Text classification in dictated radiology reports using a machine learning algorithm

Joost Timmerman

April 24, 2014

Master's thesis

Internal supervisor:
Dr. F. (Fokie) Cnossen (Artificial Intelligence, University of Groningen)

External supervisors:
Dr. ir. P.M.A. (Peter) van Ooijen (Department of Radiology, University Medical Center Groningen)

MSc. W. (Wiard) Jorritsma (Department of Radiology, University Medical Center Groningen)

university of groningen / faculty of mathematics and natural sciences

Contents

Abstract

The main goal of this study was to build a successful machine learning system for classifying texts in dictated, free-text radiology reports for use in a re-structured, standardized radiology report. At the start of the study we conducted interviews with referring clinicians to determine the ideal structure for radiology reports on the malignant lymphoma. Based on these interviews two report templates were developed, and the information in free-text radiology reports, written in Dutch, was annotated. A computational system that uses a machine learning technique specifically designed for learning sequences, called a Linear Chain Conditional Random Fields machine learner, was trained on classifying information in the annotated free-text reports. A post processing step was added to the system to correct specific tokens that were misclassified by the machine learner. The classified texts in the free-text reports were automatically re-structured into the developed templates to form standardized, structured reports. A group of five clinicians took part in a user study to evaluate the re-structured reports. The post processing step increased the system's F-score from 88.18 (micro averaged) / 87.85 (macro averaged) to 89.30 (micro averaged) / 88.60 (macro averaged). Results from the user evaluation study suggest that standardizing and improving the global structure of the radiology report increases the clinicians' impressions on clarity and organization of elements, while also decreasing impressions on report complexity. Our study shows that a computational system that uses a machine learning approach can be used to re-structure and standardize the information contained in free-text radiology reports and that the resulting re-structured reports are superior over conventional free-text reports.

Chapter 1

Introduction

Radiology is a medical speciality that uses imaging techniques to visualize and diagnose disease within the human body. For the diagnosis of a disease a physician may order the acquisition of images of a patient. The acquisition of medical imaging is usually carried out by the radiologic technologist, who have access to an array of imaging techniques to do so. The available imaging techniques are X-ray radiography, ultrasound, computed tomography (CT), positron emission tomography (PET) and magnetic resonance imaging (MRI). The radiologist receives a letter or electronic notice from the ordering physician with the request to answer one or more diagnostic questions about a patient. After the images are acquired, the radiologist retrieves the patient's images and clinical background information, and interprets the images. During or after the interpretation of images the radiologist constructs a report of his or her findings, impressions and diagnosis to answer the clinician's diagnostic question(s). When the report is finished it is made available to the ordering physician.

To retrieve and view images and to perform manipulations on these images (e.g. use tools for measurements, adjusting brightness or zooming), radiologists usually use a picture archiving and communication system (PACS). Before starting the radiological examination of newly acquired images, the radiologist retrieves the patients' dossier and the referring clinicians medical question. If required, the radiologist will also look into other information such as findings from the clinical lab technologist or findings from previous reports. The PACS system provides radiologists with a speech recognition system so they can dictate their reports as well as having the possibility of typing the information using a regular keyboard. A combination of both is also possible. When the report is completed it is checked for errors and completeness by the radiologist him-/herself. If the radiologist approves the report, it is uploaded to a database from which the ordering physician can later retrieve it.

The department of radiology in the UMCG conducts around 200,000 medical procedures and sends out 150,000 radiological reports per year (UMCG, 2014). With this large amount of reports, it is crucial that the system as a whole runs smoothly and efficiently.

1.1 Problems in free-text radiology reporting

There are several severe problems with the current situation of free-text radiology reporting. While trying to solve the problems, the core consideration should be to improve the report's ability to convey information in an effective and efficient manner without introducing new problem areas.

A first problem is the fact that the information and data in the free-text reports that are currently sent to referring clinicians is in a way 'locked in': although the information is present in the report, the data can not be automatically searched through and/or used other than searching on word level in each separate report. At first glance this may seem unimportant, but not having this ability can introduce significant workflow bottlenecks or even barriers. For example, answering the question 'How did tumour A of patient X develop in the past year?' is only possible by manually going through all reports of patient X of the past year searching for measurements of the tumour in question, and then calculating the development by hand. Beside simple problems like the fact that information can easily be overlooked or that manual calculation may contain errors, the steps needed to answer this question take valuable time. When questions become more complex, one can see why 'unlocking' the data is absolutely critical. The relatively simple question 'In the past five years, how many patients were diagnosed with a malignant neoplasm in the axillary lymph node?' becomes (almost) impossible to answer when report content is non-searchable.

A second problem with the current situation is caused by the fact that there are no obligatory regulations in the radiology department of the UMCG regarding the structure of the radiology report. This 'freedom' during dictation has its benefits, since radiologists do not have to worry over structural elements or a certain chronological order in the radiological report they are composing. Instead, the radiologists can keep their full attention focused on the interpretation of radiological images.

By allowing radiologists to keep their attention focussed on their main task of image interpretation, instead of forcing them to shift their attention constantly between the interpretation of radiological images and the radiological report, their analytical processes are not disrupted. This allows the radiologists to think more clearly about what they see in the images and may even reduce interpretation errors that are the result of task switching. However, personal preference and previous experience influence the structuring of reports, the elements they contain and the order of these elements. When examining reports of several different radiologists, one will notice that each radiologist has his or her own style. In some cases one can even determine which radiologist has written which report simply by examining the overall structure. Thus, due to personal preference, reports from various radiologists will be structured differently. This requires clinicians to adapt to the way each report structures and presents information. Not only is this annoying for the clinicians, and may lead to reports not being read as intensively, more problems may arise if the reports are not standardized. For example, critical information may get overlooked and remain unused; the need to search for information in the report arises; or information gets misinterpreted. If the report lacks structure and is incoherent, the clinical importance of the radiology report can get lost and information transmission is suboptimal. This could result in severe medical errors.

To come to well-structured radiology reports one could impose the use of structured report software for radiologists to let them report in a (highly) controlled fashion. This software would then guide the radiologist during the compilation of the report. However, the theoretical background showed evidence against using point-and-click structured reporting software as it significantly decreased report accuracy and completeness compared to conventional free-text dictation. Another argument against reporting in a structured fashion is that it would require extra steps for the radiologist, and thus would interfere with the task of image interpretation.

The solution is to develop a system that does not divert the radiologist's attention from viewing and interpreting the images while it allows radiologists to come with fully structured and standardized radiology reports. To avoid interpretive errors of the radiologist, the to-be developed system needs to be designed with the complexity of the radiologist's task in mind, and can not interfere during this task. The system also should not require the radiologist or another employee to rewrite (parts of) the report after dictation, as this is too costly to do.

To summarize, a first problem is the fact that the information and data in free-text radiology reports is unavailable outside of the report; a second problem is reports being styled and structured differently between radiologists. The solution is to develop a system that can solve the two aforementioned problems without interfering the radiologists during image interpretation and reporting.

1.2 Goal of this study

The main goal of this study is to build a successful machine learning system for classifying texts in dictated, free-text radiology reports for use in a re-structured, standardized radiology report. The system should not require the radiologist to interact with the system while the radiologists perform their main task; interpretation of the acquired images.

In order to determine the above, several subgoals were set:

- To determine, in consultation with clinicians, how a radiology report should be structured and what elements should be part of the final report.
- To automatically annotate dictated, free-text radiology reports with the use of a machine learning system trained with annotated example reports.
- To use the data from the automatically assigned annotations to compose structured reports.
- To evaluate these (re)structured reports in an empirical user evaluation study with clinicians.

For the purpose of this study the scope of dictated, free-text radiological reports was restricted to reports about malignant lymphomas. This choice was

made since the availability of these reports was high and since the reports have a relatively consistent content, therefore decreasing the amount of reports needed for training the machine learning system. The choice for using reports regarding the malignant lymphoma also complements the publication by the HOVON workgroup (the Dutch-Belgian Cooperative Trial Group for Hematology and Oncology) who proposed a set of guidelines and recommendations for the request, execution and reporting (i.e. the manner in which findings are described in the report) for patients that suffer from a malignant lymphoma (Nievelstein et al., 2012).

Chapter 2

Theoretical Background

2.1 History

Preston Hickey was an American radiologist who laid the groundwork for important developments in the field of radiology. Soon after the discovery of Röntgen-radiation (or X-radiation) by Wilhelm Röntgen in 1895, röntgen radiation was used in a medical setting. Hickey was one of the first to advocate the importance of a standardized approach to radiological reporting (Hickey, 1904). After a few years, he introduced the term 'interpretation' in the radiology report, to stress the importance of diagnosis of the radiological findings and to come to a conclusion based on these findings (Hickey, 1922). Although the purpose of the radiological report did not change much over the years, the contents of reports became more and more detailed and therefore the information about a patient's medical status became more valuable. Since the radiology report is the primary method of communication between a radiologist and referring clinician, it is important that the radiologist's findings are completely and unambiguously communicated to the referrer.

In recent years, much of the research on the radiology report has focussed on the issues and shortcomings of current radiological reports and many guidelines have been proposed to come to more useful and information-efficient reports. To determine the important elements of a high-quality written radiology report, Pool and Goergen (2010) reviewed 24 papers (1 randomized controlled trial; 1 before-and-after study of interventions; 10 observational studies, audits, or analyses; 12 surveys; and 1 narrative review of the literature) and 4 guidelines from professional bodies concerned with the radiology report (i.e. the ACR, the Canadian Association of Radiologists, the Royal College of Radiologists, and the Society of Interventional Radiology). The review showed a wide variation in the language used to describe findings, in the diagnostic certainty of these findings, and a strong preference among survey participants for structured or itemized formats. Furthermore, current radiology reporting guidelines, recommendations and standards were shown to fall short on addressing the issues concerning implementation or suggesting tools for application.

2.2 Structured reporting initiatives

Despite a strong preference among clinicians for structured formats, most radiology reporting studies and guidelines are mainly concerned with factors such as language/vocabulary, clarity or readability. Several initiatives are pursuing the standardization of radiology reports by proposing the use of predefined words and sentences called a regularized language and/or terminology (American College of Radiology, 1998; Langlotz, 2006). Up to now, these initiatives are mostly limited to the English language.

Another initiative seeking to improve the quality of the written radiology report is using template reports. Template reports typically focus on detailed itemized report content and standardized phrases and lexicon. They allow radiologists to come to full, standardized reports by filling in the blanks, for example by using point and-click software. Several organizations are working on template reports; well known are those being made made available by the Radiology Reporting Initiative of the Radiology Society of North America (Radiology Society of North America, 2013).

Johnson (2002), Johnson, Chen, Swan, et al. (2009), and Johnson, Chen, Zapadka, et al. (2010) were the first to perform a cohort study on radiology residents to evaluate the effect of using a structured reporting system (SRS) on report quality. A group of radiology residents who used free-text dictation served as the control group. Results showed a significant decrease in both accuracy and completeness scores when using the SRS (Johnson, Chen, Swan, et al., 2009). The authors suggest that this was due to constraints in the particular software used, and to the fact that the use of the software distracted the users from their main task; viewing and interpreting the radiological images. Participants in the study most commonly complained about the fact that the SRS was time-consuming to use. One year later, the authors showed that the use of a SRS did not seem to improve nor worsen attending physicians' perception of report clarity, despite two neuroradiology fellows scoring the clarity of the SRS reports significantly lower than that of free-text reports (Johnson, Chen, Zapadka, et al., 2010). The authors suggest that report clarity is likely to depend strongly on elements such as sentence structure or lexicon, which would not normally be affected when using an SRS.

A few years later, Schwartz et al. (2011) compared the content, clarity and clinical usefulness of conventional, free-text reports with template-structured (e.g. using a pre-defined template) radiology reports. Their findings differed from those found by Johnson, Chen, Swan, et al. (2009) by showing a significant increase in content satisfaction and clarity satisfaction when comparing structured reporting to conventional reporting. However, clinical usefulness, assessed by the radiology report grading scale (POCS) (Robert, Cohen, and Jennings, 2006), was not affected significantly between conventional and structured reports. The study has some limitations, namely that dissimilar reports were dictated in both the structured and conventional format, making it hard to objectively compare the results. Furthermore, the time needed for using the SRS to create the structured reports was not evaluated or compared to conventional dictation, nor were radiologists surveyed for their experience with using either method.

Thus, evaluations of reports created by using structured reporting systems show mixed results, making it unclear if this is the way forward. Future software

systems available for creating structured reports should be designed with the complexity of the radiologist's task in mind (Pool and Goergen, 2010). This thought is shared by Schwartz et al. (2011) whose study was discussed earlier. They suggested that systems for structured reporting should be non-intrusive and should not impose new distractions. It is therefore surprising that Schwartz et al. (2011) did not evaluate or discuss the influence of their own system om the already complex task of radiologists in any way.

2.3 The role of education

Beside the technological systems the radiologists use, education of the radiologist themselves plays a role in standardizing reports. Collard et al. (2014) showed that a curriculum on structured reporting can be a good means to educate radiology residents on dictation and reporting skills, instead of learning these skills informally during the years. By following a structured reporting curriculum, the variations (mostly in terms of lexicon) between institutions and among radiologists can be decreased. In the study, residents were educated using a three stage curriculum in which the first and second stage consisted of instructions and formative feedback. The third stage involved individual, biannual, written feedback on the residents' reports. The reports were also scored by specifically assessing four categories: succinctness (e.g. providing clear, precise expressions in few words), spelling/grammar, clarity, and responsible referral. Each category could be scored a minimum of zero and a maximum of three points, resulting in a possible total score of 12 being the perfect score. Residents could repeat the scoring process multiple times in order to pass an annual threshold that enabled them to proceed to the next year of residents training. Despite the fact that a single report could be reviewed/scored multiple times, only the final score given to a report was retained. Over the course of residency training for radiologists, which took four years to complete, 1500 reports were reviewed and a total of 153 reports was scored (one or more times). Mean scores (standard deviation) for first, second, third and fourth year radiology residents were 10.20 (1.06), 10.25 (0.81), 10.5 (0.74), and 10.75 (0.69). The reporting scores of radiology residents showed significant improvement over the course of their residency training. Although the use of a structured curriculum may increase report clarity, inconsistency between report formats is still possible, since each radiologists may develop his or her own global style and format over time.

2.4 Information transfer to the reader

Only two cohort studies are known to investigate the effect of report format on information transfer to the reader. A study by Sistrom and Honeyman-Buck (2005) investigated the effect of radiology report format (structured versus free-text) on the accuracy and speed at which case-specific information could be extracted from the reports. Participants (16 senior medical students) were asked to view a report and answer 10 multiple choice questions about specific medical content of that report for each of 12 cases. Participants were randomly assigned to view either structured of free-text reports. Results showed no significant differences in answers correct, time needed and the number of correctly answered

questions per minute (as an efficiency score) between the two report formats. It is important to note that there were two separate groups of participants, and that none of the participants read both types of reports. Furthermore, participants were able to switch back to view the report while answering the questions and participants received no training period to get used to the report format.

In a more extensive and more recent study, conventional free-text reports were compared with both structured text organized by organ system, and hierarchical structured text organized by clinical significance by a board-certified radiologist (Krupinski et al., 2012). Three conventional free-text reports were reformatted in the structured versions, resulting in nine reports in total. Participants in this study were internal medicine clinicians and radiologists (faculty and residents) who were shown all nine reports in a random order. After viewing a report, the participants were asked to answer 10 questions about specific medical content in the report. Overall results showed no significant effect for reading time or percentage correct scores. However, there were significant differences in both these scores for speciality and level of expertise. For the reading time, results showed that radiology faculty members took significantly more time to read the reports than the radiology residents. For the percentage correct scores a significant effect was found for radiology versus internal medicine with the radiologists (attending and residents) scoring better. The researchers also found significant differences in reading preferences: what the various groups of participants focussed on and how they read the reports (skimming versus reading the full report in detail). The overall differences in reading time or comprehension were suggested not to be the result of report format, but of individual (reading) preferences. Krupinski et al. (2012, p. 63) therefore state that "there may not be 'one-size-fits-all' radiology report format, as individual preferences differ widely".

2.5 Machine learning and radiology

Machine learning is the study of computer algorithms that can learn complex patterns in raw data and that use these patterns to make decisions about new, previously unseen data. In the field of radiology, machine learning systems have been applied for computer-aided detection/diagnosis; fusing of images from multiple modalities, angles and phases; medical image analysis; image reconstruction; and language processing of the radiology report. For this study, we are mainly interested in machine learning systems that have been applied to find and classify information in the radiology report. Research that combines machine learning/natural language processing and the radiology report is relatively scarce.

In their survey of machine learning applications on radiology, Wang and Summers (2012) devoted a brief, separate section to text analysis of radiology reports using natural language processing/natural language understanding. The main point the authors make is that these systems can enable the organization and retrieval of relevant information from the radiological reports in ways that are not feasible by human readers. The amount of data is simply too extensive for humans to meticulously work through within a reasonable amount of time, while computational systems equipped with the right algorithms can perform

this task accurately and relatively quickly. Radiology practice has filled huge databases with reports over the years, and having the ability to profit from (part of) the information stored in these databases may be of great help to support clinicians and radiologists in their tasks. Machine learning systems are able to detect global trends and patterns in vast amounts of data. These trends and patterns may provide new insights, for example when certain diseases become more likely under specific circumstances over a period of years; something that is not easily detected by humans.

Technological development in the field of radiology has provided radiologists with many advantages such as the ability to view more detailed, higher resolution images. These advantages come at a cost, since the radiologists need to request and view more and more raw data about a patient to come to more specific and more detailed descriptions of their findings and conclusion. Bhargavan et al. (2009) showed that radiologists' workload has increased considerably over the last two decades. To aid the radiologist, machine learning systems can be applied. These systems have gained interest over the last several years to provide intelligent, automated methods to process (parts of) the data into more usable information for radiologists, clinicians, and other interested parties. Although the application of machine learning systems and natural language processing to free-text clinical information is still scarce, the interest in using machine learning for the medical setting is growing. For free-text reports, machine learning systems have been developed for several purposes, such as automated registration, text analysis, computer aided diagnosis, decision support, or medical image segmentation.

MedLEE was the first natural language processing system to be actively utilized in patient care and is the best known language processing system applied to clinical texts to date (Friedman et al., 1994). The system's main purpose is to identify and encode medical information in English narrative reports for mapping in a structured representation comparable to a highly normalized representation of the report.

The MedLEE system uses three processing steps. In the first step, the text report is parsed based on grammar, and main sentence structures are identified. In the second step sentences are regularized to reduce variation. The third and final step encodes the standardized text to concepts in a controlled vocabulary. The recall and precision of the system on encoding the impression sections of 230 radiological reports were 70% and 80% respectively. Despite these relatively good results, the MedLEE system has some major limitations. First and foremost, MedLEE has to be fully customized to each new task, medical speciality or hospital. The system also has to be customized on all changing circumstances, due to the rule-based nature of the system. This is especially the case when a different type of hospital record or hospital information system is involved. Therefore, the system is extremely cost-inefficient and labour intensive to apply and maintain. Second, serious performance issues arise when text is parsed that is not highly structured or when complex structures exist in the language of the report. This is due to the fact that the system uses a semantically based text-processing system, which is effective for text that is highly structured, but not for text that lacks this property.

Demner-Fushman, Chapman, and McDonald (2009) reviewed the state of

natural language processing in computerized clinical decision support (CDS); these systems can be used to aid decision making of health care providers by matching the characteristics of a specific patient to a predefined database in order to generate intelligent assessments or recommendations. Currently reviewed systems were shown to be developed for very specific users and goals, but obtained good precision and recall scores on the tasks the systems were built for. Sparse publication and evidence prevented the authors from determining which of the systems were actually implemented and being actively used. In conclusion, the authors emphasized the fact that there is a renewed interest in using natural language processing for medical texts combined with recent local successes in language processing for CDS, which may lead to CDS becoming widely available to the community in the near future. Much seems to depend on the readiness of intended users to adopt a CDS system however.

Suominen et al. (2008) have successfully used a machine learning system to automatically assign diagnostic codes to free-text radiology reports. Their system had a modular structure with a feature engineering and text classification phase. In the feature engineering phase, the free-text radiology report are enriched and features are extracted. The second phase, that of text-classification, consisted of a cascade of two classifiers. Both classifiers perform multi-label classification (e.g. for each classifier, the task can be decomposed into 45 binary classification problems; one for each possible diagnostic code). However, this may result in impossible combinations of diagnostic codes or other irregularities. When the system recognizes such an error, it automatically triggers the cascade and the predictions of the second classifier are outputted instead. To indicate the performance of a machine learning system on a data set, one normally reports the F-score. The F-score is a measure of a test's accuracy that considers both precision and recall. The system by Suominen et al. (2008) was submitted to the Computational Medicine Center's 2007 Medical Natural Language Processing Challenge for which it was trained on a test set of 978 documents. It achieved the third place ranking with a micro-averaged F-score of 87.7. The top scorer in the challenge achieved a score of 89.1, while the mean F-score (standard deviation) over all submissions was 76.7 (13.3).

A more recent article discussed a method for using natural language processing to automatically identify and extract recommendations from radiology reports (Yetisgen-Yildiz et al., 2013). This could help in the timely execution of these follow-up recommendations, and thus help reduce errors that are the result of slacking. The authors' system consists of a series of three consecutive processing steps. First, a given radiology report is divided into its main sections with the use of a Maximum Entropy (MaxEnt) model for classification. The second step cuts the sections' contents into separate sentences using the OpenNLP sentence chunker. The third step entails extracting the recommendation sentences using another MaxEnt model and a binary classifier that labels each recommendation sentence as if it contains either a positive or a negative recommendation. Results showed an F-score of 75.8% on identifying and correctly classifying the critical recommendation sentences in radiology reports.

Esuli, Marcheggiani, and Sebastiani (2013) discussed the use of a machine learning system for information extraction from free-text radiology reports. The

machine learning algorithm in place was the conditional random fields method: a method specifically designed for learning a computational system to distinguish between tokens in sequences and for using the learned information to classify tokens in previously unseen sequences. Three approaches were compared: a standard linear-chain conditional random fields learning system that acted as a baseline, a cascaded two-stage system, and a confidence weighted ensemble of the traditional and the two-stage system. The systems were tested on a dataset of 500 mammography reports written in Italian that were annotated for nine different classes (e.g. 'Technical info', 'FollowUp Therapies' and/or 'Prosthesis Description') by two radiologists. Since the F-score's values do not change when the role of the gold standard and the automatic annotator are switched, it can also be used as a measure of agreement between annotators. Results showed that all three system scored higher than the inter-coder agreement values (the extend to which the annotators agree on the annotations of the content when applying the same set of possible annotation-tags), indicating that the systems outperformed human annotators on accuracy. Highest obtained micro averaged F-scores for the systems were 84.8, 87.3 and 85.9, respectively. Both the improvement of the cascaded system over the baseline system, and the improvement of the cascaded system over the other two systems were significant.

Other machine learning systems that have attempted to process the contents of the radiology report have mainly focussed on specific elements. An example is determining the presence of clinically important findings (i.e. findings that indicated the presence of a disease and/or had the potential to alter patient care and/or outcome) and recommendations for subsequent action (Dreyer et al., 2005). However, no one of the currently available or researched system clearly outperforms any other, making it difficult to determine which is the correct way forward.

Chapter 3

Machine learning in general

The machine learning algorithm Conditional Random Fields is used in this study. Before the details of this method are discussed, we will first introduce machine learning and machine learning problems in general, while keeping the focus on classification tasks. We will then discuss the emergence of the Conditional Random Fields approach from previously existing machine learning methods, followed by the details of this method.

3.1 Supervised or unsupervised

The two main approaches to machine learning are called supervised and unsupervised learning. In both cases you have data. The main difference is that in supervised learning one uses labelled data while in unsupervised learning one uses unlabelled data. We will discuss the differences in more detail later. Currently, the supervised learning approach is the dominant approach in machine learning. The unsupervised paradigm is much less explored, but now that datasets are growing rapidly this method is becoming more popular. This has to do with the fact that getting data is cheap nowadays, while getting labels for the data is expensive. For smaller datasets in a specific domain like the dataset used in this study this is not a real problem since they can be labelled relatively easy by hand.

3.1.1 Supervised learning

Supervised learning uses training data that consists of pairs of information, i.e. an input token and a desired output token. During training, the classifier or supervised machine learning algorithm learns the desired output for a specific input by analysing the input-output pairs in the training set and producing inferred functions. These functions can in turn be used to classify new, previously unseen input. In an optimal scenario the classifier produced generalized inferred functions that can correctly classify previously unseen input data.

3.1.2 Unsupervised learning

In the field of machine learning a task is called unsupervised when there is no training set of correctly labelled observations available. The procedure is

known as clustering or cluster analysis. Unsupervised learning tasks require the machine learning algorithm to group data into categories based on some sort of similarity. We call these similarities features. They are individual properties of the data points. In your dataset you want to use those properties that can be determined for all points in the dataset. For example, when studying energy intake, your dataset could consist of the data of a group of individuals, a single data instance would be one person and a feature would be the sex of that person (i.e. male or female). When the data is grouped based on certain features, objects in the same group share more similarity to each other (for those clustering features) than objects between groups. Since the appropriate clustering algorithm and parameter settings depend on the task at hand, many different clustering algorithms exist.

We know that a lot of different unsupervised learning algorithms exist, and the same holds true for the supervised learning algorithms. For supervised learning problems one needs to consider four major issues, namely the bias-variance trade-off, function complexity and amount of training data, dimensionality of the input space, and overfitting and noise in the output values. We will discuss each issue in more detail.

Bias-variance trade-off

The bias-variance dilemma or bias-variance trade-off is the problem of determining a balance for the model which at the same time identifies all regularities in the training data, but also generalizes over the training data to identify previously unseen data (Geman, Bienenstock, and Doursat, 1992). The bias describes how accurate a model is across different training sets while the variance is a measure of the model error or how sensitive the model is to small changes in the training data. The most widely used method of checking model error is using cross-validation. For this technique the original dataset is (randomly) partitioned into a separate training and test set, and the learning algorithm will not be trained on the test set to make sure testing of the system is always performed with previously unseen data. Multiple rounds of cross-validation are executed using different partitions of the original dataset and the validation results are averaged over the rounds in order to reduce variability.

Function complexity and amount of training data

When using a more complex classification or regression function, more training data is needed in order to learn the function. If for instance a complex classification problem needs to be learned which involves complicated interaction patterns in the data, then more data is needed than when a simpler problem needs to be learned. The bias-variance trade-off described above is generally automatically adjusted for by the learning algorithm based on the amount of available data and the function that will need to be learned.

Dimensionality of the input space

Since supervised machine learners rely on feature vectors of the input, the dimensionality of the input space is another issue to consider. When a machine learning algorithm is presented with high dimensional feature vectors this can

$$y^*(x) = \arg\max_{y \in Y} P(y|x) \tag{3.1}$$

Classification problem: predict label y given features x. In this formula $y \in Y$ while y^* indicates the final, predicted label.

lead to a higher variance in the output. The algorithm then needs to be adjusted towards a lower variance and higher bias. This can typically be done by removing features from the input and will result in a more accurate model.

Overfitting and noise in the output values

Overfitting occurs when a model describes random errors or noise instead of the underlying relationship. Overfitted models have poor predictive power, which is why one must be careful when training a model. If there is noise in the output values then the learning algorithm should not attempt to model this noisy data. In general, one can prevent overfitting by stopping the training in time or by detecting the noisy data examples and removing them from the dataset before training the algorithm.

3.2 Classification with structured labels

3.2.1 The classification problem

In a classification task the main problem is to identify to which set of categories or populations a new observation belongs. The identification is based on a dataset with correctly identified examples. Formula 3.1 shows the classification problem where y is the label that has to be predicted for a given instance based on features x of that instance.

There are multiple supervised machine learning methods that can solve classification problems as discussed earlier. The most well known methods that yield good results in general are Naive Bayes, Maximum Entropy Models and Hidden Markov Models. All three are extensively used and published about. A more recent method that is more or less an extension of the former is a method using Conditional Random Fields. We will discuss all four methods in more detail later in this section, but first we need to discuss classification problems in more detail.

3.2.2 Binary and multiclass classification problems

Many classification problems can be seen as binary classification problems, where an instance needs to be classified into one of two different classes. An example of a typical binary classification task is determining whether or not a car passes the annual test of automobile safety. In this example, the state of the vehicle can be presented by the features on which the classifier bases its output, e.g.: one feature describes the state of the brakes, another the state of the seatbelts, etc.. A binary classifier could then, after training, determine to which of two groups a new data point belongs (pass the safety test, or fail the safety test). A well known form for building a binary classifier are Support

Vector machines (Cortes and Vapnik, 1995), which forms a representation of the training data in a space in such a way that examples of the two categories that need to be distinguished are divided by as much space as possible. New input is then mapped into the space and the category is selected based on the distance to either one of the existing categories from training.

When there are more than 2 distinct classes, we talk about a multiclass classification problems in which a training point belongs to one of N different classes, where $N > 2$. Multiclass classification should not be confused with multi-label classification, in which multiple target labels or classes are to be predicted for each input. Most binary classification algorithms can also be turned into multiclass classifiers, for example by combining multiple instances. The downside of this is that it results in multiple classifiers that need to be trained. An example of a multiclass classification problem would be to determine which type of vehicle a certain transportation device is (e.g. a city bus, a bicycle, a motorcycle, a boat, a train, etc.). One can imagine that other features than the ones described for the binary classification problem are needed to distinguish between these types. As discussed in section 3.1.2 one should use only those features that are relevant to the problem at hand to reduce the dimensionality of input space.

3.2.3 Naive Bayes and Maximum Entropy Models

Naive Bayes is a supervised learning method that is based on applying Bayes' theorem (3.2) with the 'naive' assumption of independence between every pair of features given the class label. The Bayes' theorem gives the probabilities of A and B, $P(A)$ and $P(B)$, and the conditional probabilities of A given B and B given A, $P(A|B)$ and $P(B|A)$. The naive assumption means that the Naive Bayes classifier assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature given the class label. This makes Naive Bayes a generative model, meaning that it is based on a model of the joint distribution $p(y, \boldsymbol{x})$.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.2}$$

Bayes' theorem

$$\text{classify}(f_1, \ldots, f_n) = \underset{c}{\text{argmax}}\, P(C = c) \prod_{i=1}^{n} P(F_i = f_i | C = c) \tag{3.3}$$

Naive Bayes classifier

Formula 3.3 shows the Naive Bayes classifier where $P(C)$ is the class prior and $P(F_i|C)$ is the conditional probability distribution. The classifier classifies the input to whichever class is more probable than any other class. This makes the classifier robust to ignore serious flaws in the underlying naive probability model and class probabilities do not have to be estimated very accurately.

Table 3.1: Comparison of linear classification method performance on text categorization from Zhang and Oles, 2001

Classifier	F**-score**
Naive Bayes	77.0%
Linear regression	86.0%
Logistic regression	86.4%
Support vector machines	86.5%

Despite the fact that the independence assumptions are often inaccurate because they are oversimplifying the problem, Naive Bayes classifiers have been successfully used for many real world problems (Mohri, 2011).

Maximum Entropy Models are in a way related to Naive Bayes since both methods predict the probabilities of the different possible outcomes. However, Maximum Entropy Models do not assume statistical independence of the random variables (i.e. the features) on which the model bases its predictions. The main principle of maximum entropy is that when estimating the probability distribution, one should select the one that has the highest uncertainty (i.e. the maximum entropy). This may sound strange, but that way you have not introduced any additional assumptions or biases into your calculations. An example of a Maximum Entropy Model (MaxEnt Model) is logistic regression, which corresponds to the maximum entropy classifier for features.

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.4}$$

Zhang and Oles (2001) compared several known linear classification methods on text categorization problems, under which were Naive Bayes and logistic regression (a MaxEnt model). The features were the presence of each word in a document and the document class. Feature selection was performed in order to use reliable indicator words. Results for a classic Reuters data set are shown in table 3.1. The F_1 score shown in the table is a measure of a test's accuracy by considering both the precision and recall of the test to compute the score. The F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0 (Wikipedia, 2013). F_1 scores are calculated as the harmonic mean of precision and recall using the formula in equation 3.4. The scores show that that a MaxEnt classifier easily outperformed Naive Bayes. A comprehensive comparison study later showed that Bayes classification was also outperformed by several other approaches, and was actually one of the poorest performing classifiers in the test (Caruana and Niculescu-mizil, 2006).

For both Naive Bayes and Maximum Entropy models the label that has to be predicted -y in formula 3.1- is assumed atomic, while the input x can be structured (e.g., set of features).

3.2.4 Hidden Markov Models

Hidden Markov Models (HMMs) and Conditional Random Fields are extensions of Naive Bayes and Maximum Entropy models where the predicted label y is

assumed to be structured too. To incorporate this extension both HMMs and CRFs model dependencies between components y_i of label y. HMMs do this by arranging the output variables in a linear chain. A simple example of such a chain can be seen in part of speech tagging, where the input x is a sentence (i.e. a sequence of words). The output y is the corresponding sequence of parts of speech (e.g., noun, verb, etc.).

The main goal of HMMs is to model the joint distribution of observations and hidden states (output tokens), therefore they are generative models, just like the Naive Bayes classifiers. Unobserved states (output tokens) are identified on the basis of states that can be observed. For each state, a probability distribution over the possible output tokens is determined. This implies that the sequence of output tokens generated by the HMM also provides information about the sequence of states. Because HMMs generally assume time invariance, they are especially known for applications in speech, handwriting and gesture recognition.

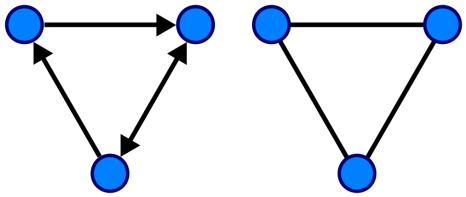
An example of a HMM would be trying to deduce the weather from a flag hanging from a shop. We know that a soggy flag means wet weather, while a dry flag means sun. If the flag is damp then we cannot be sure about the weather. Since the state of the weather is not restricted to the state of the flag, we may say on the basis of an examination that the weather is probably raining or sunny. Also, knowing what the weather was like on the preceding day could help come to a better conclusion for today.

HMMs can be considered as one of the simplest forms of Bayesian networks; a form of probabilistic networks that represents a set of variables and their conditional dependencies via a direct acyclic graph. Since HMMs assume independence between parameters and state variables, efficient iterative solutions exist. However, relaxing the independence assumption by arranging the output variables in a linear chain without dependencies may lead to a poor approximation of the real problem.

Maximum entropy Markov models Extensions to HMMs have been proposed that address the independence assumption problem among several others. The best known being the maximum entropy Markov model (MEMMs) (McCallum and Freitag, 2000), which combines features of hidden Markov models (HMMs) and maximum entropy (MaxEnt) models and assumes that the unknown values to be learnt are connected in a Markov chain rather than being conditionally independent of each other. A major advantage is that MEMMs offer an increased freedom in choosing features to represent observations such as domain-specific knowledge of the problem at hand. As an example, McCallum and Freitag (2000) wrote "... when trying to extract previously unseen company names from a newswire article, the identity of a word alone is not very predictive; however, knowing that the word is capitalized, that it is a noun, that it is used in an appositive, and that it appears near the top of the article would all be quite predictive...".

However, an important limitation of the MEMMs approach (and other non-generative finite-state models) is that these can be biased towards states with few successor states. This is caused by the fact that transitions leaving a given state compete only against each other, rather than against all other transitions in the model, and is also known as the *label bias problem*.

3.2.5 Conditional Random Fields



(a) Directed graph, arrows indicate dependencies

(b) Undirected graph

Figure 3.1: Graphical models.

Let us consider the joint probability distribution $p(\boldsymbol{y}, \boldsymbol{x})$, where $\boldsymbol{y}$ represents the attributes of the instances we want to predict and $\boldsymbol{x}$ represent the observed knowledge (represented in features) of the instances. Graphical models (see 3.1) are a commonly used technique in machine learning and statistics to indicate the conditional dependencies between random variables. Now, if we want to model the joint probability distribution, using the rich features that can occur in the data could lead to difficulties and intractable models, because it requires modelling the distribution p($\boldsymbol{x}$) which in turn can include complex dependencies (i.e. a complicated graph of arrows). If we would ignore these dependencies however, this could lead to reduced performance of the model .

To solve this problem, one could directly model the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ which is sufficient for classification. This is the approach taken by conditional random fields (Lafferty, McCallum, and F. C. N. Pereira, 2001).

CRFs were proposed by Lafferty, McCallum, and F. C. N. Pereira (2001) as a framework for building probabilistic models to segment and label sequence data and are an alternative to the HMMs and MEMMs discussed in the previous section. CRFs form a modelling method specifically used for structured prediction. Ordinary classifiers such as Naive Bayes base their prediction on a single instance without taking into account features of the neighbouring instances. CRFs can take context into account. The linear chain CRF that is popular in natural language processing can for example predict sequences of labels for sequences of input instances. CRFs offer several advantages over HMMs including the ability to relax strong independence assumptions made in those models. In short, CRFs offer all the advantages of the MEMMs approach but avoid the *label bias problem* that weakens the MEMMs approach.

In the original paper, CRFs were defined as follows:

Let $G = (V, E)$ be a graph such that
$\boldsymbol{Y} = (\boldsymbol{Y}_v)_{v \in V}$, so that $\boldsymbol{Y}$ is indexed by the vertices of G.

Then $(\boldsymbol{X}, \boldsymbol{Y})$ is a conditional random field in case, when conditioned on $\boldsymbol{X}$, the random variables $\boldsymbol{Y}_v$ obey the Markov property with respect to the graph:

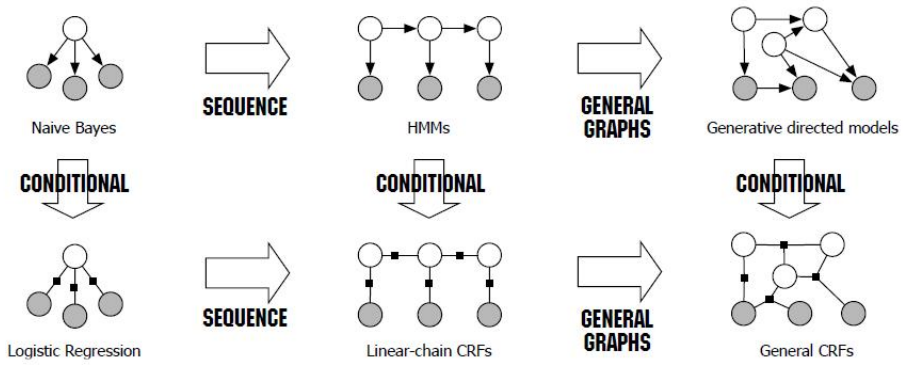


Figure 3.2: Diagram of the relationships between Naive Bayes, logistic regression (MaxEnt), HMMs, linear-chain CRFs, generative models, and general CRFs.

$p(\boldsymbol{Y}_v|\boldsymbol{X}, \boldsymbol{Y}_w, w \neq v) = p(\boldsymbol{Y}_v|\boldsymbol{X}, \boldsymbol{Y}_w, w \sim v)$, where $w \sim v$ means that w and v are neighbours in G.

This means that a CRF is an undirected graphical model whose nodes can be divided into exactly two disjoint sets $\boldsymbol{X}$ and $\boldsymbol{Y}$, which are the observed and output variables, respectively, for which the conditional distribution $p(\boldsymbol{Y}|\boldsymbol{X})$ is then modelled.

Thus, a conditional random field is simply a conditional distribution with an associated graphical structure. And because it is a conditional distribution, there is no need to explicitly represent dependencies (arrows in the graphical model) among the input variables. This enables the use of rich, global features for the input. Since CRFs model the conditional distribution, they belong to the class of discriminative models. The most important difference from generative models such as Naive Bayes and HMMs is that discriminative models do not include a model of $p(\boldsymbol{x})$, which is difficult to model because it often contains highly dependent features and which is not needed for classification anyway.

To clarify the relations between and relevance of the discussed models, figure 3.2 shows the relationships between the different general models.

3.2.6 Linear chain-Conditional Random Fields

Although CRFs can have an arbitrary graphical structure (see figure 3.2), there is a form which supports sequence modelling. This form is known as the Linear chain-Conditional Random Fields, which -as stated earlier- is popular in natural language processing. Linear chain CRFs are similar to the Maximum entropy Markov models, but they have no strong independence assumption for the model's features and they support sequence labelling which those models can not.

A linear chain CRF combines the advantages of both discriminative modelling and sequence modelling. It defines a posterior probability $p(\boldsymbol{Y}|\boldsymbol{X})$ where $\boldsymbol{Y}$ is a label sequence for a given input sequence $\boldsymbol{X}$. In a linear chain conditional random field, the label $y \in \boldsymbol{Y}$ for a given frame depends jointly on the label of the previous frame, the label of the succeeding frame, and the observed

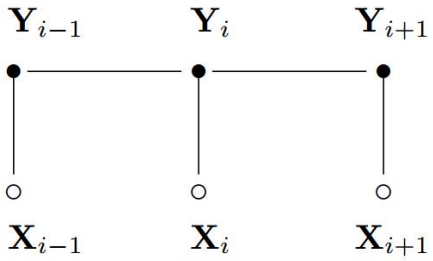


Figure 3.3: Graphical structure of a linear chain Conditional Random Fields model for sequences. An open circle indicates that the variable is not generated by the model.

data $x \in \boldsymbol{X}$. These dependencies are computed in terms of functions defined by pairs of labels and by label-observation pairs. The input sequence $\boldsymbol{X}$ is for example a series of words that together form a text. The label sequence $\boldsymbol{Y}$ then is the series of labels assigned to that observed frame sequence, which could for example be part of speech tags. During classification, each frame in $\boldsymbol{X}$ is assigned exactly one label from $\boldsymbol{Y}$. Figure 3.3 shows a graph of the structure of linear chain CRFs.

In linear chain CRFs the distribution of the label sequence $\boldsymbol{Y}$ given the observation sequence $\boldsymbol{X}$ will have the form:

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \frac{\exp \Sigma_t(\Sigma_i \lambda_i f_i(\boldsymbol{Y}, \boldsymbol{X}, t))}{Z(\boldsymbol{X})} \tag{3.5}$$

where t ranges over the indices of the observed data and $Z(\boldsymbol{X})$ is a normalizing constant over all possible label sequences of $Z(\boldsymbol{Y})$ computed as:

$$Z(\boldsymbol{X}) = \Sigma_{\boldsymbol{Y}} \exp \Sigma_t(\Sigma_i \lambda_i f_i(\boldsymbol{Y}, \boldsymbol{X}, t)) \tag{3.6}$$

The CRF is thus described by a set of *feature functions* (f_i), defined on graph cliques, with associated weights (λ_i).

Two distinct types of feature functions are defined in a linear-chain CRF, namely state feature functions and transition feature functions. The first type is related to the graph nodes and its output depends only on the observations and the label at the current time step t. The second type is related to the edges of the graph, whose output depends on the observations and both the label at the current time step t and the label at the previous time step $t-1$.

Training

Training is performed by using a set of training data to maximize the conditional likelihood function $p(\boldsymbol{Y}|\boldsymbol{X})$. For this study we used a quasi-Newton gradient descent method called the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) training algorithm (Nocedal, 1980). Since the training specifics are not the topic of interest in this study, we will only discuss the outline.

The general idea is that the L-BFGS method uses a set of algorithms to search through the variable space. It does this by using an approximation of

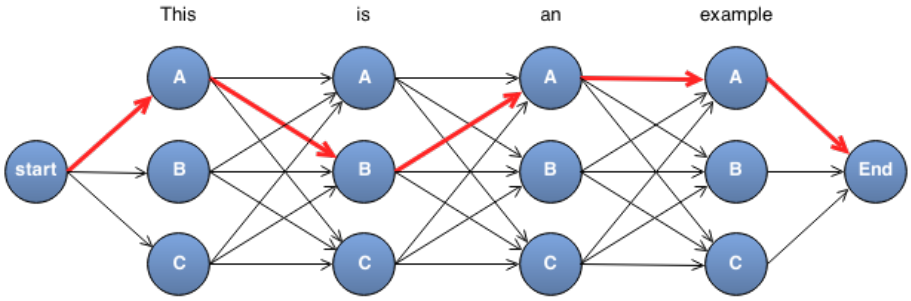


Figure 3.4: Lattice structure of a trained Conditional Random Fields model for sequences. Arrows indicate weights between the possible labels (represented as circles). Red arrows show an example of a final output path (sequence of labels).

the inverse Hessian matrix which describes variable data. To limit the use of computer memory only a few vectors that represent the approximation of the matrix are stored at each time instead of the entire nxn matrix (with n being the number of variables in the problem). First the gradient of the entire training set is computed using the current weights (λ_i). The algorithm then does batch updates in which it moves in small steps along the computed gradient to find a minimum. It is called a quasi-Newton method since it uses a set of algorithms for finding local minima or maxima based on Newton's method to find the stationary point of a function, where the gradient is 0. The L-BFGS algorithm is popular for parameter estimation in machine learning and has been shown to perform well for training CRFs (Malouf, 2002; Sha and F. Pereira, 2003).

Testing

Testing consists of finding the label sequence over the data sequence $\boldsymbol{X}$ that maximizes the conditional probability. After training, the model can be thought of as a lattice (see figure 3.4). Most CRFs make use of the Viterbi algorithm (Forney, 1973) to find the best path in the model for a sequence of input. This is a programming algorithm that is able to find the most likely sequence of states. For example, if we consider the lattice in the figure as a model trained for shallow parsing with labels A, B and C as Noun Phrase (NP), Verb Phrase (VP) and Adverb Phrase (AP) respectively, then the path indicated by the red arrows would be the Viterbi path that a good model gives as output (i.e. $[NPThis][VPis][NPan][NPexample]$).

Chapter 4

Methods

4.1 Radiology report structure

In-depth interviews with referring clinicians were conducted to determine how a radiology report on malignant lymphomas should be structured and what elements should be part of the final report. In order to prepare for the interviews, papers that have discussed the contents and elements required in a radiology report were studied (Pool and Goergen, 2010; Wallis and McCoubrie, 2011; Bosmans, Peremans, et al., 2011; Bosmans, Weyler, et al., 2011; Nievelstein et al., 2012). Open-ended questions about the current reports and radiology reports in general were prepared, i.e. questions about the first impressions on the existing reports; comprehensibility; complexity; overall overview; searching for information; reading manner (e.g. skimming or reading in detail); order when reading; the importance of items. During the interviews, participants were also asked how they felt about using a new format (e.g. using a different arrangement of elements) for the radiology report.

4.1.1 Card sorting task

Card sorting is a technique utilized frequently in the field of user experience design, where it is used for designing information structures such as workflows, menu structures, or navigation paths. In a basic card sorting task, subjects are asked to generate a category tree by arranging and/or grouping a set of cards that have terms written on them in such a way that they feel most familiar with the final arrangement. In general, participants are asked to comment on their own choices during the task, which provides valuable information on why participants place certain cards together. From the arrangements and comments, one can learn how information should be structured so that users can easily relate to the composition. For example, if the set of cards may consist of book genres, then similarities in participants' answers provide clear indicators on how to arrange books in a library.

The most frequently mentioned elements in the studied were listed and checked off by the interviewer during the card sorting task. If clinicians did not come up with the elements themselves, the interviewer made hints or suggestions from this list to check if elements were either purposefully not mentioned for discussion, or simply forgotten. Items on this list were relevant medical history,

procedure, technical details (i.e. camera type or medical contrast medium), examination quality, findings in general or normal and abnormal findings separated, medical question to answer, comparison to previous study, recommendations, diagnosis (pathological and/or differential), summary, and conclusions.

The radiology report offers variety in content and structure, but more importantly, the ideal organization of elements in a radiological report can depend greatly on subjects' preferences. This makes card sorting a great tool to investigate which items should be organized in what way in order to make sense to the target audience; in this case referring clinicians. Participants in the in-depth interviews were asked to participate in an open card sorting task. The difference of open card sorting compared to basic card sorting as described earlier, is that participants have to create their own set of cards by identifying relevant items and writing them on blank cards. This has the advantage that only relevant items end up in the final arrangement. However, the disadvantage is that participants in the task may forget to write down items that they do find important in reality.

The interviews and card sorting task resulted in two templates representing the two most preferred radiology report structures. While the overall contents is exactly the same, the difference between the two global structures is the position of the conclusion section: in the first structure (shown in figure 4.1), the conclusion section is positioned after the findings, while in the second structure the conclusion section precedes the findings. Note that the brackets and dots (e.g. in the figure depicted as '[...]') in the structure template will be replaced with the actual report content, while all other elements such as (sub-)headings will remain intact as part of the final, structured report.

4.2 The automated structuring method

4.2.1 Annotating dictated reports

For annotating the free-text radiology reports GATE or General Architecture for Text Engineering (Cunningham, Tablan, et al., 2013; Cunningham, Maynard, et al., 2011) was used. We chose GATE since it provides a simple and easy to use interface for annotating reports by hand. Furthermore, it provides clear feedback on the annotated texts by highlighting in distinct colours and annotated texts can be saved for later editing in a GATE database called a 'datastore'.

GATE is a software framework originally developed by the University of Sheffield in 1995 for the purpose of natural language processing research. It contains sets of tools for all sorts of natural language processing tasks (i.e. part-of-speech taggers, tokenizers and sentence splitters). GATE is freely available as an internet download (GATE Research Team, 1995) and is a widely used tool in the field of natural language processing.

Free-text radiology reports on the malignant lymphoma were collected in the UMCG by contacting referring clinicians and radiologists that had shown earlier interest in the project. The radiologists in question collected only on those reports that they composed themselves and clinicians collected report on

RADIOLOGICAL IMAGING REPORT

Radiologists	[…]
Visiting date	[…]
Report date	[…]

Examination

PET scan quality	[…]
PET camera	[…]
PET contrast	[…]
PET scanning protocol	[…]
CT scan quality	[…]
CT camera	[…]
CT contrast	[…]
CT scanning protocol	[…]

Comments PET
[…]

Comments CT
[…]

Application

applicant	[…]

Clinical background and question
[…]

Findings PET scan

Comparison study
[…]

Head/neck
[…]

Armpits
[…]

Thorax
[…]

Retroperitoneum
[…]

Abdomen/pelvis
[…]

Musculoskeletal
[…]

Conclusion PET scan
[…]

(a) Page 1 of 2

Figure 4.1: Page 1 of one of the two templates representing the most preferred radiology report structures based on outcomes of interviews with referring clinicians.

Findings CT scan

Comparison study
[…]

Head/neck
[…]

Armpits
[…]

Thorax
[…]

Retroperitoneum
[…]

Abdomen/pelvis
[…]

Musculoskeletal
[…]

Conclusion CT scan
[…]

Integrated conclusion

[…]

(b) Page 2 of 2

Figure 4.0: Page 2 of one of the two templates representing the most preferred radiology report structures based on outcomes of interviews with referring clinicians.

Table 4.1: Set of labels available for annotating sentences in the radiology reports.

1	Abdomen/pelvis
2	Camera
3	Conclusion
4	Contrast
5	Head/neck
6	Musculoskeletal
7	Armpits
8	Person (executing radiologist)
9	Retroperitoneum
10	Scan quality
11	Scanning method (comments)
12	Scanning protocol
13	Thorax
14	Comparison
15	Clinical background and question

their own patients only. Reports were exchanged directly with the researchers. Therefore, both parties never saw or read reports they should not have access to. Furthermore, reports were anonymous to start with, as they contained no patient names or other identifiers.

The reports were converted to plain Windows text files with the *.txt file extension. Headers that emerged during the database retrieval of the radiology reports were removed from the text files, resulting in a single, free-text radiology report per text file.

The plain text files were imported into GATE and all existing annotations were removed (i.e. tags for paragraphs and the text as a whole that were automatically annotated by GATE upon file import). New annotation labels were chosen from a predefined set of labels that was composed beforehand. The set of labels was implemented in GATE as a GATE-plugin to avoid typing mistakes in label names, since label names now could only be chosen from the predefined set. The labelset (see table 4.1) was based on the card sorting task and interviews with referring clinicians about the ideal structure and content of the radiology report.

Annotating was performed by selecting (multiple) sentences in the text and then clicking on the label related to the selected text in a pop-up that was triggered by the selection. The annotated piece of text would then be highlighted automatically with a distinct colour to provide feedback for the text's label. Figure 4.1 shows a screenshot of GATE during annotation. After annotating a report was completed, the report was exported as an Extensible Markup Language (XML) file with the annotation labels in inline format (i.e. annotation tags are embedded in the texts themselves, instead of being declared in a separate, external file). A total of 165 reports were annotated by a single annotator using GATE.

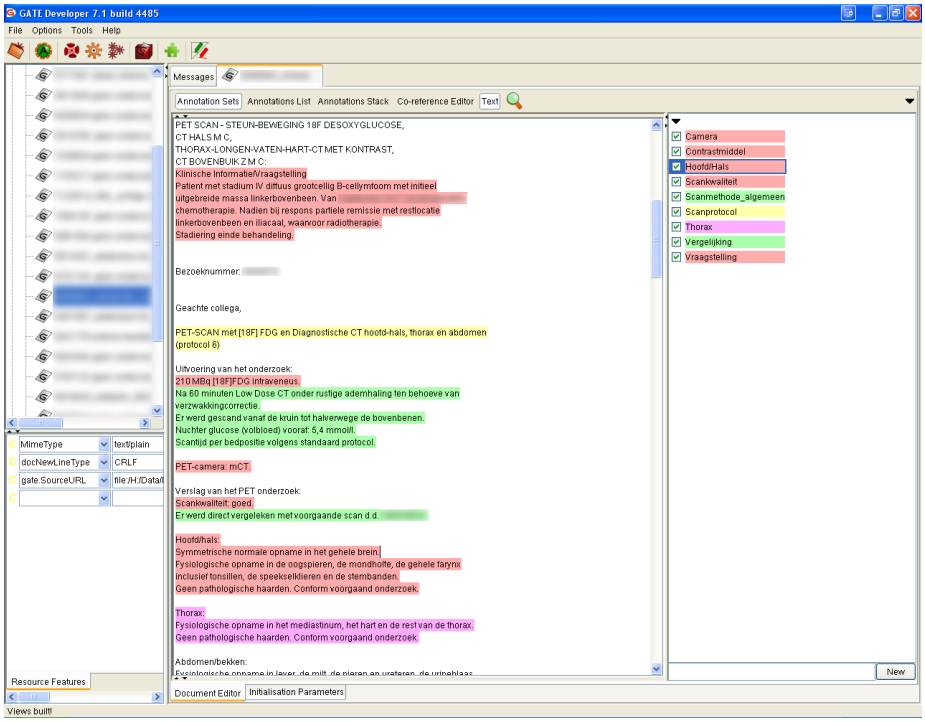


Figure 4.1: Screenshot of the GATE software framework during an annotation task.

4.2.2 The CRFs learner

For the machine learning, version 0.58 (released 2013-02-13) of the CRF++ algorithm that is available as a free internet download (Kudo, 2013) was used.

This open-source implementation of Conditional Random Fields is designed as a simple, customizable, generic purpose tool and is implemented and maintained by Taku Kudo. This implementation of a CRF was chosen because it yielded good results compared to other CRF implementations (Okazaki, 2011). Furthermore, it makes use of relatively simple to edit feature templates (more on this in section 4.2.2) for training and testing, and can encode/decode in practical time. A further advantage was that it was actively worked on, and as a result it could run on new machines with up-to-date software configurations and was able to make use of multi-threading.

Training and test sets

The inline annotations in the XML files were converted to the input file format of the CRF machine learner by a small piece of JAVA code that we wrote ourselves. The input file format of the CRF machine learner required words and punctuation signs to be on separate rows. To identify sentence boundaries, an empty line was put between two rows. Each token could furthermore be followed by zero or multiple features. The last column of the row held the label associated with the row's token. The CRF machine learner further required every row to have the exact same amount of columns. Therefore, items in the annotated text that had no assigned label after annotation (e.g. headings, addresses, etc.) got the label *noTag*.

Let's provide an example. Note the following text snippet which has inline labels $first$ and *second* as XML tags.

```
...<first>This is a full sentence,</first> it acts as
   <second>an example. This will make...</second>
```

After converting the above sentence to the CRF machine learner input, we obtain the list displayed below. Note that the text *it acts as* was labelled with the label *noTag* as a result of the above sentence not having a specific label for this part of the sentence. Also note that there were no features provided in the output, as there is no column in-between the token itself and the token's label in the last column. During the conversion from XML files to the training and test files, all tokens that were not annotated with a label from the labelset as displayed in 4.1 also were assigned the label *noTag*. Therefore, the final datasets contained not 15, but 16 labels for the CRFs learner to train on.

TOKEN	LABEL
This	first
is	first
a	first
full	first
sentence	first
,	first
it	noTag
acts	noTag
as	noTag
an	second
example	second
.	second
This	second
will	second
make	second
...	

Since the CRF++ software is unable to cope with separate data files as input (one file per report), multiple text files were concatenated to form two separate datasets; one for training and one for testing purposes, respectively.

To determine how accurate the CRF's predictions are in practice, a model validation technique called cross-validation was applied. Cross-validation is used to asses how the results of a predictive model will generalize to an independent dataset. In practice, this implies that separate datasets are used for training and testing purposes. The model is then trained on a dataset of what is called *known data*, and later tested on a dataset of *unknown data* (i.e. data that was not used during training).

For this study, the datasets were created semi-randomly to ensure a 80/20 split of data files, meaning that the training set contained approximately 80% of the data, and the test set the remaining 20% of the data. Furthermore, the data of a radiology report could only occur in either the test set or in the training set, not in both. To avoid incidental higher scores on a specific distribution of reports in the datasets, K-fold cross-validation was applied by randomly partitioning the full dataset into k equal sized partitions. Of the k partitions, a single partition is used as the test set (validation) while the remaining $k - 1$ is used for the training set. This process is then repeated k times, in this study with $k = 5$ thus resulting in five combinations of training and test sets. The CRF is trained and tested on each of the five combinations, and results are then averaged to obtain a single estimation. The advantage of using K-fold cross-validation is that each observation is used for both training and testing, and each observation is used for testing exactly once.

Features

Multiple columns of features were added to the training and test set, the first being the identification numbers of the reports that were concatenated while forming the training and test sets. By adding these numbers, the merged radiology reports could be kept apart in the merged data files and thus be retrieved

individually, or split when needed. This feature row was not used during training or testing, but it was introduced for this purpose only.

A second feature column that was added to each token held a positional feature of that token in relation to the report it was in. The hypothesis behind this is that specific words or combinations of words occur in more or less specific locations in the report. Thus, providing the CRF with the positional information about the current token is likely to improve performance. For example, the conclusion of the radiologists' findings is likely to be found in the final part of the free-text report, therefore providing the CRF with information that represents '*the current token is located in the final quarter of the report*' can help to determine the correct label for that token. For the positional features, the total amount of tokens n of each radiology reports were determined and used to divide the report into x parts, with each part consisting of n/x tokens. The positional feature would then be 0 if it fell in the first part, 1 if it fell in the second part, ..., or $n-1$ if it fell in the n^{th} part. The positional features were added to the data to improve overall scores on class labels. Since the conventional reports had a more or less fixed sequence of elements, it was important to emphasize this in the data.

The third and fourth features that were added share a similarity, as they were added based on the idea that a combination of a capital and a colon is likely to indicate a transition to another chunk of information like for example a subheading indicates. The third feature therefore holds the value *true* if the current token starts with a capital or *false* if this is not the case. The fourth feature holds the value *true* if the next token is a colon, or *false* otherwise. These features were added to the data to make it clearer for the machine learner where new sections of the report started, hoping to reduce errors where the a text section is classified with label A and then the subsequent section, that should receive class label B, is also classified under A.

The fifth to last feature columns, that is without including the annotated label column, hold characteristics about token frequencies in the labelled categories of the report. If a token is frequent in texts under a certain annotated label, it is more likely that the token needs to be assigned that label than a label under which it is not frequent at all. First, the frequencies of tokens under each specific label were determined for the data in the entire training set. A simplified example of the frequency data is shown below. Note that tokens can occur in the frequency data under multiple labels (e.g. the question mark). These frequency features were added to improve results for classes with fairly specific information that scored below average in the intermediate results.

LABEL	TOKEN	FREQUENCY
Question	?	120
Question	clinical	93
Question	questions	90
Question	lymphoma	86
Comparison	compare	99
Comparison	compared	31
Comparison	comparison	30
Comparison	previous	20
Comparison	lymphoma	11
Comparison	?	2
...		

After determining token frequencies under each label, non-unique tokens (e.g. tokens that occurred under two or more labels) were identified and iterated over. In each step only the entry of the non-unique token with the highest frequency count is retained. Entries under other labels with the same token but with a lower frequency count were removed. If two entries that shared identical tokens had an equal frequency count, both entries were retained.
The idea behind removing the non-unique entries is that although a token can occur in texts with different labels, it is likely that the token is more strongly related to the label under which the word frequency is highest. Therefore, removing the non-unique entries before rewriting the information as features provides the CRFs learner with information which better supports to correct token-label combination. Looking at the previous example again, the frequency data after removing non-unique entries would become the following.

LABEL	TOKEN	FREQUENCY
Question	?	120
Question	clinical	93
Question	questions	90
Question	lymphoma	86
Comparison	compare	99
Comparison	compared	31
Comparison	comparison	30
Comparison	previous	20
...		

Based on the frequency data, the values of the token frequency feature columns were determined. Each label has a feature column of its own. The value in this column holds *true* if the current token y is an element of the Y most frequent tokens under that label. Since non-unique entries were modified in the previous step, only one of the feature columns can hold the value *true* for a specific token. Thus, looking at the previous example again, the frequency feature columns related to the question mark token only hold the value *true* in the feature column related to the *Clinical background and question* label, and *false*

in all other columns concerned with token frequencies.

Training

Since the CRF++ software is designed as a general purpose tool, it needs a feature template file. In this file, the (combinations of) features that are used in training and testing are described. If a feature is not mentioned in any of the templates in the template file, the CRF++ software does not use it during training and testing.

Part of a simple feature template file is shown below. Each line specifies one template. In each template the macro $\%x[row, col]$ is used to specify a token in the input data. *Row* specifies the relative position from the current focusing token and *col* specifies the absolute position of the column. For unigram templates, the preposition U is used, while for bigram templates the preposition B is used.

```
# Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-1,0]/%x[0,0]
U06:%x[0,0]/%x[1,0]
...
```

Taking the following input data with simple part of speech tags as an example, the template $\%x[0,0]$ would refer to the token *is* and template $\%x[0,1]$ to the expanded feature *verb*. Negative values are also allowed, and template $\%x[-2,0]/\%x[-2,1]$ would for instance refer to *The/determiner*. Feature templates can become as complex as is needed, so that the CRF can be trained on combinations of multiple features and tokens.

```
Input data:
TOKEN   FEATURE COLUMN #1   LABEL
The     determiner          None
car     noun                Vehicle
is      verb                None       << current token
blue    adjective           Colour
.       punctuation         None
```

During training, the CRF++ software expands the templates and generates a set of feature functions ($func_1 \ldots func_n$). These describe which combinations of tokens (and features, if specified) have positive relations. For example, the macro $U03 : \%x[0,1]$ from the example above can be expanded to *adjective*, and the CRF automatically generates feature functions like the ones below to match the expanded feature. During training, the algorithm learns the associations (i.e. weights) between feature attributes and labels (e.g. if the current token's

first feature has the attribute 'noun' it is likely to have the label 'Vehicle'). The algorithm does this by searching through the variable space to find local maxima and minima (see subsection 3.2.6 for the more technical explanation of training the used CRF machine learner).

The total number of functions generated by a template is equal to $L * N$ (or $L * L * N$ for bigram features), where L is the number of output classes and N the number of unique strings expanded from the given template. The CRF++ software outputs a model file that can later be used for testing.

```
func1 = if (output = Colour and feature="U03:adjective") return 1 else return 0
func2 = if (output = None and feature="U03:adjective") return 1 else return 0
func3 = if (output = Vehicle and feature="U03:adjective") return 1 else return 0
...
```

Testing

There is no need to specify the template file explicitly for testing, because the model file that is created during training has the same information that is contained in the template. The CRF testing procedure takes this model file together with the test data as input. For the CRF to work, the test data file needs to be in the exact same format as the training file with equal amounts of columns and similar features. The CRF then outputs the information in the test file with a new column appended to the end that contains the tag or label that the CRF has assigned to the tokens. One can then evaluate the results by computing the difference between the estimated labels in the n^{th} column, and the true answer label in the $n-1^{\text{th}}$ column that was assigned during annotating.

4.2.3 Enhancement of the CRF output

The CRFs machine learner had difficulties labelling the sentences present in the texts about the *Clinical background and question* and the *Conclusion* of the reports. This was to be expected, since these texts contain sequences of tokens and transitions between these tokens that are also likely to appear in other parts of the report. Since both the *Clinical background and question* and the *Conclusion* were found in consistent parts of the free-text reports, and were succeeded by more or less predictable parts of text (e.g. a heading or a person's name), a piece of JAVA code was developed to enhance the results in these parts of the CRF's output. When the JAVA code was ran, it automatically processed the CRF's output and outputted a new file with the resulting enhancements in an appended column.

For the labels related to the *Clinical background and question*, this meant that all labels between a combination of the current token $\in$ predefined list of tokens and the CRF estimation $\equiv$ the label *Clinical background and question*, up to a CRF estimation $\equiv$ predefined list of labels, were converted to the label *Clinical background and question*, no matter what the CRF's estimate initially was.

For the part of text concerning the *Conclusion*, all labels between a combination of the current token $\in$ predefined list of tokens and the CRF estimation $\equiv$ the label *Conclusion*, up to a CRF estimation $\equiv$ predefined list of labels, were converted to the label *Conclusion*, again overruling what the CRF's estimate was beforehand.

The next section shows a simplified example in which the CRF has labelled part of the sentence 'This shows how the output can be enhanced!' with the label *body* and part of the sentence with the label *prologue*. Let's think of the part that is labelled with *prologue* as being wrongly classified. If we write a piece of code that is triggered by the combination of the token *This* AND the CRF's assigned label *body*.

After triggering the enhancer, it will output the label *body* and will continue to do so, thereby disregarding the CRF's initial output of the *prologue* label. Another trigger deactivates the enhancer. This means that from that point forwards, the CRF output is no longer overwritten. The final output file will have a similar formatting as the example below, with an extra column appended to the file that contains the enhanced label.

```
TOKEN     FEATURE   CRF LABEL     ENHANCED LABEL
          COLUMNS
This      ...       intro          intro
is        ...       intro          intro
an        ...       intro          intro
example   ...       intro          intro
sentence  ...       intro          intro
.         ...       intro          intro
Goal      ...       heading        heading
:         ...       heading        heading
This      ...       body           body      << activation trigger
shows     ...       body           body       < enhanced section
how       ...       body           body       < enhanced section
the       ...       prologue       body       < enhanced section
output    ...       prologue       body       < enhanced section
can       ...       prologue       body       < enhanced section
be        ...       prologue       body       < enhanced section
enhanced  ...       prologue       body       < enhanced section
!         ...       prologue       body       < enhanced section
Final     ...       ending         ending    << deactivation trigger
section   ...       ending         ending
of        ...       ending         ending
the       ...       ending         ending
example   ...       ending         ending
.         ...       ending         ending
...
```

4.2.4 Splitting the CRFs' output for use in templates

The CRFs' output, a single file with tokens and related label classifications, could easily be split into separate report files due to the identification numbers that were present as a feature column in the output. However, since the reports in the dataset could be combinations of both positron emission tomography (PET) and computed tomography (CT) research, or from either of those alone, an extra step was needed in post processing to further split the information that the CRF had classified under the labels. If the CRF had for example classified texts in both the PET part and the CT part of the report under the label *Head/neck*, this information needed to be separated so that it would end up in the correct position in the final structured template. The conventional, free-text reports all contained clear, reliable identifiers in the form of specific generated sub-headings to determine when the PET or when the CT part of the report began. Since all information of the input files was contained in the output, these identifiers where present in the resulting separated files and could be used to further split the CRFs' output in PET and CT parts when needed.

Not to use separate labels for the PET and CT parts of the reports during annotation was a deliberate choice, as this would have almost doubled the amount of class labels. It would not only increase the time needed for annotating the reports and the size of the dataset needed, but since texts under equal labels in the PET and CT parts

of the reports were very similar, it could also significantly influence the final results. In the worst case, the report would become unusable for clinicians when findings of the PET research would end up under the CT section of the final report.

Besides further splitting the CRFs' output in usable text blocks for the structured report, several retrieval queries were written to reliably return information such as dates and referral data from the *noText* label class. This concerned information present in the report header (which was always classified under this specific class) and could therefore easily be retrieved using rule-based queries.

4.2.5 Composing the structured report

Structured reports can be composed by ordering labelled sequences and structural elements such as (sub-)headings, page breaks and white space in any desired fashion. To form sentences from the CRFs' output after it was split, another piece of JAVA code was written. For each report that was part of the CRFs' output, this code joined successive tokens with identical labels together, while putting space characters in between. It also removed headings that were labelled by the CRF but that were no longer of any value (e.g. texts that became redundant due to the fact that template files contained new (sub-)headings). Furthermore, incorrect white space that was inside or around measurements, dates and punctuation was removed or modified to make the overall sentences more readable. Sequences of tokens were stored under the label the CRF had assigned to the sequence. The order of CRFs' output (which was in turn equal to the order of input) was fully maintained to avoid unintended sentence or paragraph transitions which in turn could result to misunderstandings of the report contents, and no texts were replaced to preserve the radiologists' intended meaning.

The blocks of text assigned under a specific label could then be individually retrieved. These text blocks could then be used to compose a structured report in any format desired (i.e. using XML tags, HTML documents, or plain text files). In this study, the CRFs' output was written to the structured templates discussed in section 4.1. For each report two formats were composed: one with the conclusion before the section communicating the findings and one with the conclusion at the end of the report, after the findings.

Finally, an iteration was performed over the entire structured report to identify whether the report contained texts on positron emission tomography (PET) and/or computed tomography (CT) information, and empty, irrelevant sections were removed. Thus, if the report did not contain any information on a PET scan at all, then all PET related sections were removed from the report to clean up the final, finished product.

4.3 User evaluation study

Two structured reports (in both formats, namely a format with the conclusion section at the end of the report, and a format with the conclusion section before the findings section) that were representative for the full set of reports were printed for evaluation purposes. The corresponding free-text reports were also printed for comparison. To avoid unintended visual differences during comparison, all reports were printed in the same font and with identical font-size.

For the user evaluation study, 14 referring clinicians at the UMCG were sent an email in which they were asked to participate in informal interviews about their impressions and judgements of the standardized report. The clinicians were contacted via one of the radiologists at the UMCG, who was asked to forward our email request to clinicians who could contribute to the project by sharing their opinion, and who were familiar with radiology reports on the malignant lymphoma. After one week the

The CONVENTIONAL report	Fully disagree						Fully agree
The report is organized well	1	2	3	4	5	6	7
The report has a clear structure	1	2	3	4	5	6	7
The report is not complex	1	2	3	4	5	6	7

Figure 4.2: Example of 7 point Likert scales in the questionnaires that were used during the evaluation interviews.

clinicians who had not yet responded to the email were called and asked to participate. A total of 5 clinicians (all with more than 20 years of experience with reading the radiology report) responded positively and were interviewed and asked to share their opinion. The interviews took place in the offices of the clinicians at the UMCG and each interview took approximately 30 minutes of the clinician's time. All interviews started with a short explanation of the study and its main goals. Clinicians were reminded of the goal of interview (e.g. evaluation the system's output in comparison to the current reports). After this short introduction the clinicians were shown the free-text reports and corresponding standardized reports simultaneously. Clinicians were asked not to focus as much on report content such as the used vocabulary or sentence structure, but more on their overall impressions. They were encouraged to freely comment on the structure of the reports. While evaluating the reports, clinicians were asked to use the think aloud protocol. During the entire evaluation, the participants could ask any question they wanted about the report.

At the end of the interview, clinicians were asked to fill out a single page questionnaire about the free-text and structured reports in order to capture their impressions in self-appointed scores. The questionnaire consisted of statements concerning the clarity, structure, complexity and problem orientation of the report, and whether the clinician was under the impression that the report allowed him/her to quickly find relevant information. At the end of the questionnaire, the clinician was asked to choose the report that had his/her preference.

All statements could be evaluated by circling a score on a Likert scale: a psychological measurement scale developed by psychologist Rensis Likert (Likert, 1932) that is very commonly used in research questionnaires. Regarding the Likert scale size, Dawes (2008) suggests that there is no real reason for favouring any one number of scale points on the Likert scale over any other. These findings were contradicted by Lozano, García-Cueto, and Muñiz (2008), who found that the use of seven point Likert scales is most optimal. An ongoing meta-analysis of scale point studies by Krosnick and Tahk (2014) supports these findings and also suggests that the seven point Likert scales are best suited for bipolar scales. Therefore, seven point bipolar Likert scales were used. The Likert scales ranged from one to seven and were end-defined with one and seven being formatted as 'I fully disagree' and 'I fully agree', respectively. Figure 4.2 shows an excerpt of the statements and Likert scales as used in the questionnaires.

Chapter 5

Results

5.1 Initial CRFs results

The CRFs' output was used to calculate the true positives, false positives, and false negatives (see contingency table 5.1) from the differences between the CRF's assigned labels in the n^{th} column in the output, and the true answer label in the $n-1^{\text{th}}$ column of the output file.

The initial results of the CRFs learner are displayed in table 5.2. The F-scores in this table are the average results of five combinations of training and test sets, which is the result from applying the K-fold cross validation technique to the dataset.

The trained feature column indicates the training condition that was used for the CRF by specifying a specific feature template. The *none, feature only* condition therefore corresponds to using a feature template in which only the current token and neighbouring tokens were taken into account during trained, without specifying any extra features to train on. The *pos 2* condition refers to training the CRF with a feature template in which a positional feature that divides the report into $x = 2$ parts is added. The corresponding *pos 4* refers to training with a positional feature that divides the report into $x = 4$ parts. Note that the *none, feature only* condition is comparable to training with a positional feature that divides the report into x=0 parts, thus *pos 0*. The *pos 4 + colon + capital* condition is an extension of the *pos 4* condition by also specifying macro templates for the colon and capital features. Finally, the most extensive conditions trained are the conditions in which the word frequency features are also specified in the feature template. They are indicated by *wordfreq*, the *wordfreq 25*, *wordfreq 50* and *wordfreq all* conditions indicate that only the top 25, top 50 or all of the frequency counts were used when creating the feature columns.

Table 5.2 shows both micro and macro averaged F-scores. The F-score considers

Table 5.1: Contingency table specifying the possible outcomes of results versus the given 'gold standard' of the annotations in the test set.

		Condition ('gold standard')	
		Condition X	Condition other than X
Test outcome	Outcome X	True positive	False positive (type I error)
	Outcome other than X	False negative (type II error)	True negative

Table 5.2: Average F-scores calculated from the CRF output over five combinations of training and test sets. For the micro averaged F-score, each classification decision is counted separately, while for macro averaged F-score, equal weight is given to each class label.

	CRF output	
Trained features	F-score (micro)	F-score (macro)
none, tokens only	86.79	87.08
pos 2	87.24	87.11
pos 4	87.70	87.45
pos 4 + colon + capital	88.18	87.85
pos 4 + colon + capital + wordfreq top 25	87.24	86.99
pos 4 + colon + capital + wordfreq top 50	87.37	87.08
pos 4 + colon + capital + wordfreq all	87.98	87.66

both the precision (i.e. the probability that the class has been predicted) and recall (i.e. the model's ability to select instances of the corresponding class; commonly called *sensitivity*) and can be interpreted as the weighted average of both.

Micro averaged F-scores are calculated by first summing up individual true positives, false positives, and false negatives of the output, and using these scores for further statistics of the F-score. Since the classification decision on each individual token counts as one, labels with higher token count will be weighted heavier in the final score. One could say that topics count proportionally to their frequency. In macro averaged scores, equal weight is given to each class or label. The effectiveness on the large classes in the test collection is therefore better represented by the micro averaged scores, while the effectiveness of smaller classes is better represented by the macro averaged scores (Manning, Raghavan, and Schütze, 2008).

5.2 Results after enhancement

The assigned labels for the texts on the *Clinical background and question* and the *Conclusion* were enhanced by applying the post processing algorithm discussed in section 4.2.3. Post processing the initial CRFs' results improved results on both precision and recall of most classes. Since the results were drastically improved, no extended analysis will be provided on the initial CRFs' results, as they were superseded by the newer, post processing results. For example, on one of the test sets the recall on the class label *Conclusion* improved from 88.53% to 98.38%.

Table 5.3 shows the F-scores after enhancement of these labels. Similar to table 5.2, the F-scores in table 5.3 are the average results of five combinations of training and test sets.

Further analysis of results on the data after the post processing step showed overall good precision and recall on class labels. In the *tokens only* condition, the lowest scoring class labels are *Armpits* and *Retroperitoneum*. The CRF obtained a precision of 88.77% and recall of 64.19% on the label *Armpits*, and a precision of 75.37% and recall of 80.74% on the label *Retroperitoneum*. Highest scoring classes are the *Person* class, with which all personal names were annotated, and the *Camera* class, which holds the information such as brand an type of the medical imaging camera used in the examination. Furthermore, the results show high precision and recall for the *noTag* class label. Scores in the other conditions were comparable, with highest and lowest scoring classes identical to the aforementioned classes.

Table 5.3: Average F-scores calculated from the enhanced CRF output over five combinations of training and test sets. For the micro averaged F-score each classification decision is counted separately, while for macro averaged F-score equal weight is given to each class label.

	Enhanced output	
Trained features	F-score (micro)	F-score (macro)
none, tokens only	89.03	88.66
pos 2	88.35	87.81
pos 4	88.68	88.11
pos 4 + colon + capital	89.30	88.60
pos 4 + colon + capital + wordfreq top 25	88.37	87.76
pos 4 + colon + capital + wordfreq top 50	88.43	87.79
pos 4 + colon + capital + wordfreq all	89.08	88.40

Table 5.4: Average precision and recall on the post-processed results from the *tokens only* condition of the CRF. Averages are calculated from five combinations of training and test sets.

Class label	**Precision**	**Recall**
Abdomen/pelvis	84.97 %	82.28 %
Camera	97.98 %	99.67 %
Conclusion	82.59 %	98.19 %
Contrast	93.89 %	85.74 %
Head/neck	87.29 %	82.57 %
Musculoskeletal	91.10 %	75.54 %
Armpits	88.77 %	64.19 %
Person (executing radiologist)	97.55 %	98.81 %
Retroperitoneum	75.37 %	80.74 %
Scan quality	94.30 %	79.82 %
Scanning method (comments)	96.81 %	88.52 %
Scanning protocol	93.17 %	87.06 %
Thorax	82.97 %	82.24 %
Comparison	92.37 %	85.32 %
Clinical background and question	93.54 %	96.87 %
noTag	99.24 %	99.46 %

Table 5.5: Likert score results from the user evaluation study. The table shows individual scores of participants (e.g. s 1,. . . , s 5) as well as the median and mode of those scores. Likert scores ranged from 1 (fully disagree) to 7 (fully agree).

CONVENTIONAL REPORT	Likert score						
Statement	s 1	s 2	s 3	s 4	s 5	**Median**	**Mode**
1. The report is organized well	5	3	5	5	4	**5**	**5**
2. The report has a clear structure	5	4	4	5	6	**5**	**5**
3. The report is not complex	1	1	2	2	1	**1**	**1**
4. The report is problem-oriented	6	3	4	6	6	**6**	**6**
5. I can quickly find what I'm looking for	6	3	5	6	3	**5**	**6**

RE-STRUCTURED REPORT	Likert score						
Statement	s 1	s 2	s 3	s 4	s 5	**Median**	**Mode**
1. The report is organized well	6	6	6	5	6	**6**	**6**
2. The report has a clear structure	6	6	5	5	6	**6**	**6**
3. The report is not complex	1	2	2	5	1	**2**	**1**
4. The report is problem-oriented	6	5	4	7	6	**6**	**6**
5. I can quickly find what I'm looking for	6	6	5	6	3	**6**	**6**

5.3 Results from the user evaluation study

Table 5.5 shows the results of the Likert scale data from the user evaluation study. It may be false to assume that the intervals between Likert scores were equidistant, since participants may have seen the intervals of the Likert scales different. Therefore, the table shows median and mode instead of mean and standard deviation.

Participants' self-assigned Likert-scores from the user evaluation study were plotted in the figures 5.1, 5.2, 5.3, 5.4, and 5.5. Each figure shows both the participants' scores on the conventional and on the re-structured report. Results show that participants scored the new, re-structured reports higher on the organization of elements, and clarity of structure. Participants furthermore scored the re-structured reports as less complex.

5.3.1 Results and recommendations from interviews

In this section, the remarks and recommendations that referring clinicians made during the evaluation interviews will be discussed. If clinicians gave suggestions for improvement, these will also be discussed.

After examining both possible templates, clinicians indicated their report preference, this was directly based on how they read the radiology report. Clinicians' reading preference (i.e. how they usually read a radiology report) was to read the report from top to bottom, left to right, with a preference for the conclusion at the end [2/5]; to jump to the conclusion, and afterwards read the rest of the report, with a preference for the conclusion before the findings [2/5]; or undetermined [1/5].

Clinicians indicated that they did not like too much white space in the report since the radiological reports are often read on (screens with) limited screen space. Therefore, unnecessary white space and new lines should be reduced as much as possible [3/5]. This remark was focussed on both the report content [2/3], as well as the

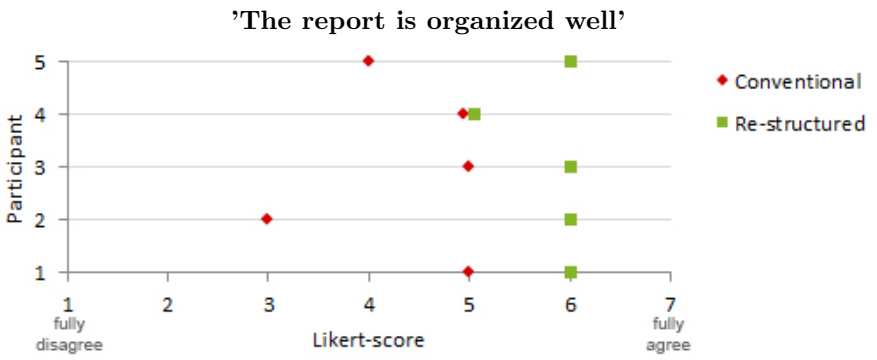


Figure 5.1: Participants' self-assigned Likert scores in the evaluation study for the statement 'The report is organized well'.

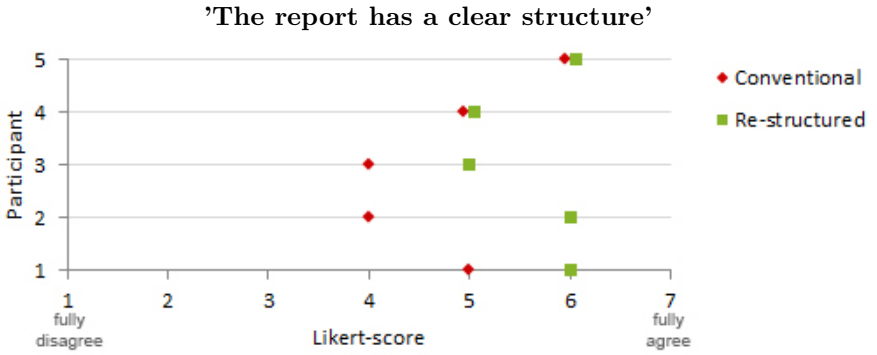


Figure 5.2: Participants' self-assigned Likert scores in the evaluation study for the statement 'The report has a clear structure'.

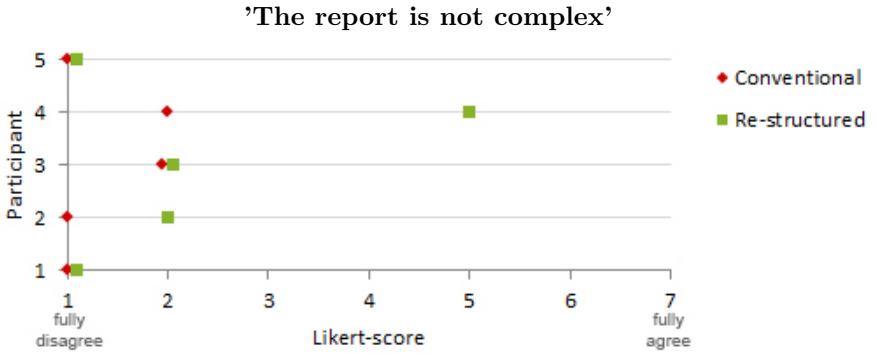


Figure 5.3: Participants' self-assigned Likert scores in the evaluation study for the statement 'The report is not complex'.

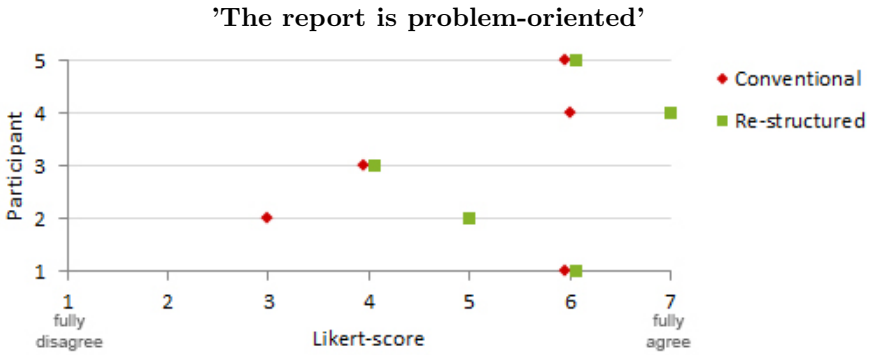


Figure 5.4: Participants' self-assigned Likert scores in the evaluation study for the statement 'The report is problem-oriented'.

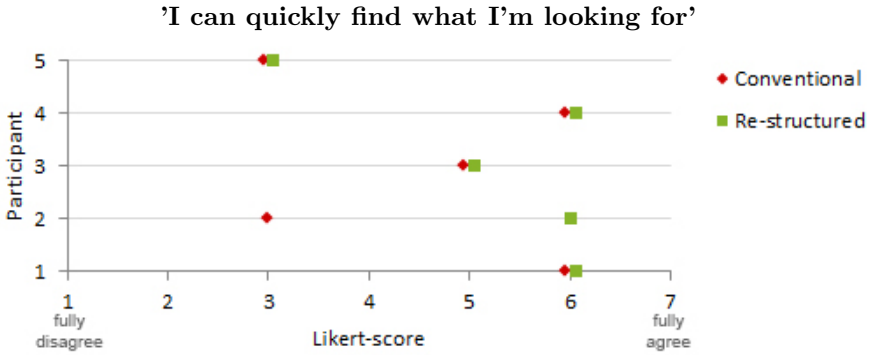


Figure 5.5: Participants' self-assigned Likert scores in the evaluation study for the statement 'I can quickly find what I'm looking for'.

global structure [2/3]. For information; the re-structured report compared to the conventional report were of approximately similar length when printed. A suggestion for improving the global report [1/5], was to automatically merge all findings that were considered normal under a single sentence. This could also be used to reduce white spaces and make the report more compact.

Another suggestion from clinicians [2/5] was to further change the order of elements and start each section of findings with abnormalities, or -if this proved to be impossible- indicate abnormalities and normalities with an indicator in the sideline or a background colour. And finally, to further standardize the report by also extending the templates with sub-level headings under the current headings of the findings section [1/5].

During the evaluation, clinicians [2/5] indicated that they found that the content of the conventional report was decent. They did had some concerns with the overall structure of the conventional report from past experiences. The clinicians indicated to be pleasantly surprised with the re-structured report. One of the clinicians even asked when a test period with the restructured reports could start to evaluate results in a real-life situation.

One clinician was not at all satisfied with the radiological report's content of reports by some radiologists, and therefore openly questioned the value of the re-structured report in those cases. The clinician in question indicated that drastic changes were needed in these cases to come to a quality level that was satisfactory again. The clinician's main concern was the fact that radiologists in question reported in such a way that the findings could not be formed into a mental picture, making the findings impossible to remember after reading the report.

Chapter 6

Discussion

We have developed an effective computational machine learning system capable of assigning class labels to texts contained in dictated, free-text radiology reports on the malignant lymphoma, together with computational post-processing steps needed to produce radiology reports with a standardized form and structure. The machine learning system in place used an approach called Linear Chain Conditional Random Fields (LC-CRFs), a technique aimed specifically at sequence learning.

Results show that the machine learning system yielded good results when trained on the sequences of words, sentences and sections in dictated, free-text radiology reports. The system was trained with relatively simple features that were directly based on the input data, without the need for expert knowledge. A combination of a positional feature and colon and capital features was shown to yield the largest increase in performance. Comparison against the output when using no features at all showed increases in the F-scores from 86.79 to 88.18 (micro averaged) and 87.08 to 87.85 (macro averaged) on the regular CRF output. The output after our post-processing step, which was aimed at eliminating potentially misclassified tokens under the class labels *Clinical background and question* and *Conclusion*, showed the overall micro averaged F-score increasing from 89.03 to 89.30 when using these features, and a slight decrease in the macro averaged F-score from 88.66 to 88.60.

The classes *Musculoskeletal* and *Armpits* score lowest on recall, but good on precision. We believe the low recall on *Musculoskeletal* can be explained by the fact that texts labelled under this class were often somewhat integrated within the context. Therefore, the texts could not always be labelled separately from its context (e.g. when the sentence 'No indication for musculoskeletal abnormalities in this area.' occurred in a text on the *Abdomen/pelvis*). In these cases, we chose to annotate the text with the context class to preserve meaning. The low recall on the *Armpits* class can be explained by the shortage of training examples. Not all annotated reports contained findings on the armpits area. The same holds for the class *Retroperitoneum*, which scores lowest on precision and reasonable on recall.

We hypothesized that the word frequency features would improve the results in these less frequent classes especially. However, the addition of feature columns that held information on word frequencies under the distinct class labels did not improve results any further. We suspect that by adding the word frequency feature columns, the extra information contained in those features does not extend beyond the increase in extra feature space that is a result of adding the new feature columns, and therefore decreases the results. Under the *Future work* section we will discuss a way to investigate this hypothesis, as we still support the idea behind it, and believe this could improve results especially in the low scoring classes *Armpits* and *Retroperitoneum* since these classes contain several high frequent tokens that may function as clear indicators

for the class labels that need to be assigned.

The results from the machine learning system support the work of Esuli, Marcheggiani, and Sebastiani (2013), who used a LC-CRFs learner for the analysis of mammography reports in the Italian language. Their baseline system is in principle comparable to the system described in the present thesis. Our study is more extensive though, as these authors trained the system only on extracting relatively short pieces of information (average length of 17.33 words). Also the amount of labels in their study was relatively low with nine distinct labels, resulting in only a few of the important pieces of information being automatically extracted.

Contrary to the MedLEE system (Friedman et al., 1994) which uses a rule based natural language processor, our computational system is based on a machine learning technique. It therefore requires only annotated reports to train on, and can easily and cost efficiently be extended to radiology reports on other topics than the malignant lymphoma.

Our system processes the radiology report as a whole, by classifying all tokens present in the report to one of 16 classes. This means that our system is capable of identifying and extracting many different types of information from the radiology report at once while many previous systems focussed only on extracting specific information. By first processing the text as a whole, we believe that further processing steps can be performed in a more informed and intelligent manner since one can benefit from the classifications. This is similar to the approach taken by Yetisgen-Yildiz et al. (2013), who used a Maximum Entropy model to first classify and process the report as a whole before extracting specific information in the form of positive or negative follow-up recommendations. It is important to note that improving results in this first classification step also may help improve the final results.

The results from the user evaluation study suggest that improvements to the global structure of the radiology report can increase scores on clarity and the organization of elements while decreasing complexity. These improvements attribute to better information transmission and improvements in patient care. While interpreting these results, it should be taken into account that the group size of the evaluation study was somewhat limited.

These results support the findings of Schwartz et al. (2011) in terms of improving the clarity satisfaction when comparing a structured report versus a conventional report. However, we did not use a structured reporting method as described in that study, but allowed the radiologist full freedom during dictation. This is in line with the view of Pool and Goergen (2010) that was discussed earlier, namely that future software systems should be designed with the complexity of the radiologist's task in mind. Machine learning systems are exceptionally well suited for situations in which the a system needs to be designed 'around' the user. The system can be trained on elements already present in the input (e.g. dictated free-text), so that no further input is required from the user and the user can keep its attention focussed on the task at hand. Furthermore, a fully trained machine learning system can process large amounts of complex input extremely fast, and therefore does not form a potential bottleneck in the users' task execution.

The study in general has some limitations. The major being that there was no dataset of annotated radiology reports available that suited our project. The only solution was to accumulate reports for a dataset and annotate these reports by hand. Time constrains limited the number of reports that could be annotated and added to the dataset. This resulted in a relatively small dataset for training and testing purposes that may not have been the most perfect representation of the reports produced by the team of radiologists at the UMCG. Although our results were shown to be stable across the five combination of training and test sets from the k-fold cross validation, the use of a larger test set would have been interesting. A larger dataset could help to better identify differences between the applied features in distinct conditions. The

final trained system would also be more likely to correctly classify the less frequent sequences of tokens in the data, as more training data is available. Furthermore, the dataset contained only radiology reports on the malignant lymphoma. The use of reports on other diseases could lead to different results.

The conditional random fields learners are a new type of machine learning tools. No previous research is available on developing features and writing powerful feature templates, making optimization of the templates used by the machine learner an act of editing and fine-tuning the features while work on the system progressed. Cases where conditional random fields learners are applied to the medical setting and medical texts are scarce. Further research using this type of machine learning technique could help to improve results in the future.

Finally, the group size of participants in our user evaluation study was relatively small, which makes it difficult to draw firm conclusions from the obtained data. During the interviews, it was also hard to keep participants focussed on the goals of the interviews as they were easily distracted by the content of the report and by providing further suggestions for improving the report. On the other hand, this also is a clear indicator that participating clinicians found it gratifying to be part of actions to improve the radiology report.

For our system or comparable systems to get widely adopted for report processing in the radiological setting, results will need to be further improved. Radiologists who have followed our research closely and who have seen the results have indicated only to let a computational system automate part of their tasks if they have a good sense of trust and confidence in the system making the correct classifications. To acquire this trust, scores on the classifications will need to be further improved. Radiologists working with the system will not accept classification errors on a regular basis, especially on word sequences or sentences that are easily identifiable (i.e. for humans who have a medical background). For the time being, this means that more work is needed.

We foresee that when the time comes that machine learning systems will be able to do text classifications on medical texts like those described in this study with near perfect results, these machine learning systems will quickly get adopted and may become the standard for real-time processing of complex (dictated) medical texts. Trained machine learning systems have the advantage over rule based solutions that they can quickly come to a classification decision based on complex input, even if the input is incomplete.

Before machine learning systems will become the standard for real-time processing of medical texts however, new systems need to be developed. Besides using a trained machine learner as the core input processor, these systems need to provide the user (i.e. the radiologist) with an interface in which a structured radiology report appears after the user provides input. The interface should also allow the user to add, remove and edit automatically classified texts by any means possible. Future systems should work in harmony with the user for optimal results and should not attempt to take over the task at hand. The final reports may be sent to referring clinicians in standardized formats that match clinicians' individual preferences, but only if this can be accomplished without changing the radiologists' intended meaning of the content.

6.1 Future work

The current research leaves various potential paths for future research. Several improvements can be made to the study as it currently is, or researchers can use this study as a starting point for developing new (support) systems.

First of all, the dataset could be extended by annotating more reports and strictly

balancing the report in order to come to a more representative set of reports as generated by the radiologists at the UMCG. Second, the used template structures could be improved by decreasing white space and evaluation tests in real life situations. The current evaluation study with five participants is likely to already address the most important aspects. However, a third improvement to the current research could be to extend the user evaluation study with more participants and a more comprehensive evaluation study. This could yield more detailed information and insights into the current template structures and its strong points and weaknesses. It would also be important to investigate the effect of our new report format on information transfer to the reader.

When one wants to further extend the current research, he or she could try to improve the conditional random fields learner's output by improving feature templates and by using more (advanced) features. New features could also improve results, and could for example be on describing grammatical structure in sentences; positional features of larger text blocks; or word roots.

When adding new features or improving the old, one must always attempt to keep the feature space as small as possible. For this study in particular, it would be interesting to determine if decreasing the amount of columns needed to add word frequency information as an extra feature (e.g. in an alternated form) could improve results after all, since then the feature space could be decreased. In the current situation this could for example be accomplished by using a single feature that holds the class label of the class under which the current token is most frequently observed, or a simple character when the token is not observed under any class label (or when it does not occur in the top x most frequent tokens). For this to work properly, one would of course also need to develop and test new feature templates that match the changed circumstances in an effective manner.

There exist several possibilities for future research that make use of the output of the discussed system. For example, one could use the labelled texts from the CRF's output for further computational processing. This means that the to be developed system can use the labels as foreknowledge to anticipate beliefs about information in the texts and/or to improve results on further processing steps. Another potential research opportunity would be to use a system similar to the one described in this study to check whether the radiologist has discussed all aspects of a study that are required for a structured template report. Only if the report contains at least some information on all the required elements, then the system may allow the radiologist to directly submit the report to the database from which the clinician can later retrieve it. If the radiologist did not discuss all elements, the system can provide a reminder to check the report for completeness.

Further information extraction systems can also help to make the radiology report into an intelligent report that specifically suits each medical specialist's needs. One could, for example, think of intelligent tools that use identified measurements in a report to plot charts and calculate developments over time. Another example could be to provide direct links between texts and related radiological images (i.e. by keeping track of viewed images and generated texts during report creation, and providing direct links between the content and images in the final report). Providing links to the most important images (or key images) could help clinicians when trying to envision the radiological findings, making them easier to remember later on.

Another research track, one that focusses more on specific report content, combines the clinical question with the conclusion and processes their actual sentences. During our research we have heard multiple clinicians mention that the clinical question sometimes was only partially or not at all answered in the conclusion. This of course raises a concern, since getting an answer to the medical question is the sole

reason of doing the radiological research in the first place. To solve this, one could use the automatically labelled sections of the medical question and conclusion and further language processing systems to determine whether an actual answer has been provided to the question at hand. This could take away frustration for the clinicians and help improve report pertinence.

A final research opportunity would be to find out what the best (and least intrusive) way would be to incorporate a computational system for composing structured radiology reports into the workflow of the radiologists. Knowing how to do this can help to develop systems with not only the radiologists' tasks in mind, but also in such a way that the interaction between the radiologists and the system will be as effective as possible.

6.2 Conclusion

In an ideal situation, radiologists have the opportunity to completely focus on image interpretation, without experiencing any distractions from their environment. For the reporting system as a whole, this would mean that the input for further computational steps would be not much more than a single block of text with findings and a conclusion. The text may even lack any form of structure or punctuation, since radiologists should not have to worry about this. The ideal systems then, would be able to efficiently and effectively process this input into a structured information format that can be shown in real time to the radiologist for direct editing when needed. The structured information could furthermore be processed in a standardized, structured report format specifically suited for (individual) referring clinicians or for other use in the medical environment.

Since medical texts are highly susceptible to abbreviations, poor sentence structure, incorrect spelling and more human-introduced complexities, they are hard to process using conventional language processing systems. The machine learning system of conditional random fields that was used in this study to assign class labels and therefore extract texts in free-text radiological reports for use in further computational systems yielded good results on both larger and smaller classes and showed that machine learning systems provide a good means for classification of information in free-text radiological reports. A user evaluation study was performed to evaluate standardized reports that were the result of arranging the machine learner's output in such a way that clinicians could easily relate to the final arrangement. Results showed that clinicians scored these re-structured reports superior over conventional free-text reports.

Our results, together with further technological advances in information extraction and support systems leave us with promising prospects for the radiology report.

Acknowledgements

I want to take this opportunity to express my gratitude to several people without whom this project would not have been possible.

First and foremost, I would like to thank my internal supervisor, Fokie Cnossen, for her valuable feedback and support throughout the project, and for all our meetings which often turned into lighthearted conversations at the end.

I would also like to thank my external supervisors Peter van Ooijen and Wiard Jorritsma for their support, feedback and input in this project, and for providing accommodation in the UMCG. Their willingness to meet at short term and to point me to the right people helped tremendously.

In addition, my thanks go out to thank Fons Bongaerts for clarifying all ambiguities I had about the radiologic report, and for putting me in contact with referring clinicians within the UMCG.

Special thanks go out to Gosse Bouma from the department of Humanities Computing at the University Groningen, for discussing the natural language processing and machine learning part of the project.

Finally, I would like to thank my friends and family for their moral support and their continuous interest in my progress in this project.

Bibliography

American College of Radiology (1998). *Breast imaging reporting and data system (BI-RADS)*. American College of Radiology.

Bhargavan, M., A. H. Kaye, H. P. Forman, and J. H. Sunshine (2009). "Workload of radiologists in United States in 2006-2007 and trends since 1991-1992." en. In: *Radiology* 252.2, pp. 458–67.

Bosmans, J., L. Peremans, A. Schepper, P. Duyck, and P. Parizel (2011). "How do referring clinicians want radiologists to report? Suggestions from the COVER survey". In: *Insights into Imaging* 2.5, pp. 577–584.

Bosmans, J., J. Weyler, A. De Schepper, and P. Parizel (2011). "The radiology report as seen by radiologists and referring clinicians: results of the COVER and ROVER surveys." In: *Radiology* 259.1, pp. 184–195.

Caruana, R. and A. Niculescu-mizil (2006). "An Empirical Comparison of Supervised Learning Algorithms". In: *Proceedings of the 23th International Conference on Machine learning*, pp. 161–168.

Collard, M. D., J. Tellier, A. I. Chowdhury, and L. H. Lowe (2014). "Improvement in Reporting Skills of Radiology Residents with a Structured Reporting Curriculum". In: *Academic Radiology* 21.1, pp. 126–133.

Cortes, C. and V. Vapnik (1995). "Support-Vector Networks". In: *Machine Learning* 20.3, pp. 273–297.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters (2011). *Text Processing with GATE (Version 6)*. GATE (April 15, 2011).

Cunningham, H., V. Tablan, A. Roberts, and K. Bontcheva (2013). "Getting more out of biomedical documents with GATE's full lifecycle open source text analytics." In: *PLoS computational biology* 9.2. Ed. by A. Prlic, pp. 1–16.

Dawes, J. G. (2008). "Do Data Characteristics Change According to the Number of Scale Points Used ? An Experiment Using 5 Point, 7 Point and 10 Point Scales". In: *International Journal of Market Research* 51.1, pp. 61–77.

Demner-Fushman, D., W. W. Chapman, and C. J. McDonald (2009). "What can natural language processing do for clinical decision support?" In: *Journal of biomedical informatics* 42.5, pp. 760–72.

Dreyer, K. J., M. K. Kalra, M. M. Maher, A. M. Hurier, B. A. Asfaw, T. Schultz, E. F. Halpern, and J. H. Thrall (2005). "Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study." en. In: *Radiology* 234.2, pp. 323–9.

Esuli, A., D. Marcheggiani, and F. Sebastiani (2013). "An enhanced CRFs-based system for information extraction from radiology reports". In: *Journal of Biomedical Informatics* 46.3, pp. 425–435.

Forney, G. (1973). "The Viterbi Algorithm". In: *Proceedings of the IEEE* 61.3, pp. 268–278.

Friedman, C., P. O. Alderson, J. H. Austin, J. J. Cimino, and S. B. Johnson (1994). "A general natural-language text processor for clinical radiology." In: *Journal of the American Medical Informatics Association* 1.2, pp. 161–74.

GATE Research Team (1995). *Gate: General Architecture for Text Engineering*. URL: `http://gate.ac.uk/` (visited on 08/22/2013).

Geman, S., E. Bienenstock, and R. Doursat (1992). "Neural Networks and the Bias/Variance Dilemma". In: *Neural Computation* 4.1, pp. 1–58.

Hickey, P. M. (1904). "The interpretation of radiographs". In: *Journal of Michigan State Medical Society* 3, pp. 496–499.

Hickey, P. M. (1922). "Standardization of Roentgen-ray reports". In: *American Journal of Roentgenology* 9, pp. 422–425.

Johnson, A. J. (2002). "Radiology report quality: a cohort study of point-and-click structured reporting versus conventional dictation." In: *Academic radiology* 9.9, pp. 1056–61.

Johnson, A. J., M. Y. Chen, J. S. Swan, K. E. Applegate, and B. Littenberg (2009). "Cohort study of structured reporting compared with conventional dictation." In: *Radiology* 253.1, pp. 74–80.

Johnson, A. J., M. Y. Chen, M. E. Zapadka, E. M. Lyders, and B. Littenberg (2010). "Radiology Report Clarity: A Cohort Study of Structured Reporting Compared With Conventional Dictation". In: *Journal of the American College of Radiology* 7.7, pp. 501–506.

Krosnick, J. A. and A. Tahk (2014). *The Optimal Length of Rating Scales to Maximize Reliability and Validity*. URL: `https://pprg.stanford.edu/krosnick-research-projects/` (visited on 01/02/2014).

Krupinski, E. A., E. T. Hall, S. Jaw, B. Reiner, and E. Siegel (2012). "Influence of radiology report format on reading time and comprehension." In: *Journal of digital imaging* 25.1, pp. 63–9.

Kudo, T. (2013). *CRF++: Yet Another CRF toolkit*. URL: `http://crfpp.googlecode.com/` (visited on 11/27/2013).

Lafferty, J. D., A. McCallum, and F. C. N. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the 18th International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., pp. 282–289.

Langlotz, C. P. (2006). "RadLex: A New Method for Indexing Online Educational Materials". In: *RadioGraphics* 26.6, pp. 1595–1597.

Likert, R. (1932). "A technique for the measurement of attitudes." In: *Archives of Psychology* 22.140, pp. 1–55.

Lozano, L. M., E. García-Cueto, and J. Muñiz (2008). "Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales". In: *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 4.2, pp. 73–79.

Malouf, R. (2002). "A comparison of algorithms for maximum entropy parameter estimation". In: *Proceedings of the 6th Conference on Natural Language Learning (CoNLL)*, pp. 49–55.

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval*. Cambridge University Press, p. 496.

McCallum, A. and D. Freitag (2000). "Maximum entropy markov models for information extraction and segmentation". In: *Proceedings of the 17th International Conference on Machine Learning*, pp. 591–598.

Mohri, M. (2011). *Course Slides for Introduction to Machine Learning*. URL: `http://www.cs.nyu.edu/~mohri/mlu/` (visited on 10/24/2013).

Nievelstein, R. A. J., C. Schaefer-Prokop, B. G. Heggelman, R. Boellaard, P. J. Lugtenburg, and J. M. Zijlstra (2012). "Aanbevelingen: Standaardisatie van aanvraag, uitvoering en verslaglegging van CT beeldvorming in het kader van FDG-PET/CT onderzoeken bij patiënten met een maligne lymfoom". In: *Nederlands Tijdschrift voor Nucleaire Geneeskunde* 34.1, pp. 724–732.

Nocedal, J. (1980). "Updating Quasi-Newton Matrices with Limited Storage". In: *Mathematics of Computation* 35.151, pp. 773–782.

Okazaki, N. (2011). *CRFsuite - CRF Benchmark test.* URL: http://www.chokkan.org/software/crfsuite/benchmark.html (visited on 01/13/2013).

Pool, F. and S. Goergen (2010). "Quality of the Written Radiology Report: A Review of the Literature". In: *Journal of the American College of Radiology* 7.8, pp. 634–643.

Radiology Society of North America (2013). *RSNA Radiology Reporting Initiative.* URL: http://reportingwiki.rsna.org/index.php?title=Main_Page (visited on 03/04/2014).

Robert, L., M. D. Cohen, and G. S. Jennings (2006). "A New Method of Evaluating the Quality of Radiology Reports". In: *Academic Radiology* 13.2, pp. 241–248.

Schwartz, L., D. Panicek, A. Berk, Y. Li, and H. Hricak (2011). "Improving communication of diagnostic radiology findings through structured reporting". In: *Radiology* 260.1, pp. 174–181.

Sha, F. and F. Pereira (2003). "Shallow parsing with conditional random fields". In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology.* Vol. 1. Morristown, NJ, USA: Association for Computational Linguistics, pp. 134–141.

Sistrom, C. L. and J. Honeyman-Buck (2005). "Free text versus structured format: information transfer efficiency of radiology reports." In: *American Journal Of Roentgenology* 185.3, pp. 804–812.

Suominen, H., F. Ginter, S. Pyysalo, A. Airola, T. Pahikkala, S. Salanter Ä, and T. Salakoski (2008). "Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: a method description". In: *System.*

Sutton, C. and A. McCallum (2006). "An Introduction to Conditional Random Fields for Relational Learning." In: *Introduction to Statistical Relational Learning.* Ed. by L. Getoor and B. Taskar. MIT Press.

UMCG (2014). *Radiologie, Over het UMCG.* URL: http://www.umcg.nl/NL/UMCG/overhetumcg/organisatie/Specialismen/Radiologie/Pages/default.aspx (visited on 03/04/2014).

Wallis, A. and P. McCoubrie (2011). "The radiology report - Are we getting the message across?" In: *Clinical Radiology* 66.11, pp. 1015–22.

Wang, S. and R. M. Summers (2012). "Machine learning and radiology." In: *Medical image analysis* 16.5, pp. 933–51.

Wikipedia (2013). *F1 score.* URL: http://en.wikipedia.org/wiki/F1_score (visited on 11/25/2013).

Yetisgen-Yildiz, M., M. L. Gunn, F. Xia, and T. H. Payne (2013). "A text processing pipeline to extract recommendations from radiology reports." In: *Journal of biomedical informatics* 46.2, pp. 354–62.

Zhang, T. and F. J. Oles (2001). "Text Categorization Based on Regularized Linear Classification Methods". In: *Information Retrieval* 4.1, pp. 5–31.