Pathology text mining - on Norwegian prostate cancer reports

Anders Dahl
Department of Computer and System Sciences (DSV)
Stockholm University
P.O Box 7003, 164 07 Kista
Sweden
Email: anda9448@student.su.se

Atilla Özkan
Department of Computer and System Sciences (DSV)
Stockholm University
P.O Box 7003, 164 07 Kista
Sweden
Email: atoz0393@student.su.se

Hercules Dalianis
Department of Computer and System Sciences (DSV)
Stockholm University
P.O Box 7003, 164 07 Kista
Sweden
Email: hercules@dsv.su.se

Abstract—Pathology reports are written by pathologists, skilled physicians, that know how to interpret disorders in various tissue samples from the human body. To obtain valuable statistics on outcome of disorders, as for example cancer and effect of treatment, statistics are collected. Therefore, cancer pathology reports interpreted and coded into databases at cancer registries. In Norway is this task carried out by the Cancer Registry of Norway (Kreftregisteret) by 25 different human coders. There is a need to automate this process.

The authors of this article received 25 prostate cancer pathology reports written in Norwegian from the Cancer Registry of Norway, each documenting various stages of prostate cancer and the corresponding correct manual coding. A rule-based algorithm was produced that processed the reports in order to prototype automation. The output of the algorithm was compared to the output of the manual coding. The evaluation showed an average F-Score of 0.94 on four of these data points namely Total Malign, Primary Gleason, Secondary Gleason and Total Gleason and a lower result with on average F-score of 0.76 on all ten data points. The results are in line with previous research.

Keywords—clinical text mining, rule based, pathology reports, prostate cancer, Norwegian.

I. INTRODUCTION

Cancer is a leading cause of death in the developed world. In the next two decades, the number of new cancer cases is predicted to rise by as much as 70% [1]. Early cancer detection and prevention is a crucial tool to counter this development. The society have undertaken cancer-preventing measures and awareness-raising campaigns, and research to fight cancer.

One example is to collect data from cancer cases in order to establish a knowledge-base for further research. The Cancer Registry of Norway (Kreftregisteret) exists for this purpose. Among other types of cancer, the Cancer Registry collects pathology data reports regarding prostate cancer.

Presently, the process of transcribing pathology data to the Cancer Registry is handled manually by 25 human coders. This includes interpreting the pathology text and code the data into a cancer registry database. Besides being time consuming and cost resources, manual coding may suffer from problems of subjective interpretation when data is coded. Natural Language Processing (NLP) could possibly reduce or even eliminate the need for manual coding. This study presents and evaluates the effectiveness of an algorithm used for this purpose.

A. Related Research

Natural Language Processing (NLP) is the science of computer interpretation of human languages in spoken and textual form, [2], [3].

Large corporations such as Apple and Google have successfully utilized artificial intelligence in their software, rendering it possible for their devices to interpret spoken human language for assistance and authenticating purposes [4], [5].

Natural language processing has also been used to process electronic patient records. For an elaborate review of the area see Meystre et al., [6] and specifically for Swedish see Dalianis, [10]. Regarding pathology reports less work has been carried out but a nice review can be seen in Spasic et al., [13].

Accurate information extraction from free text pathology reports has been reported. One study presents a four degree scale established for the output. This study used a scale from ‘excellent’ to ‘defective’ and showed a 90% score of at least ‘sufficient’ using a data extraction algorithm, see Shadow and McDonald, [6].

One other study shows the implementation of a web based tool developed to review abstracted data. The algorithm used here classifies interpretations according to its preciseness. For example, if the input data is not interpretable by the algorithm, specific outputs are flagged as appropriate for human review and filtered through the web based tool, Currie et al., [7]. Other research carried out on extracting information from pathology reports are described in Ou et al., [14] and in Nguyen et al., [15].

Aside from attaining a high precision and recall score, defining a data model organized in a manner that fulfils the requirements of the particular application is desirable. Such a requirement could be to organize cancer-staging information over time. Also, a single tissue sample may be examined by multiple physicians with different but equally valuable interpretations. Furthermore, multiple tissue sampling of a single tumor is taken over time in order to identify the cancer stage. An automatic data extraction model will need to meet these prerequisites in order to be useful. One study by Coden et al., [8], shows a model of how such prerequisite can be met.

Another study by Weegar and Dalianis investigates free text pathology reports from the Cancer Registry of Norway.
regarding breast cancers and they present an F-score of 0.92 using their rule based model of information extraction [9].

II. METHOD AND MATERIALS

The Cancer Registry of Norway provided 25 de-identified pathology reports for prostate cancer. Each report includes data regarding tissue samples, known in medical terms as biopsies. The data consists of unformatted text describing a number of biopsy analyses made by a pathologist.

The corresponding correct manual coding of each pathology report were also obtained from the Cancer Registry of Norway. The manual coding was used as a Gold Standard. Unfortunately there are no Inter Annotator Agreement (IAA), available so it is not known how difficult the task is to code prostate pathology reports.

The 25 pathology reports form the basis for our data extraction algorithm. The algorithm generates objects represented as data structures containing all the relevant information from given report. Relevant data points include the reports identification number, the total number of biopsies and the biopsies themselves from the left and right side of the prostate respectively.

SNOMED CT is available in Norway and is also used for pathology reports, however no SNOMED CT codes were provided.

Out of the 25 pathology reports received, a development set was defined and comprised the basis of our algorithm requirements. All reports in the development set stated an identification number ranging from 1 through 25. The distinction between development set and test set was made by using reports with identification number 1 through 12 as development set and the reports 13 through 25 as a test set.

The total number of tokens in the development set was 2,101. The longest report contained 496 tokens while the shortest report contained 75 tokens.

One report could contain data regarding none, one or multiple biopsy instances. A single biopsy is either malign or benign. If malign, it also includes a Gleason score, [12]. The Gleason score is expressed with:

\[ x + y = \text{Gleason Score} \]

It is important to note that the primary Gleason Score \(x\) has higher precedence than the secondary Gleason Score \(y\). Thus, \(4+2\) results in a higher value than \(3+3\). Most biopsies in the 12 examined reports also had an orientation value: left or right.

In the majority of the reports with defined orientation values, the biopsies were textually organized left and right respectively in a manner that can be seen in Figure 1.

In all cases in the development set each biopsy was prepended with a prefix. The prefix indicated a range, a sequence or a specific number of total number of biopsies present in the corresponding line. The prefix was symbolized with either Arabic or Roman numerals, or alphabetic characters. See Figure 1 for an example of this.

B. A rule-based algorithm

For the rule based algorithm we decided to use the Python programming language.

From a data-mining point of view, a single biopsy consisted of the following tokens: It is either on the left, right or not indicated side, it is either benign or malign, and - if malign - the Gleason Score, [12], for that single biopsy. The reports were exclusively written in Norwegian. Furthermore, in cases where one or more biopsies had the same values they were usually grouped together with a prefix indicating the number of biopsies for the current values. Several examples of this is illustrated in Figure 1.

A. Materials

Since there were few pathology reports in our data set it was decided not use any machine learning method for our algorithm but a rule based algorithm.

Fig. 1. A pseudonymized example of a pathology report for prostate cancer in Norwegian (Observe that the Norwegian characters, ÆØ, has been wrongly represented to Å? after the file transfer from the Cancer Registry of Norway.

From a data-mining point of view, a single biopsy consisted of the following tokens: It is either on the left, right or not indicated side, it is either benign or malign, and - if malign - the Gleason Score, [12], for that single biopsy. The reports were exclusively written in Norwegian. Furthermore, in cases where one or more biopsies had the same values they were usually grouped together with a prefix indicating the number of biopsies for the current values. Several examples of this is illustrated in Figure 1.

A. Materials

Since there were few pathology reports in our data set it was decided not use any machine learning method for our algorithm but a rule based algorithm.

Out of the 25 pathology reports received, a development set was defined and comprised the basis of our algorithm requirements. All reports in the development set stated an identification number ranging from 1 through 25. The distinction between development set and test set was made by using reports with identification number 1 through 12 as development set and the reports 13 through 25 as a test set.

The total number of tokens in the development set was 2,101. The longest report contained 496 tokens while the shortest report contained 75 tokens.

One report could contain data regarding none, one or multiple biopsy instances. A single biopsy is either malign or benign. If malign, it also includes a Gleason score, [12]. The Gleason score is expressed with:

\[ x + y = \text{Gleason Score} \]

It is important to note that the primary Gleason Score \(x\) has higher precedence than the secondary Gleason Score \(y\). Thus, \(4+2\) results in a higher value than \(3+3\). Most biopsies in the 12 examined reports also had an orientation value: left or right.

In the majority of the reports with defined orientation values, the biopsies were textually organized left and right respectively in a manner that can be seen in Figure 1.

In all cases in the development set each biopsy was prepended with a prefix. The prefix indicated a range, a sequence or a specific number of total number of biopsies present in the corresponding line. The prefix was symbolized with either Arabic or Roman numerals, or alphabetic characters. See Figure 1 for an example of this.

B. A rule-based algorithm

For the rule based algorithm we decided to use the Python programming language.

Fig. 2. A object-oriented class diagram of the algorithm used to parse pathology reports

The class diagram in Figure 2. shows the relationship between the two classes used within the scope of our algorithm. The class Report has an integer specifying its identification number and a container that is initially empty. Instances of the class named Biopsy represents a single biopsy entry. The Biopsy-class has two boolean flags - one indicating its positional alignment and the other indicating its malignity. The class also has two integer variables for the Gleason Score, representing the primary and secondary values respectively.

When the algorithm processes a pathology report it first attempts to identify the report ID with a single regular expression. In all cases in the test data was conveniently located inside an ID-tag, see Figure 1.

The next step is to search for orientation expressions; this was most commonly expressed in the development set in Norwegian with ‘venstre’ (left) and ‘høyre’ (right). From the
TABLE I. THE RESULTS OF THE ALGORITHM APPLIED ON THE TEST SET OF PROSTATE CANCER PATHOLOGY reports IN NORWEGIAN.

<table>
<thead>
<tr>
<th>Field</th>
<th>Test set</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set</td>
<td>P</td>
<td>R</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Total Left</td>
<td>1</td>
<td>0.48</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>Total Right</td>
<td>0.31</td>
<td>0.20</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.88</td>
<td>0.46</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>Primary Gleason</td>
<td>1</td>
<td>0.50</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>Secondary Gleason</td>
<td>1</td>
<td>0.50</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>Total Malign</td>
<td>1</td>
<td>0.85</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Average all</td>
<td>0.92</td>
<td>0.67</td>
<td>0.76</td>
<td></td>
</tr>
</tbody>
</table>

development set we know that it is probable that the characters following an orientation expression will contain information about at least one biopsy, so it isolates the characters starting from the orientation expression and stopping at the next orientation expression or end of file. It then proceeds to look for a prefix applying a set of regular expression rules.

In general, a prefix is an integer followed by a semi-colon (for example 5:). '5:' is interpreted by the algorithm as 'the characters following this prefix up to the following prefix is a description of one instance of a biopsy. A prefix could also indicate several biopsy instances, for example: '1-4:' (4 instances), 'A, B, D-F:' (5 instances) or even 'A, V-VII:' (4 instances). Another set of regular expression rules is applied to the string between the prefix and the next prefix (or end of file).

The algorithm assumes that one or more instances of a biopsy exists between the string following the prefix and the next prefix. The number of instances is indicated by the prefix. The string between the prefixes are then isolated and exposed to another set of regular expression rules to identify whether the biopsy (or biopsies) is malign or not. A potential Gleason Score. In a successful run, one or more biopsy objects is instantiated with no missing variables. In all the steps in the search data that has been consumed (for example a value has been interpreted from it) or deemed irrelevant is discarded.

The algorithm uses a recursive process that is executed until there is no more input. Eight regular expression rules was formulated in total.

III. RESULTS

The algorithm was executed on the test set of 13 pathology reports and was compared to the Gold Standard and evaluated using precision, recall and F-score. The results can be seen in Table I.

The test set contained 2,084 tokens where the longest report contained 384 tokens and the shortest report contained 55 tokens. In Table I, is found average scores of 0.92, 0.67 and 0.76 using precision (P), recall (R) and F-score (F) respectively. The total of ten data points were identified from the Gold Standard received from Cancer Registry of Norway.

The ID field was an incrementing integer value ranging from 13 to 25. Total Left and Right ranged from 0 to 5. The field Total was, in the cases where Total Left and Total Right were noted, the sum of Total Left and Total Right. If Total Left and Total Right were not noted it indicated the number of biopsies without indicated orientation. The integer ranged from 0 to 12.

The fields Malign Left and Malign Right indicated the number of malign biopsies on the left and right side respectively and ranged from 0 to 5. The Total Malign field was calculated in the same way as Total and ranged from 0 to 10. The field Total Gleason consisted of the sum of the fields Primary Gleason and Secondary Gleason and ranged from 0 to 8. The Gold Standard from Cancer Registry of Norway contained one manual addition error calculating the Total Gleason score.

One false positive was verified as a true positive and found by the system but apparently missed during the manual coding.

IV. DISCUSSION

The four data-points Total Gleason, Secondary Gleason, Primary Gleason and Total Malign obtained an average F-Score of 0.94 and is in line with previous research. Analyzing the test set shows that the pathologists almost consistently used the formula x+y=z. This was caught by our regular expressions for that purpose. Failures only occurred when the Gleason Score calculation was repeated outside prefixes in text that should not have been isolated and exposed to regular expressions. For example, when a block of text ranging from a prefix to the end of file, and contained free text written by the pathologist after the last prefix. This could conceivably have been avoided by identifying some kind of 'stop tag' before the end of the file.

The field Total Right obtained a precision of 0.31, a recall of 0.20 and an F-score of 0.24 in the test set. In other words, in four out of five cases where the 'Right' indicator was present, the algorithm failed to identify it. Analysis of the test set reveals a reason for this; while the development set simply indicated orientation with the words 'Left' and 'Right' ('Venstre' and 'Høyre') in most of the cases, the test set included a large variety of this indicator. In a few cases, it was simply indicated with a 'V.' for left and 'H.' for right. Other used words were 'Pdzxt' and 'Efdxt' which most likely are abbreviations for anatomical terms indicating locations in the prostate.

Variations of prefixes was also a source for failure. One report indicated prefix with the latin alphabet (e.g. 'A: Malign', 'B: Benign'). Another indicated prefix with roman numerals for ranges (for example 'III-V:') and used alphabetic characters for single biopsies ('A: Malign', 'B: Benign', 'C: Malign', etc.). This resulted in a larger number of parsed biopsies than the number of biopsies actually present in reports.

The algorithm itself was developed on the principle of applying regular expressions at the smallest possible sequence of characters. This approach was deemed appropriate because the development set without exception stated biopsies after some kind of identifying prefix. This was not always the case with the test data. In one report all the data-points was given in a free text phrase which largely caused our approach to fail in that instance. Conceivably, the above inadequacies would possibly be remedied given a larger set of development data.
V. CONCLUSION AND FURTHER RESEARCH

Comparing our results to Weegar and Dalianis, who applied a larger development set, the scores received in the current study was notably lower with an F-score of 0.76 compared to an F-score 0.86 for Weegar and Dalianis. Though it should be stated that Weegar and Dalianis performed research regarding pathological breast cancer reports and not pathology prostate reports, the discrepancy indicates a correlation between the quantity of the development data and the success rate of the extraction algorithm. Weegar and Dalianis, [9], developed an extraction algorithm utilizing a development set of 30 reports and applied it to a test set of 10 reports as compared to the our study using a development set of 12 reports and a test set of 13 reports.

Similar to Weegar’s and Dalianis’ conclusion, the validity of the this study is considered low taking into account the low number of provided reports. Utilizing a larger development set would have enabled a larger number of variations to be taken into consideration when defining regular expressions and hopefully generating a higher precision, recall and F-Score. Therefore, further research involving larger data sets is desirable.

ACKNOWLEDGMENT

We would like to thank Jan Nygård at Kreftregisteret (Cancer Registry of Norway) for access to prostate cancer pathology reports. We would also like to thank Rebecka Weegar for valuable comments and insights on our manuscript and the reviewers of HDMM who gave us valuable comments. This work was partly supported by the Nordic Information for Action eScience Center (NIAE); a Nordic Center of Excellence financed by NordForsk (Project number 62721).

REFERENCES