Mining Norwegian pathology reports: A research proposal

Rebecka Weegar
DSV/Stockholm University
P.O. Box 7003
Kista 164 07, Sweden
rebeckaw@dsv.su.se

Hercules Dalianis
DSV/Stockholm University
P.O. Box 7003
Kista 164 07, Sweden
hercules@dsv.su.se

1 Presentation

Today, most of the transfer of information from free text pathology reports into a structured format is carried out manually. This presentation proposal discusses previous approaches to automatize the extraction of structured information from free text pathology reports and then discusses an approach specifically for pathology reports written in Norwegian.

2 Introduction

Pathology reports are written by pathologists, highly skilled physicians. Pathologists study tissue samples from the human body and describe them in free text in pathology reports. Pathology reports contain careful descriptions and categorizations of each section of the tissue sample. For cancer diseases, a pathology report typically contains descriptions of the size and anatomical site of tumours, number of affected lymph nodes, hormone receptors and TNM staging information (Classification of Malignant Tumours), also called TNM staging scale. [1] [2] (Asamura et al., 2014).

The pathology report is sent to the treating physician, so that she or he can decide on how to proceed with the treatment, but the content of pathology reports is also being documented and categorized by cancer registries.

3 Related research

There has been attempts to automatize the process of interpreting the the unstructured text of pathology reports and automatically enter it into the database of cancer registries. A description and overview of various tools is available in Scharber (2009). The text mining tools for pathology reports are mostly rule based but there are also some machine learning based tools, for a nice review article on the topic, see Spasić et al. (2014) but also Weegar and Dalianis (2015). Coden et al. (2009) have written an elaborate article on how to extract nine different classes from the pathology free text describing colon cancer using both machine learning and rules. Other approaches are described by Martinez and Li (2011) and Nguyen et al. (2011).

Generally one can conclude that most efforts give around 80 per cent precision and recall in average.

4 The case of the Cancer Registry of Norway

The Cancer Registry of Norway gathers information on all cases of cancer in Norway in order to compile statistics and increase knowledge about cancer and for this aim, pathology reports serves as one important source of information. The registry collects all cancer related pathology reports from the whole of Norway. In recent years some providers have started to send in their reports in an electronic format, which makes it possible to apply Natural Language Processing (NLP) and text mining techniques to extract information from these reports. At the moment, however, the reports are manually encoded. Specialist encoders are reading the pathology reports and entering information from the reports into the registry's database. In one year about 180,000 reports are analysed at the registry, which of course is a very time consuming task.

To develop information extraction techniques, we have gained remote access to three sets of a thousand reports each from the registry. The reports in the data sets are of three types: breast cancer, prostate cancer and melanoma. The reports are defined in an XML-format with both the input free text written by the pathologist and the corresponding encoding done by the registry. It is worth

[1] Contents of a pathology report, http://ww5.komen.org /BreastCancer/ContentsofaPathologyReport.html

[2] National Cancer Institute, Pathology Reports, http://www.cancer.gov/about-cancer/diagnosis-staging/ diagnosis/pathology-reports-fact-sheet#

noting that we have the input (the free text) and the output (the encoding), but we cannot tell from our data which parts or the input that produced each value in the encoding.

5 Proposed method

Each cancer type has its own characteristics and for each cancer type, many different types of values are described in the reports. It is not feasible to manually create rules for each cancer type and all fields of interest for each type. For example, when considering breast cancer, over 40 different fields can be encoded for a pathology report describing only one tumour. Instead, a system that automatically can learn rules for information extraction or a machine learning system would be preferable, since these types of systems are capable of learning from data.

Most previous approaches focuses on a single type of cancer or a single type of information, e.g. TNM staging, but ideally an information extraction system for a cancer registry should be capable of handling many types of information and it should also be extendable to different cancer diseases with as little manual intervention as possible.

To determine which methods that are suitable for achieving this, we need to take a closer look at the information types found in the pathology reports.

Many of the values in the reports are numerical, and for these rule based methods can be well suited. Examples of such values are the *Ki67* protein for breast cancer reports (all examples translated from Norwegian to English to simplify for the reader):

```
Ki-67: Hotspot 23% positive cells
```

and tumour size:

```
Tumour diameter 15 mm
```

The second type is binary values and values with only a few classes, for example denoting whether the sample has been tested for receptors or not. Here, a machine learning based text classification can be more suitable.

There are also cases where the values of different fields interact, as for example with TNM-staging. The TNM scale is most often not explicitly expressed in the texts and to determine, for example, the T-value for breast cancer, both the size of the tumour and its location must be correctly extracted. This type of information is more complicated than numerical or binary values and will require meta rules for describing the relationships between the encoded fields.

There is also some overlap between the different information types in the data sets. Some fields can be described both textually and numerically, a value can, for example, be expressed as "negative" or with a numerical value.

Other challenges that requires special consideration include that more than one tissue sample can be described in the same report and that findings from previous exams sometimes are referred in the reports. The free text of the reports also contains expression of uncertainty, for example:

```
Histological grade 2 or 3
```

whereas the encoding does not allow for uncertainty.

So far a rule-based base line system has been developed on a small sub-set of breast cancer reports. Optimizing on F-score an average of 0.86 was achieved on ten fields from the reports, and when using automatically learned rules on four of the fields, the average F-score was 0.84 (Weegar and Dalianis, 2015). Now, with more data available, we plan to extend this work by including text classification, improving the rule learning and extending the system with meta rules. As combinations of methods will be needed and we want to investigate if we can automatically determine the best method for each field. We also plan to investigate what representations of the input texts, including n-grams, that are suitable both for rule learning and for text classification.

References

Hisao Asamura, Christian Wittekind, and Leslie H Sobin. 2014. TNM Atlas: Illustrated Guide to the TNM Classification of Malignant Tumours.

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.

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*, pages 1877–1882. ACM.

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.

Wendy Scharber. 2009. Evaluation of open source text mining tools for cancer surveillance. *CDC*, 14:4.

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

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)*, pages 73–78.