Roles for language technology and text mining for next-generation healthcare

Lawrence Cavedon[a,b], David Martinez[a,b], Hanna Suominen[a,c],
Michelle Ananda-Rajah[d,e], Graham Pittson[f], Karin Verspoor[a,b]
[a] NICTA (National ICT Australia)[1] [b] The University of Melbourne [c] Australian National University
[d] Alfred Health [e] Peter MacCallum Cancer Centre [f] Barwon Health

1. Summary

Much clinical data available in electronic health records (EHRs) are in text format. Developing text processing and mining techniques for such data is necessary for realizing the full value of this data, to support data-driven analysis, decision-making, and discovery. This abstract outlines our vision and describes case studies of text processing and text mining applications over EHRs to address challenges in Health and healthcare.

2. Introduction

New Big Data initiatives in Health and Bioinformatics, including the increasing ubiquity of EHRs and cheaper gene sequencing, are creating enormous volumes of data as well as opportunity for enormous impact in the Health industry. In particular, this increasing data offers the potential to develop data analytic technologies to predict risk of disease, to prevent conditions by detecting early indicators, to monitor symptoms, and to personalise treatments that maximise effectiveness in specific population groups.

While genomic and image data are widely considered to be sources of the "biggest" Health data from individuals, much Health data -- particularly in EHRs -- is still in unstructured or semi-structured form, as text narratives. Such data must typically be processed using Language Technologies (LT) to convert human-authored text into actionable data that can be processed by computational analytic techniques, e.g., data mining algorithms. LT can also be used to extract valuable information from a "stream" of clinical reports or other texts, enabling real-time automated monitoring. The extracted information is much more compact and problem-focused than the original text reports.

We describe research at NICTA – conducted in collaboration with clinical research partners at Alfred Health, Peter MacCallum Cancer Centre, Melbourne Health, and Barwon Health – on technology for mining text in EHRs, for developing analytic techniques to support monitoring, prediction, decision support, and biomedical discovery.

3. Description

We briefly describe two tasks as case studies of our text mining and analysis research.

In-Hospital Surveillance of Fungal Infections from Radiology Reports.
Invasive fungal diseases (IFDs) cause more than 1,000 deaths in hospitals and cost the health system more than AUD100m in Australia each year. The most common life-threatening IFD is *aspergillosis*, which has been associated with a 33-75% mortality rate [1]; a patient with this IFD incurs a median 7 days prolonged inpatient time and $AU30,957 excess hospital costs [2]. Surveillance and detection of IFDs irrespective of the stage of diagnosis (i.e., early or late in disease) is important. We have developed text mining technology, using machine learning (ML) over a range of features, to automatically detect cases of patients with IFD from the text in the reports of medical imaging performed on them. Our technology will ultimately form part of a pervasive surveillance system which also uses adjunctive forms of clinical data (e.g., microbiology, histopathology, drug-dispensing data) to produce high-accuracy detection.

Method: We used supervised sentence classification, whereby an ML algorithm predicts whether each sentence in the report is indicative of the presence of IFD. Reports are classified as IFD-positive if they contain at least one IFD-positive sentence. For the sentence-classification task, we used

[1] NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

Support Vector Machines and a variety of features, including lexical (bags-of-words), semantic (concepts from the Unified Medical Language System), as well as punctuation and context.

Results: The text mining system was developed and evaluated (using 10-fold cross-validation and held-out data to avoid training bias) over a collection of 469 reports from Alfred, Melbourne, and Peter MacCallum Hospitals. Results for classifying individual scans were: 0.95 sensitivity; 0.72 specificity; 0.96 NPV; 0.71 PPV. Moreover, we obtained 100% sensitivity for classifying *patients* as IFD-positive or not, which is the most crucial result for IFD surveillance. This technology facilitates potential real-time surveillance of fungal infections, which is not currently performed in hospitals as it is a resource-intensive task often associated with negative microbiology. Early detection would support earlier intervention, potentially reducing the health and economic costs of these infections.

Automated Information Extraction from Histopathology Reports.
Identifying cancer stage is a critical oncological task for determining management of cancer in a patient. Automatically extracting staging information from pathology reports is therefore a high-value task, both for staging individual patients and for large-scale record-processing for the purpose of analysing and detecting trends. We have developed a text mining tool that automatically extracts pertinent information about cancer and tumours from histopathology reports. Moreover, we have applied the same techniques across records from different hospital systems (Melbourne Health and Barwon Health) to evaluate the robustness of the approach.

Methods: Our tool detected concepts from the reports and assigned them to appropriate categories, including: *tumour site*, *number of nodes examined*, *number of positive nodes*, *tumour length/depth/width*, etc. We also predicted the Australian clinico-pathological stage (ACPS) code for each report. We relied on the widely used TNM Classification of Malignant Tumours, which is a cancer staging system that describes the extent of cancer in a patient's body. We again used text mining techniques based on machine learning. Nominal concepts were extracted using a single multiclass document classifier, while numeric concepts involved multiple binary sentence classifiers. Features used included bag-of-words and -lemmas, UMLS (via MetaMap [3]) and SNOMED-CT concepts, and output from NegEx [4] for identifying negative contexts.

Results: The system was first trained and tested on a collection of anonymised histopathology reports from Melbourne Health. Our best results over this dataset for the staging categories in cross-validation were 82%, 85%, 81%, 70% f-scores for T, N, M, ACPS respectively. We used the same techniques over reports obtained from Barwon Health with best results for each category being 81%, 82%, 94%, 83% f-scores for T, N, M, ACPS respectively. While there is room for improvement on these results, they indicate the potential for a system that extracts relevant information from (histo)pathology reports to provide automated support for staging and monitoring.

4. Conclusions

The text mining applications above show demonstrably reliable performance in classification and extraction of valuable information. As the size of Health data grows, automated extraction of high-value actionable data from text will provide more succinct and actionable representations of that text. This will in turn support efficient online analytics and evidence-based decision-making.

References

1. Steinbach, W. J., K. A. Marr, et al. Clinical epidemiology of 960 patients with invasive *aspergillosis* from the PATH Alliance registry. *J. Infection,* 2012: 65(5): 453--456.
2. Ananda-Rajah, M. R., A. Cheng, et al. Attributable hospital cost and antifungal treatment of invasive fungal diseases in high-risk haematology patients: an economic modeling approach. *Antimicrob Agents and Chemotherapy*, 2011: 55(5): 1953-60.
3. A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *AMIA Annual Symposium Proceedings*, Washington DC, 2001: 17—21.
4. W. W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, B. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J. Biomed Informatics*, 2001: 34(5):301—310.