Dr Lawrence Cavedon

Senior Lecturer
RMIT University

lawrence.cavedon@rmit.edu.au

Dr Lawrence Cavedon is a Senior Lecturer in the School of Computer Science and IT at RMIT University, and until recently a Senior Researcher at NICTA's Victorian Research Laboratory, where he was a member of the Biomedical Informatics team. Lawrence's current research includes text mining for biomedical applications, spoken dialogue management, and other topics in Artificial Intelligence.

Text mining for lung cancer cases over large patient admission data

David Martinez[a,e], Lawrence Cavedon[a,b,e], Zaf Alam[c], Christopher Bain[c,d], Karin Verspoor[a,e]

[a] The University of Melbourne
[b] RMIT University
[c] Alfred Health
[d] Monash University
[e] NICTA VRL

SUMMARY

We describe a text mining system running over a large clinical repository for the detection of lung cancer admissions, and evaluate its performance over different scenarios. Our results show that a Machine Learning classifier is able to obtain significant gains over a keyword-matching approach, and also that combining patient metadata with the textual content further improves performance.

INTRODUCTION

The increasing availability of linked electronic patient data creates opportunities for analysis, prediction, and automation of tasks. A challenge is that much of this data remains in text format, requiring the use of Natural Language Processing (NLP) techniques to extract actionable information. Text classification according to disease is a crucial technique for retrieving specific cases or creating patient cohorts, for enabling analytics and detection of patterns of disease occurrence, or supporting resource-planning a hospital system. It can also be a prelude to automatic ICD-coding, providing support for an extremely time-consuming manual process.

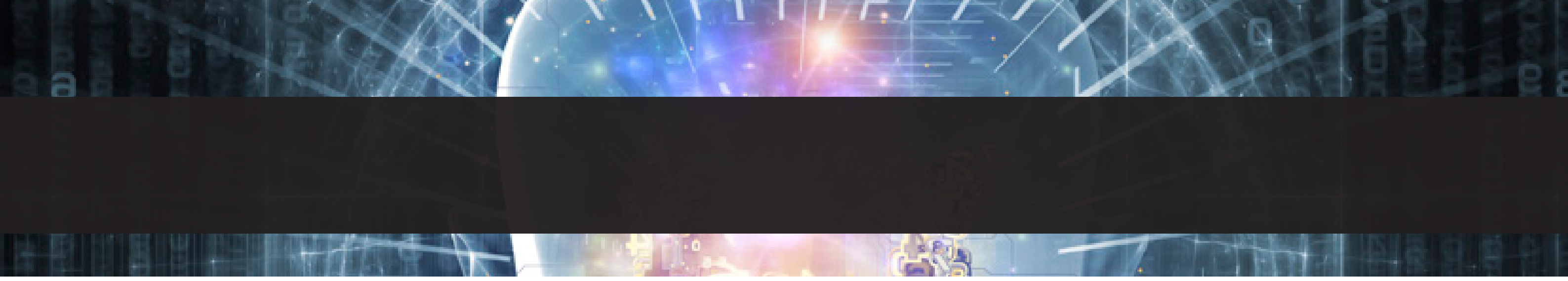
We describe initial work using data from an Informatics Platform developed at Alfred Health in Melbourne. We investigate the task of automatically assigning the ICD-10 code corresponding to lung cancer (C34, Malignant neoplasm of bronchus and lung) to a patient admission record, via application of a sophisticated text classifier using Machine Learning (ML), over two years of radiology reports from a hospital (756,520 text reports, along with associated metadata) for training and evaluation. We use manually assigned ICD codes to rigorously evaluate performance on different scenarios, using both cross-validation and time-series views of the dataset.

METHOD

The dataset for this study was extracted from the Alfred Health Informatics Platform (called REASON); it consists of all radiology reports for financial years 2011-2012 and 2012-2013. Each report is assigned an admission identifier, which is in turn linked to patient metadata, including demographics, reason for admission, etc. The metadata includes the ICD-10 codes assigned to the admission, which are used as ground truth to build a gold standard. We define the task as a binary classification problem: determine whether each admission in the test set is associated to the ICD-10 code for lung cancer: C34, Malignant neoplasm of bronchus and lung. An admission is represented by radiology scans linked to it, along with associated metadata.

Classification of lung cancer is a challenging task for automatic systems for two reasons: (i) manually-crafted keywords and phrases produce large numbers of false negatives, and also several false positives; and (ii) for our dataset only 0.8% of the admissions were positive for lung cancer: the highly-skewed nature of the data poses a specific challenge to automated ML approaches, which generally perform better over balanced class distributions.

A classifier was developed using a classical supervised learning framework. For feature representation we combined characteristics obtained from the text, along with the metadata linked to each admission, leaving out any ICD-codes since those are the target for predictions. Text in the reports was processed using the MetaMap tool[1] from the US National Library of Medicine: this identifies phrases and the polarity (negative or positive) of each, using the integrated module NegEx. We created a feature vector combining phrases obtained from MetaMap, the Bag-of-Words (BOW) representation of the text, and the metadata fields. We used the Weka Toolkit[2] implementation of the Support Vector Machine algorithm, since this has performed robustly in our previous work (e.g.[3]). We also tested the effect of applying a greedy correlation-based feature subset selection filter[4].



RESULTS

We constructed a baseline system using a simple term/phrase-matching approach, using the following (manually constructed) list of terms: “lung cancer”, “lung malignancy”, “lung malignant”, “lung neoplasm”, “lung tumour”, and “lung carcinoma”. The performance of this approach is shown at the bottom of Table 1, using the standard metrics of precision (i.e., positive predictive value), recall (i.e., sensitivity), and F-score (the harmonic mean of them). Precision in particular is low, indicating that many identified phrases were negated or neutral with respect to lung cancer. Recall is higher, but the baseline still fails to identify over one quarter of relevant admissions.

We applied the ML approach outlined above. We report here the results of the basic pipeline without use of feature selection: applying feature selection actually reduced performance, possibly because of the low proportion of positive instances in our dataset. Cross-validation was applied using random stratified 10-fold cross-validation. The results of this experiment are shown in the top two rows of Table 1 for two settings: (i) full feature set (including the metadata described above), and (ii) textual features only. There is clear improvement over the baseline in both cases, particularly in precision. The use of metadata contributes to higher performance, which illustrates the importance of linking different sources of data.

CLASSIFIER	PRECISION	RECALL	F-SCORE
Full feature set (including metadata)	0.871 (0.047)	0.820 (0.057)	0.843 (0.041)
Textual features only	0.855 (0.048)	0.800 (0.052)	0.825 (0.034)
Baseline	0.643	0.742	0.689

Table 1. Results table for the different evaluations. Standard deviation is shown between parentheses.

As a final experiment, we split the data into 3-month periods and performed two tests: (i) Test over each period using all previous history as training; and (ii) Test over each period using only the previous 3-month block as training. The results of this evaluation (using the full feature set) are shown in Figure 1, along with the keyword-matching baseline. We can see that, once we have accumulated enough training, using full history produces higher F-score than using only the previous quarter. However performance reaches a peak and then decreases over the final quarter, suggesting the possibility of changes in reporting that the model does not capture; further analysis is required to build a robust system.

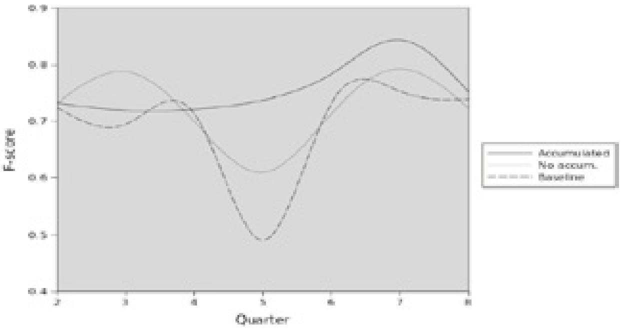


Figure 1. Time-series performance over the different classifiers

CONCLUSION

Our analysis shows promising results for automatically identifying cases of lung cancer from radiology reports, with results clearly superior to a simple keyword-matching baseline. The experiments also highlight that the model does not always improve with more data, and error analysis is required to interpret the drop in performance for the last 3-month subset of our dataset. While the techniques themselves are fairly standard, an interesting finding is the performance improvement when using metadata on top of the textual features, illustrating the importance of relying on different data sources in building more informed systems. In future work, we plan to integrate other types of clinical information in textual form, such as pathology reports, and evaluate using other disease codes.

[1] Martinez, Cavedon and Verspoor are no longer affiliated with NICTA. NICTA is funded by the Australian Government through the Dept. of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

[2] International Classification of Diseases: http://www.who.int/classifications/icd/en/

REFERENCES

1. A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA Annual Symposium Proceedings, Washington DC, 2001: 17—21.
2. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 2009, Volume 11, Issue 1.
3. D. Martinez, H. Suominen, M. Ananda-Rajah, L. Cavedon, Biosurveillance for Invasive Fungal Infections via text mining, CLEF Wshop on Cross-Language Eval of Methods, Applications, Resources for eHealth Document Analysis, Rome 2012.
4. M. Hall. Correlation-based Feature Subset Selection for Machine Learning. PhD thesis, Dept. Comp. Sci., U. Waikato, 1999.