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
© 2012 Sumithra Velupillai
Typeset by the author using LaTeX
Printed in Sweden by US-AB

Stockholm
University

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.

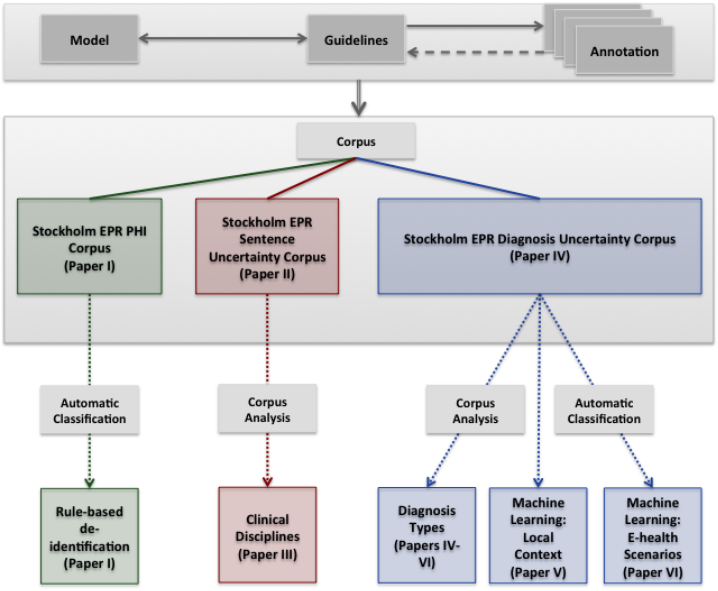

Figure 1.1: Overview: research process and contributions. By creating an annotation model along with guidelines three annotated corpora are created: the Stockholm EPR PHI Corpus, the Stockholm EPR Sentence Uncertainty Corpus, and the Stockholm EPR Diagnosis Uncertainty Corpus. These are used for corpus 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. RadReport-Miner (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.

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.

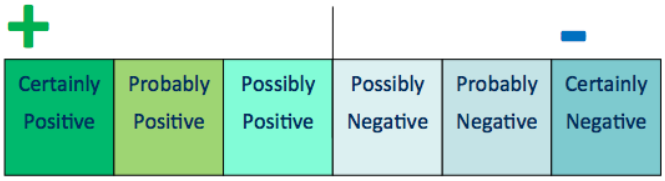

Figure 3.2: Certainty level classification of diagnostic statements: two polarities and three levels of certainty, in total six classes.

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) result ares 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).

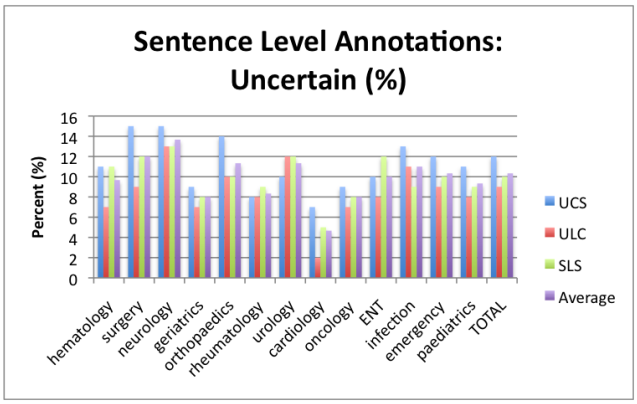

Figure 4.1: Sentence level annotation: *uncertain*, percentage per annotator and clinical practice.

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 tyd-liga 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

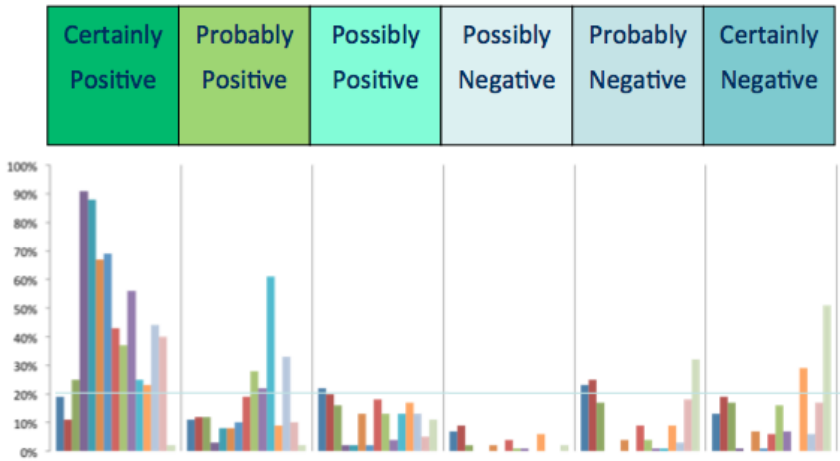

Figure 4.2: Diagnostic statement level annotation: the fifteen most common diagnostic statements and their certainty levels. *Certainly positive* is in majority. On the far right, *certainly negative*, the generic diagnostic statement *skeleton injury* is found. If a skeleton injury is confirmed, a more 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 (hypertension) or low.

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.

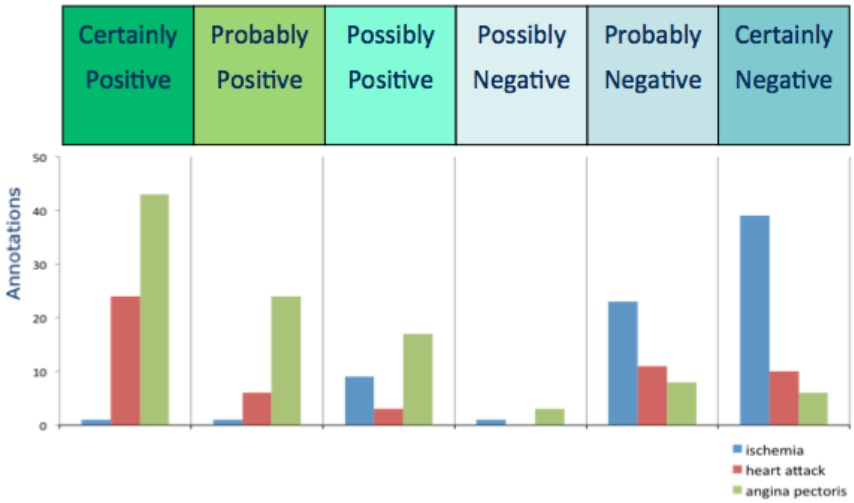

Figure 4.3: Diagnostic statement level annotation: inverted pattern, *ischaemia* in the negative polarity, *heart attack* and *angina pectoris* in the positive polarity.

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 inten-

sive 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*.