Creating a rule based system for text mining of Norwegian breast cancer pathology reports

Rebecka Weegar and Hercules Dalianis
Department of Computer and Systems Sciences
Stockholm University
P.O. Box 7003
164 07 Kista
Sweden
{rebeckaw,hercules}@dsv.su.se

Abstract

National cancer registries collect cancer related information from multiple sources and make it available for research. Part of this information originates from pathology reports, and in this pre-study the possibility of a system for automatic extraction of information from Norwegian pathology reports is investigated. A set of 40 pathology reports describing breast cancer tissue samples has been used to develop a rule based system for information extraction. To validate the performance of this system its output has been compared to the data produced by experts doing manual encoding of the same pathology reports. On average, a precision of 80%, a recall of 98% and an F-score of 86% has been achieved, showing that such a system is indeed feasible.

1 Introduction

Cancer is a common cause for death worldwide, with about 14 million new cases each year (World Health Organization, 2014). In the Nordic countries it is mandatory to report each incidence of cancer to national registries and in Norway, the reported data is handled by the Cancer Registry of Norway, (Kreftregisteret i Oslo). The registry has as its main functions to monitor the cancer prevalence in Norway by collecting data on all incidences of cancer, and also to make this data available for research (Ministry of Health and Care Services, 2001). In 2013, there were about 30,000 new cases of cancer in Norway, the most common cancer type for women being breast cancer with 3,220 new cases, and the most common type for men being prostate cancer with 4,836 new cases (Cancer Registry of Norway, 2015).

Part of the data that the Cancer Registry of Norway handles originates from pathology reports. A pathology report is written by a pathologist examining a tissue sample from a patient with known or suspected cancer and the report contains a number of test results, measurements and descriptions of the sample.

The National Cancer Registry of Norway receives about 180,000 pathology reports each year and 25 full time expert coders transfer data from the free text reports to a database via an XML template. The manual encoding of the pathology reports requires special knowledge for each cancer type and the transferal is a complicated and time consuming task where the coders have to read and interpret the content of each report.

There is therefore a need of a system capable of automatic information extraction. The system should be able to accurately extract the relevant fields for each type of cancer.

2 Related research

Several studies have been performed on information extraction in the domain of pathology reports with the aim to structure their contents (Spasic et al., 2014). Rule based systems and machine learning systems are both used, and in some cases in combination. Coden et al. (2009) built a model called Cancer Disease Knowledge Representation Model, which has nine classes including anatomical site, histology, and metastatic tumor. Evaluation found that recall was between 76% and 100% and precision was between 72% and 100% for all classes except metastatic tumor where both precision and recall were lower.

Kavuluru et al. (2013) extracted the anatomical location of neoplasms from pathology reports describing several types of cancers. They achieved an average micro F-score of 90% and an average macro F-score of 72%.

Xu et al. (2004), used the MedLee system to analyze breast cancer pathology reports and had a performance for tabular findings of 95.8% sensi-

Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), pages 73–78,
Lisbon, Portugal, 17 September 2015.

tivity (recall) and 95.4% precision. For narrative text these numbers became lower with 90.6% sensitivity (recall) and 91.6% precision.

Currie et al. (2006), constructed a rule based system to extract concepts from 5,826 breast cancer and 2,838 prostate cancer pathology reports. The authors obtained around 90-95% accuracy for most of the 80 extracted fields, using domain experts for the evaluation.

Ou and Patrick (2014) studied pathology reports concerning primary cutaneous melanomas. They used both rule and machine learning based approaches. Their system was evaluated on 97 reports and they obtained an average F-score of 85% on identifying 28 different concepts including diagnosis, size and laterality and tumor thickness.

Schadow and McDonald (2003), used 275 surgical pathology reports in their experiments. Their regular expression based parser identified around 90% of the codings correctly.

McCowan et al. (2007), Nguyen et al. (2010) and Martínez et al. (2014) use text mining to perform cancer classification according to the TNM-scale (Tumor Node Metastases) (Wittekind et al., 2014).

McCowan et al. (2007), trained on 710 pathology reports for lung cancer using the SVM algorithm and evaluated on 179 reports. They obtained an accuracy of 74% for tumor staging and 87% for node Staging. Nguyen et al. (2010), developed a rule based staging system for lung cancer using 100 lung cancer pathology reports and evaluated it on 718 reports. The authors obtained an accuracy of 72%, 78%, and 94% for tumor, node, and metastases staging, respectively. Martínez et al. (2014), obtained F-scores of 81%, 85%, and 94% for staging tumor, node, and metastases respectively for colorectal cancer pathology reports. The authors used 200 pathology reports for training and evaluation.

Although closely related and relevant to this study, these studies are all performed on pathology reports in English; therefore the systems are not directly applicable to the Norwegian reports. To the best of our knowledge, only one study of information extraction from Norwegian pathology reports exists. Singh et al. (2015) used 25 pathology reports related to prostate cancer as input data. They used SAS Institute software to extract fields and they report a percentage of correctly extracted fields of 76% for number of biopsies, 24% for number of biopsies containing tumor tissue, and 100% for Gleason score. The study focuses on system development and it is not clear if they divided the data into a development set and a test set.

3 Material and methods

The Cancer Registry of Norway has selected a set of 40 pathology reports in XML-format for this pre-study. The reports have been manually de-identified by the registry and fields identifying individual patients have been removed.

The content of a pathology report depends on the procedure that produced the tissue sample. For this study the selected report types are mastectomy, where the whole breast is removed, and breast-conserving surgery, where a smaller piece is removed. Figure 1 shows an example of a portion of free text from a pathology report. It describes a tissue sample with invasive ductal carcinoma and ductal carcinoma in situ, and the measured margins around both the invasive carcinoma and the carcinoma in situ. It also mentions the percentage of estrogen receptor positive cells, progesterone positive cells and the presence of the Ki67 marker.

A program for extracting free text fields and encoded data fields from the XML-files has been written, and the input text has been divided into tokens using a custom program. A token corresponds to a unit of text, which can be a word, a number or punctuation sign, percentage sign etc. The number of tokens in the reports is ranging from 107 to 1,203 tokens with a median of 531 tokens. There are 22,670 tokens in total in the input data.

Input text and corresponding encoding

The pathology reports used in this study consist of two parts, the free text part written by a pathologist and the encoding of the same report performed by an expert coder. Each encoded field and its possible values are described in the internal requirements defined by the registry (Kreftregisteret, 2014). The requirements do, however, not say anything about how the pathologists should write their reports; the input text is therefore not as well defined as the encoded parts of the reports.

The free text contains both macroscopic and microscopic descriptions of the tissue sample. The descriptions can include test results, size measure-

```
Mammaresektat (ve.  side) med
infiltrerende duktalt karsinom,
histologisk grad 3
Tumordiameter 15 mm
Lavgradig DCIS med utstrekning 4 mm i
kranial retning fra tumor
Frie reseksjonsrender for infiltrerende
tumor (3 mm kranialt)
Lavgradig DCIS under 2 mm fra kraniale
reseksjonsrand

ER: ca 65 % av cellene positive
PGR: negativ
Ki-67:  Hot-spot 23% positive celler.
Cold spot 8%.  Gjennomsnitt 15%
HER-2:  negativ
Tidl.  BU 13:

3 sentinelle lymfeknuter uten påviste
patologiske forandringer
```

Figure 1: Extract from the free text part of an anonymised breast cancer report in Norwegian, but the data in the figure is made up and can not be linked to any individual.

ments, the type of cancer and the possible degree of hormone receptors. Other reported findings are pre-cancers and metastases in lymph nodes.

Some of the values are explicitly stated in the text as for example tumor size in Figure 1 *Tumordiameter 15 mm* (Tumor diameter 15 mm). Other values are implicit and need to be inferred from the text.

An example of this is the pT-values. They are a kind of staging information for tumors, and in the case of breast cancer the pT-value is based on the size of the tumor and what tissues the tumor is growing in (Naume, 2015). The pT value is not explicitly stated in the text, so the human or machine encoder needs to evaluate several parts of the text to determine the value of such a field.

A small portion of values appears in the same form in the input text as in the encoding, but many of the values are translated into one of a set of predefined values. For example, estrogen receptors are reported in numerical values in the text, as in Figure 1 *ER: ca 65 % av cellene positive*. This percentage value is discretized to one of six possible values when coded.

In total there can be 83 encoded fields for a single report. There are 47 different field types and 18 of the field types can be repeated up to three times depending on the number of tumors present in the tissue sample. A majority of the fields are mandatory to encode, but an option such as *not performed* is often available.

The distribution of textual and encoded fields is presented in Table 1. The implicit type is most common in the input texts and the discretized type is most common in the encodings. There is an average of 5 different values for the discretized fields.

	Encoding type	Continuous	Discrete	True/False	*Total*
Textual type	Continuous	19%	4%		**23%**
	Discrete		11%		**11%**
	Implicit		17%	30%	**47%**
	Cont./Impl.		19%		**19%**
	Total	**19%**	**51%**	**30%**	**100%**

Table 1: The 47 encoded values sorted by type, the Cont./Impl. category contains the values that are present either as continuous or implicit values in the input texts.

4 A rule based approach for information extraction

The available pathology reports have been divided into a development set of 30 reports and a test set of ten reports. The encoding of the reports has been used for evaluation and there has not been any additional manual annotation of the free text.

The developed system is based on the idea that specific fields are identified by their form and context. There are, for example, a number of fields in the reports that are reported in the form of percentages and it is possible to distinguish them by looking at characteristic tokens appearing before and after them.

Each field therefore gets assigned one or more Regex-style rules and two optional lists containing sequences of tokens. The first list holds sequences associated with the field and appearing before it, and the second contains sequences appearing after the field. The content of the context lists was created by manual inspection of the pathology reports in the development set.

One example of a field in the reports is the Ki67 hot spot value. It is often explicitly stated in the text in the form of a percentage. Therefore, the token % has been put in the after-list, and the token sequences selected for the before-list were *hot spot, hotspot, hot spotområde, ki -* and *ki67*. A program was then used to search each sentence in the data for these tokens and a regular expression was used to extract any numerical values found between them.

An automatic approach for creating the context lists has also been tested. Each unigram, bigram, trigram and 4-gram appearing in the development set was evaluated in three steps; scoring, sorting and selecting. In the first step the individual n-gram was scored using F-scores according to its ability to extract the correct values for an investigated field. In the second step, the n-grams were sorted in descending order according to this score and in the final step a set of n-grams were selected. The selection was performed by taking each n-gram in order and putting it into the context list. If the adding of the n-gram increased the total F-score for the field, the n-gram was kept in the list.

5 Results

The system has been evaluated against the manual encoding using precision, recall and F-score. The results are presented in Table 2. The fields Sentinel Nodes and the Axillary Nodes can have two possible values, *performed* and *not performed*. The field Tumor size is encoded in millimeters and therefore has many possible values. Ki67 is a protein indicating the growing rate of tumors and the two different Ki67 fields are encoded in percent. The hormone receptors for estrogen and progesterone are also reported in percent, but encoded into five and six different values, respectively. It is also possible for these values to be encoded as *not stated* if they are not present in the reports. The pT-value can be encoded as 18 different values depending on the size of the tumor, the type of cancer and where the cancer grows.

Data set	Development set			Test set		
Field	P	R	F	P	R	F
Sentinel Nodes	83	100	91	60	100	75
Axillary Nodes	93	100	97	90	100	95
Tumor Size	77	91	83	78	100	88
Histological grade	96	96	96	100	100	100
Estrogen	77	100	87	70	100	82
Progesterone	83	100	91	70	100	82
Receptors N.R.	93	100	97	90	100	95
HotSpot Ki67	93	100	97	90	100	95
Avg. Ki67	39	100	56	100	75	86
pT	80	100	88	50	100	67
Average all	82	99	88	80	98	86

Table 2: The precision (P), recall (R) and F-score (F) achieved on the test and development data in percent. N.R. stands for not reported.

The automatic creation of context lists was tested on four of the fields, histological grade, Ki67 hot spot value, Ki67 average value and tumor size; see Table 3. The automatically created context list for tokens appearing before the Ki67 hot spot value contained *hot*, *ki67*, - and *hotspot*.

	Development set			Test set		
Field	P	R	F	P	R	F
Hist. grade	96	96	96	50	100	67
Tumor size	81	96	88	88	88	88
HotSpot Ki67	93	100	97	90	100	95
Avg. Ki67	39	100	56	100	75	86
Avg. automat.	77	98	84	82	91	84
Avg. manual	76	97	83	92	94	92

Table 3: The achieved precision (P), recall (R) and F-score (F) in percent when using the automatically created context lists. The last row shows the average scores on the same four fields when using the manual approach.

6 Conclusions and Future work

In this pre-study, the possibility a system for extracting information from pathology reports written in Norwegian has been investigated.

A number of different encoding types have been identified in the data. This suggests the need for a number of approaches for successful information extraction. One main difficulty is to determine whether a value is actually present in the report, since not all tests are preformed on all tissue samples. Here, text classification could be imagined as a useful technique. Several of the fields in the reports are explicitly stated in a limited number of possible ways. In these cases, a rule based approach as the one presented here could perform well. There is also a category of values where the encoding is more complicated. This is the case when several parts of the input text needs to be interpreted to find the correct encoding, here different machine learning techniques should be investigated. An overview of the future system is shown in Figure 2.

The manually created context lists gave a better performance than the automatically created context lists. This can be explained by the fact that a human can imagine similar contexts to the ones found in the development data and add those to the context lists. The automatic creation could, however, be useful when using more data and when expanding to other types of cancers since it requires no or little manual inspection of the input texts.

The validity of the presented precision, recall

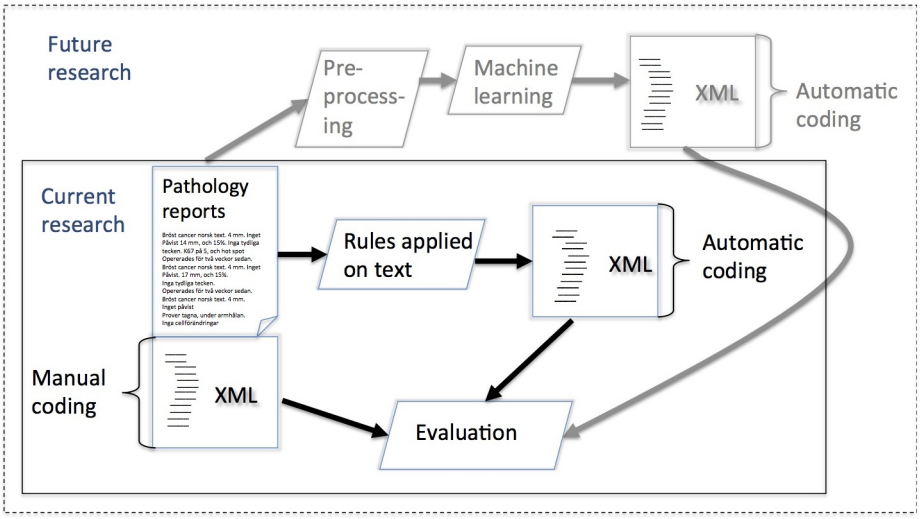

Figure 2: The pathology mining system

and F-scores for the information extraction can not be considered as very high, as too little data has been used. To make any robust claims about the performance of a future system, more test data is needed, and to properly develop the system more development data is also crucial. Ideally the performance of this system should be compared to an inter-annotator-agreement measure for the expert coders. However, the achieved results are promising and show that this system should be further developed and that a well functioning system is feasible.

Acknowledgements

The authors would like to thank Dr. Jan Nygård at Kreftregisteret Oslo, for his assistance with data access and Kjersti Østby for providing domain knowledge.

This work was supported by the Nordic Information for Action eScience Center (NIASC); a Nordic Center of Excellence financed by NordForsk (Project number 62721).

References

Cancer Registry of Norway. 2015. Cancer in Norway 2013 - Cancer incidence, mortality, survival and prevalence in Norway.

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. *J. of Biomedical Informatics*, 42(5):937–949, October.

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*.

Ramakanth Kavuluru, Isaac Hands, Durbin Eric B., and Lisa Witt. 2013. Automatic extraction of icd-o-3 primary sites from cancer pathology reports. In *AMIA Jt Summits Transl Sci Proc.*

Oslo Kreftregisteret. 2014. *Requirements specification for breast cancer reporting to the Cancer Registry of Norway, Internal document*. The Cancer Registry of Norway.

David Martínez, Graham Pitson, Andrew MacKinlay, and Lawrence Cavedon. 2014. Cross-hospital portability of information extraction of cancer staging information. *Artificial Intelligence in Medicine*, 62(1):11–21.

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*, 14(6):736–745.

Ministry of Health and Care Services. 2001. Regulations on the collection and processing of personal health data in the Cancer Registry of Norway (Cancer Registry Regulations).

Bjørn Naume. 2015. Stadier ved brystkreft, http://www.oncolex.no/bryst/bakgrunn/stadier.

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*, 17(4):440–445.

Ying Ou and Jon Patrick. 2014. Automatic population of structured reports from narrative pathology reports. In Jim Warren and Kathleen Gray, editors, *Seventh Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2014)*, volume 153 of *CRPIT*, pages 41–50, Auckland, New Zealand. ACS.

Gunther Schadow and Clement J McDonald. 2003. Extracting structured information from free text pathology reports. In *AMIA Annual Symposium Proceedings*.

Harneet Singh, Mathias Knudsen Sollie, Emil Orholm Solhøi, and Fredrik Sverre Syberg. 2015. Information Extraction: The Case of Kreftregisteret, (In Norwegian). Bachelor thesis, Westerdals Oslo ACT.

Irena Spasic, Jacqueline Livsey, John A. Keane, and Goran Nenadic. 2014. Text mining of cancer-related information: Review of current status and future directions. *I. J. Medical Informatics*, 83(9):605–623.

Christian Wittekind, Hisao Asamura, and Leslie H Sobin (eds). 2014. *TNM Atlas. Ilustrated Guide of the TNM Classification of Malignant Tumours*. Sixth edition, Wiley Blackwell.

World Health Organization. 2014. World cancer report 2014.

Hua Xu, Kristin Anderson, Victor R Grann, and Carol Friedman. 2004. Facilitating cancer research using natural language processing of pathology reports. *Medinfo*, 11(Pt 1):565–72.