Information Extraction of Multiple Categories from Pathology Reports

Yue Li
NICTA VRL
Melbourne, Australia
y.li30@ugrad.unimelb.edu.au

David Martinez
NICTA VRL
Melbourne, Australia
davidm@csse.unimelb.edu.au

Abstract

Pathology reports are used to store information about cells and tissues of a patient, and they are crucial to monitor the health of individuals and population groups. In this work we present an evaluation of supervised text classification models for the prediction of relevant categories in pathology reports. Our aim is to integrate automatic classifiers to improve the current workflow of medical experts, and we implement and evaluate different machine learning approaches for a large number of categories. Our results show that we are able to predict nominal categories with high average f-score (81.3%), and we can improve over the majority class baseline by relying on *Naive Bayes* and feature selection. We also find that the classification of numeric categories is harder, and deeper analysis would be required to predict these labels.

1 Introduction

A pathology report is the summary of the analysis of cells and tissues under a microscope, and it may also contain information of the studied specimen as it looks to the naked eye. Pathology reports play an important role in cancer diagnosis and staging (describing the extent of cancer within the body, especially whether it has spread). These reports are usually written by the pathologist in natural language, and then the relevant parts are transcribed into structured form by a different person to be stored in a database.

The use of structured information can help share the data between institutions, and can also be used to find patterns in the data. For this reason, some recent initiatives are exploring better ways to manage pathology reports. For instance, the Department of Health and Ageing of Australia is funding the project Structured Pathology Reporting of Cancer since 2008 to develop standard reporting protocols for cancer reports[1]. Another way to promote the creation of structured data is to use standard terminologies, such as SNOMED CT[2], which is a large collection of medical terminology covering most areas of clinical information such as diseases, findings, procedures, microorganisms, pharmaceuticals etc. The National E-Health Transition Authority (NEHTA) has recently launched an adapted terminology (SNOMED CT-AU) to be used by the Australian health sector[3].

These initiatives will help to increase the repositories of structured data, but they will not be a substitute to the flexibility of natural language. The relevant fields in structured reports change over time as different clinical tests are made available, and it is difficult to design a specific form to cover all the possible cases that will be observed in the pathology analysis. Clinicians need time to learn the different standards, and they prefer the flexibility of free text to record their analyses and conclusions. Ideally their natural language input would be used to automatically extract the structured data that different protocols demand.

This scenario is promising for text mining research, because tools that can perform well in this space are likely to make an impact in the way health information is stored and used. Our goal in this work is to explore this area, and develop and evaluate a text mining tool that aims to work in a real hospital setting, by predicting pieces of information to populate a database. Specifically, we focus on a system for the Royal Melbourne Hospital, where pathology reports of cancer patients are

[1] http://www.rcpa.edu.au/Publications/StructuredReporting.htm

[2] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

[3] http://www.nehta.gov.au/media-centre/nehta-news/571-snomed-ct

Yue Li and David Martinez. 2010. Information Extraction of Multiple Categories from Pathology Reports. In *Proceedings of Australasian Language Technology Association Workshop*, pages 41−48

kept in natural language, and an electronic form is manually filled with the most relevant information. We would like to predict the classes automatically in order to facilitate the process.

Our aim is to build a generic approach for different prediction categories, involving heterogeneous classes with a large set of possible values (e.g. the class "Tumour site" has 11 different values in our data, for instance "Sigmoid Colon"). We will rely on the available document-level annotations of pathology reports to build our classifiers using Machine Learning (ML) algorithms. Annotated data is difficult to obtain in this domain, and there are few works evaluating the performance of supervised classifiers for pathology reports, as we will see in Section 2. In this work we will explore how far we can get with existing annotations, and simple lexical features that can be extracted without external knowledge sources.

Thus, we present an extensive set of experiments to evaluate the ability of different models and methods to perform class-predictions over pathology reports. The problem will involve predicting nominal and numeric classes, and we test models that perform sentence-level and document-level classification. Our main challenges in this project will be the sparseness of the data, the coarseness of the annotations (document-level categories only), and the high number of heterogeneous categories. In the future, the tool resulting from this work will be integrated in the hospital workflow, and it will work interactively with the user, making predictions and allowing corrections. We will store all user interactions to continually add training data to our classifiers. We will also highlight the relevant parts of the text as predicted by our learning models, by using feature selection.

2 Related work

Related work in text mining from pathology reports has mainly relied on domain-specific lexicons and rules (Dunham et al., 1978; Schadow and McDonald, 2003; Xu et al., 2004; Hanauer et al., 2007; Coden et al., 2009; Nguyen et al., 2010); although there has been some work using ML (Nguyen et al., 2007; McCowan et al., 2007). The earliest work in this area was perfomed by Dunham et al. (1978), who built morphosyntactic rules, synonym expansion, and handcrafted rules in order to extract terms from the Systematized Nomenclature of Pathology (SNOP), which was an earlier version of the SNOMED CT terminology collection. More recent works have used SNOMED CT as the target terminology to map the raw text into. Hanauer et al. (2007) relied on custom-made lists containing approximately 2,500 terms and phrases, and 800 SNOMED codes. Their method was based on looking up relevant phrases in order to discriminate the documents of interest.

Other works have developed their own set of relevant classes instead of relying on SNOMED. This is interesting when the focus is on a specific subdomain, and this is the approach that we explored in our work. Schadow and McDonald (2003) relied on a subset of UMLS[4] (Unified Medical Language System) as target concept inventory for information extraction from surgical pathology reports. They applied a regular expression-based parser with good performance, but they also found that their target terminology was too extensive, and this caused false positives. Xu et al. (2004) also targeted surgical pathology reports, and they used a restricted set of 12 classes, referred as "types of findings". This is similar to our approach, and some of their classes are part of our relevant classes as well (e.g. "number of positive nodes"); however they do not provide the performance for each class separately, which makes comparison unfeasible. Regarding the methodology, their system is based on hand-crafted rules, and relies on a domain-specific lexicon. Our motivation is different, and we rely on ML to infer the knowledge from coarse-grained annotation for a larger set of classes.

Also in the area of information extraction from pathology reports, recent work from the Australian e-Health Research Centre[5] explored the extraction of staging information of lung cancer using Support Vector Machines (Nguyen et al., 2007). Their initial experiments showed the difficulty of the primary tumour stage detection (T), with a top accuracy of 64%. In a follow-up paper they explored richer annotation, and a combination of ML and rule-based post-processing (McCowan et al., 2007). They performed fine-grained annotation of stage details for each sentence in order to build their system, and they observed improvements over a coarse-grained (document-level) multiclass classifier. However, the authors explain that the

[4] http://www.nlm.nih.gov/research/umls
[5] http://aehrc.com/

annotation cost is high, and in their latest work they rely heavily on the SNOMED CT concepts and relationships to identify the relevant entities (Nguyen et al., 2010). They argue that this approach is more portable than fine-grained annotation, although it still requires involvement from the experts, and there is a loss in accuracy with respect to their best ML approach. These three papers evaluate their system in the prediction of staging classes (T, N, and M), which are not explicit in our dataset.

Another relevant work on this area was conducted by Coden et al. (2009), where the authors defined an extensive knowledge model for pathology reports. Their model was linked to hand-built inference rules built to process unseen data. They reported high performance over 9 target classes for a hand-annotated 300-report dataset. This system seeks to build a strong representation of the domain by relying on human experts, and its portability to a different dataset or class-set could be problematic. The classes they evaluate on are not present in our dataset.

Currently there is no dataset of pathology reports that is freely available for research, and different groups have built their own corpora. Pathology reports contain sensitive material, and even after de-identification it is not easy to make them widely available. However, initiatives as the NLP challenges leaded by the Informatics for Integrating Biology and the Bedside (i2b2)[6] illustrate that there is growing interest on text mining from clinical data, and show that the research community can collaboratively create corpora for experimentation. In 2010 they organised their fourth challenge, focused on the extraction of medical problems, tests, and treatments from patient discharge summaries[7]. Previous challenges have also focused on discharge summaries and narrative patient records for different information extraction categories. Although this data is different to pathology reports, the initiative is interesting for the future of text mining from pathology reports.

3 Experimental setting

In this section we first describe the dataset and categories we will work on, and then introduce the models and classifiers we applied. Finally we explain our feature set, and our evaluation methodology.

[6]i2b2 is a NIH-funded National Center for Biomedical Computing (NCBC), for more information see https://www.i2b2.org/about/index.html

[7]https://www.i2b2.org/NLP/Relations/

Category	Unique Values	Highest	Lowest
CAV	21	40	0
Distal Distance	48	150	0
Nodes Examined	36	73	0
Nodes Positive	12	15	0
Polyps Number	13	43	0
Radial Distance	9	80	0
Tumour Length	36	110	0
Tumour depth	22	40	0
Tumour width	35	75	0

Table 1: List of numeric categories, with the number of unique values and the full range.

3.1 Dataset

For our analysis we rely on a corpus of 203 de-identified clinical records from the Royal Melbourne Hospital. These records were first written in natural language, and then structured information about 36 fields of interest was introduced to the Colorectal Cancer Database of the hospital. The written records tend to be brief, usually covering a single page, and semantically dense. Each report contains three sections describing different parts of the intervention: macroscopic description, microscopic description, and diagnosis. All sections contain relevant information for the database.

There are two types of fields (which will be the target categories of our work), depending on the type of values they take: numeric and nominal. Numeric categories are those that take only numeric values, and they are listed in Table 1. We also show the number of different values they can take, and their value range. We can see that most categories exhibit a large number of unique values. The remaining 27 categories are nominal, and the list is shown in Table 2, where we also provide the number of unique values, and the most frequent value in the corpus for each category. Some of the categories are linked to a large number of values (e.g. *Colon Adherent To* and *Tumour Site*). During pre-processing we observed that the database had some inconsistencies, and a normalisation step was required with collaboration of the experts. For nominal categories this involved mapping empty values, “0”, and “?” into the class “N/A”; and for numeric categories we mapped empty values into “zero”.

The manual annotation is provided at document

Category	Unique Values	Most Frequent
Anastomosis Method	3	*Staple*
Anastomosis Type	4	*End-End*
Biopsy Confirmed Mata	3	*No*
Colon Adherent	3	*No*
Colon Adherent To	14	*N/A*
Differentiation	8	*Moderate*
Inflammatory Infiltrate	4	*Not reported*
Liver	3	*N/A*
Lympho Invasion	4	*No*
MLH1	4	*Not done*
MSH2	4	*Not done*
MSH6	4	*Not done*
MSI	4	*Not done*
Margins Distal	3	*Not involved*
Margins Radial	4	*N/A*
Microscopic Type	5	*Adenocarcinoma*
Mucinous	4	*Not reported*
Necrosis	4	*Not reported*
Other Meta	3	*N/A*
Pathologic Response	4	*N/A*
Peritoneal	3	*N/A*
Polyps	3	*No*
Polyps Type	6	*N/A*
Primary Tumour Rectum	4	*N/A*
Ressected Meta	3	*No*
Staging ACPS	6	*B*
Tumour Site	11	*Sigmoid Colon*

Table 2: List of nominal categories, with the number of unique values, and most frequent class.

level, and for numeric categories we automatically produce fine-grained annotation by looking up the goldstandard mentions in the text. We try to match both the string representation and the numbers, and only numbers different to zero are identified. After this automatic process, each sentence has individual annotations for each of the target categories, and this information is used to build sentence-level classifiers. Because the process is automatic, some matches will be missed, but our hypothesis is that the noisy annotation will be useful for the document-level evaluation.

3.2 Models

Our goal is to build document classifiers for each of the 36 categories with minimal hand tuning. We follow different strategies for nominal and numeric categories. For nominal categories we observed that the information can be given at different points in the document, and we decided to build a multiclass classifier for each category. This method makes a single prediction based on the class annotations in training data.

For numeric categories the information tends to be contained in a single sentence, and instead of using the full document, we relied on the sentence-level annotation that we obtained automatically. In this case the target values would be the different numeric values seen in the goldstandard. The first step is to build sentence classifiers for each class, by using the sentence-level annotations. Note that only numbers different to zero are detected, and the zero label is assigned only in cases where the sentence classifiers fail to identify any number. After the model identifies the positive sentences, the numeric values are extracted, and the number closest to the median of the class (in training data) is assigned. In the cases where no positive sentences are identified the number zero is assigned.

3.3 Classifiers

Each of our models is tested with a suite of classifiers provided by the Weka toolkit (Witten and Frank, 2005). We chose a set of classifiers that has been widely used in the text mining literature in order to compare their performances over our dataset:

- Naive Bayes (*Naive Bayes*): A simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) to obtain the conditional probability of each class given the features in the context. It assumes independence of the features, which in real cases can be a strong (naive) assumption.

- Support Vector Machines (*SVM*): They map feature vectors into a high-dimensional space and construct a classifier by searching for the hyperplane in that space that gives the greatest separation between the classes.

- AdaBoost (*AdaBoost*): This is a meta-learning algorithm where an underlaying classifier is used to update a distribution of weights that indicates the importance of the training examples. Adaboost is an adaptive algorithm, and the prediction hits and misses in each iteration are used to build the final weight distribution for the model.

We use the default parameter settings of Weka (version 3-6-2) for each of the classifiers. As underlaying classifier for *AdaBoost* we rely on simple Decision Stumps (one-level decision trees).

We also explore the contribution of feature selection to the classification performance. We apply a correlation-based feature subset selection method, which considers the individual predictive

Category	*Majority Class*			*Naive Bayes*			*SVM*			*AdaBoost*		
	Prec.	Rec.	F-sc	Prec.	Rec.	F-sc	Prec.	Rec.	F-sc	Prec.	Rec.	F-sc
Tumour site	3.9	19.7	6.5	23.3	40.4	27.7	28.5	38.4	**32.2**	12.4	34.0	18.1
Staging ACPS	12.9	36.0	19.0	39.2	43.8	36.7	44.1	48.3	**45.1**	33.2	49.3	37.5
Anastomosis type	22.4	47.3	30.4	46.6	52.7	44.9	53.5	59.1	**55.0**	32.7	50.2	39.3
Colon adherent	23.3	48.3	31.4	63.2	67.0	62.0	67.4	70.4	**67.0**	47.8	58.6	51.4
Lympho invasion	23.8	48.8	32.0	48.4	51.2	42.9	53.4	55.7	**53.5**	32.3	51.2	39.6
Polyps	27.8	52.7	36.4	69.6	72.9	69.3	83.1	84.2	**83.3**	74.9	77.8	74.3
Colon adherent to	28.3	53.2	37.0	59.9	70.0	63.9	62.6	73.4	**67.4**	40.6	47.3	43.7
Margins radial	31.0	55.7	39.8	65.4	73.4	**68.6**	65.2	70.9	67.4	59.9	67.5	62.0
MlH1	31.5	56.2	40.4	41.4	50.7	45.4	46.7	49.8	**46.9**	35.6	58.1	43.4
MSH6	31.5	56.2	40.4	41.5	50.7	45.5	43.0	48.8	**45.7**	35.0	57.6	42.7
MSH2	31.5	56.2	40.4	41.5	50.7	45.5	43.0	48.8	**45.7**	35.0	57.6	42.7
MSI	32.7	57.1	41.6	44.5	53.7	**48.4**	45.8	50.7	47.9	32.7	57.1	41.6
Mucinous	34.9	59.1	43.9	42.6	58.6	46.7	70.2	72.4	**68.8**	57.8	73.4	64.5
Anastomosis method	41.6	64.5	50.6	57.6	70.4	62.0	62.4	71.9	**65.8**	56.9	69.5	60.4
Necrosis	42.9	65.5	51.9	64.5	74.4	68.7	73.5	77.3	**74.6**	62.8	69.5	62.7
Polyps type	49.6	70.4	58.2	49.6	70.4	58.2	66.9	72.9	**63.4**	49.6	70.4	58.2
Differentiation	51.0	71.4	59.5	51.2	70.9	59.5	70.3	78.3	72.1	66.0	80.8	**72.7**
Inflammatory infiltrate	51.0	71.4	59.5	66.5	72.4	62.6	68.4	76.8	**70.9**	68.9	70.9	66.2
Liver	53.2	72.9	61.5	53.0	68.0	59.5	59.5	64.5	**61.7**	52.8	70.9	60.5
Other meta	53.2	72.9	61.5	53.0	68.0	59.5	59.5	64.5	**61.7**	53.1	71.4	60.9
Primary tumour rectum	53.2	72.9	61.5	56.5	72.4	62.7	70.9	80.3	**75.1**	67.5	73.9	70.4
Peritoneal	53.2	72.9	61.5	53.0	68.0	59.5	59.5	64.5	**61.7**	52.7	70.4	60.3
Ressected meta	63.7	79.8	70.8	68.6	79.8	72.1	70.4	79.3	**73.2**	66.2	77.8	70.8
Margins distal	76.0	87.2	81.2	76.0	87.2	81.2	86.1	92.1	**89.0**	86.1	92.1	**89.0**
Biopsy confirmed mata	80.4	89.7	84.8	80.4	89.7	84.8	85.0	89.7	**86.5**	80.3	88.7	84.3
Pathologic response	81.3	90.1	**85.5**	81.3	90.1	**85.5**	81.3	90.1	**85.5**	81.3	90.1	**85.5**
Microscopic type	87.6	93.6	**90.5**	87.6	93.6	**90.5**	88.0	93.1	**90.5**	87.6	93.6	**90.5**
Macro-average	43.5	63.8	51.0	56.5	67.1	59.8	63.3	69.1	**65.1**	54.1	67.8	59.0

Table 3: Performances of multiclass document classifiers for nominal categories without feature selection. Results sorted by baseline f-score performance. Best f-score per category is given in bold.

ability of each feature and the redundancy of each subset (Hall, 1999). We relied on Weka's implementation of this technique, and used Best-First search, with a cache-size of one element, and 5 levels of backtracking.

3.4 Features

Pathology reports tend to be short and dense, and the selection of words tries to precisely specify the relevant pieces of information. For this reason we rely on a bag-of-words (BOW) approach for our feature representation, without any lemmatisation. We built a simple tokeniser based on regular expressions to separate words, numbers, and punctuation. We also use regular expressions to convert the textual mentions of numbers into their numeric representation. Finally, we include the binary feature "NUMBER" to indicate whether there is a numeric reference in the text.

3.5 Evaluation

In order to evaluate the different models and classifiers we use precision, recall, and f-score by micro-averaging the results over the different class values. The macro-averaged scores over all categories are also provided to compare different systems. 10-fold cross-validation is used in all our experiments.

As a baseline we rely on the *Majority Class* classifier, which assigns the most frequent class from training data to all test instances. In case of ties the value is chosen randomly among those tied.

4 Results

We first present our result over the nominal categories, and then show the performances over numeric categories.

4.1 Nominal categories

Our first experiment applies the multiclass document classifier to nominal categories. The results are given in Table 3. We can see that the best performance is achieved by *SVM*, with a large improvement over the majority class baseline. *Naive Bayes* and *AdaBoost* also perform above the baseline, and attain similar results. However, a maximum f-score of 65.1% seems insufficient to be of use for an application. Regarding the different categories, as expected these with lowest baseline

Category	*Naive Bayes*			*SVM*		
	Prec.	Rec.	F-sc	Prec.	Rec.	F-sc
Tumour site	53.0	54.2	**51.1**	43.9	44.8	43.8
Staging ACPS	71.1	71.9	**70.7**	58.0	59.1	58.3
Anastomosis type	74.1	71.9	**71.3**	69.4	69.5	69.4
MSH6	75.7	70.9	**71.4**	65.3	65.0	64.7
MSH2	75.7	70.9	**71.4**	65.3	65.0	64.7
Colon adherent to	68.9	74.9	**71.6**	64.4	70.4	66.5
MSI	77.6	72.9	**73.3**	71.5	72.4	71.3
Lympho invasion	74.3	75.4	**74.4**	69.8	70.9	70.3
MlH1	78.1	75.4	**75.2**	71.2	71.4	70.6
Anastomosis method	77.8	78.8	**77.8**	75.1	76.4	75.5
Margins radial	79.1	79.8	**78.5**	76.8	76.4	75.5
Colon adherent	78.2	80.3	78.8	80.0	80.3	**79.7**
Other meta	83.6	83.7	**83.7**	75.6	77.3	76.4
Peritoneal	83.9	84.2	**84.0**	79.0	80.3	78.6
Inflammatory infiltrate	82.9	86.2	**84.0**	83.3	84.2	82.9
Polyps type	84.8	85.2	**84.4**	80.5	82.3	81.2
Necrosis	84.6	85.7	**85.1**	79.9	81.3	80.5
Mucinous	84.5	86.7	**85.4**	82.6	83.7	82.8
Liver	86.9	86.2	**86.4**	76.1	76.8	76.4
Primary tumour rectum	87.8	86.2	**86.6**	84.4	84.7	83.8
Ressected meta	89.8	90.1	**89.7**	86.6	86.7	86.2
Margins distal	90.0	92.6	**90.3**	88.4	91.6	89.5
Polyps	90.3	91.6	**90.9**	90.1	90.1	90.1
Differentiation	91.8	92.6	**91.9**	90.0	92.1	90.7
Biopsy confirmed mata	93.8	94.1	**93.9**	94.3	94.6	**93.9**
Microscopic type	96.6	96.6	**96.4**	94.2	95.6	94.6
Pathologic response	97.7	97.5	**97.5**	91.4	93.1	91.9
Macro-average	81.9	82.1	**81.3**	77.3	78.4	77.4

Table 4: Performances of multiclass document classifiers for nominal categories using feature selection. Results sorted by baseline f-score performance. Best f-score per category is given in bold.

performance are the ones most benefited from our classifier, and the categories with highest baseline score are the only ones that do not get any improvement.

Our next experiment applies feature selection over the initial classifiers. The results are given in Table 4 for *Naive Bayes* and *SVM*[8]. We can see that the scenario changes when we add feature selection, with *Naive Bayes* achieving the highest performance in all cases. The performance for the hardest category (which is again *Tumour site*) raises to above 50% f-score, clearly beating the baseline. The highest-performing category is now *Pathologic response*, and *Naive Bayes* almost reaches perfect scores over this category, improving the baseline again. The macro-averaged results show that our best classifier is able to reach an f-score of 81.3% over the 27 nominal categories, with an improvement of 30.3% over the majority class baseline.

[8] *AdaBoost* obtains the same results with and without feature selection.

4.2 Numeric categories

In this section we present the results of our numeric classifiers in Table 5. In this case the results of *Naive Bayes* are worse than the baseline, and *AdaBoost* and SVM only achieve small improvements. One of the reasons for the low performance seems to be the strong bias of the categories towards the majority value. On these conditions, the baseline obtains the best result for 6 of the 9 categories. The macro-averaged performances show that the performance is insufficient for a real application.

For our next experiment we applied feature selection to the numeric classifiers, and the results are presented in Table 6. We can see that the overall performance goes down when applying feature selection, and the main cause for this seems the low number of features that are left for each instance.

5 Discussion

Our results over nominal categories show that our classifiers can achieve high performance (above 80% f-score in average) by relying on feature

Category	*Majority Class*			*Naive Bayes*			*SVM*			*AdaBoost*		
	Prec.	Rec.	F-sc	Prec.	Rec.	F-sc	Prec.	Rec.	F-sc	Prec.	Rec.	F-sc
Nodes examined	7.4	7.4	7.4	83.1	58.1	**68.4**	82.1	58.6	68.4	81.8	57.6	67.6
Tumour length	42.9	42.9	**42.9**	41.7	24.6	31.0	50.3	37.0	42.6	44.9	39.4	42.0
Tumour width	47.8	47.8	**47.8**	41.4	26.1	32.0	51.8	42.4	46.6	46.5	42.9	44.6
Distal distance	52.2	52.2	52.2	63.3	34.0	44.2	70.1	53.2	**60.5**	52.0	51.7	51.8
Polyps number	62.1	62.1	**62.1**	48.9	43.4	46.0	57.0	54.2	55.6	58.9	58.6	58.8
Nodes positive	64.0	64.0	64.0	66.5	58.6	62.3	79.0	75.9	**77.4**	67.5	67.5	67.5
Tumour depth	70.0	70.0	**70.0**	70.1	49.8	58.2	74.7	61.1	67.2	70.0	70.0	**70.0**
Cav	72.4	72.4	**72.4**	72.1	71.4	71.8	73.1	69.5	71.2	72.4	72.4	**72.4**
Radial distance	94.1	94.1	**94.1**	94.1	94.1	**94.1**	95.4	92.1	93.7	94.1	94.1	**94.1**
Macro-average	57.0	57.0	57.0	64.6	51.1	56.4	70.4	60.4	**64.8**	65.3	61.6	63.2

Table 5: Performances for numeric categories without feature selection. Results sorted by baseline f-score performance. Best f-score per category is given in bold.

Category	*Majority Class*			*Naive Bayes*			*SVM*			*AdaBoost*		
	Prec.	Rec.	F-sc	Prec.	Rec.	F-sc	Prec.	Rec.	F-sc	Prec.	Rec.	F-sc
Nodes examined	7.4	7.4	7.4	35.3	32.0	**33.6**	22.7	21.7	22.2	26.7	25.1	25.9
Tumour length	42.9	42.9	**42.9**	39.4	31.0	34.7	42.3	37.9	40.0	43.7	40.9	42.2
Tumour width	47.8	47.8	47.8	53.5	49.3	51.3	53.3	51.7	**52.5**	49.0	48.8	48.9
Distal distance	52.2	52.2	**52.2**	53.8	31.5	39.8	52.5	51.7	52.1	52.2	52.2	**52.2**
Polyps number	62.1	62.1	62.1	52.8	51.2	52.0	64.4	64.0	**64.2**	60.2	59.6	59.9
Nodes positive	64.0	64.0	64.0	69.0	67.0	**68.0**	68.5	67.5	68.0	68.2	67.5	67.8
Tumour depth	70.0	70.0	70.0	70.2	62.6	66.1	70.8	70.4	**70.6**	70.0	70.0	70.0
Cav	72.4	72.4	**72.4**	71.1	65.5	68.2	72.4	72.4	**72.4**	72.4	72.4	**72.4**
Radial distance	94.1	94.1	**94.1**	94.1	94.1	**94.1**	94.1	94.1	**94.1**	94.1	94.1	**94.1**
Macro-average	57.0	57.0	57.0	59.9	53.8	56.4	60.1	59.1	**59.6**	59.6	58.9	59.3

Table 6: Performances for numeric categories with feature selection. Results sorted by baseline f-score performance. Best f-score per category is given in bold.

selection. These results have been attained using BOW features, and this indicates that pathology reports tend to use similar lexical elements to refer to the relevant classes. The results show promise to incorporate an extraction prototype into the medical workflow for nominal classes, which would aid the collection of structured information, and benefit from the interaction with the user.

One of the most interesting findings has been the effect of the feature selection step to achieve high performance. Apart from the increment of the f-score, feature selection would allow us to highlight the relevant terms in the document, and present them to the user for a better interaction.

Regarding the results for numeric categories, our strategy has not been successful, and the increments over the majority class baseline have been small. The baseline for these categories is higher than for nominal categories, and there is a strong bias towards the "zero" value. We observed that the main difficulty was to discriminate between "zero" and other classes, and a 2-step classifier would have been a better option to build upon. Our results over numeric categories also indicate that the generic BOW approach successfully evaluated over nominal categories may not be enough, and deeper analysis of the feature space may be required for these categories.

6 Conclusion

We have presented the results of a set of supervised text classification systems over different prediction categories in the domain of pathology records. Our results show that we are able to predict nominal labels with high average f-score (81.3%) and improve the majority class baseline by relying on Naive Bayes and feature selection. These results are positive for the integration of automatic aids in the medical workflow, and they illustrate that pathology reports contain repetitive lexical items that can be captured by a bag-of-words model. Our experiments also show that this is not the case for numeric labels, and richer features would be required in order to improve the baselines.

For future work one of our goals is to improve numeric classifiers by adding an initial classifier that identifies zero-valued instances before looking for the final value. We observed that

lexical items expressing negation may be relevant for this category (e.g. "No positive nodes were found"), we plan to incorporate the negation-classifier Negex (Chapman et al., 2001) to the feature extraction.

Finally, we want to combine our classifiers with a user interface that will allow clinicians to upload structured information into the database with the help of automatic predictions. The users will be able to copy the pathology reports, and the database fields will be pre-filled with the categories from the predictors. We will also highlight the top features from the selection process, and the user will be able to correct the automatic predictions before saving. All interactions will be kept and used to improve our classifiers.

Acknowledgments

We would like to thank Henry Gasko and his colleagues at the Royal Melbourne Hospital for providing anonymised pathology reports for our research.

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

Wendy W. Chapman, Will Bridewellb, Paul Hanburya, Gregory F. Cooperb, and Bruce G. Buchananb. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, October.

Anni Coden, Guergana Savova, Igor Sominsky, Michael Tanenblatt, James Masanz, Karin Schuler, James Cooper, Wei Guan, and Piet C de Groen. 2009. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. *Journal of Biomedical Informatics*, 42:937–949.

G. S. Dunham, M. G. Pacak, and A. W. Pratt. 1978. Automatic indexing of pathology data. *Journal of the American Society for Information Science*, 29(2):81–90, Mar.

Mark Hall. 1999. *Correlation-based Feature Subset Selection for Machine Learning*. Ph.D. thesis, Department of Computer Science, University of Waikato, New Zealand.

David A Hanauer, Gretchen Miela, Arul M Chinnaiyan, Alfred E Chang, and Douglas W Blayney. 2007. The registry case finding engine: an automated tool to identify cancer cases from unstructured, free-text pathology reports and clinical notes. *Journal of the American College of Surgeons*, 205(5):690–697, Nov.

Iain A McCowan, Darren C Moore, Anthony N Nguyen, Rayleen V Bowman, Belinda E Clarke, Edwina E Duhig, and Mary-Jane Fry. 2007. Collection of cancer stage data by classifying free-text medical reports. *Journal of the American Medical Informatics Association (JAMIA)*, 14:736–745.

Anthony Nguyen, Darren Moore, Iain McCowan, and Mary-Jane Courage. 2007. Multi-class classification of cancer stages from free-text histology reports using support vector machines. *Proceedings of the IEEE Engineering in Medicine and Biology Society Conference*, 2007:5140–5143.

Anthony N Nguyen, Michael J Lawley, David P Hansen, Rayleen V Bowman, Belinda E Clarke, Edwina E Duhig, and Shoni Colquist. 2010. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *Journal of the American Medical Informatics Association (JAMIA)*, 17:440–445.

Gunther Schadow and Clement J McDonald. 2003. Extracting structured information from free text pathology reports. *AMIA Annual Symposium Proceedings*, pages 584–588.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, USA.

Hua Xu, Kristin Anderson, Victor R Grann, and Carol Friedman. 2004. Facilitating cancer research using natural language processing of pathology reports. *Studies in health technology and informatics*, 107(Pt 1):565–572.