Automatic Structuring of Breast Cancer Radiology Reports for Quality Assurance

Shreyasi Pathak
University of Twente
Enschede, Netherlands
s.pathak@student.utwente.nl

Jorit van Rossen
Hospital Group Twente (ZGT)
Hengelo, Netherlands
j.vrossen@zgt.nl

Onno Vijlbrief
Hospital Group Twente (ZGT)
Hengelo, Netherlands
o.vijlbrief@zgt.nl

Jeroen Geerdink
Hospital Group Twente (ZGT)
Hengelo, Netherlands
J.Geerdink@zgt.nl

Christin Seifert
University of Twente
Enschede, Netherlands
c.seifert@utwente.nl

Maurice van Keulen
University of Twente
Enschede, Netherlands
m.vankeulen@utwente.nl

***Abstract*—Hospitals often set protocols based on well defined standards to maintain quality of patient reports. To ensure that the clinicians conform to the protocols, quality assurance of these reports is needed. Patient reports are currently written in free-text format, which complicates the task of quality assurance. In this paper, we present a machine learning based natural language processing system for automatic quality assurance of radiology reports on breast cancer. This is achieved in three steps: we i) identify the top level structure of the report, ii) check whether the information under each section corresponds to the section heading, iii) convert the free-text detailed findings in the report to a semi-structured format. Top level structure and content of report were predicted with an F_1 score of 0.97 and 0.94 respectively using Support Vector Machine (SVM). For automatic structuring, our proposed hierarchical Conditional Random Field (CRF) outperformed the baseline CRF with an F_1 score of 0.78 vs 0.71. The third step generates a semi-structured XML format of the free-text report, which helps to easily visualize the conformance of the findings to the protocols. This format also allows easy extraction of specific information for other purposes such as search, evaluation and research.**

***Index Terms*—Quality Assurance, Automatic Structuring, Radiology Reports, Conditional Random Field**

I. Introduction

Medical reports are essential for communicating the findings of imaging procedures with referring physicians, who further treat the patients by considering these reports. Thus, medical reports are very important for diagnosis of diseases, which brings forward the need of their quality assurance.

To maintain the quality of reports, hospitals often set well-defined protocols for reporting. For example, for breast cancer radiology reporting, hospitals generally use the "Breast Imaging-Reporting And Data System" (BI-RADS) [1], which is a classification system proposed by American College of Radiology (ACR), to represent the malignancy risk of breast cancer of the patient. It was implemented to standardize reporting and quality control for mammography. The BI-RADS lexicon provides specific terms to be used to describe findings. Along with that, it also describes the desired report structure, for example, a report should contain breast composition and a clear description of findings. The rate of compliance with these reporting standards can be used for quality assurance and also to further measure clinical performance [2].

Conformance to reporting standards can be seen as a part of assessing report clarity, organization, and accuracy [3], [4]. Quality assurance is currently mainly a manual process. Peer review is used to assess report quality, mainly geared towards accuracy of reports [5]. Yang et al. [6] used psychometric assessment to measure report quality and analyzed parameters like report preparation, organization, readability. Making quality assurance systems automatic would reduce workload of radiologists and make the process more efficient. To the best of our knowledge, no system exists to automate this process.

Quality assurance is complicated due to the fact that reporting is done in free-text, narrative format. The inaccessibility of narrative structure for computers makes it hard to analyze if all the necessary information are present in the report. Structured reporting templates can be introduced to force the radiologists to stick to the reporting standards and improve the quality of reports [7], [8]. However, a study [9] shows that this type of system resulted in lower quality reports, as it restricts the style and format of writing. Another method can be automatic structuring of free-text reports after they have been written, without additional technical burden on the radiologists. Thus, the radiologists can concentrate more on the task of interpreting images rather than structure of writing, which helps in maintaining accuracy of the report content.

Thus, in this work, we follow the post-structuring paradigm. We present an approach for automatic structuring of radiology reports for quality assurance using machine learning. We define quality of the report by how well the reports conform to the reporting standards as set by ACR BIRADS. Concretely, we (i) identify the top-level structure from the reports (henceforth, referred to as heading identification), (ii) verify if the report contents are placed under the correct top-level headings (referred to as content identification), and, (iii) automatically convert the free-text report findings to a structured format for making the task of comparison to well-defined protocols easier

(referred to as automatic structuring). For visualization and further use, we generate a semi-structured XML format for the automatic structuring (Table I). We focus on Dutch radiology reports on breast cancer; for automatic structuring we focus on findings from mammography imaging modality.

In the remainder of this paper, we first review structured reporting initiatives and application of natural language processing to radiology reports (Section II). Section III describes the dataset. Our approach to heading and content identification, and automatic structuring is detailed in Section IV. We describe our experimental setup in Section V followed by experimental results in Section VI. We discuss the implication of our results and some future work in Section VII.

II. Related Work

In this section, we will discuss structuring initiatives for radiology reporting, followed by various natural language processing techniques applied in radiology.

A. Structured Reporting Initiatives

Accuracy, clarity, timeliness, readability, organization are some of the important factors for good quality of radiology reporting [3], [4]. Sistrom and Langlotz [7] identified i) language, ii) format as two key attributes for improving the quality of a radiology report. *Standardizing the language* of the report promotes common interpretation of the reports by the radiologists through out the world. Breast Imaging-Reporting and Data System (BI-RADS) is a very successful attempt by ACR at standardizing the language for breast cancer reporting [1]. RadLex [10] is another attempt at standardizing disease terminology, observation and radiology procedure. *Structured reporting* further increases efficiency of information transfer and referring clinicians can extract the relevant information easily. Sistrom and Langlotz [7] clarified that structured reporting does not mean having a point-and-click interface for data capture, rather a simple report format that reflects the way radiologist and referring physician sees the report and should not impose any restriction on the radiologists. Radiological Society of North America (RSNA) highlighted that structured reporting would improve clinical quality and help in addressing *quality assurance* [4].

Though there has been a lot of discussion about the effect of structuring on the quality of radiology report, not much actual assessment was done until 2005. In 2005, Sistrom and Honeyman-Buck [11] tested information extraction from free-text and structured reports. It was found that both the free-text and structured report resulted in *similar accuracy and efficiency* in information extraction, but a post-experimental questionnaire expressed clinicians' opinion in favour of structured report format. Schwartz, Panicek, Berk, Li and Hricak [8] reported that referring clinicians and radiologists found *greater satisfaction with content and clarity* in structured reports, but the clinical usefulness did not vary significantly between the two formats. Whereas, a study by Johnson, Chen, Swan, Appelgate and Littenberg [9], concluded that structured reporting resulted in a *decrease in report accuracy and completeness*. The subjects were asked to use commercially available structured reporting system (SRS), a point-and-click menu driven software, to create the structured reports and they found it to be *overly constraining* and *time-consuming*.

To summarize, past works have shown that firstly, structured reporting and standard language are important for quality of report. But structured reporting should be such that it should not impose restriction on the radiologist. Secondly, structuring reporting can help in addressing quality assurance.

B. Natural Language Processing in Radiology

Electronic health records (EHRs), like radiology reports, increases the use of digital content and thus generates new challenges in the medical domain. It is not possible for humans to analyze this huge amount of data and extract relevant information manually, so automated strategies are needed. There are two types of techniques used in natural language processing for processing data: *i) rule-based* and *ii) machine learning-based* approaches.

In *rule-based approaches*, rules are manually created by experts to match a specific task. Various rule-based systems have been used for information extraction tasks in radiology reports on breast cancer. Nassif et al. [12] developed a rule-based system in 2009 to extract BI-RADS related features from a mammography study. The system was tested on 100 radiology reports manually tagged by radiologists, resulting in a precision of 97.7% and a recall of 95.5%. Sippo et al. [13] developed a rule-based NLP system in 2013 to extract the BI-RADS final assessment category from radiology reports. They tested their system on $>$220 reports for each type of study – diagnostic and screening mammography, ultrasound etc. achieving a recall of 100% and a precision of 96.6%.

Machine learning (ML) approaches can learn the patterns from data automatically given the input text sequence and some labeled text samples. *Hidden Markov Model*, *Conditional Random Field (CRF)* [14] are some of the ML approaches used for sequence labeling. Hassanpour and Langlotz [15] compared dictionary-based (a type of rule-based) model, Conditional Markov Model and CRFs on the task of information extraction from chest radiology reports, finding that ML approaches (F_1: 85.5%) performed better than rule-based (F_1: 57.8%). Torii, Wagholikar and Liu [16] investigated the performance of CRF taggers for extracting clinical concepts and also tested the portability of the taggers on different datasets. Esuli, Marcheggiani and Sebastiani [17] developed a cascaded 2-stage Linear Chain CRF model (one CRF for identifying entities at clause level and another one at word level) for information extraction from breast cancer radiology reports. The cascaded system (F_1: 0.873) outperformed their baseline model of standard one level LC-CRF (F_1: 0.846) on 500 mammography reports.

Hybrid approaches combine rule-based and machine learning-based approaches. For example, Taira, Sodrland and Jakobovits [18] developed a automatic structuring of free-text thoracic radiology reports using some rule-based and some statistical and machine learning methods like maximum

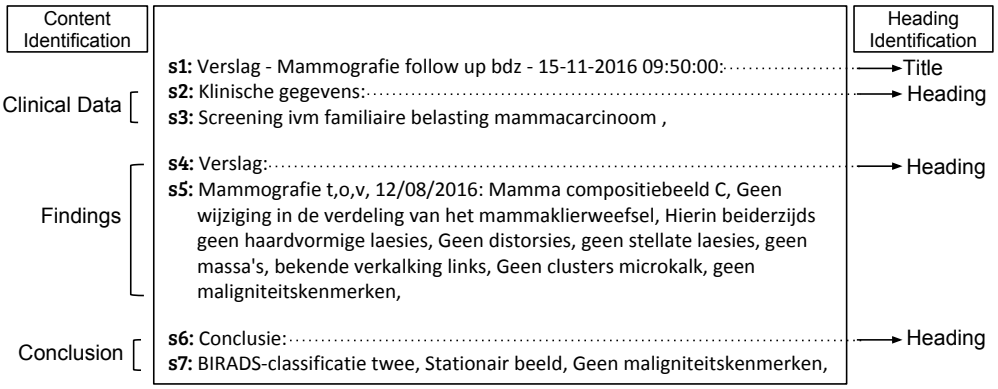


Fig. 1: Example of a breast cancer radiology report

entropy classifier. We want to develop a fully automated system without any rule creation involved from experts, which is why we will not follow hybrid approach.

In this work, we apply machine learning-based approaches to avoid manual rule construction and use CRFs which have been shown to provide high performance on sequence labeling.

III. Corpus: Radiology Reports on Breast Cancer

According to BI-RADS [19], a breast cancer radiology report should contain an indication of examination (clinical data), a breast composition, a clear description of findings, and a conclusion with the BI-RADS assessment category. For our purpose of quality assurance of a report, we will consider these things and annotate the reports accordingly.

We used a dataset of 180 Dutch radiology reports on breast cancer from 2012 to 2017 (30 reports per year). Thus, the dataset contains variation in reports over the years. The reports were gathered from Hospital Group Twente (ZGT) in The Netherlands. The reports are produced by dictation from trainee or consultant radiologist, into an automatic speech recognition system. These automatically generated reports are further cross-checked with the dictation, by radiologists or secretary. The reports were anonymized such that they do not contain patient identity data like patient id, name, data of birth and address. A sample report is shown in Fig. 1. The report has 3 sections, namely *Clinical Data*, *Findings* and *Conclusion*. *Clinical Data* contains clinical history of the patient including any existing disease or symptoms. *Findings* consists of noteworthy clinical findings (abnormal, normal) observed from imaging modalities like mammography, MRI and ultrasound. *Conclusion* provides a summary of the diagnosis and follow-up recommendations and should necessarily contain a BI-RADS category. In the report, these sections start with a heading describing the name of the section, for example, *Klinische gegevens* (Clinical Data), *Verslag* (Findings) and *Conclusie* (Conclusion) (see Fig. 1). Reports from 2017 and 2016 (60 reports) additionally contain a *title*. The dataset consists of both male and female breast cancer reports; for automatic structuring, we focus on female breast cancer reports.

For the first two sub-tasks of heading identification and content identification, 180 reports were manually annotated at the sentence-level by a trained expert. The reports were split into sentences, where a sentence means start of a new line, resulting in 1591 sentences in total. In Fig. 1, sentences are indicated by the labels s1 to s7. For the first sub-task of heading identification, sentences were labeled as *heading* (e.g. s2, s4, s6), *not heading* (e.g. s3, s5, s7) and *title* (e.g. s1). For the second sub-task of content identification, sentences were labeled as *title*, *clinical data* (e.g. s2, s3), *findings* (e.g. s4, s5) and *conclusion* (e.g. s6, s7). For the third sub-task of automatic structuring, we manually extracted the mammography imaging modality findings from the *findings* section of the report, which generated 108 mammography findings. These were manually annotated by two radiologists – a trainee (2 years of experience) and a consultant. Out of 108 reports, 18 reports were labeled collaboratively by both, 45 reports by the trainee and 47 by the consultant. After labeling, these 45 reports and 47 reports were analyzed to highlight any inter-annotator discrepancy, which were further resolved by the annotators.

A 3-level annotation scheme at word-level was followed for automatic structuring as shown in Fig. 2. CA-n in the diagram will be explained in the approach (Section IV-C). At the first level, the reports were annotated as:

- *positive finding* (PF): something suspicious was detected about the lesion in the breast, which might indicate cancer.
- *negative finding* (NF): nothing bad was found or absence of specific abnormalities.
- *breast composition* (BC): density of the breast.
- *other* (O): text not belonging to the above.

After this first level of annotation, the PF were further annotated into second level classes – *mass* (MS), *calcification* (C), *architectural distortion* (AD), *associated features* (AF) and *asymmetry* (AS). At the third level, mass was further annotated as *location* (L), *size* (SI), *margin* (MA), *density* (DE), AF and *shape* (SH). Calcification was further annotated as *morphology* (MO), *distribution* (DI), SI, L and AF. Similar third level annotation was done with AD, AF and AS. The same scheme of second and third level annotation was followed for NF,

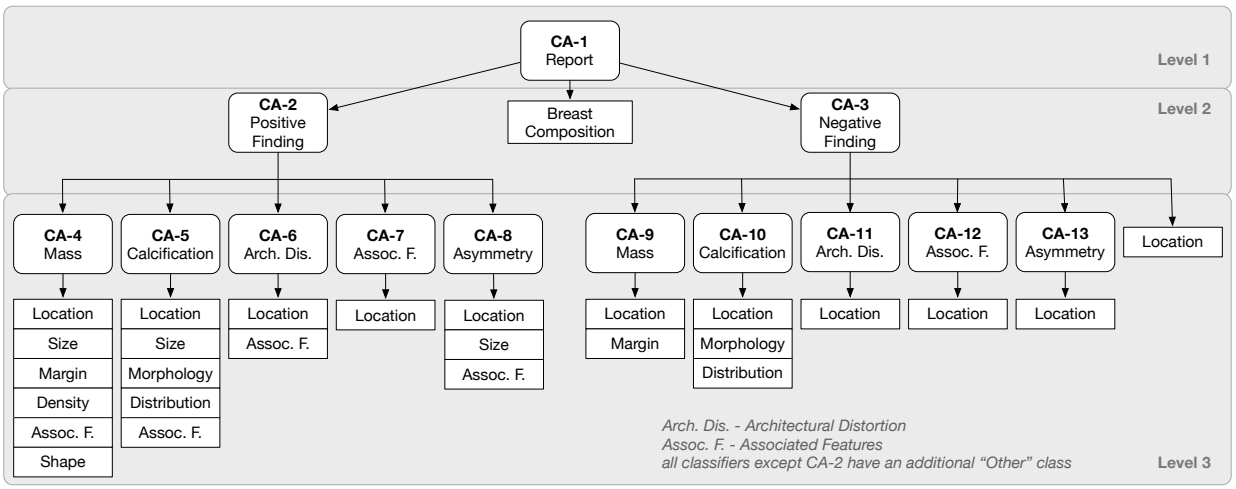


Fig. 2: 3-level annotation scheme for automatic structuring of mammography findings (Hierarchical Conditional Random Field Model A (Section IV-C2))

though they have different combination of classes (as shown in Fig. 2). BC does not have any further levels of annotation. Thus, complete label (global) of a token is a concatenation of the labels at the 3 levels, resulting in 39 different labels. Our dataset only had data for 34 labels. Our model can also be applied to findings from other imaging modalities but it needs to be trained on manually labeled data for those modalities. Due to absence of labeled data from other modalities, we only performed automatic structuring of mammography findings.

IV. Approach

In this section, we describe our approach for the three sub-goals – heading identification, content identification, and automatic structuring of findings from mammography study.

A. Heading Identification

a) Feature extraction: Reports were separated into sentences as explained in Section III. The sentences were separated into word-level tokens using regular expression $\backslash b \backslash w \backslash w + \backslash b$, which means tokens with at least 2 alphanumeric characters. Punctuations are always ignored and treated as token separator. For example, a sentence like *"Mammografie t,o,v, 12/08/2016: Mamma compositiebeeld C"* will generate $\{$*mammografie, 12, 08, 2016, mamma, compositiebeeld*$\}$ as tokens. Only unigrams were taken as tokens and converted to lowercase. The maximum document frequency was set such that the terms occurring in more than 60% of the documents will be ignored. Increasing the maximum document frequency did not improve the performance, so most probably high frequency non-informative words were removed.

Word List feature: A vocabulary was built using the unique words generated after preprocessing. Each sentence is represented by a term vector, where TF-IDF score is used for the tokens present in the sentence and a zero for absent tokens.

The length of the sentence and the symbol at the end of sentence were also tested as features but they did not improve performance and were not considered further.

b) Classifiers: Heading identification is a multiclass classification problem, where the sentences are to be classified into one of the following classes: *heading*, *not heading* and *title*. We trained a Multinomial Naive Bayes (NB), a linear Support Vector Machine (SVM) and a Random Forest (RF) classifier [1]. For NB, Laplace smoothing was used. SVM was trained using stochastic gradient descent and L2 loss. We used a maximum tree depth of 10 and bootstrap sampling for RF classifier.

B. Content Identification

Content identification is a multiclass classification problem, where the sentences are to be classified into *title*, *clinical data*, *findings* and *conclusion*. We followed the same approach as explained in Section IV-A.

C. Automatic Structuring

Our goal is to convert the free-text mammography findings into a semi-structured XML format. An example of this is shown in Table I, where the first column shows a free-text mammography finding report and the second column shows the semi-structured XML version. Let $\mathbf{X}$ be a mammography finding, consisting of a sequence of tokens, $\mathbf{x}$=(x_1,x_2,..x_t,..,x_n) and the task is to determine a corresponding sequence of labels $\mathbf{y}$= (y_1,y_2,..y_t,..,y_n) for $\mathbf{x}$. This task can be seen as *sequence labeling*, which is a task of predicting the most probable label for each of the tokens in the sequence. In this task, the context of the token, which means labels of immediately preceding or following tokens, is taken into account for label prediction. To achieve our goal, we used a Linear-Chain Conditional Random Field (LC-CRF)[2] [14], a supervised classification algorithm for sequence labeling. In our models, LC-CRF considers the label y_{t-1} of the immediately preceding token x_{t-1} for predicting the label y_t of the current token x_t.

[1]Classifiers were built using Python scikit-learn package

[2]We have used scikit-learn Python package, sklearn-crfsuite, implementation of LC-CRF

TABLE I: Example of structuring of free-text mammography finding

Free-text Report	Structured Report
Mammografie t,o,v, 22/09/2016: Mamma compositiebeeld C, Geen wijziging in de verdeling van het mammaklierweefsel, Hierin beiderzijds geen haardvormige laesies, Geen distorsies, geen stellate laesies, geen massa's, bekende verkalking links, Geen clusters kalk, geen maligniteitskenmerken,	⟨report⟩ ⟨O⟩Mammografie t,o,v, 12/08/2016:⟨/O⟩ ⟨breast_composition⟩Mamma compositiebeeld C,⟨/breast_composition⟩ ⟨O⟩Geen wijziging in de verdeling van het mammaklierweefsel,⟨/O⟩ ⟨negative_finding⟩ ⟨mass⟩Hierin ⟨location⟩beiderzijds⟨/location⟩ geen haardvormige laesies⟨/mass⟩ ⟨architectural_distortion⟩Geen distorsies,⟨/architectural_distortion⟩ ⟨mass⟩geen ⟨margin⟩stellate⟨margin⟩ laesies, geen massa's, ⟨/mass⟩ ⟨/negative_finding⟩ ⟨positive_finding⟩ ⟨calcification⟩bekende verkalking ⟨location⟩links⟨/location⟩ ⟨/calcification⟩ ⟨/positive_finding⟩ ⟨negative_finding⟩ ⟨calcification⟩Geen ⟨distribution⟩clusters⟨/distribution⟩ ⟨morphology⟩microkalk,⟨/morphology⟩ ⟨/calcification⟩⟨/negative_finding⟩ ⟨O⟩geen maligniteitskenmerken⟨/O⟩ ⟨/report⟩

a) Data Preprocessing: Each report from the dataset of 108 mammography findings was split at punctuations {,().?:-} (retaining them as tokens after splitting) and space, to generate tokens, **x**, which were transformed according to the IOB tagging scheme [20]. Here, B means beginning of an entity, I means inside (also including end) of an entity and O means not an entity. For example, as shown in Table I, *"Mamma compositiebeeld C,"* labeled as *breast_composition* was transformed to [(mamma, B-breast_composition), (compositiebeeld, I-breast_composition), (C, I-breast_composition), (',' , I-breast_composition)], where each entry stands for (token, label IOB scheme). Each digit was replaced by *#NUM* for the purpose of reducing the vocabulary size without removing any important information.

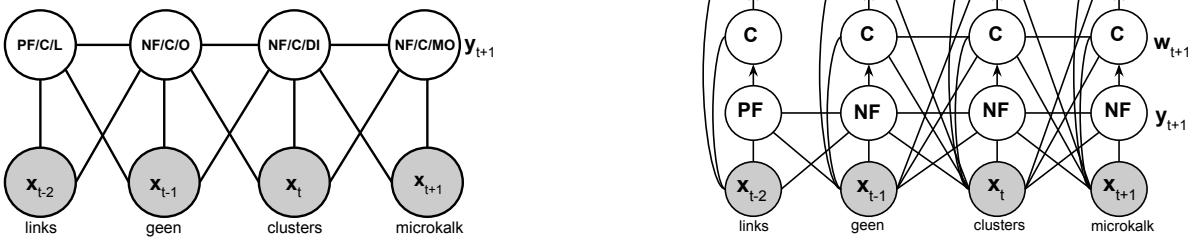
b) Feature Extraction: Each extracted token, x_t, is represented by a feature vector $\mathbf{x}_t$ for LC-CRF, including linguistic features of the current token, x_t, and also features of the previous token, x_{t-1}, and the next token, x_{t+1}. A feature vector $\mathbf{x}_t$ consists of the following 10 features for x_t and the same 10 features for x_{t-1} and x_{t+1} (a total of 30 features):

- The token x_t itself in lowercase, its suffixes (last 2 and 3 characters) and the word stem.
- Features indicating if x_t starts with a capital letter, is uppercase, is a Dutch stop word or is punctuation. The part-of-speech (POS) tag of x_t and its prefix (first 2 characters).

Below, we describe the 3 models for automatic structuring:

1) Baseline Model: As baseline, we used one LC-CRF classifier, as described at the starting of Section IV-C, to predict the complete label (concatenation of labels at the 3 levels) of a token and as input to the classifier, we used the feature vectors described in *Feature Extraction* (Section IV-Cb). For example, the LC-CRF classifier will predict the tokens *clusters* and *microkalk* as *NF/C/DI* and *NF/C/MO* respectively (see Table I). The graphical representation of this model is shown in Fig. 3a. Here, $\mathbf{x}_{t-1}$, $\mathbf{x}_t$, $\mathbf{x}_{t+1}$ are feature vectors of the tokens in a sequence and their corresponding labels are y_{t-1}, y_t, y_{t+1}, shown as NF/C/O, NF/C/DI, NF/C/MO. The lines indicate dependency on feature vectors $\mathbf{x}_{t-1}$, $\mathbf{x}_t$, $\mathbf{x}_{t+1}$ and preceding label y_{t-1} for prediction of the label y_t. Thus, in this model, only one classifier is used to predict 34 labels.

2) Hierarchical CRF: We built a model using a three-level hierarchy of LC-CRF classifiers, called Model A, as shown in Fig. 2. The model has 13 LC-CRF classifiers and all the classifiers perform token-level prediction. One classifier (CA-1) is at level 1 for classifying the tokens into the first level classes. At level 2, there are 2 classifiers – one (CA-2) for further classifying the tokens predicted as *positive finding* by CA-1, another (CA-3) for *negative finding* tokens. At level 3, there are 10 classifiers for further classification of tokens into third level classes. For example, the tokens classified as PF by CA-1 at level 1 and as MS by CA-2 at level 2, will be



(a) (b)

Fig. 3: Graphical representation of a) baseline CRF model and b) hierarchical CRF model, for input feature vectors $\mathbf{x}_{t-2}$ to $\mathbf{x}_{t+1}$={links geen clusters microkalk}

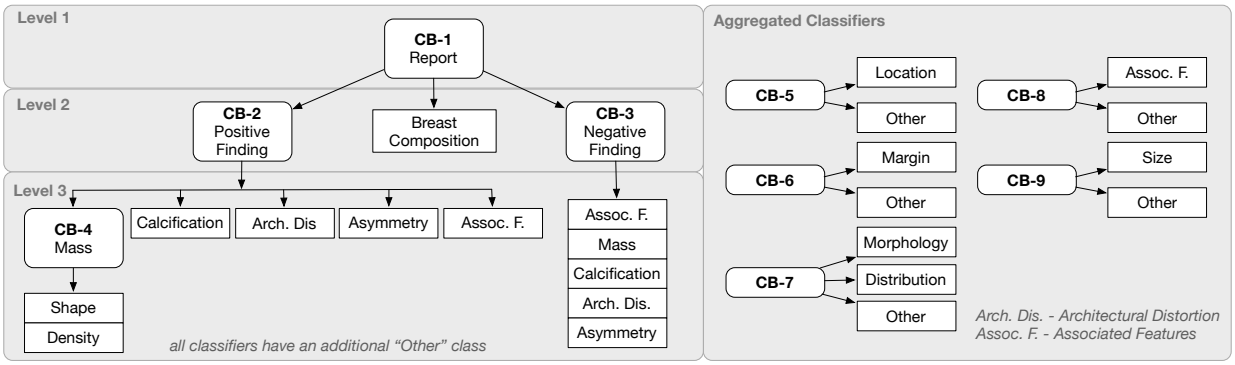


Fig. 4: Hierarchical Conditional Random Field Model B

sent to CA-4 classifier to further get classified as either L, SI, MA, DE, SH or AF. The complete predicted label for each token is the concatenation of its predicted classes at the three levels. The graphical representation of this model is shown in Fig. 3b. For example, for given feature vectors $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ of the tokens *clusters* and *microkalk* respectively and for given classes at the same-level of the immediately preceding token, the first level class predictions for both the tokens are NF. The feature vector of these tokens are sent to NF classifier, CA-3, for second level prediction, where they get classified as C. Consequently, they are sent to the *calcification* classifier, CA-10, where they get classified as MO and DI respectively. Labels at each level are combined resulting in NF/C/DI and NF/C/MO labels for the two tokens. The undirected lines are dependency lines and directed lines are flow between the 3 levels (y, w, z). There is no dependency line between the first two columns at the second level (w) as *links* goes to PF and *geen* to NF classifier and two different classifiers are independent of each other's feature vectors and predicted class.

3) Hierarchical CRF with Combined Classes: As can be seen in Fig. 2, every classifier at level 3, predicts *location* as one of its classes. All the *location* classes describe similar tokens like *rechts, links, beide mamma*. Thus, we build one classifier for the similar classes instead of having different classifiers. This will provide us with more training data for a classifier. Fig. 4 shows the modified model with combined classes having 9 classifiers. Henceforth, this is referred to as Model B and all classifiers in this model are referred to as CB-n ($n = 1, \ldots, 9$). We can see instead of having 11 classifiers that predict *location* (CA-n, $n = 3, \ldots, 13$) in Model A, we have only one classifier CB-5 in Model B. Analogously, classifiers were aggregated for MA, MO, DI, AF and SI. All the classifiers use LC-CRF and perform token-level prediction. When classifying a token, classifiers might contradict each other. Consider for example NF/MS: CB-5 and CB-6 are the two classifiers predicting *location, margin* or *other* for the same token. If the predictions are *location* by CB-5 and *other* by CB-6, then *location* is selected (no contradiction). Similarly, if both classifiers predict *other*, then the resulting class is *other* (no contradiction). If the predicted class is *location* by CB-5 and *size* by CB-6 (contradiction), then the class with the highest a-posteriori probability is selected.

V. Experimental Setup

We used the F_1 score to evaluate the performance of a classifier on predicting different classes. The F_1 score of a class is the harmonic mean of precision and recall of that class and is defined as

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

with TP being the number of true positives, FP - false positives and FN - false negatives. As our problem is a multiclass problem, the TP, FN, FP of a class are calculated according to one-vs-rest binary classification, where the class in consideration is positive and all other classes are negative.

We also measured F_1 score of the models on the entire test set using *micro-averaged* and *weighted macro-averaged* F_1 (F_1^μ and F_1^M). F_1^μ was computed by calculating the TP as sum over the TP of all the classes (same for FN, FP). F_1^M was calculated by computing the F_1 scores of each class separately and then averaging it. As, averaging gives equal weight to all the classes, the fact that our classes have unequal number of instances, is not taken into account. Thus, we used weighted averaging for F_1^M. F_1^M and F_1^μ gave similar results, so we only report F_1^M scores in the rest of the paper.

We evaluated our classifiers at 3 levels: i) token-level (TL), ii) partial phrase-level (PP), and iii) complete phrase-level (CP). At the token-level, we consider all the token labels in the dataset to calculate the TP, TN, FP, FN scores of a class. At the partial phrase-level and the complete phrase-level, we measure how well the classifier is performing in identifying multi-token phrases. A complete match requires all the tokens of the phrase to be correctly labeled. We consider a match with Dice's coefficient greater than 0.65 as a partial match. For similarity calculation, we take the phrase from the ground truth and match with the corresponding predicted labels. Phrase-level scores are important from the radiologists' point of view. They care about how well their phrases are matching. Table IIIa shows 6 tokens, with their token-level labels (B-PF, I-PF

TABLE II: Heading and content identification and automatic structuring performance in terms of F_1 scores

(a) Heading identification

Classes	NB	SVM	RF	#Instances (Sentences)
Heading	**0.96**	**0.96**	0.88	540
Not Heading	**0.98**	**0.98**	0.94	991
Title	0.97	0.98	**0.99**	60
Avg (F_1^M)	0.97	0.97	0.92	1591

(b) Content identification

Classes	NB	SVM	RF	#Instances (Sentences)
Conclusion	0.89	**0.92**	0.90	413
Clinical Data	0.86	**0.94**	0.70	405
Title	0.89	**0.99**	0.91	60
Findings	0.88	**0.94**	0.82	678
Avg (F_1^M)	0.88	0.94	0.81	1556

(c) Automatic structuring

Measures	Baseline	Model A	Model B	#Instances (Tokens)
F_1^M(all)	0.71	0.78	0.78	4230
F_1^M(w/o O)	0.67	0.73	0.74	2813
F_1^M(w/o<10&O)	0.70	0.76	0.76	2649

etc). A PF phrase starts at the B-PF and ends at the last I-PF. For the NF phrase, the Dice's coefficient is calculated as $2 * 2/(3 + 3) = 0.66 > 0.65$, resulting in a partial match. For each class, we calculate the number of partial matches called partial phrase accuracy (PP-Acc); how well the partial phrases match by averaging the Dice's coefficient for each match (PP-Sim); the number of complete matches (CP-Acc); and the F_1 scores for token-level matches (TL F_1).

For heading and content identification, we evaluated NB, SVM and RF models, using 5-fold cross validation on 180 reports. For automatic structuring, we built three different LC-CRF models: the baseline model, Model A and Model B. We evaluated our models using 4-fold cross validation on 108 mammography findings. For automatic structuring, we evaluated the models on different combinations of classes (Table IIc). 'All' means evaluation on all the 34 classes. 'w/o O' means all the classes except the *other (O)* class at the first level (33 classes). 'w/o<10&O' means classes excluding O class and classes with instances<10. All codes associated with this paper are available as open source[3].

VI. Results

In this section, we describe the results of heading and content identification and automatic structuring.

A. Heading and Content Identification

Table IIa shows that classes *headings* and *not headings* were identified with an F_1 score of 0.96 and 0.98 respectively both by SVM and NB. For these classes, SVM and NB performed better than RF but for *title*, RF performed better. Table IIb shows that the SVM performed better for predicting the classes *conclusion, clinical data, title* and *findings* with an F_1 score of 0.92, 0.94, 0.99 and 0.94 respectively.

B. Automatic Structuring

Table IIc compares the performance of our baseline model to the hierarchical Models A and B. Both, Model A and B (F_1^M= 0.78) outperformed the baseline model (F_1^M= 0.71). No difference in performance was observed between Model A and B. Without the not important *other (O)* class, the Model B has F_1^M= 0.74. On further removing classes with instances<10, the F_1^M score improves from 0.74 to 0.76 for Model B. This means that the classes having instances<10 were not predicted well enough. If we would have at least 10 instances for each class, then the F_1^M score could be expected to be around 0.76.

Table IIIb shows the performance of the classifier (CA-1 and CB-1) at the first level in predicting *breast composition*, *negative finding*, *positive finding*. BC (TL F_1= 0.94) and NF (TL F_1= 0.95) were identified better than PF (TL F_1= 0.87). This is because PF contains varied vocabulary for describing an abnormality, while NF contains specific terms like no presence of mass, calcification. BC is also described using specific terms like "mamma compositiebeeld". Token-level measure is always higher than complete phrase-level measure. PP-Acc is at least as good as CP-Acc. All the partial phrase matches in BC and PF are complete matches except for NF. But even for NF, the partial phrases have similarity of 0.99 (PP-Sim) with the ground truth.

Table IV shows the performance obtained for the some of the global classes. Overall, it can be seen that NF sub-classes were predicted better than PF sub-classes, as most of the NF sub-classes are described using specific tokens. Generally, Model A and B predicted PF sub-classes better than baseline. BC, NF/AF/O, NF/C/DI, NF/MS/MA and NF/C/MO were predicted very well in all the models. Some classes were predicted better in baseline – NF/MS/O, NF/MS/MA and PF/C/O. This indicates that for these classes, the neighbouring global classes of the baseline model may be informative during prediction. Also, multi-level prediction increased the number of false positives for a class, specially for classes with greater number of instances. The effect of aggregated classifiers in model B can be seen in NF/C/DI, NF/C/MO, PF/C/L, PF/MS/L and PF/C/SI. As the aggregated classifiers were trained on all L, DI, MO and SI in the dataset, it resulted in better prediction of third level classes like L, SI, even with few instances (14 tokens of PF/C/SI). But aggregating classifiers also resulted in loss of information about the context, which is reflected

[3]https://www.dropbox.com/sh/y4czin4llue2t6w/AACqHRcC2pxg0zzg42JuPtQna?dl=0

TABLE III: Token level and phrase level measures

(a) Tokens and phrases

	bekende	verkalking	links	geen	clusters	microkalk
true	B-PF	I-PF	I-PF	B-NF	I-NF	I-NF
predicted	B-PF	I-PF	I-PF	O	B-NF	I-NF
true	PF phrase			NF phrase		
predicted	PF complete phrase match			NF partial phrase match		

(b) Token and phrase level scores

Classes	TL F_1	PP-Acc	CP-Acc	PP-Sim	#Tokens	#Phrases
BC	0.94	0.93	0.93	1.00	622	99
NF	0.95	0.97	0.91	0.99	1101	118
PF	0.87	0.87	0.87	1.00	1090	87

TABLE IV: F_1 measures of global classes for the 3 models of automatic structuring

Models	BC	NF/AF/O	NF/C/O	NF/C/DI	NF/C/MO	NF/MS/O	NF/MS/MA	PF/C/O	PF/C/SI	PF/C/L	PF/MS/L	PF/MS/MA	PF/C/AF	PF/AS/O
Baseline	0.89	**0.96**	**0.81**	0.98	0.95	**0.93**	**1.00**	**0.45**	0.00	0.50	0.30	0.53	0.00	0.00
Model A	**0.94**	**0.96**	0.76	0.98	0.91	0.88	0.96	0.37	0.00	0.44	0.40	**0.72**	**0.18**	**0.58**
Model B	**0.94**	**0.96**	**0.81**	**0.99**	**0.97**	0.89	0.97	0.37	**0.22**	**0.60**	**0.47**	0.70	0.00	0.56
#Instances	622	397	148	54	56	210	35	138	14	68	139	59	33	172

TABLE V: Error propagation through classifiers at the 3 levels

Measures	Level2_A	Level2_B	Level3_A	Level3_B
ΔF_1^M	0.05	0.04	0.17	0.16
#Instances	2191	2191	2093	2093

through slightly lower performance in Model B for classes PF/MS/MA, PF/C/AF and PF/AS/O. Aggregating AF classifier (CB-8) did not help in predicting any third level AF classes in PF due to not much similarity in their descriptions.

Table V gives an indication of error propagation through the classifiers at the 3 levels for Model A and B. ΔF_1^M at a level indicate the difference in F_1^M of that level of classifiers on predicted classes when given true classes from previous level and when given predicted classes from previous level. This can be interpreted as error made by the classifiers at the previous level. Error made by level 1 (ΔF_1^M at level 2) is not much significant as compared to error by level 2 (ΔF_1^M at level 3) as the latter is a combination of errors from both level 1 and level 2 classifiers, while the former only considers error from level 1.

VII. Conclusion and Future Work

We have addressed three tasks for the purpose of quality assurance of radiology reports: heading identification, content identification and automatic structuring using BI-RADS standard. Heading and content were identified with an F_1^M score of 0.97 and 0.94 respectively using SVM. For automatic structuring, hierarchical CRF (F_1^M= 0.78) performed better than baseline CRF (F_1^M= 0.71), while Model A and B did not show any significant difference.

From the point of view of quality assurance, heading and content contribute to identification of the presence of indication of examination, findings and conclusion. A post-processing step can be performed to check if the content corresponds to the correct heading. Automatic structuring is used to check the presence of clear description of findings. According to BI-RADS, findings should contain mass, calcification, asymmetry, architectural distortion and associated features. Our model structures the findings automatically into these concepts, further generating a semi-structured XML format. This provides a platform to check the presence of important concepts. Another important information that must be present in reports is breast composition. Our model predicts breast composition with F_1= 0.94.

As future work, the presence and quality of BI-RADS category can be evaluated. Based on findings, BI-RADS category can be predicted to check how well it was assigned. More reports can be labeled to get more training data. Development of a prototype and actual trial in clinical practice can be done. The approach taken in this research can also be extended to reports for other conditions, written in other languages.

References

[1] *Breast imaging reporting and data system*. BI-RADS Committee, American College of Radiology, 1998.

[2] H. H. Abujudeh, R. Kaewlai, B. A. Asfaw, and J. H. Thrall, "Quality initiatives: key performance indicators for measuring and improving radiology department performance," *Radiographics*, vol. 30, no. 3, pp. 571–580, 2010.

[3] A. J. Johnson, J. Ying, J. S. Swan, L. S. Williams, K. E. Applegate, and B. Littenberg, "Improving the quality of radiology reporting: a physician survey to define the target," *Journal of the American College of Radiology*, vol. 1, no. 7, pp. 497–505, 2004.

[4] C. E. Kahn Jr, C. P. Langlotz, E. S. Burnside, J. A. Carrino, D. S. Channin, D. M. Hovsepian, and D. L. Rubin, "Toward best practices in radiology reporting," *Radiology*, vol. 252, no. 3, pp. 852–856, 2009.

[5] N. Strickland, "Quality assurance in radiology: peer review and peer feedback," *Clinical radiology*, vol. 70, no. 11, pp. 1158–1164, 2015.

[6] C. Yang, C. J. Kasales, T. Ouyang, C. M. Peterson, N. I. Sarwani, R. Tappouni, and M. Bruno, "A succinct rating scale for radiology report quality," *SAGE open medicine*, vol. 2, p. 2050312114563101, 2014.

[7] C. L. Sistrom and C. P. Langlotz, "A framework for improving radiology reporting," *Journal of the American College of Radiology*, vol. 2, no. 2, pp. 159–167, 2005.

[8] L. H. Schwartz, D. M. Panicek, A. R. Berk, Y. Li, and H. Hricak, "Improving communication of diagnostic radiology findings through structured reporting," *Radiology*, vol. 260, no. 1, pp. 174–181, 2011.

[9] A. J. Johnson, M. Y. Chen, J. S. Swan, K. E. Applegate, and B. Littenberg, "Cohort study of structured reporting compared with conventional dictation," *Radiology*, vol. 253, no. 1, pp. 74–80, 2009.

[10] C. P. Langlotz, "Radlex: a new method for indexing online educational materials," 2006.

[11] C. L. Sistrom and J. Honeyman-Buck, "Free text versus structured format: information transfer efficiency of radiology reports," *American Journal of Roentgenology*, vol. 185, no. 3, pp. 804–812, 2005.

[12] H. Nassif, R. Woods, E. Burnside, M. Ayvaci, J. Shavlik, and D. Page, "Information extraction for clinical data mining: a mammography case study," in *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*. IEEE, 2009, pp. 37–42.

[13] D. A. Sippo, G. I. Warden, K. P. Andriole, R. Lacson, I. Ikuta, R. L. Birdwell, and R. Khorasani, "Automated extraction of bi-rads final assessment categories from radiology reports with natural language processing," *Journal of digital imaging*, vol. 26, no. 5, pp. 989–994, 2013.

[14] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[15] S. Hassanpour and C. P. Langlotz, "Information extraction from multi-institutional radiology reports," *Artificial intelligence in medicine*, vol. 66, pp. 29–39, 2016.

[16] M. Torii, K. Wagholikar, and H. Liu, "Using machine learning for concept extraction on clinical documents from multiple data sources," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 580–587, 2011.

[17] A. Esuli, D. Marcheggiani, and F. Sebastiani, "An enhanced crfs-based system for information extraction from radiology reports," *Journal of biomedical informatics*, vol. 46, no. 3, pp. 425–435, 2013.

[18] R. K. Taira, S. G. Soderland, and R. M. Jakobovits, "Automatic structuring of radiology free-text reports," *Radiographics*, vol. 21, no. 1, pp. 237–245, 2001.

[19] E. A. Sickles, C. J. D'Orsi, L. W. Bassett, and et al, *ACR BI-RADS Mammography. In*, Reston, VA, 2013.

[20] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Natural language processing using very large corpora*. Springer, 1999, pp. 157–176.