Artificial Intelligence in Medicine 62 (2014) 11–21

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Cross-hospital portability of information extraction of cancer staging information

David Martinez[a,*], Graham Pitson[b], Andrew MacKinlay[a], Lawrence Cavedon[c]

[a] Department of Computing and Information Systems, The University of Melbourne, Doug McDonell Building, Parkville, 3010 VIC, Australia
[b] Barwon Health, Geelong Hospital, 1/75 Bellerine Street, Geelong, 3220 VIC, Australia
[c] School of Computer Science and IT, RMIT University, 124 Latrobe St, Melbourne, 3000 VIC, Australia

ARTICLE INFO

Article history:
Received 18 November 2013
Received in revised form 14 June 2014
Accepted 16 June 2014

Keywords:
Machine learning
Text mining
Information extraction
Cancer staging detection
Colorectal cancer

ABSTRACT

Objective: We address the task of extracting information from free-text pathology reports, focusing on staging information encoded by the TNM (tumour-node-metastases) and ACPS (Australian clinico-pathological stage) systems. Staging information is critical for diagnosing the extent of cancer in a patient and for planning individualised treatment. Extracting such information into more structured form saves time, improves reporting, and underpins the potential for automated decision support.
Methods and material: We investigate the portability of a text mining model constructed from records from one health centre, by applying it directly to the extraction task over a set of records from a different health centre, with different reporting narrative characteristics. Other than a simple normalisation step on features associated with target labels, we apply the models from one system directly to the other.
Results: The best *F*-scores for in-hospital experiments are 81%, 85%, and 94% (for staging T, N, and M respectively), while best cross-hospital *F*-scores reach 84%, 81%, and 91% for the same respective categories.
Conclusions: Our performance results compare favourably to the best levels reported in the literature, and—most relevant to our aim here—the cross-corpus results demonstrate the portability of the models we developed.

1. Introduction

As new technologies for health care are deployed, increasing access to electronic information opens opportunities for improved productivity and decision support. *Pathology reports* are one rich source of valuable patient information: these contain cell and tissue data and are often critical in determining presence of certain diseases and performing diagnosis. Pathology reports are typically semi-structured, containing distinguishable components but with most information in free text (though often abbreviated or terse). One specific important use of information in certain pathology reports is in determining *cancer staging*, i.e., describing the extent of cancer within a patient. This is most usually represented via the *TNM (tumour-node-metastases)* scale (described below), a global standard defined by the American Joint Committee on Cancer (AJCC) and International Union Against Cancer (UICC) [1]. Information to determine cancer staging is typically contained within the text of pathology reports but needs to be extracted, in particular for conversion to the TNM scale. While there is a proposed move towards *synoptic* (highly structured) reports, few reports (historically, virtually none) are in this format and text processing is required to automatically access the required information.

Identifying cancer stage is a critical clinical and analytic task. For an individual patient, staging information is essential for clinical decision making and determining optimal treatment [2]. For population-based cancer registries, staging data is crucial in determining overall treatment outcomes and planning research and resource allocation [3]. However extracting data from free-text pathology reports is a resource-intensive activity requiring skilled staff to manually process each report. Even when protocols for collection exist, different studies have found serious problems with completeness [4–6], along with manual encoding errors [4,5,7,8]. These problems are considered sufficiently large to be part of the Cancer Australia's National Cancer Data Strategy, where it is noted that "Traditional methods of clinical cancer registration are likely to

* Corresponding author. Tel.: +61 0390358809.
E-mail address: david.martinez.iraola@gmail.com (D. Martinez).

http://dx.doi.org/10.1016/j.artmed.2014.06.002
0933-3657/

be too labour-intensive to be sustainable".[1] Consequently, the ability to automatically extract staging data from pathology reports will assist both individual patient care and population-based analysis of outcomes. We explore the application of *text mining*, a combination of natural language processing (NLP) and machine learning (ML) techniques, to this task.

Text mining tools that perform well on the task of extracting staging information from pathology reports are likely to have important impact on clinical practice and collection of data for cancer registries and population-based studies. They can also be the building blocks of data mining methods that exploit large data for cancer prognosis [9,10]. The main challenges for developing widely-applicable text mining tools are the need for expert domain knowledge (by means of hand-coded rules or manual annotation), and the portability to different environments. However a recent systematic literature review [11] has shown limitations of existing text mining tools for the biomedical domain, which tend to be very context-dependent and not readily portable. The question of portability has been little studied, but is central to developing systems robust enough for wide clinical implementation.

In this paper, we describe a text mining tool based on using machine learning with light training-data annotation needs, for the task of extracting cancer-staging information for colorectal cancer. Our tool relies on manually annotated pathology reports for learning and for evaluation, where domain experts provide the cancer staging codes that correspond to the textual information in each pathology report. Our annotation requirement is at document level, i.e., labels are assigned to full documents (e.g., a full report is given the "staging N1" category) rather than finer linguistic levels (e.g., the phrase "3 positive lymph nodes" is given the "staging N1" category). This considerably alleviates the annotation effort (cf. [12]) and makes the techniques more scalable, given that we can address portability across different styles of pathology reports.

We specifically investigate the issue of portability of our system across distinct data sets from different hospitals. We extend our previous work on this topic [13], and apply an information extraction model that was trained from a collection of reports obtained from one health precinct—Melbourne Health—directly to a collection of reports from another—Barwon Health: these are two distinctly different health centres covering different geographical regions in the state of Victoria in Australia. The data involves differences in pathology reporting formats and authoring patterns, but also different linguistic characteristics in the authored reports. Initial performance was improved by adding a simple term-mapping approach, directed by a feature selection algorithm, resulting in clear performance stability.

Initial work over this multi-site dataset was previously reported in [14]. In the current article, we extend the analysis of the performance of our system, and also compare our results with the use of an open-source tool—*MedKAT/P* [15]—for automatically processing pathology reports. Our analysis demonstrates the limitations of off-the-shelf systems, and highlights the importance of using feature selection for normalising different datasets.

2. Background

Our specific focus is on extracting staging information for *colorectal cancer* tumours from pathology reports. In particular, our main task is to extract values for categories in the widely-used TNM cancer-staging system, which describes the extent of cancer in a patient's body. As well as providing a structure for planning a specific patient's individualised treatment, the use of a globally-adopted scale facilitates the collection of information from cancer-treatment sites and registries around the world for use in analysis.

Table 1
TNM and ACPS staging subcategorisation.

ACPS	T	N	M
0	Tis	N0	M0
I	T1	N0	M0
	T2	N0	M0
IIA	T3	N0	M0
IIB	T4a	N0	M0
IIC	T4b	N0	M0
IIIA	T1-T2	N1	M0
	T1	N2a	M0
IIIB	T3-T4a	N1	M0
	T2-T3	N2a	M0
	T1-T2	N2b	M0
IIIC	T4a	N2a	M0
	T3-T4a	N2b	M0
	T4b	N1-N2	M0
IV	Any T	Any N	M1

The categories in the TNM system are as follows: *T (size of original (primary) tumour)*; *N (nearby lymph nodes that are involved)*; and *M (distant metastasis (spread to another part of the body))*. We also apply the *Australian clinico-pathological stage (ACPS)* [16] code for each of the reports: ACPS is a scheme specific to colorectal cancer, based on the original Dukes classification for colorectal cancer, that uses a letter A–C to designate the depth of tumour through bowel wall and the letter D to denote metastatic disease. Each of these categories is assigned a score which can be combined to determine cancer staging, as illustrated in Table 1. The description of each of the categories is given in Fig. 1, as applied by the Royal Melbourne Hospital for encoding colorectal cancer. Our main task, then, is to extract the values for T, N, M, and ACPS from pathology reports.

Work on extracting staging information has previously been performed by Nguyen et al. [17–19], with a focus on *lung cancer*. Their initial approach used machine learning (ML) techniques, specifically support vector machines, and illustrated the difficulty of primary tumour stage detection (T), with a best accuracy of 64%. In a follow-up paper they explored richer annotation and a combination of ML and rule-based post-processing [18]. They performed fine-grained annotation of stage details for each sentence in order to build their system (e.g., phrases such as *chest wall invasion*, *diaphragm invasion*, etc.), and observed improvements over a coarse-grained (document-level) multiclass classifier. However, the authors explain that the annotation cost is high, and the performance for "staging T" was still low (65% accuracy). In more recent work [20], they use a lightweight NLP pipeline to identify and map entities to *SNOMED-CT* concepts.[2] After mapping to concepts, they then apply hand-authored rules to those concepts to extract TNM information, reporting accuracies of 78%, 89%, 95% for extracting T, N, M respectively. Evaluation was performed over a set of pathology reports that were not examined during system development. Given the use of a general-purpose NLP processing pipeline, their performance may transfer better to another corpus of reports, but this has not been explicitly verified.

Patrick et al. [12] outline an end-to-end system for detecting and summarising cancer reports for inclusion in a cancer registry. Their system both classifies reports from radiology departments

[1] canceraustralia.gov.au/publications-resources/cancer-australia-publications/national-cancer-data-strategy-australia (accessed 17 April 2014).

[2] *SNOMED-CT* (systematized nomenclature of medicine-clinical terms) is a large electronic collection of clinical terminology, including terms, codes, and related items; it is maintained by the International Health Terminology Standards Development Organisation (IHTSDO): http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html (accessed 17 April 2014).