Automated Collection of Cancer Stage Data from Free-text Pathology Reports

Iain McCowan[a], Darren Moore[a], Anthony Nguyen[a], Mary-Jane Courage[b]

[a] *CSIRO e-Health Research Centre, Brisbane, Australia*
[b] *Queensland Cancer Control Analysis Team (QCCAT), Queensland Health, Brisbane, Australia*

Keywords: Cancer Staging; Lung Cancer; Machine Learning.

Overview

This poster presents a prototype software system to assist in collecting retrospective cancer stage data using automatic text categorisation of free-text pathology reports. Cancer staging provides a basis for planning clinical management, but also allows for meaningful analysis of cancer outcomes and evaluation of cancer care services. Despite this, stage data in cancer registries is often incomplete, inaccurate or simply not collected.

A prototype system for T and N staging was developed on 710 lung cancer cases using free-text pathology reports stored in clinical information systems. The system uses automatic text categorisation techniques to detect relevant statements within reports, and then automatically assign a stage. In a trial on a further 179 cases, system output was compared to that of two clinical experts. Inter-expert agreement was also studied. Such a system could be used to obtain stage data for patients without a formally recorded stage, allowing more comprehensive population-level analysis of cancer outcomes.

Method

The input to the system is the set of lung cancer related pathology reports for a patient. All report text is first standardised and sentence boundaries are identified. Sequences of words are then transformed into codes from the UMLS Specialist Lexicon and phrases implying negation are identified and associated with surrounding terms. Each transformed report is then classified for relevance to T and N staging by support vector machines. Reports deemed irrelevant are omitted from further processing.

Each sentence is then input to a series of classifiers corresponding to factors from the staging guidelines. Classifiers use support vector machines for factors that were sufficiently well represented in the development data set. For less common factors, a rule-based classifier analyses the proximity of tumour, invasion and body part key-phrases. The output stage is the highest associated with any factors classified as positive across all sentences for that patient.

Results

The two main objectives of the clinical trial were 1) to study the level of agreement in expert staging decisions, and 2) to evaluate the reliability of the automatic staging system.

The main findings were:

1. On the 179 trial cases, inter-expert agreement was 89.9% and 97.8% for T and N staging, respectively. Disagreement was due to ambiguity in the reporting, resulting in different assumptions and interpretations in the expert decisions.
2. The automatic system was evaluated against the expert-assigned stage. T staging performance was 75.0% and N staging performance was 87.4%.

Conclusion

The automatic system for staging lung cancer was validated against expert decisions in a trial setting with promising results. Ongoing work is focussing on adapting the system for staging other cancer types.