Artificial Intelligence in Medicine 70 (2016) 77–83

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction

Giulio Napolitano[a,*], Adele Marshall[b], Peter Hamilton[c], Anna T. Gavin[d]

[a] Institut für Medizinische Biometrie, Informatik und Epidemiologie (IMBIE), Universität Bonn, Haus 325/11/1.OG/Raum 620, Sigmund-Freud-Straße 25, 53105 Bonn, Germany
[b] Queen's University Belfast, School of Mathematics and Physics, University Road, Belfast BT7 1NN, United Kingdom
[c] Queen's University Belfast, School of Medicine, Dentistry and Biomedical Sciences, 97 Lisburn Road, Belfast BT9 7BL, United Kingdom
[d] NICR—Centre for Public Health, The Queen's University of Belfast, Mulhouse Building, Grosvenor Road, Belfast BT12 6DP, United Kingdom

ARTICLE INFO

Article history:
Received 27 July 2015
Received in revised form 3 June 2016
Accepted 7 June 2016

Keywords:
Natural language processing
Information extraction
Supervised machine learning
Surgical pathology report
Cancer staging

ABSTRACT

Background and aims: Machine learning techniques for the text mining of cancer-related clinical documents have not been sufficiently explored. Here some techniques are presented for the pre-processing of free-text breast cancer pathology reports, with the aim of facilitating the extraction of information relevant to cancer staging.
Materials and methods: The first technique was implemented using the freely available software RapidMiner to classify the reports according to their general layout: 'semi-structured' and 'unstructured'. The second technique was developed using the open source language engineering framework GATE and aimed at the prediction of chunks of the report text containing information pertaining to the cancer morphology, the tumour size, its hormone receptor status and the number of positive nodes. The classifiers were trained and tested respectively on sets of 635 and 163 manually classified or annotated reports, from the Northern Ireland Cancer Registry.
Results: The best result of 99.4% accuracy – which included only one semi-structured report predicted as unstructured – was produced by the layout classifier with the *k* nearest algorithm, using the binary term occurrence word vector type with stopword filter and pruning. For chunk recognition, the best results were found using the PAUM algorithm with the same parameters for all cases, except for the prediction of chunks containing cancer morphology. For semi-structured reports the performance ranged from 0.97 to 0.94 and from 0.92 to 0.83 in precision and recall, while for unstructured reports performance ranged from 0.91 to 0.64 and from 0.68 to 0.41 in precision and recall. Poor results were found when the classifier was trained on semi-structured reports but tested on unstructured.
Conclusions: These results show that it is possible and beneficial to predict the layout of reports and that the accuracy of prediction of which segments of a report may contain certain information is sensitive to the report layout and the type of information sought.

1. Introduction

In a recent review, Spasić et al. [1] have analysed published efforts for the text mining of cancer-related information from a range of clinical documents. From their analysis, they conclude that machine learning (ML) techniques should be more widely investigated as an alternative to the commonly used rule-based methods. This paradigm shift, they argue, might help resolve the many difficulties faced by traditional methods in dealing with unconventional or erroneous spelling and grammar of clinical documents. A recent natural language processing (NLP) challenge for clinical records [2] has also shown that, while rule-based systems appear to dominate for clinical information extraction tasks, hybrid systems [3] combining ML algorithms and rule-based engines may outperform them. While rule-based approaches do achieve high levels of performance [4] and are usually adopted by commercial tools (such as AutoCode[1]), their application may be less successful in some contexts [5,6]. More recent projects and contests have consistently shown that ML is a promising methodology in the context of text

* Corresponding author.
E-mail address: g.napolitano@imbie.uni-bonn.de (G. Napolitano).

[1] http://www.aim.on.ca/products/autoCode.jsp (Accessed: 6 August, 2008).

http://dx.doi.org/10.1016/j.artmed.2016.06.001
0933-3657/

mining of clinical narratives [7]. In the cancer domain, this is especially true for both for the detection of reportable cases [8,9] and the actual extraction of relevant information [10,11]. Those studies also seem to confirm that ML techniques are best deployed alongside rule-based methodologies.

In this study, the performance of off-the-shelf freely available ML tools and techniques which may be used as an aid to the tasks of information extraction from breast cancer pathology reports was explored. Pathology reports are produced by trained clinicians after the macroscopic and microscopic examination of surgically resected tissue specimens and are considered the most authoritative source of cancer diagnosis information. The underlying idea, here, is that the reports can be classified into different types according to their overall structure and that, depending on their classification, fragments of the reports can be identified. These fragments most likely contain the information to be found. The isolation of relevant fragments may both reduce the complexity and improve the performance of subsequent rule-based techniques for the actual extraction of information.[2] Document structure and chunk recognition are active areas of research in several fields of application. However, their application to surgical pathology reports has never been explored. In particular, the value of automatically recognising the structure and relevant sections of reports in the context of information extraction has been neglected.

The remainder of this introduction will provide some background to the problem. The following Section 2 will outline the methods for the two main tasks—layout classification and chunk recognition. Section 3 illustrates the various tests and evaluation tasks performed for layout classification (3.1) and chunk recognition (3.2), followed by a discussion (Section 4). Finally, a brief summary and some reflections on further research that might follow from this work are provided in Section 5. Further details of the software tools used, of the textual features of the corpus and the pre-processing operations that were required on it are collected in Supplementary materials, together with a Glossary of medical terms, sample documents and novel source code.

1.1. Background

One of the available routes to reducing years of life lost due to cancer illness, at the point of care, is to ensure that epidemiological research is based upon high quality data on the incidence, prevalence and survival rates of cancer: high quality data on cancer episodes constitutes vital support for epidemiological cancer research, cancer care auditing and assessment [12]. In particular, improved completeness of staging information at diagnosis (i.e. the extent of spread of the cancer) allows for improved treatment planning and assessment of treatment effectiveness [13] and provides data for survival analysis closer to the true outcomes of the diseases [14].

The present study was conducted in the Northern Ireland Cancer Registry (NICR), which has received in electronic format all cancer-related pathology reports from all laboratories in Northern Ireland since 1993.[3] We concentrated on the reports stored in the NICR for breast cancers diagnosed in 2006.[4] Although the cancer registration dataset in the UK has been approved [15] and hopefully a higher level of integration of IT systems within the National Health Service [16] will be achieved, currently the values of many clinical test results, staging information and other pathology-related data items are not captured automatically by the main database system of the registry. Because such information is not recorded at the source (laboratory computer systems), it is not received from the data providers as a specific item of the dataset. As a result, all such data are either not available or have to be obtained by human inspection of the free text pathology reports or by the application of ad-hoc techniques. The aim of the NICR is to achieve 70% completeness in staging information. In other words, 70% of all cancer registrations should eventually be associated with some acceptable form of staging information in the NICR database. This is in alignment with the same target agreed by the UK and Ireland Association of Cancer Registries for all their registries [17]. At the same time, particular emphasis has also been placed on the importance of collecting data on some specific hormone receptor protein expression for breast cancer. In some cases the data provided from the employment of these biomarkers can indicate the effectiveness of particular treatment strategies [18] and can be used to achieve accurate patient stratification and deliver personalised medicine [19]. For these reasons, it was decided to focus the study on breast cancer reporting and on this specific information contained in the pathology reports, in addition to another essential piece of information about cancer, namely its morphology [20]:

- TNM staging [21]. This is the most commonly used cancer staging classification system. It includes information on the extent of the primary tumour (T), the absence or presence of lymph node metastasis (N) and the absence or presence of distant metastasis (M).
- Cancer morphology, which is the type of cancer cells and affects the behaviour of the disease [22].
- Hormone receptors (oestrogen, herceptin, progesterone).

1.2. Previous work

Document classification is a vast topic. Automatic document classification systems have already been developed as layers within more complex systems, for example aiming at information filtering [23]. However, in the domain of document *layout* classification, available research is usually devoted to the classification of documents under one of a number of already known, strictly well-defined semi-structured layouts. Tresch et al. [24], for instance, show how vector space classification can be used to determine the type of semi-structured documents (e.g. LaTeX document or XML) on the basis of features of their content. However, we are not aware of research exploring the possibility to detect the presence of structure per se.

Similarly, *chunk recognition* has already been used as an intermediate step towards the extraction of information, by isolating the phrases or arbitrary sections of text which may contain the information sought [25]. This, however, has never been applied to surgical pathology reports or in combination with layout classification.

2. Materials and methods

2.1. Pathology report general layout and content

Preliminary analysis showed that the reports could be roughly classified into two main categories, referred to here as *semi-structured* and *unstructured*. Semi-structured reports display several sections, each with a heading comprising one or more paragraphs. These headings originate from document templates that the clerical staff in the laboratories use to write down the dictated reports. These templates, however, can be fully edited

[2] In some cases, information extraction precision may increase by up to 19% (11%) for structured (unstructured) reports (publication in preparation).

[3] U.S. cancer registries also have increasing access to full text pathology reports, with several of them achieving >90% population-based coverage (Eric Durbin, personal communication, June 2015).

[4] From now on these will simply be referred to as "the reports".