the majority of the instances were affirmed with the highest level of certainty. However, through the fine-grained model, a broader certainty level picture was gained: a disease that *might not* be is very different from one that *might* be, and the prevalence of these is significantly high for a distinction to be important.

Disagreements in annotations also reflected subjective interpretations. Uncertainties can be expressed in subtle ways, and the context plays an important role. Linguistic modifiers can be ambiguous and background knowledge may influence judgements. For some cases, such issues could be clarified through refining guidelines. However, it is impossible to reach perfect agreement, as this phenomenon to a large extent is inherently subjective. Specifically, boundaries in intermediate certainty levels are a source for different interpretations, which is also found in the studies presented by e.g. Khorasani et al. (2003) and Hobby et al. (2000).

How can a corpus annotated for uncertainty on a diagnostic statement level be used for automatic classification?

Applying a Conditional Random Fields classifier using simple local context features yielded promising results, showing that the corpus can be used for automatic classification. Certainty levels are mostly expressed by lexical markers preceding the diagnostic statement, which is shown in Paper V. Best results were obtained using a window of ±4, i.e. the four preceding and posterior words, lemmas and part-of-speech features. Although applied on a different certainty level distinction, on a different language and on a different data set, results are in line with those presented in Uzuner et al. (2009) in that local context features in a window of ±4 yield best results. This classification model is to be seen as a first step in improving automatic classification of certainty levels, and the results are useful for future developers of systems incorporating such a model. In particular, it is clear that local context features are very important, and that certainty levels to a high extent are indicated through lexical markers, which is in line with the findings from the corpus analysis. However, other features may also be important, such as syntactic information for e.g. conjunctional phrases, and higher-level features such as test results.

The choice of machine learning algorithm and the overall setup could, of course, be studied further. A majority class baseline together with a simple classifier baseline only gives a limited perspective on how such a corpus could be used optimally for automatic classification. Moreover, as discussed above, discrepancies in the annotations may influence classification results negatively, where a refined corpus might yield better results.

How can a corpus annotated for uncertainty on a diagnostic statement level be used for automatic classification of different information needs (i.e. real-world scenarios)?

The fine-grained certainty level annotation model applied on a diagnostic statement level can be used for real-word e-health scenarios by identifying different boundaries on the certainty level scale and creating new, coarser-grained certainty level groups for automatic classification. The features obtaining best classification results in Paper V were used and promising results were obtained for each scenario, compared to majority class baselines and baseline classifiers not using context features.

The important take home message is not only that the building of one fine-grained model and resource can save time and manual labor (as opposed to creating separate annotation tasks for each scenario), but also that real-world scenarios should be kept in mind when creating corpora and classifiers.

5.2 CONTRIBUTIONS

This thesis contributes resources that are valuable for further research, and knowledge about the characteristics of Swedish medical records when it comes to identifiable information and medical uncertainty. Two annotation models of certainty are provided, which are the first in their kind applied on Swedish medical records. A deeper understanding of the language use linked to conveying medical certainty levels is gained, from both a layman and domain expert perspective. Most importantly, through a broad coverage approach, knowledge has been gained as to how uncertainties are expressed when looking at different clinical disciplines and different diagnostic statement types.

Three annotated resources that can be used for further research have been created: the Stockholm EPR PHI Corpus, the Stockholm EPR Sentence Uncertainty Corpus, and the Stockholm EPR Diagnosis Uncertainty Corpus. One lexicon containing Swedish diagnostic expressions is also produced. Moreover, one of the corpora has been successfully applied for building automatic classifiers, and for classification tasks reflecting real-world scenarios.

5.3 POSSIBLE WAYS FORWARD

The overall goal of building more accurate information extraction systems that can aid clinicians, researchers and other professions in their daily work, and the long-term goal of improving health care in general, are still future dreams. However, through the contributions of this thesis, that future is not as distant as before. The steps taken in the presented research serve as sub-modules in a larger picture that will develop further in the near future.

As with all research projects, the proposed answers and contributions are accompanied with at least as many questions and ideas for future endeavors. With the knowledge gained from pursuing this journey, many insights have been reached.

Interdisciplinarity and domain knowledge

Working with medical records requires understanding of the contents of the data. Building efficient tools requires understanding of technical issues. Building efficient tools that can handle (noisy) text written in natural language requires understanding of languages. Working in an interdisciplinary research group facilitates this combination and provides invaluable literacy for solving difficult tasks. Collaborations across different disciplines are emerging, which is evident when looking at publications in influential conferences.

Clinical language

The findings from this research raises many questions. What importance does context play? Would different results evolve if more context from the medical records were used? Are there similarities in other languages that could be exploited? We have already seen that medical records are very similar in content even if they are written in different languages such as Finnish and Swedish (Allvin et al., 2011). As was shown in Chapter 2, a large amount of research in this area exists for English; ongoing research on comparing English and Swedish[1] uncertainty expressions in medical records will perhaps reveal similarities and differences that are useful for building multilingual tools and for discovering whether existing resources can be used across languages.

Furthermore, there are other crucial aspects that need to be taken into account when building more accurate information extraction systems from medical records. For instance, *time* is a critical component in health care. When was a disease confirmed? When was it first suspected? How long is the time in between? Moreover, *perspective* is important: *who* is the owner of a suspicion, the patient, the treating clinician, or someone else?

[1] Collaboration with researchers at the Division of Biomedical Informatics, University of California, San Diego, USA.

Language modeling and automatic classification

The approaches taken in this work for modeling the language of uncertainty have focused on characterizing the phenomenon on a surface level. Results indicate that syntactic information is important for some cases, e.g. conjunctional phrases, which should be studied further.

Moreover, the classification approaches have been employed using simple features; higher-level features should be studied as well. Whether to build a rule-based system or a machine learning based system depends on many factors. With rule-based systems, training data is not needed. On the other hand, such methods require extensive tailoring - lexical resources, pattern definitions, etc. Machine learning methods are easier to maintain, but creating training data is time-consuming.

Defining when results are 'good' is not trivial. Care needs to be taken in defining desired thresholds for how well an automated system is to work, and in defining the task itself. What is the purpose of the classification, in which setting is it to be used, and what requirements are there?

Uncertainty from a philosophical and psychological perspective

What are the philosophical implications of certainty levels expressed in natural language? From a psychological perspective, what role do these expressions play, and how are they interpreted by different actors, e.g. clinicians and patients, in the case of medical records? Is it feasible to, instead of grouping certainty levels into discrete categories, whether or not at a scale, represent them in some other way? Are there *shades* of certainty, or not, or does this depend on different situations? Is the task of defining these better suited as a two-way approach, first assigning polarity (*positive* or *negative*)?

Real-world scenarios

As stated in Chapter 1, this research is positioned in a pragmatic framework. The results are intended to serve practical use, at least in the long run. By creating a fine-grained annotation model of certainty levels, it is possible to address different information needs, e.g. the hospital administrator who needs to identify adverse events requires a strict distinction between affirmed and negated, the physician who misses to take action needs a coarser yes-no distinction, and the physician who would benefit from an automated overview requires a yes-no-maybe, which is possible to obtain through the created model and could be implemented in an information extraction system.

The presented scenarios are, of course, only examples of use cases where certainty level distinctions play an important role. Other examples include medical education; can medical students benefit from learning about the implications of uncertainties expressed in medical records? Biosurveillance is another example, is it possible to detect severe diseases in real time by analysis of medical records? This is, for instance, studied in ongoing research with the National Institute for Information and Communications Technology Australia's (NICTA) Health Business Team and Machine Learning Research Group in the case of detecting invasive fungal infections from radiology reports[2]. Patients are also playing an increasingly active role in their own health process, and want to read what is written about them by medical professionals; what type of tools would be useful for them, and how do certainty levels play a role in these?

[2] Initial findings from this work were presented as a poster at the NICTA Techfest 2012; Sydney, NSW, Australia; 23 February 2012, entitled *Bio-Surveillance via Text Mining – Improved Safety for Patients and Hospitals*.

REFERENCES

Allvin, H., Carlsson, E., Dalianis, H., Danielsson-Ojala, R., Daudaravicius, V., Hassel, M., Kokkinakis, D., Lundgren-Laine, H., Nilsson, G., Nytrø, Ø., Salanterä, S., Skeppstedt, M., Suominen, H., & Velupillai, S. (2011). Characteristics of Finnish and Swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. *Journal of Biomedical Semantics*, *2(Suppl 3):S1, doi:10.1186/2041-1480-2-S3-S1*.

Alpaydin, E. (2010). *Machine Learning*. Cambridge, Massachusetts: The MIT Press, 2nd ed.

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, *34*(4), 555–596.

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern Information Retrieval – the concepts and technology behind search*. Pearson Education Ltd., second ed.

Bayerl, P. S., & Paul, K. I. (2011). What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation. *Computational Linguistics*, *34*(4), 699–725.

Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, *33*, 23–60.

Boytcheva, S., Nikolova, I., Paskaleva, E., Angelova, G., Tcharaktchiev, D., & Dimitrova, N. (2009). Extraction and exploration of correlations in patient status data. In *Proceedings of the Workshop on Biomedical Information Extraction*, (pp. 1–7). Borovets, Bulgaria: Association for Computational Linguistics.

Boytcheva, S., Nikolova, I., Paskaleva, E., Angelova, G., Tcharaktchiev, D., & Dimitrova, N. (2010). Structuring of Status Descriptions in Hospital Patient Records. In *Proceedings of the 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2010)*. Malta.

Campbell, D., & Johnson, S. B. (2001). Comparing Syntactic Complexity in Medical and non-Medical Corpora. In *Proceedings of the AMIA Annual Symposium*, (pp. 90–95).

Cartoni, B., & Zweigenbaum, P. (2010). Semi-Automated Extension of a Specialized Medical Lexicon for French. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta.

Chapman, B. E., Lee, S., Kang, H. P., & Chapman, W. W. (2011). Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. *Journal of Biomedical Informatics*, *44*, 728–737.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform.*, *34*, 301–310.

Chapman, W. W., & Dowling, J. N. (2006). Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports. *Journal of Biomedical Informatics*, *39*(2), 196–208.

Christopher, M. M., & Hotz, C. S. (2004). Cytologic diagnosis: expression of probability by clinical pathologists. *Veterinary Clinical Pathology*, *33*(2), 84–95.

Clark, D. A. (1990). Verbal Uncertainty Expressions: A Critical Review of Two Decades of Research. *Current Psychology: Research & Reviews*, *9*(3), 203–235.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Collier, N., Park, H. S., Ogata, N., Tateishi, Y., Nobata, C., Ohta, T., Sekimizu, T., Imai, H., Ibushi, K., & Tsujii, J. (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)*, (pp. 271–272). Bergen, Norway.

Craggs, R., & Wood, M. M. (2005). Evaluating Discourse and Dialogue Coding Schemes. *Computational Linguistics*, *31*(3), 289–296.

Dalianis, H., & Åström, E. (2001). SweNam – a Swedish Named Entity Recognizer, Its Construction, Training and Evaluation. Tech. Rep. TRITA-NA-P0113, IPLab-NADA, KTH.

Dalianis, H., Hassel, M., & Velupillai, S. (2009). The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proceedings of ISHIMR 2009, Evaluation and implementation of e-health and health information initiatives: international perspectives. 14th International Symposium for Health Information Management Research*. Kalmar, Sweden.

Dalianis, H., & Velupillai, S. (2010). De-identifying Swedish Clinical Text - Refinement of a Gold Standard and Experiments with Conditional Random Fields. *Journal of Biomedical Semantics*, *1*(6).

de Bruijn, B., Cherry, C., Kiritchenko, S., Martin, J., & Zhu, X. (2011). Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, *18*, 557–562.

Di Eugenio, B. (2000). On the Usage of Kappa to Evaluate Agreement on Coding Tasks. In *In Proceedings of the Second International Conference on Language Resources and Evaluation*, (pp. 441–444). Athens, Greece.

Di Eugenio, B., & Glass, M. (2004). The kappa statistic: A second look. *Comp. Ling.*, *30*(1), 95–101.

Ejerhed, E., Källgren, G., & Brodda, B. (2006). Stockholm Umeå Corpus Version 2.0, SUC 2.0.

Farkas, R., Vincze, V., Móra, G., Csirik, J., & Szarvas, G. (2010). The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, (pp. 1–12). Uppsala, Sweden: Association for Computational Linguistics.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook – Advanced Approaches in Analyzing Unstructured Data*. New York, USA: Cambridge University Press.

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

Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc.*, *1*(2), 161–174.

Friedman, C., Hripcsak, G., DuMouchel, W., Johnson, S. B., & Clayton, P. D. (1995). Natural language processing in an operational clinical information system. *Natural Language Engineering*, *1*(1), 83–108.

Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated Encoding of Clinical Documents Based on Natural Language Processing. *Journal of the American Medical Informatics Association*, *11*(5), 392–402.

Goldkuhl, G. (2004). Meanings of Pragmatism: Ways to conduct information systems research. In *Proceedings of the 2nd International Conference on Action in Language, Organisations and Information Systems (ALOIS-2004)*.

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

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., J. Padó, S., Štěpánek, Straňák, P., Surdeanu, M., Xue, N., & Zhang, Y. (2009). The conll-2009 shared task: syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, (pp. 1–18). Stroudsburg, PA, USA: Association for Computational Linguistics.

Harkema, H., Dowling, J. N., Thornblade, T., & Chapman, W. W. (2009). ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, *42*, 839–851.

Haverinen, K., Ginter, F., Laippala, V., & Salakoski, T. (2009). Parsing Clinical Finnish: Experiments with Rule-Based and Statistical Dependency Parsers. In *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*. Odense, Denmark.

Hevner, A., & Chatterjee, S. (2010). *Integrated Series in Information Systems 22*, chap. Design Science Research in Information Systems. Springer-Verlag New York, Inc.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105.

Hewson, M., Kindy, P., Kirk, J. V., Gennis, V., & Day, R. (1996). Strategies for managing uncertainty and complexity. *Journal of General Internal Medicine*, *11*, 481–485.

Hobby, J. L., Tom, B. D. M., Todd, C., Bearcroft, P. W. P., & Dixon, A. K. (2000). Communication of doubt and certainty in radiological reports. *The British Journal of Radiology*, *73*, 999–1001.

Hripcsak, G., & Rothschild, A. S. (2005). Technical brief: Agreement, the f-measure, and reliability in information retrieval. *JAMIA*, *12*(3), 296–298.

Hyland, K. (1998). *Hedging in scientific research articles*. Philadelphia: Benjamins.

i2b2 (2012). Informatics for Integrating Biology & the Bedside. Partners Healthcare. Available at: https://www.i2b2.org/. Accessed on March 21, 2012.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher*, *33*(7), 14–26.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.*. Pearson Education International – Prentice-Hall.

Kaplan, B., & Duchon, D. (1988). Combining qualitative and quantitative methods information systems research: a case study. *Manage. Inf. Syst. Q.*, *12*, 571–586.

Khorasani, R., Bates, D. W., Teeger, S., Rotschild, J. M., Adams, D. F., & Seltzer, S. E. (2003). Is terminology used effectively to convey diagnostic certainty in radiology reports? *Academic Radiology*, *10*, 685–688.

Kilicoglu, H., & Bergler, S. (2008). Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, *9*(S-11).

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

Kundel, H. L., & Polansky, M. (2003). Measurement of Observer Agreement. *Radiology*, *208*(2), 303–308.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, (pp. 282–289).

Lakoff, G. (1973). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, *2*, 458–508.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174.

Lester, H., & Tritter, J. Q. (2001). Medical error: a discussion of the medical construction of error and suggestions for reforms of medical education to decrease error. *Medical Education*, *35*, 855–861.

Light, M., Qiu, X. Y., & Srinivasan, P. (2004). The language of bioscience: Facts, speculations, and statements in between. In L. Hirschman, & J. Pustejovsky (Eds.) *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*, (pp. 17–24). Boston, Massachusetts, USA: Association for Computational Linguistics.

Lingard, L., Garwood, K., Schryer, C. F., & Spafford, M. M. (2003). A certain art of uncertainty: case presentation and the development of professional identity. *Social science medicine*, *56*(3), 603–616.

Lovis, C., Baud, R. H., & Planche, P. (2000). Power of expression in the electronic patient record: structured data or narrative text? *International Journal of Medical Informatics*, *58-59*, 101–110.

Mani, I., Hu, Z., Jang, S. B., Samuel, K., Krause, M., Phillips, J., & Wu, C. H. (2005). Protein name tagging guidelines: lessons learned. *Comp. Funct. Genomics*, *6*(1-2), 72–76.

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

Medlock, B., & Briscoe, T. (2007). Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, (pp. 992–999). Prague, Czech Republic: Association for Computational Linguistics.

Mendonca, E. A., Haas, J., Shagina, L., Larson, E., & Friedman, C. (2005). Extracting Information on Pneumonia in Infants Using Natural Language Processing of Radiology Reports. *Journal of Biomedical Informatics*, *38*(4), 314–321.

Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., & Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, *10*(1), 70.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. E. (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*.

Mingers, J. (2001). Combining IS Research Methods: Towards a Pluralist Methodology. *Information Systems Research*, *12*(3), 240–259.

Morante, R., Asch, V. V., & Daelemans, W. (2010). Memory-based resolution of in-sentence scopes of hedge cues. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, (pp. 40–47). Uppsala, Sweden: Association for Computational Linguistics.

Morante, R., & Daelemans, W. (2009). Learning the scope of hedge cues in biomedical texts. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, (pp. 28–36). Morristown, NJ, USA: Association for Computational Linguistics.

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

Nilsson, I. (2007). *Medicinsk dokumentation genom tiderna*. Enheten för medicinens historia, Lunds universitet. In Swedish.

Nuyts, J. (2001). *Epistemic modality, language, and conceptualization: a cognitive-pragmatic perspective*. Human cognitive processing. J. Benjamins.

Ogren, P. V. (2006). Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, (pp. 273–275). Morristown, NJ, USA: Association for Computational Linguistics.

OHNLP (2012). Open Health Natural Language Processing Consortium. Available at: https://wiki.nci.nih.gov/display/VKC/
Open+Health+Natural+Language+Processing+%28OHNLP%29+Consortium. Accessed on March 21, 2012.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, *31*(1), 71–106.

Petersson, G., & Rydmark, M. (Eds.) (1996). *Medicinsk informatik*. Almqvist & Wiksell Medicin, Liber Utbildning. In Swedish.

Proux, D., Marchal, P., Segond, F., Kergourlay, I., Darmoni, S., Pereira, S., Gicquel, Q., & Metzger, M. H. (2009). Natural language processing to detect risk patterns related to hospital acquired infections. In *Proceedings of the Workshop on Biomedical Information Extraction*, (pp. 35–41). Borovets, Bulgaria: Association for Computational Linguistics.

Renooij, S., & Witteman, C. (1999). Talking probabilities: communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, *22*, 169–194.

Rubin, V. L., Liddy, E. D., & Kando, N. (2006). Certainty identification in texts: Categorization model and manual tagging results. In *Computing Affect and Attitutde in Text: Theory and Applications*. Springer.

Ruch, P., Baud, R., & Geissbühler, A. (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.

Sahlgren, M. (2006). *The Word-Space Model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University.

Saurí, R. (2008). *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Brandeis University.

Saurí, R., & Pustejovsky, J. (2009). FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, *43*(3), 227–268–268.

Savova, G., Masanz, J., Ogren, P., Zheng, J., Sohn, S., Kipper-Schuler, K., & Chute, C. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, *17*(5), 507–513.

Strauss, A. L., & Corbin, J. (1990). *Basics of qualitative research: grounded theory procedures and techniques*. Sage.

Suominen, H. (2009). *Machine Learning and Clinical Text –Supporting Health Information Flow*. Ph.D. thesis, University of Turku, Turku Centre for Computer Science (TUCS), Department of Information Technology.

Szarvas, G. (2008). Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of ACL-08: HLT*, (pp. 281–289). Columbus, Ohio: Association for Computational Linguistics.

Szarvas, G., Farkas, R., & Busa-Fekete, R. (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.

Tang, B., Wang, X., Wang, X., Yuan, B., & Fan, S. (2010). A cascade method for detecting hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, (pp. 13–17). Uppsala, Sweden: Association for Computational Linguistics.

Tange, H. (1996). How to approach the structuring of the medical record? Towards a model for flexible access to free text medical data. *International Journal of BioMedical Computing*, *42*(1-2), 27–34.

Tange, H. J., Hasman, A., Robbé, P. F. D. V., & Schouten, H. C. (1997). Medical narratives in electronic medical records. *International Journal of Medical Informatics*, *46*(1), 7–29.

Timmermans, D. (1994). The Roles of Experience and Domain of Expertise in Using Numerical and Verbal Probability Terms in Medical Decisions. *Medical Decision Making*, *14*, 146–156.

Uzuner, Ö. (2009). Recognizing Obesity and Comorbidities in Sparse Data. *Journal of American Medical Informatics Association*, *16*, 561–570.

Uzuner, Ö., Luo, Y., & Szolovits, P. (2007). Evaluating the State-of-the-Art in Automatic De-identification. *Journal of American Medical Informatics Association*, *14*, 550–563.

Uzuner, Ö., Sibanda, T. C., Luo, Y., & Szolovits, P. (2008). A De-identifier for Medical Discharge Summaries. *Artificial Intelligence in Medicine*, *42*(1), 13–35.

Uzuner, Ö., Solti, I., & Cadag, E. (2010). Extracting medication information from clinical text. *Journal of American Medical Informatics Association*, *17*, 514–518.

Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *JAMIA*, *18*(5), 552–556.

Uzuner, Ö., Zhang, X., & Sibanda, T. (2009). Machine Learning and Rule-based Approaches to Assertion Classification. *Journal of the American Medical Informatics Association*, *16*(1), 109–115.

Vincze, V., Szarvas, G., Farkas, R., Móra, G., & Csirik, J. (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, *9*(S-11).

Wayne, G. R. (2010). Design science research and the grounded theory method: Characteristics, differences, and complementary uses. In *Proceedings of the 18th European Conference on Information Systems*.

Wiebe, J., Bruce, R., Bell, M., Martin, M., & Wilson, T. (2001). A corpus study of evaluative and speculative language. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16*, SIGDIAL '01, (pp. 1–10). Stroudsburg, PA, USA: Association for Computational Linguistics.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, *39*, 165–210.

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

Wu, A. S., Do, B. H., Kim, J., & Rubin, D. L. (2009). Evaluation of negation and uncertainty detection and its impact on precision and recall in search. *J Digit Imaging*.

CHAPTER 6

INCLUDED PAPERS

PAPER I

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.* International journal of medical informatics 78:12 (2009) 19–26.

Author contributions I was responsible for developing the annotation guidelines and preparing the annotation work as well as extracting a subset for the annotation work and analyzing the results on the gold standard. Both Hercules Dalianis, Gunnar Nilsson and I was involved in the annotation work. Martin Hassel was responsible for creating external resources for the automatic classifier. The article was written jointly.

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, Karolinska Institutet, 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

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

(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

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 manual evaluation on a small subset of the gold corpus and found

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

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

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].

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.
- 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.

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.

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.

PAPER II

HOW CERTAIN ARE CLINICAL ASSESSMENTS? ANNOTATING SWEDISH CLINICAL TEXT FOR (UN)CERTAINTIES, SPECULATIONS AND NEGATIONS

Hercules Dalianis and Sumithra Velupillai. 2010. *How Certain are Clinical Assessments? Annotating Swedish Clinical Text for (Un)certainties, Speculations and Negations.* In Proceedings of LREC '10 – 7th International Conference on Language Resources and Evaluation, Valletta, Malta, May 19–21, 2010.

Author contributions I was responsible for extracting a subset for the annotation work and analyzing the results on the gold standard. Both Hercules Dalianis and I were involved in developing the annotation guidelines and preparing the annotation work. Hercules Dalianis was responsible for recruiting annotators. The article was written jointly.

How Certain are Clinical Assessments? Annotating Swedish Clinical Text for (Un)certainties, Speculations and Negations

Hercules Dalianis, Sumithra Velupillai
Department of Computer and Systems Sciences (DSV)
Stockholm University
Forum 100, 164 40 Kista, Sweden
E-mail: {hercules, sumithra}@dsv.su.se

Abstract

Clinical texts contain a large amount of information. Some of this information is embedded in contexts where e.g. a patient status is reasoned about, which may lead to a considerable amount of statements that indicate uncertainty and speculation. We believe that distinguishing such instances from factual statements will be very beneficial for automatic information extraction.

We have annotated a subset of the Stockholm Electronic Patient Record Corpus for certain and uncertain expressions as well as speculative and negation keywords, with the purpose of creating a resource for the development of automatic detection of speculative language in Swedish clinical text. We have analyzed the results from the initial annotation trial by means of pairwise Inter-Annotator Agreement (IAA) measured with F-score. Our main findings are that IAA results for certain expressions and negations are very high, but for uncertain expressions and speculative keywords results are less encouraging. These instances need to be defined in more detail. With this annotation trial, we have created an important resource that can be used to further analyze the properties of speculative language in Swedish clinical text. Our intention is to release this subset to other research groups in the future after removing identifiable information.

1. Introduction

The use of electronic patient records (EPRs) is increasing in the healthcare sector, which leads to a growing amount of digitalized data. Automatic methods for accessing information from such data is an important research area.

The Stockholm Electronic Patient Corpus (Stockholm EPR Corpus) is a clinical corpus containing over one million patient records, encompassing 2 000 clinics from the Stockholm area stretching over the years 2006 to 2008 (Dalianis et al., 2009). The Stockholm EPR Corpus contains both structured information and unstructured information (free text). The free text entries are semi-structured, since the free text is entered under several free text categories, for example *Bedömning (Assessment), Aktuell status (Current status), Social Bakgrund (Social Background)*.

In EPRs, the patient status is described and reasoned about. We believe that this leads to a considerable amount of statements that indicate uncertainty and speculation, where clinicians describe situations that are difficult to confirm. Distinguishing such instances from factual, or certain, instances is important if the information is to be extracted automatically, since the former alters the meaning of the expression. In the long run, systems for Information Extraction, Information Retrieval or Knowledge Discovery may be improved by including such distinctions, where, for instance, a clinician would benefit from accessing information about previous, similar cases when faced with a difficult situation.

We have annotated a subset of the Stockholm EPR corpus for certain and uncertain expressions as well as speculative and negation keywords, with the purpose of creating a resource for the development of automatic detection of speculative language in Swedish clinical text.

Our aim is to analyze the results from the initial annotation trial by means of pairwise Inter-Annotator Agreement (IAA) measured with F-score. Our intention is to release this subset to other research groups in the future, after ensuring that no identifiable information is included in the subset.

2. Previous research

Research on the identification of speculative language, or "hedging", has gained a large amount of interest lately, especially for scientific articles and abstracts in the biomedical domain. Research findings often contain tentative results, where further analysis might be needed. Distinguishing such findings from factual statements is crucial for information extraction systems. Several research groups have analyzed the characteristics of speculative language in biomedical scientific writings.

Light et al. (2004) found 11 percent speculative language in Medline abstracts from scientific articles in Biomedicine. Here, four annotators annotated 891 sentences each as either highly speculative, low speculative, or definite. Their Inter-Annotator Agreement (IAA) results, measured with kappa, ranged between 0.54 and 0.68. They also found that the majority of the speculative sentences appeared towards the end of the abstract. Finally they also annotated a larger set of sentences (the last two sentences in all annotated data sets, (i.e. the last two sentences in the abstracts)) containing in total 2 093 sentences and found 18 percent speculative sentences and 82 percent definite sentences.

In the BioScope corpus (Vincze et al., 2008), both medical (clinical) free texts, biological full papers and biological scientific abstracts have been annotated,

This research has been carried out after approval from the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2007/1625-31/5.

encompassing more than 20 000 sentences, where over 10 percent of the sentences were either speculative or negated. We are specifically interested in the results for the clinical sub-corpus where 13.55 percent of the clinical texts contained negation and 13.99 percent contained speculative keywords However, the authors do not report any results of whether negation and speculation keywords co-occur. In Kilicoglu & Bergler (2008), non-lexical features for identifying speculative language are used (as well as lexical cues). Some of these are defined as negated non-speculative ("unhedging") cues, such as *no evident*. They report promising results on the automatic identification of speculative language in biomedical research articles.

The IAA results in the clinical sub-corpus of BioScope for negation keywords ranged between 91 and 96 percent F-score, and for speculative keywords the results ranged between 84 and 92 percent F-score. These results indicate that negation keywords seem to be easier to identify than speculation keywords. In Light et al. (2004), methods for automatic identification of speculative language by using annotated corpora have also been developed. Using Support Vector Machines (SVM), and evaluating with 10 fold cross evaluation, 84 percent precision and 39 percent recall was obtained.

Özgür & Radev (2009) used two parts of the annotated BioScope corpus, namely the biological full papers and biological scientific abstracts (9 full text papers and 1 273 abstracts) for automatic identification of speculative language. They also used SVM for two classification tasks: identifying keywords used in speculative context, and determining the scope of these keywords. For scientific abstracts they obtained 88.16 percent recall, 95.21 percent precision and 91.50 F-score. They also found that speculative keywords co-occur and that they are more common in the *Conclusion* and *Discussion* parts of the articles.

Morante & Daelemans (2009) describe work on the same two classification tasks on all three BioScope subcorpora. Here, different machine learning methods are used for the different tasks, including SVM, Memory-based learning and Conditional Random Fields (CRF). Overall, the results for abstracts and papers are considerably higher than for clinical text for the first classification task, which influences results on the second classification task. These results show that differences in text type are important to consider.

3. Method

We have annotated 6 740 randomly extracted sentences from the Stockholm EPR corpus, from the free text category *Bedömning* (Assessment). Three annotators with no prior knowledge of the content worked on the task; one senior level student (SLS), one undergraduate computer scientist (UCS), and one undergraduate language consultant (ULC).

In order to make the corpus comparable, we developed guidelines similar to those for the BioScope corpus (Vincze et al., 2008). However, in the BioScope corpus, certain expressions, as well as expressions containing question marks (*?*), were not annotated. The following annotation classes were used in the work presented here: *Certain_expression, Uncertain_expression, Negation, Speculative_words, Undefined_expression* and *Undefined_speculative_words*.

For each randomly extracted sentence, the full free text entry was shown to the annotators, in order for them to see the context of the sentence. (See Examples 1 and 2 below). Sentences were extracted using a simple tokenizing strategy based on regular expressions. Each sentence had to be judged either as a certain, uncertain or undefined expression. In cases where a sentence contained both, for instance through subordinate clauses, a sentence could be broken into sub-expressions. Within these expressions, negated or speculative keywords were annotated if present. By doing this, both sentence level and token level annotations were captured. We did not, however, in this annotation trial, include annotations for the scope of a token level speculative or negated keyword, i.e. those syntactic units that are modified by the keyword In even intervals (in total seven), during the three working weeks, the group of annotators met to discuss the task. This was carried out in order to measure IAA results over time and after resolving problems, similar to Haverinen et al. (2009).

4. Results

We have measured IAA by pair wise F-score, treating one set of annotations as the gold standard for each combination of annotator pairs. As a final result, we give the average result. We have measured both partial and exact matching. Exact matching is at token level while partial matching is at character level. In Tables 1, 2, 3, 4 and 5, results are shown.

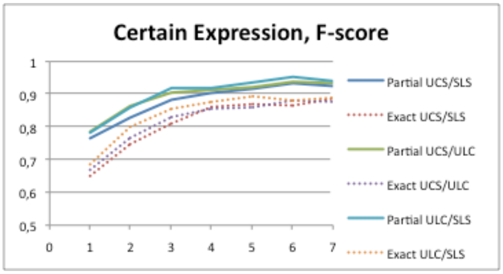


Table 1. Results for Certain Expression, pairwise IAA (F-score) over time both for partial and exact matching.

Looking at sentence level, the IAA results for *Certain_expression* were in general very high (0.84 F-score for exact matches) while considerably lower for *Uncertain_expression*, see Tables 1 and 2.

Having discussions among the annotators in time intervals yields an improvement for results on

Certain_expression, we also see a convergence between partial and exact matching, see Table 1. For *Uncertain_expression*, results over time are very disparate. However, we also see a tendency for convergence between partial and exact matching, specifically between annotation intervals 5 to 7, which is probably due to the discussions among the annotators, see Table 2.

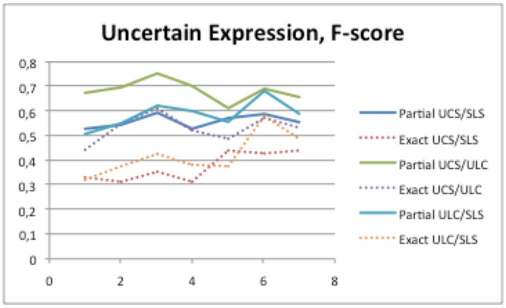


Table 2. Results for Uncertain Expression, pairwise IAA (F-score) over time both for partial and exact matching.

The annotation class *Speculative_words* obtains low IAA results in general, see Table 3. However, we also see a tendency for convergence between partial and exact matching and discrepancies between exact and partial matches are lower here, which shows that larger scopes for annotating uncertain expressions, see Table 2, are more difficult to define.

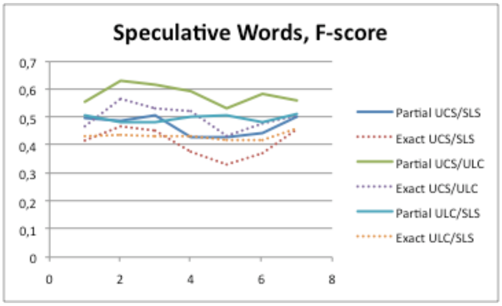


Table 3. Results for Speculative Words, pairwise IAA (F-score) over time both for partial and exact matching.

Negation keywords obtained very high IAA results (over 0.80 F-score), for these there was also an improvement in results over time. The total average results, see Table 4, show an overall improvement over time.

These results are, however, heavily influenced by the dominance of the annotation class *Certain_expression*.

Looking at the actual contents of the annotations, from the 6 740 sentences, we find an average total amount of 6 996 annotated expressions. On average, 13.5 percent of these are annotated as uncertain expressions (ranging between 11.8 and 15.7 percent). The average amount of annotated speculative words was 1 624 and the average amount of negation keywords was 1 008.

Looking at the token level annotations *Speculative_words* and *Negations*, the average amount of unique keywords was 538 and 13, respectively. The most common speculative keywords for all three annotators were unigrams such as *sannolikt* (*likely*) and *möjligen* (*possibly*).

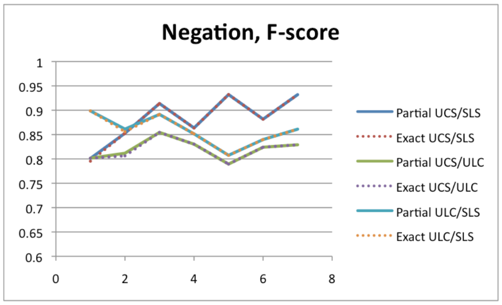


Table 4. Results for Negation Words, pairwise IAA (F-score) over time both for partial and exact matching.

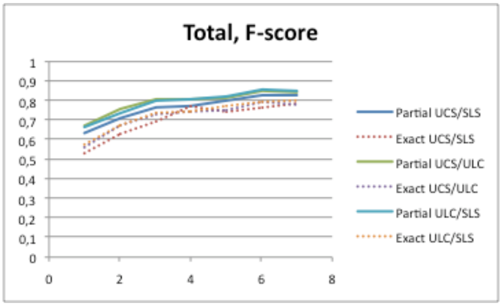


Table 5. Total average results, pairwise IAA (F-score) over time both for partial and exact matching.

However, 52 percent (on average) of the speculative keywords were unigrams, the rest being *n*-grams of varying length. Moreover, several annotations of speculative keywords included negations, such as *ingen misstanke* (*no suspicion*) and *inga tydliga tecken* (*no clear signs*). Many of these conform well to those listed as indicative features of speculative language in Kilicoglu & Bergler (2008), where negated “unhedgers” form speculative cues. Notable is also that the negation keywords only included evident negation words such as *inte* (*not*) and *inga* (*none*). In Swedish, it is also possible to negate words with prefixes such as *o*-, as in *oklar* (*indistinct*). No such words were annotated as negation keywords by the annotators.

When looking at the sentences and the contexts in which they were annotated, there was a great variety in how large the context was, how long the sentences were, and in what setting they were written. In Example 1 we see a sentence that is annotated as an *uncertain expression*. This sentence contains a negation combined with a non-speculative keyword, which together form a multi-word annotation (*Speculative_words*), making the whole expression uncertain. In this example, we also see that the simple sentence tokenization created an annotation instance that had to be broken down into two sub-expressions.

```
Bedömning:
<sentence_1>
<Uncertain_expression>Statusmässigt
<Speculative_words><Negation>inga
</Negation> säkra</Speculative_words>
artriter</Uncertain_expression>.
<Certain_expression>Lungrtg Huddinge ua
</Certain_expression>.</sentence>
Leverprover ua.
```

Translation to English:

```
Assessment:
<sentence_1>
<Uncertain_expression>Status-wise
Speculative_words><Negation>no
</Negation> certain</Speculative_words>
arthritis</Uncertain_expression>.
<Certain_expression>cxr Huddinge woco
</Certain_expression>.</sentence>
Liver samples woco
```

Example 1. An annotated sentence containing a negation and a certain expression making the whole expression uncertain.

Example 2 shows an annotated sentence within a context that contains a relatively large amount of reasoning, concerning several issues regarding the patient status, giving a more thorough account of the level of certainty (please cf. with the reasoning processes in Groopman (2007)).

```
Bedömning:
<sentence_2><Uncertain_expression>
Har lite <Speculative_words> undringar
</Speculative_words> om brakyterapi
<Speculative_words> kunde vara
</Speculative_words> aktuellt i hans fall
</Uncertain_expression>.</sentence> Har
haft den diskussionen uppe med Bengt
Karlsson. Jag har svårt och tro det eftersom
han går på Onkologen och rimligtvis hade man
tänkt på den behandlingen om man hade ansett
att det finns möjlighet men jag lovar att
skriva ett brev till Lars Olof Svensson om
detta. Vad det gäller pricken på mandibeln
verkar det mest som ett lite aterom tycker
jag men det är klart att hudmetastas är ju
inte uteslutet. Jag lämnar dock den frågan
helt till Onkologen.
```

Translation to English:

```
Assessment:
<sentence_2><Uncertain_expression>
I have some <Speculative_words> concerns
</Speculative_words> about whether
brachytherapy <Speculative_words> could be
</Speculative_words> considered in his case
</Uncertain_expression>.</sentence> I have
had that discussion with Bengt Karlsson. I
have difficulties believing this since he is
treated at the Oncology clinic and they must
have considered this treatment if they
thought this was possible, but I promise to
write a letter to Lars Olof Svensson
regarding this. Regarding the mark on the
mandible, I think it mostly seems to be a bit
of aterom but of course a Cutaneous
metastasis can not be excluded. However, I
leave that question entirely to the Oncology
clinic.
```

Example 2. An annotated sentence within a context that contains a relatively large amount of reasoning.

Regarding sensitive information that can reveal the identity of a patient, the annotators identified in total 15 personal names (from the total amount of 290 085 tokens). One half consisted of personal names of clinical personnel, and the other half consisted of patient first personal names. Moreover, seven social security numbers were found. This indicates that personal names are extremely rare in the Assessment field (0.02 per thousand). In the Stockholm EPR PHI Corpus (another subset of the Stockholm EPR Corpus), consisting of 380 000 tokens (containing all the free text entry fields), 0.19 per thousand patient first personal names were found. However, in this corpus, no social security numbers were identified (Dalianis & Velupillai 2010). Although identifiable information seems to be very infrequent, it is crucial to ensure that no identifiable information about an individual is kept if a corpus is to be released for further research.

5. Discussion

We have presented initial results on an annotation trial for speculative language in Swedish clinical texts. Our main findings are that IAA results for certain expressions and negations are very high, but for uncertain expressions and speculative words results are less encouraging. These instances need to be defined in more detail.

Our results are comparable to those presented in Light et al. (2004). However, our annotations of certainties and negations obtain high IAA results and the training effect is significant. Our IAA results are lower than Vincze et al. (2008), but this may be due to differences in corpora. In the clinical sub-corpus presented in Vincze et al. (2008), radiology reports are annotated, while the annotations presented here were randomly extracted from all clinics in the Stockholm EPR corpus. Moreover, the sentences extracted for this annotation trial were extracted from the free-text entries under the heading *Bedömning* (*Assessment*). This heading may be used differently in different health care units, and may hence contain diverse types of statements. The sentences could, for instance, contain descriptions of the current, overall status of a patient or a short-term plan for medication. It was also evident that expressions of speculations may differ greatly between clinical disciplines.

During discussions among the annotators, some specific properties were pointed out. One was the question of perspective; the patient's and the physician's, especially for uncertain expressions. After annotation interval 2 it was decided to only annotate the physician's perspective. Another point was the level of (un)certainty; many expressions were more or less (un)certain. A grading of four scales was proposed: *Completely certain*, *Quite certain*, *Quite uncertain* and *Completely uncertain*. Such a distinction would probably have a great effect on the sentences currently annotated as *Certain_expression*, which, in the current set, in the majority of cases, merely indicate that the sentence is *not* uncertain. Furthermore, vagueness was often difficult to distinguish from uncertainty.

5. Conclusions and future work

The research presented here is to our knowledge the first work carried out on annotating speculations in clinical text written in Swedish. It is also the first time that both certain and uncertain expressions have been explicitly annotated.

Although IAA results for speculative words and uncertain expressions were low, we believe that the identification of such language is important for future Information Access research. However, further definitions are needed. In particular, the distinction between different perspectives in uncertain expressions is important and needs to be handled. This distinction is probably a specific property of EPRs and probably not present to the same extent in scientific text.

Moreover, looking at different health care disciplines, there may be great differences in how uncertainties and speculations are expressed. This is particularly interesting when looking at specific diagnoses, e.g. speculations about certain diagnoses such as brain tumors are probably very rare, while speculations about for instance psychiatric diagnoses may be much more common. We will analyze the annotated set by dividing it into different health care units, in order to analyze whether such differences are apparent.

We plan to analyze the annotations further, by looking in more detail at the speculative words, investigating their characteristics, analyzing the multi-word expressions, finding out to what extent they are combined with negations and what implications this has, as well as analyzing in which part of the text the uncertain expressions are present. When it comes to negation keywords, we plan to analyze them in particular for finding which constructions where they, combined with non-speculative keywords, form a speculative expression. We will also analyze the scopes of the annotated keywords, in order to identify what expressions they modify. Moreover, we plan to create a consensus corpus from the annotated set presented here, to use for training and testing a machine learning system on our annotations, to investigate the possibilities of automatic classification. For such a system, we will look at syntactic patterns as well as word-level features.

6. Acknowledgements

We would like to thank our three annotators, Helen Allvin, Freja Dalianis and Aron Henriksson for their endurance and accuracy during their work and also for the interesting and fruitful discussions during the annotation sessions

7. References

Dalianis, H., M. Hassel and S. Velupillai. 2009. The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proceedings of, the 14th International Symposium for Health Information Management Research (ISHIMR 2009),* Kalmar, Sweden, 14-16 October, 2009, pp 243-249. Awarded Best Paper.

Dalianis, H. and S. Velupillai. 2010. De-identifying Swedish Clinical Text - Refinement of a Gold Standard and Experiments with Conditional Random Fields. To be published in Journal of Biomedical Semantics..

Groopman, J. 2007. *How doctors think*, Houghton Mifflin Company, New York.

Kilicoglu, H. and S. Bergler 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 9(S-11).

Haverinen, K., F. Ginter, V. Laippala and T. Salakoski. 2009. Parsing Clinical Finnish: Experiments with Rule-Based and Statistical Dependency Parsers. In *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*.

Light, M., X. Y. Qiu, and P. Srinivasan. 2004. The Language of Bioscience: Facts, Speculations, and Statements in Between. In *BioLINK 2004: Linking Biological Literature, Association for Computational Linguistics Ontologies, and Databases*, pp. 17-24.

Morante, R. and Daelemans, W. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on BioNLP*, pp 28-36, Boulder, Colorado, June 2009.

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.

PAPER III

TOWARDS A BETTER UNDERSTANDING OF UNCERTAINTIES AND SPECULATIONS IN SWEDISH CLINICAL TEXT – ANALYSIS OF AN INITIAL ANNOTATION TRIAL

Sumithra Velupillai. 2010. *Towards A Better Understanding of Uncertainties and Speculations in Swedish Clinical Text – Analysis of an Initial Annotation Trial* Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, Uppsala, Sweden, July 10, 2010. ACL.

Towards A Better Understanding of Uncertainties and Speculations in Swedish Clinical Text – Analysis of an Initial Annotation Trial

Sumithra Velupillai
Department of Computer and Systems Sciences (DSV)
Stockholm University
Forum 100
SE-164 40 Kista, Sweden
sumithra@dsv.su.se

Abstract

Electronic Health Records (EHRs) contain a large amount of free text documentation which is potentially very useful for Information Retrieval and Text Mining applications. We have, in an initial annotation trial, annotated 6 739 sentences randomly extracted from a corpus of Swedish EHRs for sentence level (un)certainty, and token level speculative keywords and negations. This set is split into different clinical practices and analyzed by means of descriptive statistics and pairwise Inter-Annotator Agreement (IAA) measured by F_1-score. We identify *geriatrics* as a clinical practice with a low average amount of uncertain sentences and a high average IAA, and *neurology* with a high average amount of uncertain sentences. Speculative words are often n-grams, and uncertain sentences longer than average. The results of this analysis is to be used in the creation of a new annotated corpus where we will refine and further develop the initial annotation guidelines and introduce more levels of dimensionality. Once we have finalized our guidelines and refined the annotations we plan to release the corpus for further research, after ensuring that no identifiable information is included.

1 Introduction

Electronic Health Records (EHRs) contain a large amount of free text documentation which is potentially very useful for Information Retrieval and Text Mining applications. Clinical documentation is specific in many ways; there are many authors in a document (e.g. physicians, nurses), there are different situations that are documented (e.g. admission, current status). Moreover, they may often be written under time pressure, resulting in fragmented, brief texts often containing spelling errors and abbreviations. With access to EHR data, many possibilities to exploit documented clinical knowledge and experience arise.

One of the properties of EHRs is that they contain reasoning about the status and diagnoses of patients. Gathering such information for the use in e.g. medical research in order to find relationships between diagnoses, treatments etc. has great potential. However, in many situations, clinicians might describe uncertain or negated findings, which is crucial to distinguish from positive or asserted findings. Potential future applications include search engines where medical researchers can search for particular diseases where negated or speculative contexts are separated from asserted contexts, or text mining systems where e.g. diseases that seem to occur often in speculative contexts are presented to the user, indicating that more research is needed. Moreover, laymen may also benefit from information retrieval systems that distinguish diseases or symptoms that are more or less certain given current medical expertise and knowledge.

We have, in an initial annotation trial, annotated 6 739 sentences randomly extracted from a corpus of Swedish EHRs for sentence level (un)certainty, and token level speculative keywords and negations[1]. In this paper, a deeper analysis of the resulting annotations is performed. The aims are to analyze the results *split into different clinical practices* by means of descriptive statistics and pairwise Inter-Annotator Agreement (IAA) measured by F_1-score, with the goal of identifying a) whether specific clinical practices contain higher or lower amounts of uncertain expressions, b)

[1] This research has been carried out after approval from the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2009/1742-31/5

whether specific clinical practices result in higher or lower IAA - indicating a less or more difficult clinical practice for judging uncertainties, and c) identifying the characteristics of the entities annotated as speculative words, are they highly lexical or is a deeper syntactic and/or semantic analysis required for modeling? From this analysis, we plan to conduct a new annotation trial where we will refine and further develop the annotation guidelines and use domain experts for annotations in order to be able to create a useful annotated corpus modeling uncertainties, negations and speculations in Swedish clinical text, which can be used to develop tools for the automatic identification of these phenomena in, for instance, Text Mining applications.

2 Related Research

In recent years, the interest for identifying and modeling speculative language in natural language text has grown. In particular, biomedical scientific articles and abstracts have been the object of several experiments. In Light et al. (2004), four annotators annotated 891 sentences each as either highly speculative, low speculative, or definite, in biomedical scientific abstracts extracted from Medline. In total, they found 11 percent speculative sentences, resulting in IAA results, measured with kappa, between 0.54 and 0.68. One of their main findings was that the majority of the speculative sentences appeared towards the end of the abstract.

Vincze et al. (2008) describe the creation of the BioScope corpus, where more than 20 000 sentences from both medical (clinical) free texts (radiology reports), biological full papers and biological scientific abstracts have been annotated with speculative and negation keywords along with their scope. Over 10 percent of the sentences were either speculative or negated. In the clinical sub-corpus, 14 percent contained speculative keywords. Three annotators annotated the corpus, and the guidelines were modified several times during the annotation process, in order to resolve problematic issues and refine definitions. The IAA results, measured with F_1-score, in the clinical sub-corpus for negation keywords ranged between 0.91 and 0.96, and for speculative keywords between 0.84 and 0.92. The BioScope corpus has been used to train and evaluate automatic classifiers (e.g. Özgür and Radev (2009) and Morante and Daelemans (2009)) with promising results.

Five qualitative dimensions for characterizing scientific sentences are defined in Wilbur et al. (2006), including levels of certainty. Here, guidelines are also developed over a long period of time (more than a year), testing and revising the guidelines consecutively. Their final IAA results, measured with F_1-score, range between 0.70 and 0.80. Different levels of dimensionality for categorizing certainty (in newspaper articles) is also presented in Rubin et al. (2006).

Expressions for communicating probabilities or levels of certainty in clinical care may be inherently difficult to judge. Eleven observers were asked to indicate the level of probability of a disease implied by eighteen expressions in the work presented by Hobby et al. (2000). They found that expressions indicating intermediate probabilities were much less consistently rated than those indicating very high or low probabilities. Similarly, Khorasani et al. (2003) performed a survey analyzing agreement between radiologists and non-radiologists regarding phrases used to convey degrees of certainty. In this study, they found little or no agreement among the survey participants regarding the diagnostic certainty associated with these phrases. Although we do not have access to radiology reports in our corpus, these findings indicate that it is not trivial to classify uncertain language in clinical documentation, even for domain experts.

3 Method

The annotation trial is based on sentences randomly extracted from a corpus of Swedish EHRs (see Dalianis and Velupillai (2010) for an initial description and analysis). These records contain both structured (e.g. measure values, gender information) and unstructured information (i.e. free text). Each free text entry is written under a specific heading, e.g. *Status*, *Current medication*, *Social Background*. For this corpus, sentences were extracted only from the free text entry *Assessment* (Bedömning), with the assumption that these entries contain a substantial amount of reasoning regarding a patient's diagnosis and situation. A simple sentence tokenizing strategy was employed, based on heuristic regular expressions[2]. We have used Knowtator (Ogren, 2006) for the annotation

[2] The performance of the sentence tokenizer has not been evaluated in this work.

work.

One senior level student (SLS), one undergraduate computer scientist (UCS), and one undergraduate language consultant (ULC) annotated the sentences into the following classes; on a sentence level: *certain*, *uncertain* or *undefined*, and on a token level: *speculative words*, *negations*, and *undefined words*.

The annotators are to be considered *naive* coders, as they had no prior knowledge of the task, nor any clinical background. The annotation guidelines were inspired by those created for the BioScope corpus (Vincze et al., 2008), with some modifications (see Dalianis and Velupillai (2010)). The annotators were allowed to break a sentence into subclauses if they found that a sentence contained conflicting levels of certainty, and they were allowed to mark question marks as speculative words. They did not annotate the linguistic scopes of each token level instance. The annotators worked independently, and met for discussions in even intervals (in total seven), in order to resolve problematic issues. No information about the clinic, patient gender, etc. was shown. The annotation trial is considered as a first step in further work of annotating Swedish clinical text for speculative language.

Clinical practice	# sentences	# tokens
hematology	140	1 494
surgery	295	3 269
neurology	351	4 098
geriatrics	142	1 568
orthopaedics	245	2 541
rheumatology	384	3 348
urology	120	1 393
cardiology	128	1 242
oncology	550	5 262
ENT	224	2 120
infection	107	1 228
emergency	717	6 755
paediatrics	935	8 926
total, clinical practice	4 338	43 244
total, full corpus	6 739	69 495

Table 1: Number of sentences and tokens per clinical practice (#sentences $>$ 100), and in total. ENT = Ear, Nose and Throat.

3.1 Annotations and clinical practices

The resulting corpus consists of 6 739 sentences, extracted from 485 unique clinics. In order to be able to analyze possible similarities and differences across clinical practices, sentences from clinics belonging to a specific practice type were grouped together. In Table 1, the resulting groups, along with the total amount of sentences and tokens, are presented[3]. Only groups with a total amount of sentences $>$ 100 were used in the analysis, resulting in 13 groups. A clinic was included in a clinical practice group based on a priority heuristics, e.g. the clinic ”Barnakuten-kir” (*Paediatric emergency surgery*) was grouped into paediatrics.

The average length (in tokens) per clinical practice and in total are given in Table 2. Clinical documentation is often very brief and fragmented, for most clinical practices (except urology and cardiology) the minimum sentence length (in tokens) was one, e.g. ”basal”, ”terapisvikt” (*therapy failure*), ”lymfödem” (*lymphedema*), ”viros” (*virosis*), ”opanmäles” (*reported to surgery*, compound with abbreviation). We see that the average sentence length is around ten for all practices, where the shortest are found in rheumatology and the longest in infection.

As the annotators were allowed to break up sentences into subclauses, but not required to, this led to a considerable difference in the total amount of annotations per annotator. In order to be able to analyze similarities and differences between the resulting annotations, all sentence level annotations were converted into *one* sentence class only, the primary class (defined as the first sentence level annotation class, i.e. if a sentence was broken into two clauses by an annotator, the first being *certain* and the second being *uncertain*, the final sentence level annotation class will be *certain*). The sentence level annotation class *certain* was in clear majority among all three annotators. On both sentence and token level, the class *undefined* (a sentence that could not be classified as *certain* or *uncertain*, or a token which was not clearly speculative) was rarely used. Therefore, all sentence level annotations marked as *undefined* are converted to the majority class, *certain*, resulting in two sentence level annotation classes (*certain* and *uncertain*) and two token level annotation classes (*speculative words* and *negations*, i.e. to-

[3] White space tokenization.

kens annotated as *undefined* are ignored).

For the remaining analysis, we focus on the distributions of the annotation classes *uncertain* and *speculative words*, per annotator and annotator pair, and per clinical practice.

Clinical practice	Max	Avg	Stddev
hematology	40	10.67	7.97
surgery	57	11.08	8.29
neurology	105	11.67	10.30
geriatrics	58	11.04	9.29
orthopaedics	40	10.37	6.88
rheumatology	59	8.72	7.99
urology	46	11.61	7.86
cardiology	50	9.70	7.46
oncology	54	9.57	7.75
ENT	54	9.46	7.53
infection	37	11.48	7.76
emergency	55	9.42	6.88
paediatrics	68	9.55	7.24
total, full corpus	120	10.31	8.53

Table 2: Token statistics per sentence and clinical practice. All clinic groups except urology (min = 2) and cardiology (min = 2) have a minimum sentence length of one token.

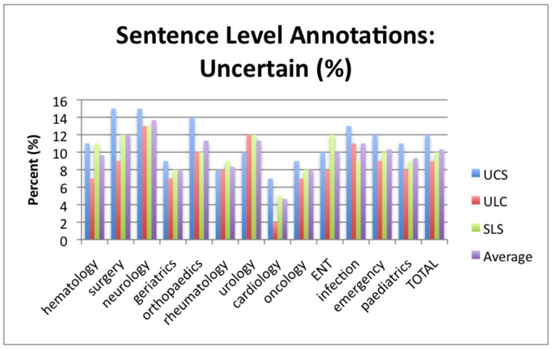


Figure 1: Sentence level annotation: *uncertain*, percentage per annotator and clinical practice.

4 Results

We have measured the proportions (in percent) per annotator for each clinical practice and in total. This enables an analysis of whether there are substantial individual differences in the distributions, indicating that this annotation task is highly subjective and/or difficult. Moreover, we measure IAA by pairwise F_1-score. From this, we may draw conclusions whether specific clinical practices are harder or easier to judge *reliably* (i.e. by high IAA results).

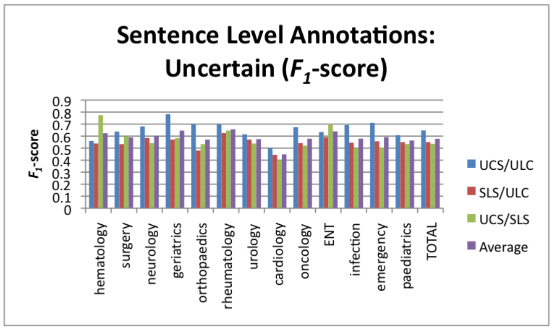


Figure 2: Pairwise F_1-score, sentence level annotation class *uncertain*.

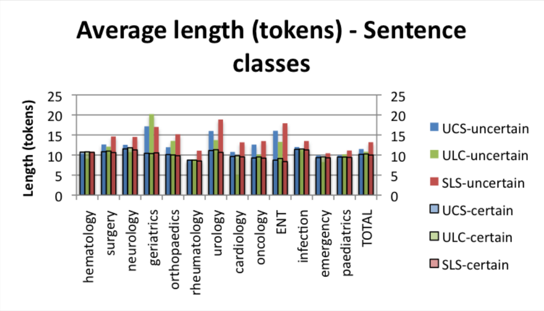


Figure 3: Average length in tokens, per annotator and sentence class.

In Figure 1, we see that the average amount of uncertain sentences lies between 9 and 12 percent for each annotator in the full corpus. In general, UCS has annotated a larger proportion of uncertain sentences compared to ULC and SLS.

The clinical discipline with the highest average amount of uncertain sentences is *neurology* (13.7 percent), the lowest average amount is found in *cardiology* (4.7 percent). Surgery and cardiology show the largest individual differences in proportions (from 9 percent (ULC) to 15 percent (UCS), and from 2 percent (ULC) to 7 percent (UCS), respectively).

However, in Figure 2, we see that the pairwise IAA, measured by F_1-score, is relatively low, with an average IAA of 0.58, ranging between 0.54 (UCS/SLS) and 0.65 (UCS/ULC), for the entire corpus. In general, the annotator pair UCS/ULC have higher IAA results, with the highest for *geriatrics* (0.78). The individual proportions for uncertain sentences in *geriatrics* is also lower for all annotators (see Figure 1), indicating a clinical practice with a low amount of uncertain sentences, and a slightly higher average IAA (0.64 F_1-score).

4.1 Sentence lengths

As the focus lies on analyzing sentences annotated as *uncertain*, one interesting property is to look at sentence lengths (measured in tokens). One hypothesis is that uncertain sentences are in general longer. In Figure 3 we see that in general, for all three annotators, uncertain sentences are longer than certain sentences. This result is, of course, highly influenced by the skewness of the data (i.e. uncertain sentences are in minority), but it is clear that uncertain sentences, in general, are longer on average. It is interesting to note that the annotator SLS has, in most cases, annotated longer sentences as uncertain, compared to UCS and ULC. Moreover, *geriatrics*, with relatively high IAA but relatively low amounts of uncertain sentences, has well above average sentence lengths in the *uncertain* class.

4.2 Token level annotations

When it comes to the token level annotations, *speculative words* and *negations*, we observed very high IAA for *negations* (0.95 F_1-score (exact match) on average in the full corpus, the lowest for *neurology*, 0.94). These annotations were highly lexical (13 unique tokens) and unambiguous, and spread evenly across the two sentence level annotation classes (ranging between 1 and 3 percent of the total amount of tokens per class). Moreover, all negations were unigrams.

On the other hand, we observed large variations in IAA results for *speculative words*. In Figure 4, we see that there are considerable differences between exact and partial matches[4] between all annotator pairs, indicating individual differences in the interpretations of what constitutes a speculative word and how many tokens they cover, and the lexicality is not as evident as for negations. The highest level of agreement we find between UCS/ULC in *orthopaedics* (0.65 F_1-score, partial match) and *neurology* (0.64 F_1-score, partial match), and the lowest in *infection* (UCS/SLS, 0.31 F_1-score).

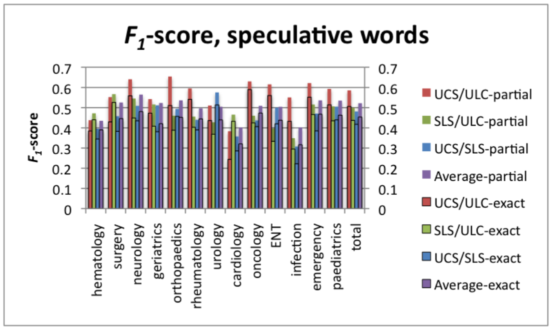


Figure 4: F_1-score, speculative words, exact and partial match.

4.2.1 Speculative words – most common

The low IAA results for *speculative words* invites a deeper analysis for this class. How is this interpreted by the individual annotators? First, we look at the most common tokens annotated as *speculative words*, shared by the three annotators: ”?”, ”sannolikt” (*likely*), ”ev” (*possibly*, abbreviated), ”om” (*if*). The most common speculative words are all unigrams, for all three annotators. These tokens are similar to the most common speculative words in the clinical BioScope subcorpus, where *if*, *may* and *likely* are among the top five most common. Those tokens that are most common per annotator and not shared by the other two (among the five most frequent) include ”bedöms” (*judged*), ”kan” (*could*), ”helt” (*completely*) and ”ställningstagande” (*standpoint*).

Looking at *neurology* and *urology*, with a higher overall average amount of uncertain sentences, we find that the most common words for *neurology* are similar to those most common in total, while for *urology* we find more n-grams. In Table 3, the five most common speculative words per annotator for neurology and urology are presented.

When it comes to the unigrams, many of these are also *not* annotated as speculative words. For instance, ”om” (*if*), is annotated as speculative in only 9 percent on average of its occurrence in the neurological data (the same distribution holds, on average, in the total set). In Morante and Daelemans (2009), *if* is also one of the words that are subject to the majority of false positives in their automatic classifier. On the other hand, ”sannolikt” (*likely*) is almost always annotated as a speculative word (over 90 percent of the time).

[4] Partial matches are measured on a character level.

	UCS	**ULC**	**SLS**
neurology	? sannolikt (*likely*) kan (*could*) om (*if*) pröva (*try*) ter (*seem*)	? kan (*could*) sannolikt (*likely*) om (*if*) verkar (*seems*) ev (*possibly*, abbr)	? sannolikt (*likely*) ev (*possibly*, abbr) om (*if*) ställningstagande (*standpoint*) möjligen (*possibly*)
urology	kan vara (*could be*) tyder på(*indicates*) ev (*possibly*, abbr) misstänkt (*suspected*) kanske (*perhaps*) planeras tydligen (*apparently planned*)	mycket (*very*) inga tecken (*no signs*) kan vara (*could be*) kan (*could*) tyder (*indicates*) misstänkt (*suspected*)	tyder på(*indicates*) i första hand (*primarily*) misstänkt (*suspected*) kanske (*perhaps*) skall vi försöka (*should we try*) kan vara (*could be*)

Table 3: Most common speculative words per annotator for *neurology* and *urology*.

4.2.2 Speculative words – n-grams

Speculative words are, in Swedish clinical text, clearly not simple lexical unigrams. In Figure 5 we see that the average length of tokens annotated as *speculative words* is, on average, 1.34, with the longest in *orthopaedics* (1.49) and *urology* (1.46). We also see that SLS has, on average, annotated longer sequences of tokens as *speculative words* compared to UCS and ULC. The longest n-grams range between three and six tokens, e.g. ”kan inte se några tydliga” (*can't see any clear*), ”kan röra sig om” (*could be about*), ”inte helt har kunnat uteslutas” (*has not been able to completely exclude*), ”i första hand” (*primarily*). In many of these cases, the strongest indicator is actually a unigram (”kan” (*could*)), within a verb phrase. Moreover, negations inside a *speculative word* annotation, such as ”inga tecken” (*no signs*) are annotated differently among the individual annotators.

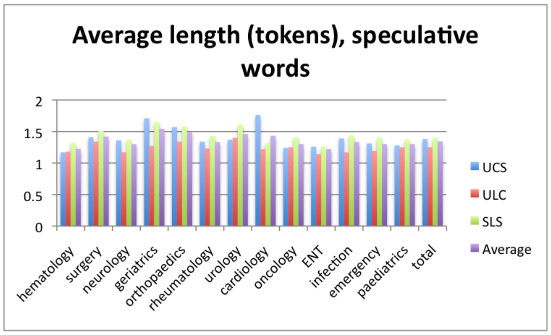


Figure 5: Average length, speculative words.

4.3 Examples

We have observed low average pairwise IAA for sentence level annotations in the *uncertain* class, with more or less large differences between the annotator pairs. Moreover, at the token level and for the class *speculative words*, we also see low average agreement, and indications that *speculative words* often are n-grams. We focus on the clinical practices *neurology*, because of its average large proportion of uncertain sentences, *geriatrics* for its high IAA results for UCS/ULC and low average proportion of uncertain sentences, and finally *surgery*, for its large discrepancy in proportions and low average IAA results.

In Example 1 we see a sentence where two annotators (ULC, SLS) have marked the sentence as *uncertain*, also marking a unigram (”ospecifik” (*unspecific*) as a *speculative word*. This example is interesting since the utterance is ambiguous, it can be judged as certain as in *the dizziness is confirmed to be of an unspecific type* or uncertain as in *the type of dizziness is unclear*, a type of utterance which should be clearly addressed in the guidelines.

<C> Yrsel av ospecifik typ. </C>

<U> Yrsel av <S> ospecifik </S> typ. </U>

<U> Yrsel av <S> ospecifik </S> typ. </U>

Dizziness of unspecific type

Example 1: Annotation example, *neurology*. Ambiguous sentence, *unspecific* as a possible speculation cue. C = Certain, U = Uncertain, S = Speculative words.

An example of different interpretations of the minimum span a *speculative word* covers is given in Example 2. Here, we see that ”inga egentliga märkbara” (*no real apparent*) has been annotated in three different ways. It is also interesting to note the role of the negation as part of amplifying speculation. Several such instances were marked by the annotators (for further examples, see Dalianis and Velupillai (2010)), which conforms well with the findings reported in Kilicoglu and Bergler (2008), where it is showed that explicit certainty markers together with negation are indicators of speculative language. In the BioScope corpus (Vincze et al., 2008), such instances are marked as speculation cues. This example, as well as Example 1, is also interesting as they both clearly are part of a longer passage of reasoning of a patient, with no particular diagnosis mentioned in the current sentence. Instead of randomly extracting sentences from the free text entry *Assessment*, one possibility would be to let the annotators judge all sentences in an entry (or a full EHR). Doing this, differences in where speculative language often occur in an EHR (entry) might become evident, as for scientific writings, where it has been showed that speculative sentences occur towards the end of abstracts (Light et al., 2004).

<U> <S><N> Inga </N> egentliga </S> <S> märkbara</S> minnessvårigheter under samtal. </U>.

<U> <N> Inga </N> <S> egentliga </S> märkbara minnessvårigheter under samtal. </U>.

<U> <S><N> Inga </N> egentliga märkbara </S> minnessvårigheter under samtal. </U>.

No real apparent memory difficulties during conversation

Example 2: Annotation example, *neurology*. Different annotation coverage over negation and speculation. C = Certain, U = Uncertain, S = Speculative words, N = Negation

In *geriatrics*, we have observed a lower than average amount of uncertain sentences, and high IAA between UCS and ULC. In Example 3 we see a sentence where UCS and ULC have matching annotations, whereas SLS has judged this sentence as certain. This example shows the difficulty of interpreting expressions indicating possible speculation – is ”ganska” (*relatively*) used here as a marker of certainty (as certain as one gets when diagnosing this type of illness)?

The word ”sannolikt” (*likely*) is one of the most common words annotated as a speculative word in the total corpus. In Example 4, we see a sentence where the annotators UCS and SLS have judged it to be *uncertain*, while UCS and ULC have marked the word ”sannolikt” (*likely*) as a *speculative word*. This is an interesting example, through informal discussions with clinicians we were informed that this word might as well be used as a marker of high certainty. Such instances show the need for using domain experts in future annotations of similar corpora.

<U> Både anamnestiskt och testmässigt <S> ganska </S> stabil vad det gäller Alzheimer sjukdom. </U>.

<U> Både anamnestiskt och testmässigt <S> ganska </S> stabil vad det gÏler Alzheimer sjukdom. </U>.

<C> Både anamnestiskt och testmässigt ganska stabil vad det gäller Alzheimer sjukdom. </C>.

Both anamnesis and tests relatively stabile when it comes to Alzheimer's disease.

Example 3: Annotation example, *geriatrics*. Different judgements for the word ”ganska” (*relatively*). C = Certain, U = Uncertain, S = Speculative words.

<C>En 66-årig kvinna med <S>sannolikt</S> 2 synkrona tumörer vänster colon/sigmoideum och där till levermetastaser.</C>.

<U>En 66-årig kvinna med <S>sannolikt</S> 2 synkrona tumörer vänster colon/sigmoideum och där till levermetastaser.</U>.

<C>En 66-årig kvinna med sannolikt 2 synkrona tumörer vänster colon/sigmoideum och där till levermetastaser.</C>.

A 66 year old woman likely with 2 synchronous tumours left colon/sigmoideum in addition to liver metastasis.

Example 4: Annotation example, *surgery*. Different judgements for the word ”sannolikt” (*likely*). C = Certain, U = Uncertain, S = Speculative words.

5 Discussion

We have presented an analysis of an initial annotation trial for the identification of uncertain sentences as well as for token level cues (*speculative words*) across different clinical practices. Our main findings are that IAA results for both sentence level annotations of uncertainty and token level annotations for speculative words are, on average, fairly low, with higher average agreement in *geriatrics* and *rheumatology* (see Figures 1 and 2). Moreover, by analyzing the individual distributions for the classes *uncertain* and *speculative words*, we find that *neurology* has the highest average amount of uncertain sentences, and *cardiology* the lowest. On average, the amount of uncertain sentences ranges between 9 and 12 percent, which is in line with previous work on sentence level annotations of uncertainty (see Section 2).

We have also showed that the most common *speculative words* are unigrams, but that a substantial amount are n-grams. The n-grams are, however, often part of verb phrases, where the head is often the speculation cue. However, it is evident that speculative words are not always simple lexical units, i.e. syntactic information is potentially very useful. Question marks are the most common entities annotated as *speculative words*. Although these are not interesting indicators in themselves, it is interesting to note that they are very common in clinical documentation.

From the relatively low IAA results we draw the conclusion that this task is difficult and requires more clearly defined guidelines. Moreover, using *naive* coders on clinical documentation is possibly not very useful if the resulting annotations are to be used in, e.g. a Text Mining application for medical researchers. Clinical documentation is highly domain-specific and contains a large amount of internal jargon, which requires judgements from clinicians. However, we find it interesting to note that we have identified differences between different clinical practices. A consensus corpus has been created from the resulting annotations, which has been used in an experiment for automatic classification, see Dalianis and Skeppstedt (2010) for initial results and evaluation.

During discussions among the annotators, some specific problems were noted. For instance, the extracted sentences were not always about the patient or the current status or diagnosis, and in many cases an expression could describe (un)certainty of someone other than the author (e.g. another physician or a family member), introducing aspects of perspective. The sentences annotated as *certain*, are difficult to interpret, as they are simply *not uncertain*. We believe that it is important to introduce further dimensions, e.g. explicit certainty, and focus (*what* is (un)certain?), as well as time (e.g. *current* or *past*).

6 Conclusions

To our knowledge, there is no previous research on annotating Swedish clinical text for sentence and token level uncertainty together with an analysis of the differences between different clinical practices. Although the initial IAA results are in general relatively low for all clinical practice groups, we have identified indications that *neurology* is a practice which has an above average amount of uncertain elements, and that *geriatrics* has a below average amount, as well as higher IAA. Both these disciplines would be interesting to continue the work on identifying speculative language.

It is evident that clinical language contains a relatively high amount of uncertain elements, but it is also clear that naive coders are not optimal to use for interpreting the contents of EHRs. Moreover, more care needs to be taken in the extraction of sentences to be annotated, in order to ensure that the sentences actually describe reasoning about the patient status and diagnosis. For instance, instead of randomly extracting sentences from within a free text entry, it might be better to let the annotators judge all sentences within an entry. This would also enable an analysis of whether speculative language is more or less frequent in specific parts of EHRs.

From our findings, we plan to further develop the guidelines and particularly focus on specifying the minimal entities that should be annotated as *speculative words* (e.g. ”kan” (*could*)). We also plan to introduce further levels of dimensionality in the annotation task, e.g. cues that indicate a high level of certainty, and to use domain experts as annotators. Although there are problematic issues regarding the use of *naive* coders for this task, we believe that our analysis has revealed some properties of speculative language in clinical text which enables us to develop a useful resource for further research in the area of speculative language. Judging an instance as being certain or uncertain is, perhaps, a task which can never exclude subjective interpretations. One interesting way of exploiting this fact would be to exploit individual annotations similar to the work presented in Reidsma and op den Akker (2008). Once we have finalized the annotated set, and ensured that no identifiable information is included, we plan to make this resource available for further research.

References

Hercules Dalianis and Maria Skeppstedt. 2010. Creating and Evaluating a Consensus for Negated and Speculative Words in a Swedish Clinical Corpus. To be published in the proceedings of the Negation and Speculation in Natural Language Processing Workshop, July 10, Uppsala, Sweden.

Hercules Dalianis and Sumithra Velupillai. 2010. How Certain are Clinical Assessments? Annotating Swedish Clinical Text for (Un)certainties, Speculations and Negations. In *Proceedings of the of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta, May 19-21.

J. L. Hobby, B. D. M. Tom, C. Todd, P. W. P. Bearcroft, and A. K. Dixon. 2000. Communication of doubt and certainty in radiological reports. *The British Journal of Radiology*, 73:999–1001, September.

R. Khorasani, D. W. Bates, S. Teeger, J. M. Rotschild, D. F. Adams, and S. E. Seltzer. 2003. Is terminology used effectively to convey diagnostic certainty in radiology reports? *Academic Radiology*, 10:685–688.

Halil Kilicoglu and Sabine Bergler. 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 9(S-11).

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In Lynette Hirschman and James Pustejovsky, editors, *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*, pages 17–24, Boston, Massachusetts, USA, May 6. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 28–36, Morristown, NJ, USA. Association for Computational Linguistics.

Philip V. Ogren. 2006. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–275, Morristown, NJ, USA. Association for Computational Linguistics.

Arzucan Özgür and Dragomir R. Radev. 2009. Detecting speculations and their scopes in scientific text. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1398–1407, Singapore, August. Association for Computational Linguistics.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *HumanJudge '08: Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 8–16, Morristown, NJ, USA. Association for Computational Linguistics.

Victoria L. Rubin, Elizabeth D. Liddy, and Noriko Kando. 2006. Certainty identification in texts: Categorization model and manual tagging results. In *Computing Affect and Attitutde in Text: Theory and Applications*. Springer.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S-11).

J. W. Wilbur, A. Rzhetsky, and H. Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7:356+, July.

PAPER IV

FACTUALITY LEVELS OF DIAGNOSES IN SWEDISH CLINICAL TEXT

Sumithra Velupillai, Hercules Dalianis and Maria Kvist. 2011. *Factuality Levels of Diagnoses in Swedish Clinical Text.* In Proceedings of the XXIII International Conference of the European Federation for Medical Informatics (MIE), Oslo, Norway, August 28–31 2011. IOS Press.

Author contributions I was responsible for extracting a subset for the annotation work and marking them for diagnostic statements. Both Hercules Dalianis, Maria Kvist and I were involved in developing the annotation guidelines and preparing the annotation work. Both Maria Kvist and I analyzed the results of the gold standard. The article was written jointly.

User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011

doi:10.3233/978-1-60750-806-9-559

Factuality Levels of Diagnoses in Swedish Clinical Text

Sumithra VELUPILLAI [a,1], Hercules DALIANIS [a], Maria KVIST[a, b]
[a] *Dept. of Computer and Systems Sciences (DSV),*
Stockholm University, Forum 100, SE-164 40 Kista, Sweden
[b] *Dept. of clinical immunology and transfusion medicine, Karolinska University Hospital, SE-171 76 Stockholm, Sweden*

Abstract. Different levels of knowledge certainty, or factuality levels, are expressed in clinical health record documentation. This information is currently not fully exploited, as the subtleties expressed in natural language cannot easily be machine analyzed. Extracting relevant information from knowledge-intensive resources such as electronic health records can be used for improving health care in general by e.g. building automated information access systems. We present an annotation model of six factuality levels linked to diagnoses in Swedish clinical assessments from an emergency ward. Our main findings are that overall agreement is fairly high (0.7/0.58 F-measure, 0.73/0.6 Cohen's κ, Intra/Inter). These distinctions are important for knowledge models, since only approx. 50% of the diagnoses are affirmed with certainty. Moreover, our results indicate that there are patterns inherent in the diagnosis expressions themselves conveying factuality levels, showing that certainty is not only dependent on context cues.

Keywords. Diagnosis reasoning, factuality levels, annotation, Swedish, clinical text, electronic health records.

1. Introduction

The process of diagnosing a patient is not trivial, and involves making decisions based on many diverse criteria. Clinicians are documenting reasoning processes and decisions in free-text, information that is currently not fully exploited for further knowledge management or research. Accurate and situation-specific information access is extremely important, especially in the clinical domain. This will provide clinicians with tools for information retrieval, using extracted information to produce relevant summaries, aggregating extracted information for knowledge discovery and further clinical research [1].

In order to create information access solutions that utilize the knowledge documented in free-text, it is necessary to be able to model subtleties expressed in natural language. One important aspect to consider is the level of certainty expressed in the reasoning and decision context. For instance, a likely scenario is the incorporation of a search engine in an electronic health record system, where clinicians can search for previous mentions of diagnoses for a particular patient. However, some of these diagnoses are written in a negated or speculative context, e.g. *this is definitely not*

[1] Corresponding author

diabetes or *angina pectoris cannot be excluded*. It is crucial that such distinctions are observed, as they convey different levels of knowledge certainty.

Research on modeling factuality levels, or degrees of certainty, in textual data, has increased in recent years. In the BioScope corpus [2], which contains biomedical texts, certainty levels are annotated at a sentence level, while negation and speculation cues are annotated at a token (word) level. In FactBank, factuality levels in newspaper articles are instead annotated on an event level [3]. In the clinical domain, agreement on probability expressions in radiology reports has been studied. Two studies analyzed phrases indicating different levels of certainty with respect to diagnoses [4, 5]. Both studies show that intermediate probabilities are more difficult to agree on while phrases indicating very high or low probabilities result in higher agreement. In automatic information retrieval settings, these issues have also been addressed in the research community lately. RadReportMiner [6] is a context-aware search engine, taking into account negations and uncertainties, achieving improved precision results (81%) compared to a generic search engine (27%).

In this paper, we present a model for annotating factuality distinctions in clinical documentation. Our aim is to develop automated systems that distinguish factuality levels of diagnoses in Swedish. Two clinicians annotate diagnoses in free-text entries for factuality levels. We analyze and evaluate the annotations with Intra- and Inter-Annotator Agreement (IAA). To our knowledge, this is the first attempt at modeling these distinctions and creating such a resource in Swedish.

2. Methods

Work process: we (1) assembled a list of diagnoses and created a resource for annotation, (2) developed annotation guidelines and annotated the created set, (3) evaluated Inter- and Intra-Annotator Agreement and did a qualitative analysis. We used the Knowtator plugin in the Protégé tool [7] for all annotation work. All documents were extracted randomly. Two senior physicians, A1 and A2, performed all annotation tasks, both accustomed to reading and writing medical records.

We extracted free-text entries from an emergency ward included in the Stockholm EPR Corpus [8]. Only entries documented under the category *Bedömning* (Assessment) were used in the annotation task. This field was chosen since it is the documentation entry containing most reasoning.

2.1. Creating a set of Documents Marked with Diagnoses

Instead of using diagnoses from Swedish medical terminology resources, we wanted to capture many diagnosis variants (e.g. inflections, misspellings, abbreviations). A collection of Swedish diagnoses was produced through a manual analysis of a subset of 150 assessment fields. A diagnosis was defined as a medical condition with a known cause, prognosis or treatment. All different variants and inflections of the same diagnosis expression were annotated.

A simple string matching procedure was employed to automatically mark diagnoses from the created diagnosis collection. A general language automatic lemmatizer for Swedish[2] was used for capturing further inflections. Each diagnosis was marked with brackets, e.g. *Patient with <Diagnosis>diabetes</Diagnosis>*.

2.2. Annotation Classes and Guidelines

Factuality levels were modeled in two polarities: Positive and Negative. These were further graded: Certain, Probable or Possible. Each extracted diagnosis expression was annotated as belonging to one polarity and gradation, e.g. Certainly Positive, resulting in six annotation classes. Furthermore, the class Not Diagnosis was included for cases where the current context was not a diagnosis (e.g. infektion – short for clinic), and the class Other, for cases where e.g. the diagnosis referred to someone other than the patient, or where the annotator was uncertain. A first annotation task was performed in order to create detailed guidelines for the remaining task3.

2.3. Evaluation Metrics

The results were evaluated with IAA: F-measure, and Cohen's κ. IAA (Intra) results were measured on documents annotated twice by annotator A1, the second time in a new, randomized order. IAA (Inter) results were measured on documents annotated by two annotators; A1 and A2, treating A1 as the gold standard.

3. Results

In total, the number of annotated diagnosis instances was 2 182 (A1 vs A1) and 2 070 (A1 vs A2)4, extracted from 1 297 Assessment fields (approx. 51% of the total amount of Assessment fields). From the collection of 337 diagnoses, 227 were found.

3.1. Intra- and Inter-Annotator Agreement

A confusion matrix over the number of instances assigned to each class is shown in Table 1. *Certainly Positive* was in clear majority, almost 50% of the total number of instances. *Possibly Negative* and *Not Diagnosis* were very rare. The main discrepancies between the two annotators were in cases of assigning intermediate factuality levels. A1 generally assigned higher levels of factuality. Intra- and Inter-Annotator Agreement was very high for the majority class *Certainly Positive* (0.9 F-measure, respectively), while very low for *Possibly Negative* (0.35/0.03 F-measure, respectively), being a rare class. It is interesting to note that the classes *Not Diagnosis* and *Other*, both relatively rare, resulted in fairly high agreement results (0.82/0.62 and 0.69/0.65 F-measure, respectively). Overall IAA measured by Cohen's κ is: 0.73 (Intra), and 0.60 (Inter).

[2] http://www.cst.dk/online/lemmatiser/

[3] Annotation guidelines, including examples, can be found at http://www.dsv.su.se/hexanord/guidelines/ (guidelines_stockholm_epr_diagnosis_factuality_corpus.pdf)

[4] The discrepancy between the two sets is caused by mismatches and missed instances

Table 1. Confusion matrix, Intra- and Inter-Annotator Agreement.

	CP	**PrP**	**PoP**	**PoN**	**PrN**	**CN**	**ND**	**O**	**Σ**
CP Intra	**990**	78	4	0	3	4	2	19	1100
Inter	***834***	*59*	*7*	*0*	*4*	*5*	*1*	*20*	*930*
PrP Intra	20	**236**	55	1	1	0	1	0	314
Inter	*66*	***134***	*10*	*1*	*0*	*0*	*2*	*1*	*214*
PoP Intra	4	38	**127**	25	9	0	0	2	205
Inter	*11*	*149*	***180***	*41*	*45*	*1*	*1*	*10*	*438*
PoN Intra	0	0	6	**14**	7	1	0	1	29
Inter	*0*	*0*	*0*	***1***	*5*	*1*	*0*	*0*	*7*
PrN Intra	1	1	1	10	**118**	25	0	5	161
Inter	*0*	*0*	*0*	*2*	***35***	*18*	*0*	*1*	*56*
CN Intra	2	0	4	0	51	**195**	0	1	253
Inter	*2*	*0*	*0*	*4*	*99*	***193***	*1*	*3*	*302*
ND Intra	0	0	0	0	0	0	**26**	0	26
Inter	*13*	*5*	*3*	*2*	*1*	*3*	***30***	*4*	*61*
O Intra	8	1	4	1	7	0	8	**65**	94
Inter	*1*	*1*	*1*	*1*	*5*	*3*	*1*	***49***	*62*
Σ Intra	1025	354	201	51	196	225	37	93	**2182**
Inter	*927*	*348*	*201*	*52*	*194*	*223*	*36*	*88*	***2070***

Columns: A1, first annotation iteration. Rows: Intra: A1, second annotation iteration (same set randomized), Inter: A2. CP = Certainly Positive, PrP = Probably Positive, PoP = Possibly Positive, PoN = Possibly Negative, PrN = Probably Negative, CN = Certainly Negative, ND = Not Diagnosis, O = Other, **Σ** = Total

3.2. Qualitative Analysis

We also performed a manual, qualitative analysis of the resulting class assignments. We found that Certainly Positive dominated where a) diagnoses show overtly, e.g. skin diseases (eczema, urthicaria, skin infection) and general conditions (overweight, asystolia, fainting), or b) diagnosis was made by an apparatus (auricular fibrillation/ ECG). Probably Positive dominates for diagnoses with medical reasons for not securing certainty, e.g. virosis, gasthritis. Linguistic reasons seem to direct the following for some diagnoses: 1) an inverted pattern with a complementary vocabulary, e.g. ischemia (Certainly/Probably Negative in majority), heart attack or angina pectoris (Certainly/Probably Positive in majority), 2) a lack of negative annotation classes when normality was not expressed as negation (hypertension), 3) for lunginflammation (pneumonia), speculation was expressed in Swedish while we saw certainty expressed in Greek.

4. Discussion

In this study we present a model for knowledge certainty classification. This is used for the creation of an annotated set of Assessment entries from a Swedish emergency ward for factuality levels assigned to diagnoses. The model was functional and agreeable to the domain expert annotators. Our IAA results suggest that this model and resource can be used for developing automated systems. We also show, through a qualitative analysis, that factuality levels for different diagnoses are dependent on diagnosis type as well as inherent linguistic factors. This demonstrates that factuality and speculation in clinical text resides not only in linguistic context cues.

4.1. Limitations

The study design has some limitations that lowered the recall of diagnoses to be annotated. By employing a strict matching approach, yielding high precision, possible variants in form of misspellings, compounding and other formulations were missed. Fuzzier matching techniques could increase recall, at the cost of lower precision. The use of a limited list of diagnoses will inevitably result in a skewed distribution of diagnosis types. As a result, the model may not catch enough numbers and types of expressions of subtleties in conveying levels of factuality. How this in turn limits the created resources' ability to be used for machine learning is yet to be seen. The main limitation of this model for future work is the low numbers of annotations in some annotation classes. Intermediate probability assignments are clearly not self-evident (e.g. [4] and [5]). It can be argued that factuality levels *Possibly* and *Probably* may be fused, or even two *Possibly* classes, to lower the number of factuality levels, and increaseing training instances for machine-learning tasks. Such fusion was not agreeable to the involved physicians, as it would be a less accurate description of reality.

4.2. Significance of Study

Our results have important implications on the creation of intelligent information access from electronic health records. Without factuality analysis, uncertain or negated diagnoses would be identified as factual diagnoses. We have chosen a broad context-aware approach, in order to receive a wide perspective on how factuality levels are expressed concerning diagnoses. To our knowledge, no other studies have used a similar approach in this domain. Studies in the biomedical field (e.g. [3]) use hedge cues to detect uncertainty. We hope our approach will reveal inherent and previously unknown features that will aid in future machine-learning and text-mining studies.

Acknowledgments: This research has been carried out after approval from the Regional Ethical Review Board, Stockholm (Etikprövningsnämnden i Stockholm), permission no 2009/1742-31/5

References

[1] Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JE. Extracting Information from Textual Documents in the Electronic Health Record, *IMIA Yearbook of Medical Informatics 2008* **47** Suppl. 1 (2008), 138–154.

[2] Vincze V, Szarvas G, Farkas R, Móra G, Csirik J. The Bioscope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes, *BMC Bioinformatics* **9(S-11)** (2008)

[3] Saurí R, Pustejovsky J. FactBank: a corpus annotated with event factuality, *Language Resources & Evaluation* **43** (2009), 227–268

[4] Khorasani R, Bates DW, Teeger S, Rothschild JM, Adams DF, Seltzer SE. Is Terminology Used Effectively to Convey Diagnostic Certainty in Radiology Reports?, *Academic Radiology* **10** (2003), 685–688.

[5] Hobby JL, Tom BDM, Todd C, Bearcroft PWP, Dixon AK. Communication of Doubt and Certainty in Radiology Reports, *The British Journal of Radiology* **73** (2000), 999–1001.

[6] Wu AS, Do BH, Kim J, Rubin DL. Evaluation of Negation and Uncertainty Detection and its Impact on Precision and Recall in Search, *Journal of Digital Imaging*

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

[8] Dalianis H, Hassel M, Velupillai S. *The Stockholm EPR Corpus – Characteristics and some Initial Findings*, in Proc. 14th ISHIMIR, Kalmar, Sweden, 2009.

PAPER V

AUTOMATIC CLASSIFICATION OF FACTUALITY LEVELS – A CASE STUDY ON SWEDISH DIAGNOSES AND THE IMPACT OF LOCAL CONTEXT

Sumithra Velupillai. 2011. *Automatic Classification of Factuality Levels – A Case Study on Swedish Diagnoses and the Impact of Local Context.* In Proceedings of The Fourth International Symposium on Languages in Biology and Medicine (LBM 2011), Singapore, Dec 14–15 2011.

Automatic Classification of Factuality Levels – A Case Study on Swedish Diagnoses and the Impact of Local Context

Sumithra Velupillai
Dept. of Computer and Systems Sciences (DSV)
Stockholm University
Forum 100, SE-16440 Kista, Sweden
sumithra@dsv.su.se

Abstract

Clinicians express different levels of knowledge certainty when reasoning about a patient's status. Automatic extraction of relevant information is crucial in the clinical setting, which means that factuality levels need to be distinguished. We present an automatic classifier using Conditional Random Fields, which is trained and tested on a Swedish clinical corpus annotated for factuality levels at a diagnosis statement level: the Stockholm EPR Diagnosis-Factuality Corpus. The classifier obtains promising results (best overall results are 0.699 average F-measure using all classes, 0.762 F-measure using merged classes), using simple local context features. Preceding context is more useful than posterior, although best results are obtained using a window size of ±4. Lower levels of certainty are more problematic than higher levels, which was also the case for the human annotators in creating the corpus. A manual error analysis shows that conjunctions and other higher-level features are common sources of errors.

1 Introduction and Background

Decision-making is a central task in clinical work, which involves complex reasoning based on information at hand. Clinicians are faced with new patients and need to be able to assess the patient's status according to several criteria, depending on situation, clinical expertise, previous history, patient descriptions, etc. Clinicians document their findings and reasoning in words, either through dictation or directly in written form. Today, most documentation is inserted in digitized systems, where information is stored both in structured and unstructured (free text) forms. One of the central activities in clinical work is the process of diagnosing. A clinician needs to classify what (possible) problem(s) a patient suffers from. This process involves much reasoning. Since clinicians document a large amount in free text, there is a lot of information to be extracted that could be of use in the decision-making process, e.g. similar cases and overviews. For this, accurate information extraction techniques are needed.

In many situations, it is not clear what disease a patient actually suffers from. A physician might receive insufficient background information, symptoms might be unclear or there might be several alternative possibilities to a patient's status. Moreover, it may also be the case that a disease is excluded as a possibility. Such reasoning is documented in free text, and these distinctions are crucial to model if an information extraction system is to be built for retrieving diagnostic information from clinical documentation.

The importance of modeling modality and negation for information extraction and information access purposes has been recognized in several different research areas lately, e.g. in the biomedical domain, for opinion mining and subjectivity analysis, summarization, text mining. Different models for representing modality and negation have been proposed, ranging from analyzing sentence levels to event levels, exploiting specific surface markers (keywords and phrases) or more complex linguistic constructions. When it comes to building automatic systems for distinguishing factuality levels, we see two general approaches: rule-based or machine learning models exploiting annotated corpora.

Annotation models: Wilbur et al. (2006) present a model of five qualitative dimensions for characterizing scientific articles: focus, polarity, certainty, evidence and directionality. The aim is to be able to identify reliable scientific facts, or informative fragments, along these dimensions. This model is applied on a sentence level (or subsentential if the sentence is complex). Polarities are modeled on a positive and negative axis, and certainty levels are modeled on a scale of 0 – 3, where 0 indicates complete uncertainty. The highest degree (3) represents complete certainty. Similarly, Rubin et al. (2006) create an annotation scheme where degree (certainty level), perspective (whose certainty), focus (object of certainty) and time is modeled. Certainty levels are modeled on four levels: absolute, high, moderate and low. Here, polarity is not included in the model. This model is applied on newspaper articles.

A different approach is presented in FactBank (Saurí, 2008), where factuality levels are annotated on an event level. Moreover, factuality is modeled on two different polarities: *positive* and *negative*, followed by certainty levels *certainly*, *probably* and *possibly*. Linguistically motivated markers are discussed in detail. For cases where polarity cannot be ascertained, *underspecified* is used. This corpus consists of newspaper articles, as a second layer on top of TimeBank (Pustejovsky et al., 2006).

Automatic systems: The BioScope corpus (Vincze et al., 2008) is a manually annotated corpus containing biomedical texts as well as clinical free-text (radiology reports), annotated for negation and speculation cues (token level) along with their linguistic scope (sentence level). This corpus has been used for the development of supervised learning classifiers, and was used in the CoNLL 2010 Shared task (Farkas et al., 2010), where the top performing system obtained an F-measure of 0.864 for detecting uncertain sentences (Tang et al., 2010), and 0.573 for detecting in-sentence hedge cues (Morante et al., 2010).

In the clinical domain, rule-based systems for distinguishing negations and uncertainties have been successfully developed, e.g. Harkema et al. (2009) and Friedman et al. (2004). ConText (Harkema et al., 2009) is an extension of the NegEx algorithm (Chapman et al., 2001), where negated, historical, hypothetical conditions, and conditions not experienced by the patient are automatically identified in emergency department reports. RadReportMiner (Wu et al., 2009) is a context-aware search engine, taking into account negations and uncertainties, achieving improved precision results (0.81) compared to a generic search engine (0.27) using a modified version of the NegEx algorithm, including expanded sets of negation and uncertainty keywords.

Studies on uncertainty expressions in the clinical domain: Verbal and numerical uncertainty expressions and their role in communicating clinical information have been studied from many perspectives and for different purposes, e.g. decision-making, interpretation, impact on physicians, patients and information systems. Most often, studies have used direct and indirect scaling procedures, where a fixed number of verbal expressions are given for judgment, and evaluating results by inter- and intra-subject agreement (see e.g. Clark (1990) for a critical review). In general, intra-evaluator agreement is found to be high, and inter-evaluator agreement to be low. Intermediate probabilities are often more difficult to agree on, while very high or low probabilities result in higher agreement (see Khorasani et al. (2003), Hobby et al. (2000), Christopher and Hotz (2004)). In many cases, the main conclusion is to recommend the use of controlled vocabularies for expressing different levels of certainty. The verbal expressions range from one word expressions such as *definite*, *likely*, *possible* to longer expressions such as *cannot be excluded*. The relationship between expressing probabilities verbally or numerically has also been studied (e.g. Timmermans (1994) and Renooij and Witteman (1999)), where findings suggest that verbal expressions are found to be more vague than numerical, and hence more difficult to use in decision-making.

The work presented here is divided into the following parts: 1) automatically classifying factuality levels using the Stockholm EPR Diagnosis-Factuality Corpus (Velupillai et al., 2011) with local context features and 2) evaluating by measuring precision, recall and F-measure and 3) performing a qualitative, manual error analysis. To our knowledge, no previous research have modeled factuality levels in clinical assessment documentation on a diagnostic statement level, nor on Swedish clinical documentation.

2 Methods

Our work process is: (1) automatic classification of the Stockholm EPR Diagnosis-Factuality Corpus using local context features and (2) evaluating the classification results quantitatively (precision, recall and F-measure) and qualitatively by manual error analysis[1].

2.1 The Stockholm EPR Diagnosis-Factuality corpus

> Låg sannolikhet för <D>dvt</D> pga frånvaro av riskfaktorer och blygsamma klin. fynd.
> *Low probability for <D>dvt</D> (abbr.) due to lack of risk factors and modest clinical (abbr.) findings.*.

Example 1: Example sentence from the Stockholm EPR Diagnosis-Factuality Corpus, D = Diagnostic statement. In this case, the diagnostic statement *dvt* (deep venous thrombosis) was to be annotated for factuality level, e.g. *possibly positive*.

The Stockholm EPR Diagnosis-Factuality corpus consists of documents that have been extracted from a university hospital emergency ward included in the Stockholm EPR Corpus (Dalianis et al., 2009). The documents are extracted from a medical emergency ward, since this is a type of clinic where several different types of diseases can be encountered. Only entries documented under the category *Bedömning* (Assessment) have been used. This entry type was chosen since it is where most reasoning, speculation and discussion regarding the patient status is documented. Each assessment entry is saved as one document, i.e. no other information from the patient record is used in the annotation task. Two domain experts (A1 and A2); senior physicians, both accustomed to reading and writing Swedish medical records, annotated the diagnostic statements for factuality levels. The largest set was annotated by A1, which is used in the presented work. Inter- and Intra-Annotator Agreement (IAA) results are 0.7/0.58 F-measure and 0.73/0.6 Cohens κ, respectively. The corpus is further described in (Velupillai et al., 2011).

	Total (#)	With diagnoses (#)
Documents	3 846	
Sentences	26 232	5 741
Tokens	283 007	69 355
Types (lemmas)	14 834	6 077
Diagnoses	6 483	
Diagnosis types (lemmas)	302	

Table 1: General statistics: Stockholm EPR Diagnosis-Factuality Corpus. Total set annotated by annotator A1. Each assessment entry is one document. Each diagnostic statement is one annotation instance. Punctuation is included in tokens and types.

2.1.1 Corpus characteristics

In the Stockholm EPR Diagnosis-Factuality Corpus, sentence and keyword level annotations are not used. Instead, only diagnostic statements are annotated for factuality levels. A manually created list of diagnostic statements was used, including different inflections, variants and abbreviations. The diagnostic statements in this list were automatically marked in brackets for the annotators to assign factuality levels. The whole assessment entry was shown to the annotators. An example sentence is shown in Example 1. General statistics of the Stockholm EPR Diagnosis-Factuality Corpus are shown in Table 1.

Following the factuality modeling presented in (Saurí, 2008), factuality levels are first defined in two polarities: *Positive* and *Negative*. Each of these were also graded: *Certain*, *Probable* or *Possible*. In total six annotation classes are used for marking factuality levels. Furthermore, the annotation class *Not Diagnosis* is used for cases where, e.g. the diagnostic statement in its context was something else (e.g. infektion (*infection*, short for clinical department)), or kol (*coal* in its meaning medical coal, not the diagnosis *COPD*). The annotation class *Other* is also included for cases where e.g. the diagnostic statement referred to someone other than the patient, or where the annotator could not assess the diagnostic statement

[1] This research has been carried out after approval from the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2009/1742-31/5.

according to any of the other classes[2]. The resulting annotation classes were the result of thorough discussions between the annotators and the research group. Guidelines for the annotation task are publicly available[3].

2.1.2 Class distributions

The distribution of factuality level annotation classes is shown in Table 2. *Certainly positive* is in clear majority, almost 50%. *Possibly negative*, *Not diagnosis* and *Other* are very rare, with less than 3%, respectively. The negative polarity amounts to 21.7% in total, and intermediate positive factuality levels (probably and possibly) amount to 26.2%, which means that a fair amount of diagnostic statements are speculative or negated. Thus, distinguishing factuality levels is very important for accurate information extraction.

Annotation Class	*n*	%
Certainly Positive	3 088	47.6
Probably Positive	1 039	16.0
Possibly Positive	663	10.2
Possibly Negative	139	2.2
Probably Negative	546	8.4
Certainly Negative	711	11.0
Not Diagnosis	117	1.8
Other	180	2.8
Σ	6 483	100.0

Table 2: Class distributions.

As a broad coverage approach was chosen, several different diagnostic statements are present in the annotated set. In Table 3, we see example distributions per class for some of the most frequent diagnostic statements. We observe that some diagnostic statements are more commonly used only in one class, e.g. förmaksflimmer (*atrial fibrillation*) and hypertoni (*hypertension*): *certainly positive* (93% and 89%, respectively). On the other hand, dvt (*deep venous thrombosis*, abbreviated), and infektion (*infection*) are more spread out and can be discussed in all factuality levels and polarities. Infektion (*infection*) is also sometimes used for mentioning a clinic, which is why it can be annotated as *not diagnosis*. Ischemi (*ischaemia*) is almost always assigned a negative polarity annotation class (28% *probably negative*, 58% *certainly negative*).

2.2 Automatic classification

For automatic classification we have used Conditional Random Fields (CRF) (Lafferty et al., 2001), as implemented in CRF++ [4], a classification algorithm that has been successful for similar Natural Language Processing (NLP) classification tasks. We use default settings, with no added parameter tuning. As there are cases where there are several diagnostic statements in one sentence, we do not treat this as a sentence level classification task. Instead, each token in all sentences containing an annotation instance (the assigned factuality level class for the marked diagnostic statement[5]) is classified. We have, in this work, used local features surrounding each annotation instance.

Many previous studies on expressions of probabilities in the clinical domain have used specific keywords and phrases within a small context window (e.g. Khorasani et al. (2003), Hobby et al. (2000)). Although these studies have been used in English settings, we found similar patterns in our Swedish clinical corpus. We limit the context window to ±4. For expanding the language model, we also use Part-of-Speech (PoS) tags and lemmas, extracted from a general language tagger for Swedish (Knutsson et al., 2003). We use simple features: word, lemma and PoS tag.

All results from the automatic classification experiments were measured on a test set containing 20% of the total amount of annotations. 80% of the total set is used for training. Approximately the same proportions of annotation class distributions are used in both sets. Results were measured with precision, recall and F-measure, using the CoNLL 2010 Shared task evaluation script conlleval.pl[6]. 95% confidence intervals were calculated for precision and recall.

[2] This class can be considered as a *neutral* class, for cases where no polarity and factuality level can be assessed (underspecified in Saurí (2008)).

[3] http://www.dsv.su.se/hexanord/guidelines/guidelines_stockholm_epr_diagnosis_factuality_corpus.pdf

[4] http://crfpp.sourceforge.net/#source

[5] diagnostic statements that are multiword tokens, such as *angina pektoris* are concatenated into one token.

[6] http://www.cnts.ua.ac.be/conll2000/chunking/output.html

Diagnosis	CP	PrP	PoP	PoN	PrN	CN	ND	O
deep venous thrombosis								
dvt	83 (21)	36 (9)	89 (23)	25 (6)	91 (23)	55 (14)	0	12 (3)
infection								
infektion	74 (25)	41 (14)	40 (13)	8 (3)	49 (16)	55 (18)	18 (6)	13 (4)
atrial fibrillation								
förmaksflimmer	241 (93)	6 (2)	5 (2)	0	0	3 (1)	0	5 (2)
hypertension								
hypertoni	213 (89)	16 (7)	5 (2)	0	0	0	0	4 (2)
ischaemia								
ischemi	2 (2)	1 (1)	11 (9)	2 (2)	32 (28)	67 (58)	0	1 (1)

Table 3: Example distributions diagnostic statements vs factuality level classes, *n* (%). CP = Certainly Positive, PrP = Probably Positive, PoP = Possibly Positive, PoN = Possibly Negative, PrN = Probably Negative, CN = Certainly Negative, ND = Not Diagnosis and O = Other

3 Results

The training set consists of 4 583 sentences, 5 171 annotation instances, and the test set of 1 158 sentences, 1 312 annotation instances. In these initial experiments, we are interested in looking at the local context, which is why we use only those sentences that contain annotated instances. For evaluating the automatic classification results, we use as a baseline the word itself as the only feature. Following the IAA-results (see Velupillai et al. (2011)), where the intermediate factuality levels often were a source of lower results, we also perform automatic classification where we merge the two intermediate factuality level classes per polarity, i.e. *probably/possibly positive/negative* are merged into *probably possibly positive* and *probably possibly negative*. We also merge *other* and *not diagnosis* into one class in order to increase the number of instances. Majority class baseline is also used for evaluating results.

All instances that are not annotated are assigned the class *NONE*. Baseline results are shown in Table 4. The majority class *certainly positive* obtains relatively high results (0.742 (all classes) and 0.758 (merged classes) F-measure). Overall average results for all classes is 0.561 F-measure and 0.605 for merged classes, an improvement over the majority class baseline (47.6% for all classes as well as merged classes).

3.1 Local context features

Using the closest context (window ± 1) improves results considerably compared to the baseline (using only the word itself) for all classes and settings (0.659 F-measure, all annotation classes, 0.704 F-measure, merged classes), using only words and lemmas. Intermediate classes in the positive polarity gain from merging, while *not diagnosis* obtains lower results. Increasing the window size step by step improves results further, and best results are obtained using a window size of ± 4, with words, lemmas and PoS information (Table 5). Using only words, lemmas and PoS information in a four-span window *preceding* the word itself yields similar results (0.69 F-measure for all classes, 0.736 for merged classes), indicating that preceding context is extremely valuable. Contrasting with posterior features (±4) yields lower results: 0.599 (all classes) and 0.649 (merged classes). PoS information is useful in combination with words and/or lemmas, not as a feature on its own. A considerable improvement is seen when increasing the window size from ±2 to ±3 (0.67 to 0.69 all classes, 0.716 to 0.737 merged classes). The greatest gain is seen for *certainly negative*, with an increase in F-measure from 0.546 to 0.674 (all classes) and 0.55 to 0.676.

3.2 Error analysis

The erroneous classification results from using window ± 4 with CRF classification have been analyzed (semi-)manually. The most frequent errors are misclassifications within the same polarity, or missed instances. We observed some general trends:

- Conjunctions: in many cases, conjunctions such as *och* (and), *eller* (or), cause errors, indicating that surface level features are problematic; these instances might have been captured if syntactic information was used, e.g. *Inga hållpunkter i lab och ekg för pågående* **ischemi** (No basis in lab and ecg for ongoing **ischaemia**).

	P_a (95% CI)	R_a (95% CI)	F_a	P_m (95% CI)	R_m (95% CI)	F_m	Merged
CP	0.657 ± 0.04	0.853 ± 0.03	0.742	0.736 ± 0.03	0.781 ± 0.03	0.758	CP
PrP	0.5 ± 0.07	0.3202 ± 0.06	0.390	0.478 ± 0.05	0.611 ± 0.05	0.536	PrPoP
PoP	0.464 ± 0.08	0.09 ± 0.05	0.151				
PoN	0.0 ± 0.0	0.0 ± 0.0	0.0	0.377 ± 0.08	0.153 ± 0.06	0.217	PrPoN
PrN	0.273 ± 0.08	0.296 ± 0.09	0.284				
CN	0.393 ± 0.08	0.487 ± 0.08	0.435	0.454 ± 0.08	0.351 ± 0.07	0.396	CN
O	0.0 ± 0.0	0.0 ± 0.0	0.0	1.0 ± 0.0	0.2545 ± 0.11	0.40	O-ND
ND	1.0 ± 0.0	0.433 ± 0.18	0.605				
Avg_a	0.565 ± 0.03	0.557 ± 0.03	0.561	0.609 ± 0.03	0.601 ± 0.03	0.605	Avg_m

Table 4: CRF++, Baseline, i.e. only using the word itself as feature, without surrounding context. $_a$ = all annotation classes: CP = Certainly Positive, PrP = Probably Positive, PoP = Possibly Positive, PoN = Possibly Negative, PrN = Probably Negative, CN = Certainly Negative, ND = Not Diagnosis and O = Other. $_m$ = merged classes: PrPoP = Probably and Possibly positive, PrPoN = Probably and Possibly Negative, O-ND = Other and Not diagnosis. P = Precision, R = Recall, F = F-measure, CI = Confidence Interval.

- Lab results: in some cases, lab results (or similar) seem to be highly indicative for specific diagnoses, but are not frequent enough in the training set or captured well in this model.
- Short sentences: in some cases, the sentence only contained the diagnostic statement itself, where the reasoning was documented in the remaining document. Here, it is evident that larger contexts may be very important.
- Longer discussions: for some diagnostic statements, a long discussion preceded the diagnostic statement itself, with many modifiers and speculations. In these cases, the local window did not model the factuality level well.

4 Discussion

In this study we present experiments on the impact of local features for an automatic factuality level classifier of Swedish diagnostic statements using the Stockholm EPR Diagnosis-Factuality Corpus. Using local context features improves results, in particular for annotation classes in the positive polarity, as well as for *certainly negative*. Preceding features are very valuable, both on their own and in combination with posterior features. Posterior features are not useful on their own. PoS information in combination with words and/or lemmas contributes to slight improvements. More complex language models are probably needed for improving results in the infrequent classes where context plays a larger role, as shown in the error analysis. Using syntactic features such as dependency parses and rules for linguistic constructions might be useful here (see e.g. Kilicoglu and Bergler (2008) and Velldal et al. (2010)). Moreover, in some cases we observe the need for including features at a cross-sentence level, and the inclusion of other types of features such as laboratory results. Some phrases might reflect ambiguous uses; for some diagnostic statements they are used for indicating high levels of certainty while for other diagnostic statement types they are used for indicating speculation. This is worth investigating further.

Merging annotation classes is fruitful for obtaining improved results, especially in the positive polarity. The same trends are not evident for the negative polarity, which might be due to the fact that the number of instances is much lower. Moreover, the two *possibly* levels were often confused even for the same annotator. These very low certainty levels might instead be merged into one *neutral* or *very low certainty* class, where polarity is not as important. The classes *not diagnosis* and *other* are too different to merge. Successful classification of the annotation class *other* probably needs more sophisticated language modeling, such as co-reference resolution, in the cases where instances are diagnostic statements referring to someone other than the patient.

In this corpus, we have a large amount of different diagnostic statement types. Grouping these and classifying factuality levels according to diagnostic statement type might lead to the insight that different types of features are indicative for different types of diagnostic statements. Moreover, the different annotation classes might also benefit from class-specific feature modeling, as was seen for *certainly negative*, where using the preceding context as features gave the best results.

	P$_a$ (95% CI)	R$_a$ (95% CI)	F$_a$	P$_m$ (95% CI)	R$_m$ (95% CI)	F$_m$	Merged
CP	**0.826** ± 0.03	**0.814** ± 0.03	0.82	**0.839** ± 0.03	**0.818** ± 0.03	0.828	CP
PrP	0.64 ± 0.07	0.576 ± 0.07	0.604	**0.825** ± 0.04	0.72 ± 0.05	0.769	PrPoP
PoP	0.643 ± 0.08	0.437 ± 0.08	0.521				
PoN	0.636 ± 0.20	0.304 ± 0.18	0.412	0.58 ± 0.08	0.55 ± 0.08	0.564	PrPoN
PrN	0.504 ± 0.09	0.528 ± 0.09	0.516				
CN	**0.789** ± 0.06	0.584 ± 0.08	**0.716**	0.79 ± 0.06	0.604 ± 0.08	0.686	CN
O	0.444 ± 0.19	0.16 ± 0.14	0.25				
ND	1.0 ± 0.0	0.6 ± 0.18	0.75	0.885 ± 0.08	0.418 ± 0.13	0.568	O-ND
Avg	0.744 ± 0.02	0.66 ± 0.03	0.699	0.805 ± 0.02	0.723 ± 0.02	0.762	Avg$_m$

Table 5: CRF++, window ± 4, word, lemma and PoS. $_a$ = all annotation classes: CP = Certainly Positive, PrP = Probably Positive, PoP = Possibly Positive, PoN = Possibly Negative, PrN = Probably Negative, CN = Certainly Negative, ND = Not Diagnosis and O = Other. $_m$ = merged classes: PrPoP = Probably and Possibly positive, PrPoN = Probably and Possibly Negative, O-ND = Other and Not diagnosis. P = Precision, R = Recall, F = F-measure, CI = Confidence Interval.

4.1 Limitations

The study design has some limitations. The concept of a diagnostic statement is not trivial and given the limited collection of diagnostic statements created in this work, the distribution of diagnostic statements might not reflect a real-world scenario. However, with this corpus, we have material for analyzing differences and similarities in how different diseases and diagnosis types are expressed with regards to factuality levels. We have shown that there are patterns among diagnostic statements, these should be analyzed further.

A further limitation of this model and the resulting corpus is the low number of annotations in some annotation classes. Merging intermediate probability levels improved results in the positive polarity, but in the negative polarity the same trend could not be observed. Here, we also had a much lower amount of instances. *Possibly negative* was a difficult class even for the same annotator, and might need further definitions. Moreover, the annotation class *other* is very complex, as it can be used for diagnostic statements referring to someone other than the patient. For these types of instances, co-reference resolution is needed, and adding further levels to the annotation model such as *perspective* or *source* might be useful (see e.g. Saurí (2008), Wilbur et al. (2006) and Rubin et al. (2006)). As the overall IAA results are relatively low Velupillai et al. (2011), further refinements in guidelines and resolving conflicting annotations to build a consensus corpus would be useful.

There are also limitations in the classification design; we have not tuned any parameters, nor have we compared with other learning algorithms. This should be further studied. Moreover, in order to increase the number of annotations and extending the corpus, active learning techniques could be very useful. The factuality level model with in total six levels of certainty could be considered as a continuum or scale, not necessarily as mutually independent classes. From this point of view, the factuality classification might be modeled differently, for instance through treating factuality as a continuous variable.

4.2 Significance of study

To our knowledge, no other studies have approached the study of factuality levels on a diagnosis basis in clinical Swedish. Our results show that the created model is feasible for an annotation task, resulting in a corpus that can be used for automatic classification. We see that speculative expressions in Swedish clinical assessments to a large extent are fairly consistent within a small context window, but that for improving results further, deeper language and feature models and might be needed. Automatic factuality level classification could be integrated in an information extraction system for clinicians and clinical researchers, where different factuality levels are distinguished. Choosing a broad approach gives further knowledge in how similarities and differences between different factuality levels among diagnostic statements in Swedish are expressed.

Acknowledgments

The author wishes to thank the two domain expert annotators Mia Kvist, M. D., PhD, and Prof. Gunnar Nilsson, M. D. for their valuable work on the creation of the annotation guidelines and the annotated corpus.

References

W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. 2001. Evaluation of negation phrases in narrative clinical reports. In *AMIA Symp*, pages 105–109.

M. M. Christopher and C. S. Hotz. 2004. Cytologic diagnosis: expression of probability by clinical pathologists. *Veterinary Clinical Pathology*, 33(2):84–95.

D. A. Clark. 1990. Verbal Uncertainty Expressions: A Critical Review of Two Decades of Research. *Current Psychology: Research & Reviews*, 9(3):203–235.

H. Dalianis, M. Hassel, and S. Velupillai. 2009. The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proc. ISHIMR 2009*, Kalmar, Sweden, October 14-16. Awarded best paper.

R. Farkas, V. Vincze, G. Móra, J. Csirik, and G. Szarvas. 2010. The conll-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proc. 14th CoNLL*, pages 1–12, Uppsala, Sweden, July. Association for Computational Linguistics.

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

H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman. 2009. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomedical Informatics*, 42:839–851.

J. L. Hobby, B. D. M. Tom, C. Todd, P. W. P. Bearcroft, and A. K. Dixon. 2000. Communication of doubt and certainty in radiological reports. *The British Journal of Radiology*, 73:999–1001, September.

R. Khorasani, D. W. Bates, S. Teeger, J. M. Rotschild, D. F. Adams, and S. E. Seltzer. 2003. Is terminology used effectively to convey diagnostic certainty in radiology reports? *Academic Radiology*, 10:685–688.

H. Kilicoglu and S. Bergler. 2008. Recognizing Speculative Language in Biomedical Research Articles: A Linguistically Motivated Perspective. In *Proc. BioNLP 2008*, pages 38–45, June.

O. Knutsson, J. Bigert, and V. Kann. 2003. A robust shallow parser for Swedish. In *Proc. Nodalida 2003*, Reykavik, Iceland.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289.

R. Morante, V. Van Asch, and W. Daelemans. 2010. Memory-based resolution of in-sentence scopes of hedge cues. In *Proc. 14th CoNLL*, pages 40–47, Uppsala, Sweden, July.

J. Pustejovsky, M. Verhagen, R. Saurí, J. Littman, R. Gaizauskas, G. Katz, I. Mani, R. Knippen, and A. Setzer. 2006. Timebank 1.2. Linguistic Data Consortium (LDC). Philadelphia, PA.

S. Renooij and C. Witteman. 1999. Talking probabilities: communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, 22:169–194.

V. L. Rubin, E. D. Liddy, and N. Kando. 2006. Certainty identification in texts: Categorization model and manual tagging results. In *Computing Affect and Attitutde in Text: Theory and Applications*. Springer.

R. Saurí. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Brandeis University.

B. Tang, X. Wang, X. Wang, B. Yuan, and S. Fan. 2010. A cascade method for detecting hedges and their scope in natural language text. In *Proc. 14th CoNLL*, pages 13–17, Uppsala, Sweden, July.

D. Timmermans. 1994. The Roles of Experience and Domain of Expertise in Using Numerical and Verbal Probability Terms in Medical Decisions. *Medical Decision Making*, 14:146–156.

E. Velldal, L. Øvrelid, and S. Oepen. 2010. Resolving speculation: Maxent cue classification and dependency-based scope rules. In *Proc. 14th CoNLL*, pages 48–55, Uppsala, Sweden, July.

S Velupillai, H. Dalianis, and M. Kvist. 2011. Factuality levels of diagnoses in swedish clinical text. In *Proceedings of MIE 2011*, pages 559–563. IOS Press, August.

V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S-11).

J. W. Wilbur, A. Rzhetsky, and H. Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7:356+, July.

A. S. Wu, B. H. Do, J. Kim, and D. L. Rubin. 2009. Evaluation of negation and uncertainty detection and its impact on precision and recall in search. *J Digit Imaging*.

Paper VI

Fine-grained Certainty Level Annotations Used for Coarser-grained E-health Scenarios – Certainty Classification of Diagnostic Statements in Swedish Clinical Text

Sumithra Velupillai and Maria Kvist. 2012. *Fine-grained Certainty Level Annotations Used for Coarser-grained E-health Scenarios – Certainty Classification of Diagnostic Statements in Swedish Clinical Text.* In In A. Gelbukh (Ed.): CICLing 2012 Part II, LNCS 7182, pp. 450–461. Springer-Verlag Berlin Heidelberg 2012.

Author contributions I was responsible for setting up the classification experiments and analyzing these results. Maria Kvist developed the scenarios and performed the qualitative error analysis. The article was written jointly.

Fine-grained Certainty Level Annotations Used for Coarser-grained E-health Scenarios

Certainty Classification of Diagnostic Statements in Swedish Clinical Text

Sumithra Velupillai[1] and Maria Kvist[1,2]

[1]Dept. of Computer and Systems Sciences (DSV)
Stockholm University, Forum 100, SE-164 40 Kista, Sweden
[2]Dept. of clinical immunology and transfusion medicine
Karolinska University Hospital, SE-171 76 Stockholm, Sweden
sumithra@dsv.su.se, maria.kvist@karolinska.se

Abstract. An important task in information access methods is distinguishing factual information from speculative or negated information. Fine-grained certainty levels of diagnostic statements in Swedish clinical text are annotated in a corpus from a medical university hospital. The annotation model has two polarities (positive and negative) and three certainty levels. However, there are many e-health scenarios where such fine-grained certainty levels are not practical for information extraction. Instead, more coarse-grained groups are needed. We present three scenarios: *adverse event surveillance*, *decision support alerts* and *automatic summaries* and collapse the fine-grained certainty level classifications into coarser-grained groups. We build automatic classifiers for each scenario and analyze the results quantitatively. Annotation discrepancies are analyzed qualitatively through manual corpus analysis. Our main findings are that it is feasible to use a corpus of fine-grained certainty level annotations to build classifiers for coarser-grained real-world scenarios: 0.89, 0.91 and 0.8 F-score (overall average).

Key words: Clinical documentation, Certainty level classification, Annotation granularity, Automatic Summary, Decision Support Alerts, Adverse Event Surveillance, E-health

1 Introduction

A challenging Natural Language Processing (NLP) task is to accurately extract relevant facts from clinical documentation. Speculative and negated information need to be distinguished from asserted information. Electronic health records are rich in factual and speculative opinions about a patient's clinical conditions, often expressed in free-text. This information is valuable for many e-health information access situations.

Certainty level classification in corpora is a growing research area in the domain of computational linguistics and information access, in particular for domain-specific purposes.

1.1 Related Work

In the interdisciplinary area of clinical natural language processing, several studies have targeted the issue of accurate information extraction by including negations and speculations in the information extraction model. In [1], assertion classification (present, absent or uncertain) is performed on medical problems. Rule-based and machine-learning techniques are used and compared. The machine-learning method, using features in a window of ± 4, outperforms the rule-based method. Contextual features, including negation, are used for classifying clinical conditions in [2]. In this study, uncertainties are, however, not modeled. The BioScope corpus contains annotations for negation and uncertainty [3] on a sentence level, with a subset of clinical radiology reports (the remaining corpus contains biomedical research articles and abstracts). The 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text [4] included a subtask for classifying assertion levels of medical problems. The top performing system on the assertion task obtained an F-score of 0.94 [5]. However, certainty levels are not modeled on a fine-grained level in these studies. In other domains, more fine-grained certainty levels are proposed, e.g. [6], [7] and [8]. The above-mentioned studies are performed on English.

1.2 Aim and Objective

In this work, we use a Swedish clinical corpus with diagnostic statements annotated at a fine-grained certainty level [9] to build coarser-grained classifications reflecting three e-health scenarios where this distinction differs for each scenario: *adverse event surveillance*, *decision support alerts* and *automatic summaries*. Creating annotation models is costly. Using fine-grained models for several purposes might be an efficient approach. Our aim is to study whether an existing corpus with fine-grained certainty level annotations can be used for creating multiple scenario-specific certainty level groups, and to study whether limitations in the existing corpus are transferred as limitations in the chosen scenarios. We build automatic classifiers for each scenario, and analyze the results quantitatively. Annotation discrepancies in the corpus are scrutinized and analyzed qualitatively. To our knowledge, no previous research has used fine-grained certainty level annotations for building several use cases with coarse-grained certainty level groups, nor has this been performed on Swedish clinical text.

2 Method

A Swedish clinical corpus annotated for fine-grained certainty levels on a diagnostic statement level was used[1]. The fine-grained classification was collapsed into groups for three different coarse-grained e-health scenarios. Automatic classifiers for each scenario were built, using Conditional Random Fields and simple

[1] Approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm) permission number 2009/1742-31/5.

local context features. Results were evaluated quantitatively through precision, recall and F-score. Annotation discrepancies were analyzed qualitatively through manual corpus analysis.

2.1 Corpus Characteristics

The corpus consists of assessment entries from a medical emergency ward in the Stockholm area. In these entries, reasoning about the patient's status and diseases is documented. Diagnostic statements were automatically tagged in the clinical notes and the annotators judged their certainty levels [9]. An example entry is shown in Figure 1.

Oklart vad pats symtom kan komma av. Ingen säker <D>infektion</D>. Inga tecken till inflammatorisk sjukdom eller <D>allergi</D>. Reflux med irritation av luftrör och således hosta? Dock har pat ej haft några symtom på <D>refluxesofagit</D>. Ingen ytterligare akut utredning är befogad. Hänvisar till pats husläkare för fortsatt utredning.

Unclear what patient's (abbr.) symptoms arise from. No certain <D>infection</D>. No signs of inflammatory disease or <D>allergy</D>. Reflux with irritation of airways and therefore cough? But pat has not had any symptoms of <D>refluxoesophagitis</D>.No further urgent investigation required. Refer to pats GP for continued investigation..

Fig. 1. Example assessment entry. D = Diagnostic statement. Each marked diagnostic statement was judged for certainty levels. In this case, the diagnostic statements *infektion* (infection), *allergi* (allergy) and *refluxesofagit* (refluxoesophagitis) were to be assigned one of the six certainty level annotation classes.

The annotators were shown the entire assessment entry and were asked to annotate each marked diagnostic statement into one of the six certainty level annotation classes[2]. The certainty levels are modeled in two polarities: *positive* and *negative*, as well as certainty level: *certain*, *probable* or *possible*, see Figure 2. Overall Inter- and Intra Annotator (IAA) results, measured on a subset of the total amount of annotations, were 0.7/0.58 and 0.73/0.6 F-measure/Cohens κ, respectively. This subset was used for the qualitative error analysis. The corpus along with guidelines and further analysis are presented in [9][3].The full corpus consists of 5 473 assessment entries, 6 186 annotated diagnostic statements and 64 832 tokens (7 464 types) annotated by one annotator. Common error types in the annotations are shown in Table 1. We see similarities in both inter- and intra-annotator discrepancies, the most common error type is *1-step* (66% and 69%).

[2] Other classes were also included, but are not analyzed in this work.

[3] The annotators were two senior physicians, accustomed to reading and writing medical records.

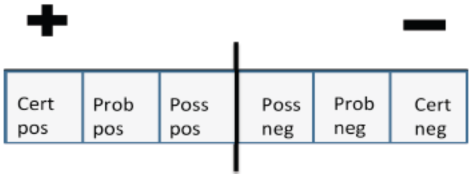


Fig. 2. Fine-grained certainty level classification of diagnostic statements into two polarities and three levels of certainty, in total six classes.

Table 1. The most common error types in the annotated corpus. 1-step = discrepancy in one step, e.g. *certainly negative* vs *probably negative*. Certain/Uncertain = discrepancy between the highest level of certainty and intermediate certainty level classes (*probably* or *possibly*). Polarity = discrepancy in *positive* vs *negative*. n_{inter} = inter-annotator analysis. n_{intra} = intra-annotator analysis

Type	n_{inter}	%	n_{intra}	%
1-step	408	66	284	69
Certain/Uncertain	270	44	191	46
Polarity	99	16	58	14
Total	614	100	411	100

2.2 E-health Scenarios

We define three tentative e-health scenarios: *adverse event surveillance*, *decision support alerts* and *automatic summaries*. These scenarios reflect different needs when it comes to distinguishing and defining the boundaries between certainty levels. The different coarse-grained certainty level groups for the chosen scenarios relate to the original fine-grained classification model as shown in Figure 3. The fine-grained classes *certainly positive*, *probably positive*, *possibly positive*, *possibly negative*, *probably negative* and *certainly negative* are included and excluded in different ways for each scenario. The scenarios are further described below.

Adverse event surveillance One instrument used for surveillance of adverse events in hospital care is the Global Trigger Tool [10]. Here, a number of triggers are defined and used for extraction of records which are subsequently manually scrutinized for adverse events. Automation of the trigger identification procedure and extraction of records saves manual labor, and is presently employed at Karolinska University Hospital for triggers in the structured parts of medical records. Further development of this system would be automatic identification of some of these triggers found in the free-text part of health records, and to this add trigger negation detection. Only cases that are negated with the highest possible level of certainty should be excluded in a potential trigger extraction system. Accurate exclusion of negated cases would lower the overall manual work load. Hence, in this scenario, we get a binary grading: *existence* (at some level of certainty) or *no existence* (at the most certain level). All five annotation classes except *certainly negative* are collapsed into the *existence* grade.

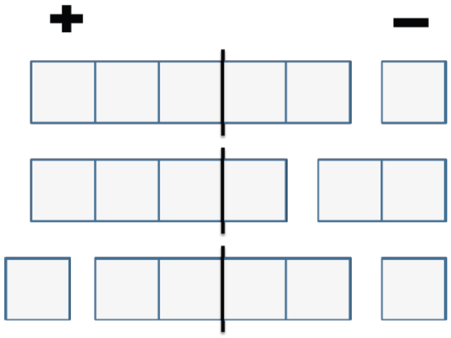


Fig. 3. Modeling e-health use cases by utilizing fine-grained certainty level annotations for coarser-grained classifications, reflecting scenario-specific needs. Top: *adverse event surveillance*. Middle: *decision support alerts*. Bottom: *automatic summaries*.

Decision support alerts In this scenario, the important distinction in an information access setting, is to flag whenever there is a plausible diagnosis [11]. An example of an automated application would be a decision support: if a plausible case is identified, guidelines or other similar recommendations are automatically shown to the clinician in order to take suitable action. Another potential application would be alerting the clinician who is medically responsible for a patient: a nurse documenting a plausible condition produces an automatic alert to the responsible clinician to take action. Separating positive (or near positive) cases from negative cases is important here. Using the fine-grained certainty level annotation classes, we collapse all positive classes as well as *possibly negative*[4] to one group: *plausible existence*. At the negative polarity *probably negative* and *certainly negative* are collapsed into: *no plausible existence*.

Automatic summaries When presented with a new patient, an overview, e.g. textual summary, would help the clinician to get an overall impression of earlier diagnoses and health history. A presentation of diagnoses that have been affirmed, excluded, or discussed as a possibility need to be processed by an automatic information extraction system that can distinguish such cases [12]. Moreover, from a different perspective, patients might be interested in obtaining an overview of their own health records in a similar manner, in order to

[4] The two classes *possibly positive* and *possibly negative* are in this case judged together as a joint middle class.

understand and participate in her or his clinical situation. In this scenario, we use *affirmed* and *negated* as two separate groups, and the remaining intermediate, speculative classes are collapsed into one *speculated* group. Hence, we get a multi-class classification problem with three class labels.

2.3 Automatic Classification and Evaluation

We have used Conditional Random Fields [13], as implemented in CRF++ [5] with default parameter settings for building token level classifiers. All sentences containing diagnostic statements annotated for certainty levels were tokenized[6], and local context features (word, lemma and Part-of-Speech (PoS) tags[7]) with a window of ±4 were used for each token, as this setting produces best results [15]. Each diagnostic statement token was assigned exactly one certainty level class, all other tokens were assigned the class *NONE*.

The corpus was divided into a training set (80%, 4 367 sentences, 4 929 diagnostic statements, 51 523 tokens) and a test set (20%, 1 106 sentences, 1 257 diagnostic statements, 13 309 tokens), with a stratified distribution of annotation class labels, see Table 2.

Table 2. Coarser-grained certainty level annotation class labels, training and test set: number of class instances and percentages in parentheses. S-1 = *adverse event surveillance.* S-2 = *decision support alerts.* S-3 = *automatic summaries.*

		Training set			Test set		
Scenario	Group	S-1 (%)	S-2 (%)	S-3 (%)	S-1 (%)	S-2 (%)	S-3 (%)
S-1	existence	4 372 (89)			1 103 (88)		
	no existence	557 (11)			154 (12)		
S-2	plausible existence		3 934 (80)			995 (80)	
	no plausible existence		995 (20)			262 (20)	
S-3	affirmed			2 463 (50)			625 (50)
	speculated			1 909 (39)			478 (38)
	negated			557 (11)			154 (12)
Total		4 929 (100)	4 929 (100)	4 929 (100)	1 257 (100)	1 257 (100)	1 257 (100)

Results were measured with precision, recall and F-measure, using the CoNLL 2010 Shared task evaluation script conlleval.pl[8]. 95% confidence intervals were calculated for precision and recall. Two baselines were used: majority class baseline and a classifier with no local context features, i.e. the diagnostic statement itself is used as the only feature.

[5] http://crfpp.sourceforge.net/#source

[6] multi-word diagnostic statements such as *heart attack* were concatenated and treated as one token

[7] using a general Swedish tagger [14]

[8] http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt

3 Results

In this section we present automatic classification results for each e-health scenario, as well as a qualitative error analysis based on the annotated corpus. In the error analysis, we find that difficulties in the distinction between the fine-grained classes *probably negative* and *certainly negative* seem to be the source of most errors in the corpus, and Inter- and Intra-Annotator Agreement (IAA) problems are therefore reflected differently in the three scenarios. We also find that results in the error analysis for the coarse-grained grades are correlated with the distribution of diagnostic statements along the scale of the fine-grained certainty levels. Some diagnostic statements are evenly distributed along this scale, while others are more frequent in the positive polarity (e.g. hypertension, different types of arrythmias, hyperventilation, allergies, different skin diseases) or negative polarity (e.g. thrombosis and ischemia), as shown in [9]. This reflects the clinical need to negate certain disorders in the documentation, but not others. The discrepancies reflect difficulties in judging certainty for different types of diagnostic statements at the respective polarities, with different types of linguistic and clinical assessment problems arising at the respective polarities accordingly.

3.1 Adverse Event Surveillance

In this scenario, we have a binary classification problem: *existence* and *no existence*. This could also be considered similar as a negation detection task.

Classification results In Table 3, results for the baseline (without context features) and for the classifier using a local context window of ±4 is shown. A majority class baseline is 88%. In general, using local context features improves results compared to both baselines (0.89 F-score), but compared to the majority class baseline only a slight improvement is seen. For the minority class *no existence*, context features increase results considerably, in particular for precision (from 0.54 to 0.83), although recall is low (0.51).

Table 3. Classification results for the scenario *adverse event surveillance*. Binary classification: *existence* and *no existence*. P = Precision, R = Recall, F = F-score. 95% confidence intervals are given (±). Majority class baseline = 88%. Baseline = no context features, Local context = word, lemma and PoS-tag, window ±4.

	Baseline			Local context		
Class label	P	R	F	P	R	F
existence	0.53±0.03	0.98±0.01	0.68	0.93±0.01	0.91±0.02	0.92
no existence	0.54±0.08	0.14±0.05	0.23	0.83±0.06	0.51±0.08	0.63
Total	0.53±0.03	0.88±0.02	0.66	0.92±0.01	0.86±0.02	0.89

Error analysis The lower results for *no existence* in the automatic classification for this scenario appears to be connected to known difficulties in the distinction between *probably negative* and *certainly negative* in the annotated corpus. There are not many errors in assigning polarity (see Table 1), i.e. the diagnostic statements are clearly in the negative polarity, but the strength of the negation has been judged differently in many cases. Part of the errors are due to the lexical context surrounding the diagnostic statement. For instance, the phrase *inga hållpunkter för* (no indicators of), has been inconsistently interpreted. These cases are also a source of many errors in the automatic classification. Moreover, these inconsistencies are often related to diagnostic statements belonging to diagnosis types that are difficult to exclude, such as *DVT* (deep venous thrombosis), where complete exclusion is clinically difficult. Speculations arise around these diagnosis types because of important severe consequences if missed or misjudged. There are also inconsistencies that depend on whether the annotator(s) have judged the local or global context (i.e. the whole assessment entry, or only the current sentence). Modifiers such as *liten*, e.g. *liten misstanke* (small suspicion), are an interesting source of errors: these can be interpreted differently depending on whether emphasis is put on *misstanke* (suspicion), or *liten* (small), and would need to be defined further in the guidelines.

3.2 Decision Support Alerts

In this scenario we need two groups. The classification task is hence modeled with binary class labels: *plausible existence* and *no plausible existence*.

Classification results In Table 4, results are shown for the classification baseline as well as for using local context features. A majority class assignment is 80%. Overall results are improved using local context features (from 0.61 F-score to 0.91), and are also improved compared to the majority class baseline. For the minority class *no plausible existence*, results are considerably improved both for precision (from 0.72 to 0.92) and recall (from 0.22 to 0.79).

Table 4. Classification results for the scenario *alerts for decision support*. Binary classification: *plausible existence* and *no plausible existence*. P = Precision, R = Recall, F = F-score. 95% confidence intervals are given (±). Majority class baseline = 80%. Baseline = no context features, Local context = word, lemma and PoS-tag, window ±4.

	Baseline			Local context		
Class label	P	R	F	P	R	F
plausible existence	0.48±0.03	0.97±0.01	0.64	0.95±0.01	0.90±0.02	0.92
no plausible existence	0.72±0.05	0.22±0.05	0.34	0.92±0.03	0.79±0.05	0.85
Total	0.49±0.03	0.82±0.02	0.61	0.94±0.01	0.88±0.02	0.91

Error analysis The boundary in the fine-grained classification model is shifted towards the positive polarity, as compared to the *adverse event surveillance* scenario. The main source of errors lies in cases where certain clinical exclusion is very difficult, due to the nature of the diagnosis itself (e.g. *DVT*). Another source of errors lies in cases where tests have been performed in order to exclude a specific diagnosis. These cases are difficult since performing a test in itself is an indication that there is a risk of this diagnosis, but from the surrounding context it can be evident that the diagnosis is highly unlikely.

3.3 Automatic Summaries

In this scenario, we need three grades, resulting in a multi-class classification problem: *affirmed*, *speculated*, and *negated*.

Classification results A majority class assignment (*affirmed*) is 50%. In Table 5 results for the classifiers (baseline, and context window ±4) are shown. Using local context features result in a considerable improvement for all classes (0.8 F-score, overall average, compared to 0.5, both baselines). Recall for *negated* is, however, relatively low (0.55).

Table 5. Classification results for the scenario *automatic summary*. Multi-class classification: *affirmed*, *speculated* and *negated*. P = Precision, R = Recall, F = F-score. 95% confidence intervals are given (±). Majority class baseline = 50%. Baseline = no context features, Local context = word, lemma and PoS-tag, window ±4.

	Baseline			Local context		
Class label	P	R	F	P	R	F
affirmed	0.79±0.03	0.72±0.03	0.75	0.87±0.03	0.81±0.03	0.84
speculated	0.25±0.02	0.77±0.02	0.38	0.81±0.02	0.77±0.02	0.79
negated	0.50±0.08	0.18±0.08	0.27	0.81±0.06	0.55±0.08	0.66
Total	0.40±0.03	0.67±0.03	0.50	0.84±0.02	0.76±0.02	0.80

Error analysis In this scenario, we focus on an error analysis in the positive polarity, which is not covered in the other two scenarios. These errors mostly reflect difficulties in distinguishing between *probably positive* and *certainly positive* in the annotated corpus. A majority of the cases are due to linguistic markers such as *misstänkt <D>x</D>* (suspected <D>x</D>) or *kliniska tecken på <D>x</D>* (clinical signs of <D>x</D>). We see more discrepancies in the annotations concerning diagnosis types determined by subjective judgement, e.g. *hyperventilering* (hyperventilation) and *panikångest* (panic disorder) than diagnosis types that are measured objectively, e.g. *hypertoni* (hypertension). A difference in the judgments made by the human annotators lies in whether they have based their judgments on clinical knowledge or linguistic markers, e.g. *Urinprov pos. därför troligen urinvägsinf.* (Urine sample pos. thus probably urinary tract inf.) We observe some difficult cases for chronic diseases. For instance, the example *troligen stressutlöst astma* (probably stress triggered asthma), could be interpreted as *certainly positive* in the sense that the patient is diagnosed with asthma, or as *probably positive* in the sense that this particular event of an asthma attack is probably triggered by stress.

4 Analysis and Discussion

In this study we present work using a corpus annotated with fine-grained certainty classes on a diagnostic statement level, for coarser-grained e-health scenarios. We present three scenarios: *adverse event surveillance*, *decision support alerts* and *automatic summaries.* These scenarios are real-world situations where computerized support is beneficial [12], and where Natural Language Processing techniques involving negation handling may be useful [11]. Each scenario requires different certainty level models, and we collapse classes from the fine-grained classification model into three different coarser-grained groups. We build classifiers using local context features for each scenario. A qualitative analysis on annotation errors deepens the understanding of problems in the boundaries between certainty level classes. We observe promising results by the automatic classifiers for all three scenarios (0.89 F-score (*adverse event surveillance*), 0.91 F-score (*decision support alerts*) and 0.8 F-score (*summaries*), overall average). Our main findings are that it is feasible to use a fine-grained certainty level classification model of diagnostic statements for building coarser-grained e-health scenarios. Although overall IAA is relatively low for the fine-grained model [9], most errors are found in the 1-step borders between the fine-grained levels, thus yielding higher IAA for coarser-grained situations. Annotation discrepancies in intermediate certainty level classes do not pose problems when classes are collapsed into coarser-grained certainty level groups. However, there are some problematic issues, in particular in the distinction between *probably negative* and *certainly negative* in the fine-grained classification model, which need to be further defined in the annotation guidelines. This problem becomes evident when looking at the results for the automatic classifier for the scenario *adverse event surveillance*, where recall in the minority class *no existence* is 0.51. Whether the fine-grained model is considered a sliding scale, or a two-step decision (polarity followed by certainty level) by the annotators is also a factor that should be studied further and need to be clarified when creating fine-grained certainty level annotation tasks.

Previous work (e.g. [1], [2], [4], [5]), on similar tasks are difficult to compare for several reasons. For instance, the certainty level models, annotation tasks, corpora and classification approaches are different to those employed in this work. However, some general trends are observed, such as the problem of skewed class distributions and ambiguity of context cues. Interestingly, local context features in a window of ±4 are shown to be useful also for English [1], as well as for Swedish [15]. Cross-lingual studies would be a very interesting continuation of this work. Moreover, the fine-grained certainty levels might also be useful as features for other (higher-level) classification tasks.

Qualitative studies on terminologies used for expressing diagnostic certainties reveal that intermediate probabilities are more often difficult to agree on among human (clinical) evaluators ([16] and [17]), which is in line with our observations. This is an inherently subjective task, and it is not trivial to define what upper performance bounds would be for classifiers.

4.1 Limitations

The automatic classifiers have been built on annotations by one annotator only, not on a consensus set by several annotators. Overall results are also affected by skewed class distributions, results for minority classes need to be further analyzed. Moreover, other classification algorithms should be tested. We treat this task as a token level classification problem, using Conditional Random Fields for classification. Other classification algorithms or representations might be better suited for this task, this should be studied further and compared. More detailed feature analysis is also needed, as well as under- or oversampling data for dealing with the problem of skewed class distributions. For instance, no global context features have been used, nor any clinical domain-knowledge based features, such as test results.

Moreover, the qualitative error analysis is performed on annotations by two annotators, and only on a subset of the original corpus. A correlation between inter-annotator discrepancies and the errors resulting from the classifiers should be analyzed in future studies.

4.2 Significance of Study

Our results are valuable for further work on creating accurate information extraction methods for clinical real-world cases. In health care, there is a constant need for quick decisions based on earlier documentation. This is often complicated by the accumulating mass of text surrounding every patient case. Automatic text processing for applications such as decision support and summaries or overviews, adapted to natural language, would facilitate the clinical workday. Also, automation of surveillance tools for adverse events can assist in improvement of hospital care. This study indicates that it is possible to use a general resource for specific scenario solutions. Instead of creating, in this case, three coarse-grained annotation tasks and subsequent corpora, one fine-grained model can be used for several purposes successfully. To our knowledge, no previous research has used fine-grained certainty level annotations for building several coarse-grained use cases, nor has this been studied on Swedish clinical text.

Acknowledgments We would like to express our appreciations to the anonymous and known reviewers for invaluable comments and suggestions for this paper.

References

1. Uzuner, Ö., Zhang, X., Sibanda, T.: Machine Learning and Rule-based Approaches to Assertion Classification. JAMIA **16** (2009) 109–115
2. Harkema, H., Dowling, J.N., Thornblade, T., Chapman, W.W.: ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics **42** (2009) 839–851
3. Vincze, V., Szarvas, G., Farkas, R., Móra, G., Csirik, J.: The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics **9** (2008)
4. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. JAMIA **18** (2011) 552–556
5. de Bruijn, B., Cherry, C., Kiritchenko, S., Martin, J., Zhu, X.: Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. JAMIA **18** (2011) 557–562
6. Wilbur, J.W., Rzhetsky, A., Shatkay, H.: New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinformatics **7** (2006) 356+
7. Rubin, V.L., Liddy, E.D., Kando, N.: Certainty identification in texts: Categorization model and manual tagging results. In: Computing Affect and Attitutde in Text: Theory and Applications. Springer (2006)
8. Saurí, R.: A Factuality Profiler for Eventualities in Text. PhD thesis, Brandeis University (2008)
9. Velupillai, S., Dalianis, H., Kvist, M.: Factuality Levels of Diagnoses in Swedish Clinical Text. In Moen, A., Andersen, S.K., Aarts, J., Hurlen, P., eds.: Proc. XXIII Intl. Conf. of the European Federation for Medical Informatics, Oslo, IOS Press (2011) 559 – 563
10. F.A., G., Resar, R.: IHI Global Trigger Tool for Measuring Adverse Events (Second Edition). IHI Innovation Series white paper. Cambridge, Massachusetts: Institute for Healthcare Improvement (2009)
11. Denny, J.C., Miller, R.A., Waitman, L.R., Arrieta, M.A., Peterson, J.F.: Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor. IJMI **78** **S** **1** (2009) S34–S42
12. Kvist, M., Skeppstedt, M., Velupillai, S., Dalianis, H.: Modeling human comprehension of Swedish medical records for intelligent access and summarization systems, a physician's perspective. In: Proc. 9th Scandinavian Conf. on Health Informatics, SHI, Oslo (2011)
13. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML. (2001) 282–289
14. Knutsson, O., Bigert, J., Kann, V.: A robust shallow parser for Swedish. In: Proceedings of Nodalida 2003, Reykavik, Iceland (2003)
15. Velupillai, S.: Automatic Classification of Factuality Levels – A Case Study on Swedish Diagnoses and the Impact of Local Context. In: Proc. 4th Intl. Symp. on Languages in Biology and Medicine (LBM 2011), Singapore (2011)
16. Khorasani, R., Bates, D.W., Teeger, S., Rotschild, J.M., Adams, D.F., Seltzer, S.E.: Is terminology used effectively to convey diagnostic certainty in radiology reports? Academic Radiology **10** (2003) 685–688
17. Hobby, J.L., Tom, B.D.M., Todd, C., Bearcroft, P.W.P., Dixon, A.K.: Communication of doubt and certainty in radiological reports. The British Journal of Radiology **73** (2000) 999–1001

DEPARTMENT OF COMPUTER AND SYSTEMS SCIENCES
Stockholm University/KTH
www.dsv.su.se
Ph.D. theses:

No 91-004 **Olsson, Jan**
An Architecture for Diagnostic Reasoning Based on Causal Models
No 93-008 **Orci, Terttu**
Temporal Reasoning and Data Bases
No 93-009 **Eriksson, Lars-Henrik**
Finitary Partial Definitions and General Logic
No 93-010 **Johannesson, Paul**
Schema Integration Schema Translation, and Interoperability in Federated Information Systems
No 93-018 **Wangler, Benkt**
Contributions to Functional Requirements Modelling
No 93-019 **Boman, Magnus**
A Logical Specification for Federated Information Systems
No 93-024 **Rayner, Manny**
Abductive Equivalential Translation and its Application to Natural-Language Database Interfacing
No 93-025 **Idestam-Almquist, Peter**
Generalization of Clauses
No 93-026 **Aronsson, Martin**
GCLA: The Design, Use, and Implementation of a Program Development
No 93-029 **Boström, Henrik**
Explanation-Based Transformation of Logic programs
No 94-001 **Samuelsson, Christer**
Fast Natural Language Parsing Using Explanation-Based Learning
No 94-003 **Ekenberg, Love**
Decision Support in Numerically Imprecise Domains
No 94-004 **Kowalski, Stewart**
IT Insecurity: A Multi-disciplinary Inquiry
No 94-007 **Asker, Lars**
Partial Explanations as a Basis for Learning
No 94-009 **Kjellin, Harald**
A Method for Acquiring and Refining Knowledge in Weak Theory Domains
No 94-011 **Britts, Stefan**
Object Database Design
No 94-014 **Kilander, Fredrik**
Incremental Conceptual Clustering in an On-Line Application
No 95-019 **Song, Wei**
Schema Integration: - Principles, Methods and Applications
No 95-050 **Johansson, Anna-Lena**
Logic Program Synthesis Using Schema Instantiation in an Interactive Environment
No 95-054 **Stensmo, Magnus**
Adaptive Automated Diagnosis
No 96-004 **Wærn, Annika**
Recognising Human Plans: Issues for Plan Recognition in Human - Computer Interaction
No 96-006 **Orsvärn, Klas**
Knowledge Modelling with Libraries of Task Decomposition Methods
No 96-008 **Dalianis, Hercules**
Concise Natural Language Generation from Formal Specifications
No 96-009 **Holm, Peter**
On the Design and Usage of Information Technology and the Structuring of Communication and Work
No 96-018 **Höök, Kristina**
A Glass Box Approach to Adaptive Hypermedia
No 96-021 **Yngström, Louise**
A Systemic-Holistic Approach to Academic Programmes in IT Security

No 97-005 **Wohed, Rolf**
A Language for Enterprise and Information System Modelling
No 97-008 **Gambäck, Björn**
Processing Swedish Sentences: A Unification-Based Grammar and Some Applications
No 97-010 **Kapidzic Cicovic, Nada**
Extended Certificate Management System: Design and Protocols
No 97-011 **Danielson, Mats**
Computational Decision Analysis
No 97-012 **Wijkman, Pierre**
Contributions to Evolutionary Computation
No 97-017 **Zhang, Ying**
Multi-Temporal Database Management with a Visual Query Interface
No 98-001 **Essler, Ulf**
Analyzing Groupware Adoption: A Framework and Three Case Studies in Lotus Notes Deployment
No 98-008 **Koistinen, Jari**
Contributions in Distributed Object Systems Engineering
No 99-009 **Hakkarainen, Sari**
Dynamic Aspects and Semantic Enrichment in Schema Comparison
No 99-015 **Magnusson, Christer**
Hedging Shareholder Value in an IT dependent Business society - the Framework BRITS
No 00-004 **Verhagen, Henricus**
Norm Autonomous Agents
No 00-006 **Wohed, Petia**
Schema Quality, Schema Enrichment, and Reuse in Information Systems Analysis
No 01-001 **Hökenhammar, Peter**
Integrerad Beställningsprocess vid Datasystemutveckling
No 01-008 **von Schéele, Fabian**
Controlling Time and Communication in Service Economy
No 01-015 **Kajko-Mattsson, Mira**
Corrective Maintenance Maturity Model: Problem Management
No 01-019 **Stirna, Janis**
The Influence of Intentional and Situational Factors on Enterprise Modelling Tool Acquisition in Organisations
No 01-020 **Persson, Anne**
Enterprise Modelling in Practice: Situational Factors and their Influence on Adopting a Participative Approach
No 02-003 **Sneiders, Eriks**
Automated Question Answering: Template-Based Approach
No 02-005 **Eineborg, Martin**
Inductive Logic Programming for Part-of-Speech Tagging
No 02-006 **Bider, Ilia**
State-Oriented Business Process Modelling: Principles, Theory and Practice
No 02-007 **Malmberg, Åke**
Notations Supporting Knowledge Acquisition from Multiple Sources
No 02-012 **Männikkö-Barbutiu, Sirkku**
SENIOR CYBORGS- About Appropriation of Personal Computers Among Some Swedish Elderly
People
No 02-028 **Brash, Danny**
Reuse in Information Systems Development: A Qualitative Inquiry
No 03-001 **Svensson, Martin**
Designing, Defining and Evaluating Social Navigation
No 03-002 **Espinoza, Fredrik**
Individual Service Provisioning
No 03-004 **Eriksson-Granskog, Agneta**
General Metarules for Interactive Modular Construction of Natural Deduction Proofs
No 03-005 **De Zoysa, T. Nandika Kasun**
A Model of Security Architecture for Multi-Party Transactions
No 03-008 **Tholander, Jakob**
Constructing to Learn, Learning to Construct - Studies on Computational Tools for Learning

No 03-009 **Karlgren, Klas**
Mastering the Use of Gobbledygook - Studies on the Development of Expertise Through Exposure to Experienced Practitioners' Deliberation on Authentic Problems
No 03-014 **Kjellman, Arne**
Constructive Systems Science - The Only Remaining Alternative?
No 03-015 **Rydberg Fåhræus, Eva**
A Triple Helix of Learning Processes - How to cultivate learning, communication and collaboration among distance-education learners
No 03-016 **Zemke, Stefan**
Data Mining for Prediction - Financial Series Case
No 04-002 **Hulth, Anette**
Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction
No 04-011 **Jayaweera, Prasad M**.
A Unified Framework for e-Commerce Systems Development: *Business Process Patterns Perspective*
No 04-013 **Söderström, Eva**
B2B Standards Implementation: Issues and Solutions
No 04-014 **Backlund, Per**
Development Process Knowledge Transfer through Method Adaptation, Implementation, and Use
No 05-003 **Davies, Guy**
Mapping and Integration of Schema Representations of Component Specifications
No 05-004 **Jansson, Eva**
Working Together when Being Apart – An Analysis of Distributed Collaborative Work through ICT from an Organizational and Psychosocial Perspective
No 05-007 **Cöster, Rickard**
Algorithms and Representations for Personalised Information Access
No 05-009 **Ciobanu Morogan, Matei**
Security System for Ad-hoc Wireless Networks based on Generic Secure Objects
No 05-010 **Björck, Fredrik**
Discovering Information Security Management
No 05-012 **Brouwers, Lisa**
Microsimulation Models for Disaster Policy Making
No 05-014 **Näckros, Kjell**
Visualising Security through Computer Games
Investigating Game-Based Instruction in ICT Security: an Experimental approach
No 05-015 **Bylund, Markus**
A Design Rationale for Pervasive Computing
No 05-016 **Strand, Mattias**
External Data Incorporation into Data Warehouses
No 05-020 **Casmir, Respickius**
A Dynamic and Adaptive Information Security Awareness (DAISA) approach
No 05-021 **Svensson, Harald**
Developing Support for Agile and Plan-Driven Methods
No 05-022 **Rudström, Åsa**
Co-Construction of Hybrid Spaces
No 06-005 **Lindgren, Tony**
Methods of Solving Conflicts among Induced Rules
No 06-009 **Wrigstad, Tobias**
Owner-Based Alias Management
No 06-011 **Skoglund, Mats**
Curbing Dependencies in Software Evolution
No 06-012 **Zdravkovic, Jelena**
Process Integration for the Extended Enterprise
No 06-013 **Olsson Neve, Theresia**
Capturing and Analysing Emotions to Support Organisational Learning:
The Affect Based Learning Matrix
No 06-016 **Chaula, Job Asheri**
A Socio-Technical Analysis of Information Systems Security Assurance
A Case Study for Effective Assurance

No 06-017 **Tarimo, Charles N.**
ICT Security Readiness Checklist for Developing Countries:
A Social-Technical Approac**h**

No 06-020 **Kifle Gelan, Mengistu**
A Theoretical Model for Telemedicine
- Social and Value Outcomes in Sub-Saharan Africa

No 07-001 **Fernaeus, Ylva**
Let’s Make a Digital Patchwork
Designing for Children’s Creative Play with Programming Materials

No 07-003 **Bakari, Jabiri Kuwe**
A Holistic Approach for Managing ICT Security in Non-Commercial Organisations
A Case Study in a Developing Country

No 07-004 **Sundholm, Hillevi**
Spaces within Spaces: The Construction of a Collaborative Reality

No 07-005 **Hansson, Karin**
A Framework for Evaluation of Flood Management Strategies

No 07-007 **Aidemark, Jan**
Strategic Planning of Knowledge Management Systems
- A Problem Exploration Approach

No 07-009 **Jonsson, Martin**
Sensing and Making Sense
Designing Middleware for Context Aware Computing

No 07-013 **Kabilan, Vandana**
Ontology for Information Systems (O4IS) Design Methodology:
Conceptualizing, Designing and Representing Domain Ontologies

No 07-014 **Mattsson, Johan**
Pointing, Placing, Touching
- Physical Manipulation and Coordination Techniques for Interactive Meeting Spaces

No 07-015 **Kessler, Anna-Maria**
A Systemic Approach Framework for Operational Risk
- SAFOR

No 08-001 **Laaksolahti, Jarmo**
Plot, Spectacle and Experience: Contributions to the design and evaluation of Interactive Storytelling

No 08-002 **Van Nguyen Hong**
Mobile Agent Approach to Congestion Control in Heterogeneous Networks

No 08-003 **Rose-Mharie Åhlfeldt**
Information Security in Distributed Healthcare
- Exploring the Needs for Achieving Patient Safety and Patient Privacy

No 08-004 **Sara Ljungblad**
Beyond users:
Grounding technology in experience

No 08-005 **Eva Sjöqvist**
Electronic Mail and its Possible Negative Aspects in Organizational Contexts

No 08-006 **Thomas Sandholm**
Statistical Methods for Computational Markets
– Proportional Share Market Prediction and Admission Control

No 08-007 **Lena Aggestam**
IT-supported Knowledge Repositories:
Increasing their Usefulness by Supporting Knowledge Capture

No 08-008 **Jaana Nyfjord**
Towards Integrating Agile Development and Risk Management

No 08-009 **Åsa Smedberg**
Online Communities and Learning for Health
- The Use of Online Health Communities and Online Expertise for People with Established Bad Habits

No 08-010 **Martin Henkel**
Service-based Processes
- Design for Business and Technology

No 08-012 **Jan Odelstad**
Many-Sorted Implicative Conceptual Systems

No 09-001 **Marcus Nohlberg**
Securing Information Assets
- Understanding, Measuring and Protecting against Social Engineering Attacks

No 09-002 **Maria Håkansson**
Playing with Context
- Explicit and Implicit Interaction in Mobile Media Applications
No 09-003 **Petter Karlström**
Call of the Wild
Using language technology in the second language classroom
No 09-009 **Ananda Edirisurya**
Design Support for e-Commerce Information Systems using Goal, Business and Process Modelling
No 10-005 **Moses Niwe**
Organizational Patters for Knowledge Capture in B2B Engagements
No 10-007 **Mats Wiklund**
Perception of Computer Games in Non-Gaming Contexts
No 10-008 **Petra Sundström**
Designing Affective Loop Experiences
No 10-009 **Tharaka Ilayperuma**
Improving E-Business Design through Business
No 11-002 **David Sundgren**
The Apparent Arbitrariness of Second-Order Probability Distributions
No 11-003 **Atelach Argaw**
Resource Lenient Approaches to Cross Language Information Retrieval using Amharic
No 11-004 **Erik Perjons**
Model-Driven Networks, Enterprise Goals, Services and IT Systems
No 11-005 **Lourino Chemane**
ICT Platform Integration – A MCDM Based Framework for the Establishment of Value Network
Case Study: Mozambique Government Electronic Network (GovNet**)**
No 11-010 **Christofer Waldenström**
Supporting Dynamic Decision Making in Naval Search and Evasion Tasks
No 11-012 **Gustaf Juell-Skielse**
Improving Organizational Effectiveness through Standard Application Packages and IT Services
No 12-001 **Edephonce Ngemera Nfuka**
IT Governance in Tanzanian public sector organisations