Efficient Encoding of Pathology Reports Using Natural Language Processing

Rebecka Weegar
Dept. of Computer and Systems Sciences
Stockholm University
rebeckaw@dsv.su.se

Jan F Nygård
The Cancer Registry of Norway
jfn@kreftregisteret.no

Hercules Dalianis
Dept. of Computer and Systems Sciences
Stockholm University
hercules@dsv.su.se

Abstract

In this article we present a system that extracts information from pathology reports. The reports are written in Norwegian and contain free text describing prostate biopsies. Currently, these reports are manually coded for research and statistical purposes by trained experts at the Cancer Registry of Norway where the coders extract values for a set of predefined fields that are specific for prostate cancer. The presented system is rule based and achieves an average F-score of 0.91 for the fields Gleason grade, Gleason score, the number of biopsies that contain tumor tissue, and the orientation of the biopsies. The system also identifies reports that contain ambiguity or other content that should be reviewed by an expert. The system shows potential to encode the reports considerably faster, with less resources, and similar high quality to the manual encoding.

1 Introduction

A cancer diagnosis is often based on an examination of a biopsy, a small tissue sample taken from a patient with a suspected cancer disease. These samples are visually examined by a pathologist, using a microscope. To document the examination, the pathologist writes a report describing findings and a diagnosis.

Pathology reports are primarily a tool used for communicating findings done by the pathologist to the physician treating the patient, but if the findings in the reports are encoded and registered in a systematic way, they can also be used for research purposes.

In Norway, all pathology reports concerning cancer diseases are reported to The Cancer Registry of Norway. The contents of each report is read and encoded by a trained coder. This is an area where an efficient information extraction system could prove very useful, since about 180,000 reports are sent yearly to the registry where they are coded by 25 full time coders. The coding of a single report takes between two and ten minutes. (Observe that Norway has a population of 5.2 million inhabitants).

In this study we have focused on pathology reports written in Norwegian describing results from prostate biopsies. The goal of the system presented in this paper is to accurately extract information from the unstructured textual content of pathology reports, so that the extracted information can be stored in a structured data format suitable for a cancer registry.

2 Previous Research

A number of studies have applied information extraction techniques to pathology reports for several types of cancer, including breast cancer, colorectal cancer, lung cancer, and prostate cancer. Scharber (2007) provides an overview of various available tools.

Most of the work in this field has been done for English text, and both rule based and machine learning methods as well as combinations thereof have been applied. For a review of the research area, see Spasić et al. (2014).

Coden et al. (2009) extracted information from pathology reports for colon cancer. A combination of rules and machine learning were used to extract nine different classes from the reports.

Ou and Patrick (2014) extracted 28 different concepts from pathology reports for primary cutaneous melanoma (skin cancer).

Martinez and Li (2011) classified colorectal cancer according to the *TNM (Tumor, Node and*

Proceedings of Recent Advances in Natural Language Processing, pages 778–783,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_100

Metastases) scale using Naïve-Bayes and Support Vector Machines.

Nguyen et al. (2011) applied rule based methods to pathology reports for lung cancer.

The mentioned studies present results in the terms of F-score ranging from 0.7 to 0.9.

Currie et al. (2006) used rules to extract concepts in 5,826 breast cancer and 2,838 prostate cancer pathology reports. The extracted around 80 fields and obtained 90-95 percent accuracy. The evaluation was carried out by domain experts.

Two studies have applied rule based methods to Norwegian pathology reports. Dahl et al. (2016) extracted values for nine concepts from 25 pathology reports describing prostate biopsies. They obtained F-scores ranging from 0.24 to 0.94. Weegar and Dalianis (2015) extracted values for ten concepts related to breast cancer with an F-score ranging from 0.67 to 1.0 using 40 reports. Both studies were done on small data sets, but the results show that rule based methods are a promising approach for information extraction from pathology reports written in Norwegian.

3 Materials

Each document in the data set consists of the report written by the pathologist and the corresponding manual encoding of the report. There are no additional annotations of the reports, meaning that the documents contain no information about which parts of the text that an encoded value is based on. The full text of each report is therefore used as input for each encoded value. The reports do not contain names or other identifiers and have been securely stored and remotely accessed to ensure privacy protection.

The data was divided into a development set containing 70 percent of the documents and a test set with the remaining 30 percent. After removing duplicate files, there were 388 documents in the development set and 176 in the test set.

The reports in the development set contain 276 tokens on average, and each report describes between one and 21 biopsies, the average number of biopsies per report is 8.25. An example of a pathology report can be seen in Figure 1.

A specific set of fields is associated with and encoded for each type of cancer. For prostate cancer biopsies, 9 fields are extracted and each of them is encoded as an integer value. The fields are:

- Primary Gleason grade, a numerical value ranging from 1 to 5.
- Secondary Gleason grade, a numerical value ranging from 1 to 5.
- Gleason score, the sum of the primary and secondary Gleason grade.
- The number of biopsies.
- The number of malign biopsies.
- The number of biopsies with orientation right/left.
- The number of malign biopsies with orientation right/left.

```
Biopsier fra  venstre prostatalapp.
2:        Prostatakarsinom, Gleason
score  3+4=7(utbredelse 4/13 mm)
4:        Prostatakarsinom, Gleason
score 3+3=6(utbredelse 0,5/12 mm)
1,3:Ikke påvist malignitet

Biopsier fra høyre prostatalapp:
5-7,9:   HPIN og adenokarsinom, Gleason
score 3+3=6(utbredelse 1/13 mm)
8:   Prostatakarsinom, Gleason
score 3+4=7(utbredelse 5/15,4/15 mm)

Perinevral infiltrasjon: ikke påvist
Infiltrasjon i fettvev: ikke påvist
```

Figure 1: A pseudonymized example of text from a pathology report describing prostate biopsies. The text contains descriptions of 9 biopsies, four from the left side and five from the right side.

A biopsy is malign if it contains cancerous cells and the Gleason grades and score are a type of cancer staging specific to prostate cancer. Gleason score is calculated as primary grade + secondary grade = score, for example:

```
Gleason score 4+3=7
```

gives that primary grade is 4, secondary grade is 3 and that Gleason score is 7. Primary here means the dominant grade seen in the biopsy. Different grades and score can be observed in different biopsies, and this means that there can be several different values for Gleason grades and score given in the same report. Of these values, the most prominent is selected and encoded.

The biopsies are typically indexed by numbers or letters and grouped together if they have the same characteristics. For example, the sentence

```
A-E, G: Biopsies from left side,
benign samples
```

	gleason	L	**prostatacarcinom**	L	**adenocarcinom**	L	**venstre**	L	**høyre**	L
1	***gleaoson***	1	***prostatakarsinom***	2	***adenokarsinom***	1	***vesntre***	2	***høye***	1
2	***glelason***	1	***prostatakarinom***	3	***adenocarcnom***	1	vekst	3	***h?yre***	1
3	***glerason***	1	***prostatakarsionom***	4	***adenokarsinom***	2	minste	3	***hlyre***	1
4	***gleaosn***	2	***protatakarsinom***	4	***adneocarcinom***	2	høystre	3	høre	1
5	***gelason***	2	***medprostatakarsinom***	5	***denokarsinom***	3	lengste	3	høyt	2
6	glass	3	***prostatatakarsinom***	6	***adenokarsinomet***	4	meste	3	nøye	2
7	glasa	3	prostatabiopiser	7	carcinom	5	tettere	4	høystre	2
8	glas	3	prostatakjertler	8	karsinom	7	hentet	4	sørre	2
9	glemt	4	adernocarcinom	8	derimot	8	beskr	4	høy	2
10	reaksjon	4	prostatasylindre	8	prostatacarcinom	8	nesten	4	score	3

Table 1: Spelling variants of key concepts *Gleason*, *Prostate carcinoma*, *Adenocarcinoma*, *Left* (venstre), and *Right* (høyre) identified by and ranked using Levenshtein distance (L) . Relevant variants are in boldface.

describes six biopsies (as indicated by the indices A-E, G) without tumor tissue with *left* orientation, and the sentence

```
3: Biopsy with prostate carcinoma,
Gleason grade 3+3=6
```

indicates one biopsy with tumor tissue and a Gleason score of six.

4 Methods

4.1 Value Extraction

A rule based solution has been implemented since the texts in reports are relatively structured. Using rules also has the benefit of transparency, the user always knows why a specific value was given by the system.

The rules included in the system were manually written using regular expressions and string matching and the system is implemented in Java. Firstly, the system reads the texts and the corresponding encoding from the documents. The text is preprocessed and the extraction rules are applied. The results of the extractions are evaluated and finally encoded.

The nine fields that are encoded for the current task can be divided into two groups, Gleason fields and Biopsy fields. The values of the fields in each group are highly dependent on each other and a set of rules have been written for each of the groups.

The Gleason grades and score are in most cases extracted together, as they are typically written as *primary grade* + *secondary grad* = *score*, with a number of minor variations. The exception is when only one or a few biopsies are reported, in those cases there is a larger variation in the reports, which requires additional rules.

For the biopsy group, six values are extracted, and the correct extraction of each value is necessary to get correct values for the subsequent fields. As a first step, each biopsy in the reports needs to be correctly identified to get the correct value for the field *Number of biopsies*. Then each identified biopsy is classified as benign or malignant and as having either the orientation left or right. The performance of the the extraction of the number biopsies that are malignant and the orientation each biopsy is limited by the performance of the first step.

4.2 Spelling Variations of Key Concepts

We identified a set of key concepts that are central to the encoding process, these concepts are *Gleason*, *Prostate carcinoma*, *Adenocarcinoma*, *Left* and *Right*. The texts contain a number of spelling variants of the key concepts and it is essential to correctly identify each of them, both standard variants, such as the two spellings of prostate carcinoma: *prostatakarsinom* and *prostatacarcinom*, and misspellings, in order to correctly encode the reports. Using the concept representation reduces the number of rules needed, since individual rules are not needed for each spelling variation.

The alternative spellings of the key concepts were found using Levenshtein distance (Levenshtein, 1966). The Levenshtein distance can be used as a measurement for how similar two strings are, and the distance is calculated by counting the number of substitutions, deletions, and insertions of characters that are required to make two strings equal.

To find the variations, the texts in the reports

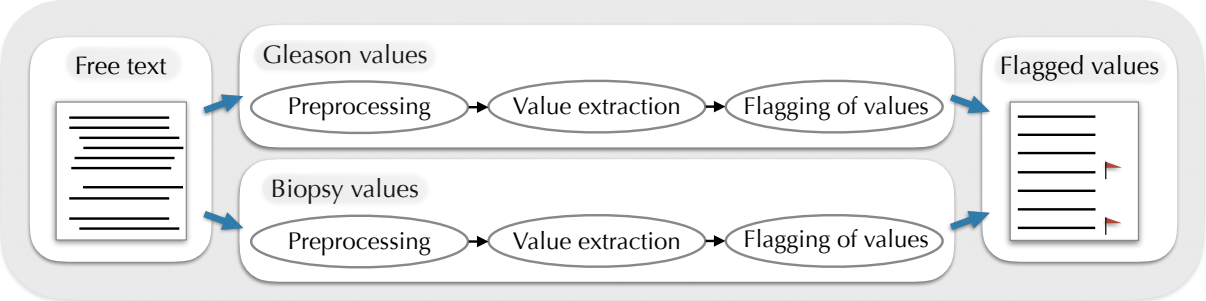

Figure 2: The contents of each report is processed separately for the two groups of fields, Gleason and Biopsy. After the values are extracted any inconsistencies in the values are flagged for manual review.

were tokenized and the Levenshtein distances between the tokens and the concept strings was calculated. Strings that are highly similar receive a low score, and the ten lowest scoring tokens for each concept are shown in Table 1. The relevant variants were manually selected from lowest scoring tokens and these tokens were included in the set of key concepts. For the concepts *Left* and *Right*, a number of abbreviations were also included.

4.3 Identifying Reports for Manual Review

The texts in the reports are relatively structured, but there are exceptions where the system might fail to extract the correct values. It would therefore be beneficial if the system itself could identify the reports that it is incapable of handling correctly. This would increase the precision of the system and allow the difficult cases to be manually reviewed.

To achieve this, a module was added that evaluates the extracted values. If inconsistencies are found, the values are flagged for review, see Figure 2. In total, three indicators for flagging were identified, and for the Gleason group, there is one such flag; a report gets marked for manual review if the concept *Gleason* is mentioned, but none of the extraction rules matches the contents of the text.

For the Biopsy group, the main source of error is that many of the reports lack information on the orientation of biopsies in the text. This information is instead often located in a sketch accompanying the reports. Since the current system is not able to process these images, the first mechanism is to flag any report that mentions neither the concept *left* nor the concept *right*.

The system contains a set of heuristic rules for inferring the orientation of the biopsies when the orientation is not explicitly stated for each biopsy. For example, if a report only mentions the concept *Left*, all biopsies in that report are considered as having the orientation left. These heuristics improve performance when applied to all files, but in some cases they also introduce errors. These errors are partly due to the fact that the orientation of the malign biopsies is more often mentioned than the orientation of the benign biopsies.

The second mechanism for the Biopsy group therefore flags the files where the sum of the biopsies identified as having orientation left or right does not match the total number of biopsies identified, or when the total sum of malign biopsies does not match the sum of malign biopsies identified as left or right.

5 Results

The system has been evaluated against the test set containing 176 reports, using precision, recall and F-score. The results for are presented in Table 2.

Field	P	R	F
Gleason grade 1	1.0	0.98	0.99
Gleason grad 2	1.0	0.99	0.99
Gleason score	1.0	0.99	0.99
Number of biopsies	0.96	0.98	0.97
Biopsies w. tumor tissue	0.92	1.0	0.96
Biopsies, right	0.69	1.0	0.82
Biopsies, left	0.69	1.0	0.82
With tumor tissue, right	0.68	1.0	0.81
With tumor tissue, left	0.69	1.0	0.82

Table 2: Precision (P), recall (R) and F-score (F) for the nine extracted fields.

The next step was then to separate out the reports which the system determines should be manually reviewed. This procedure was applied at

group level, first to the Gleason group and next to the Biopsy Group.

The system identified and flagged three reports in the test set for which the values in the Gleason group should be manually reviewed. When excluding these reports (< 2 percent of the test set), the performance was improved, see Table 3.

Field	P	R	F
Gleason grade 1	1.0	1.0	1.0
Gleason grad 2	1.0	1.0	1.0
Gleason score	1.0	1.0	1.0

Table 3: Precision (P), recall (R) and F-score (F) when excluding the three reports that the system flagged for manual review

The same process was applied to the Biopsy group. The first method for discovering challenging reports was to exclude all reports not mentioning the orientation of the biopsies. This step marked and excluded 67 reports in the test set (43 percent), and the results for remaining reports are shown in Table 4.

Field	P	R	F
Number of biopsies	0.98	0.99	0.99
Biopsies w. tumor tissue	0.95	1.0	0.97
Biopsies, right	0.89	1.0	0.94
Biopsies, left	0.9	1.0	0.95
With tumor tissue, right	0.90	1.0	0.95
With tumor tissue, left	0.91	1.0	0.95

Table 4: Precision (P), recall (R) and F-score (F) for the extracted fields in the Biopsy group, when only including the 110 reports mentioning the concepts right or left.

The second flagging mechanism for the Biopsy group marks 25 additional reports, meaning that the system is confident in correctly determining the values for all the fields in the Biopsy group for 48 percent of the reports in the test data. This step also improves the precision of the system, as shown in Table 5.

Field	P	R	F
Number of biopsies	0.99	0.99	0.99
Biopsies w. tumor tissue	0.97	1.0	0.98
Biopsies, right	0.95	1.0	0.98
Biopsies, left	0.97	1.0	0.98
With tumor tissue, right	0.97	1.0	0.98
With tumor tissue, left	0.95	1.0	0.98

Table 5: Precision (P), recall (R) and F-score (F) for the extracted fields in the Biopsy group, when only including the 85 reports mentioning the concepts right or left and reports where the sum of the fields left and right is equal to the total number of biopsies found

6 Error Analysis

The errors produced by the system are either due to the system not being able to handle previously unseen text structure or text content, or due to the data lacking information or containing noise. Lack of information is mostly regarding orientation, and noise can for example be typos, such as when two different biopsies are indexed with the same number.

System errors can be corrected by updating the rule set, and errors due to noise or lack of information is most often caught by the flagging mechanisms

The data also contains a small number of cases where there is a mistake in the encoding.

A manual error analysis has been performed on the reports that are not marked by any flagging mechanism but still contain errors. There are no such reports for the Gleason group and eight reports for the Biopsy group.

Two of the reports contain one biopsy each that is erroneously classified as malign by the system. The system fails on correctly identifying the orientation of biopsies for two reports (one because of an unusual file structure and one because of some of the biopsies actually having the orientation "center"). Three reports contain an error in the encoding and one report is correctly encoded by the system based on the actual contents of the text, but where there likely is a typo in the text corrected during the manual encoding.

7 Conclusions and Future Work

We have demonstrated the possibility of automatically extracting and encoding information from free text pathology reports with a high level of accuracy. The developed system is not designed to be implemented as fully automatic, but to reduce the amount of manual work currently needed for the encoding of the reports. The results in this study in terms of precision and recall were high for a majority of the extracted fields, and will en-

able the Cancer Registry to encode the reports considerably faster, with less resources. A vital part of the system is marking the cases which should be manually reviewed, and notifying the coding experts to be extra vigilant in the coding of the flagged reports. Thus, contributing to a more consistent encoding and further improving the quality of the data.

The results in term of precision, recall and F-score are similar to the ones described in the studies in Section 2, but though the studies all share the domain of free text pathology reports, the actual task depends on the cancer type, the number of extracted fields, availability of annotations, and language, making a fair comparison difficult. The study by Dahl et al. (2016) was developed for a similar, but much smaller, data set and achieved an average F-score of 0.73 for the nine fields, whereas the current system has a significantly higher performance with an average F-score of 0.91.

The fields concerning orientation of the biopsies are the most challenging for the system, and the encoding produced by the system for these fields are somewhat difficult to evaluate. This is largely due to the fact that the values of these fields often are based on sketches not available to the system. Excluding the reports flagged by the system as not containing the concepts *Left* and *Right* improves the results for the orientation fields, but also reduces the number of reports that the system is able to handle automatically. A high precision is prioritised over a high recall in this case since it is necessary to produce data of a high enough quality for the registry.

This study focuses only on prostate cancer, but each cancer type that is encoded by the registry is associated with a specific set of fields. Future work therefore includes to extend the system to other cancer types as well as to investigate methods for automatic rule creation.

Acknowledgments

This work was supported by the Nordic Center of Excellence in Health-Related e-Sciences (NIASC); financed by NordForsk (Project number 62721).

References

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(5):937–949. https://doi.org/10.1016/j.jbi.2008.12.005.

Anne-Marie Currie, Travis Fricke, Agnes Gawne, Ric Johnston, John Liu, and Barbara Stein. 2006. Automated Extraction of Free-Text from Pathology Reports. In *AMIA Annual Symposium Proceedings*.

Anders Dahl, Atilla Özkan, and Hercules Dalianis. 2016. Pathology text mining on norwegian prostate cancer reports. In *Data Engineering Workshops (ICDEW), 2016 IEEE*. IEEE, pages 84–87. https://doi.org/10.1109/ICDEW.2016.7495622.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8):707–710.

David Martinez and Yue Li. 2011. Information extraction from pathology reports in a hospital setting. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, pages 1877–1882. https://doi.org/10.1145/2063576.2063846.

Anthony Nguyen, Michael Lawley, David Hansen, and Shoni Colquist. 2011. Structured pathology reporting for cancer from free text: Lung cancer case study. *Electronic Journal of Health Informatics* 7(1):8.

Ying Ou and Jon Patrick. 2014. Automatic population of structured reports from narrative pathology reports. In *Proceedings of the Seventh Australasian Workshop on Health Informatics and Knowledge Management*. Australian Computer Society, Inc., HIKM '14, pages 41–50. http://dl.acm.org/citation.cfm?id=2667680.2667685.

Wendy Scharber. 2007. Evaluation of Open Source Text Mining Tools for Cancer Surveillance. *CDC* 24:28. https://www.cdc.gov/cancer/npcr/pdf/aerro/text_mining_tools.pdf.

Irena Spasić, Jacqueline Livsey, John A. Keane, and Goran Nenadić. 2014. Text mining of cancer-related information: Review of current status and future directions. *International Journal of Medical Informatics* 83(9):605–623. https://doi.org/10.1016/j.ijmedinf.2014.06.009.

Rebecka Weegar and Hercules Dalianis. 2015. Creating a rule based system for text mining of Norwegian breast cancer pathology reports. In *Sixth International Workshop in Health Text Mining and Information Analysis (LOUHI), in conjunction with EMNLP 2015, Portugal*. pages 73–78. https://doi.org/10.18653/v1/W15-2609.