18th Annual Health Informatics Conference, HIC 2010
D.P. Hansen, L.K. Schaper & D. Rowlands (Eds.)

Structured pathology reporting for cancer from free text: lung cancer case study

Anthony N. Nguyen[1], Michael J. Lawley[1], David P. Hansen[1], Shoni Colquist[2]
[1]*The Australian E-Health Research Centre, CSIRO ICT Centre, Brisbane, Australia*
[2]*Queensland Cancer Control Analysis Team, Queensland Health, Brisbane, Australia*

Abstract.

Objective

To automatically generate structured reports for cancer, including TNM (Tumour-Node-Metastases) staging information, from free-text (non-structured) pathology reports.

Method

A symbolic rule-based classification approach was proposed to identify symbols (or clinical concepts) in free-text reports that were subsumed by items specified in a structured report. Systematised Nomenclature of Medicine – Clinical Terms (SNOMED CT) was used as a base ontology to provide the semantics and relationships between concepts for subsumption querying. Synthesised values from the structured report such as TNM stages were also classified by building logic from relevant structured report items. The College of American Pathologists' (CAP) surgical lung resection cancer checklist was used to demonstrate the methodology.

Results

Checklist items were identified in the free text report and used for structured reporting. The synthesised TNM staging values classified by the system was evaluated against explicitly mentioned TNM stages from 487 reports and achieved an overall accuracy of 78%, 89% and 95% for T, N and M stages respectively.\

Conclusion

A system to generate structured cancer case reports from free-text pathology reports using symbolic rule-based classification techniques was developed and shows promise. The approach can be easily adapted for other cancer case structured reports.

Introduction

Surgical pathology cancer case reporting involves the communication of an extensive amount of scientifically validated clinical information for each tumour and tumour site (Qu *et. al.*, 2007). To assist pathologists with the consistent reporting of cancer specimens, the United Kingdom through the Royal College of Pathologists (RCP) and the United States through the College of American Pathologists (CAP) have developed and reviewed processes for defining structured (or synoptic) reporting protocols. In line with these developments the Royal College of Pathologists of Australasia (RCPA) has initiated the development of protocols for the structured pathology reporting of cancer (Royal College of Pathologists of Australasia, 2009).

In particular, CAP has produced checklists containing a list of tumour site specific items for structured reporting (College of American Pathologists, 2009). The value of the checklists have been recognised by the American College of Surgeons Commission on Cancer (ACS CoC) and has mandated, as a minimum requirement, the documentation of checklist items in pathology reports at CoC-approved cancer programs (College of American Pathologists, 2009). Although, the ACS does not require a specific format for pathology reports, the cancer checklist provides a structured and standardised framework for cancer pathology reporting. Major cancer centres and institutions in USA and Canada have moved towards structured cancer checklist data entry systems (e.g. Qu et. al. (2007)).

Structured reporting provides many advantages compared to traditional free-text reports such as providing a summary of reportable clinical findings and decreased variation in the content of cancer-related pathology reports (commonly caused by individual and institutional variations, transcription errors during dictation, and insufficient and omitted clinical data in free text) (College of American Pathologies, 2009). Despite the benefits of structured reporting, a large portion of historical data and free text practice still exists.

Motivated by the fact that retrospective structured reporting (and staging) is important for clinical management and treatment planning of individual patients, cancer notification and registration, and outcomes analysis of cancer management and intervention programs, we have identified that automatically extracting structured report items from free text would help realise these outcomes with reduced (or limited) manual intervention.

It is hypothesised that items from structured reports such as the CAP cancer checklist can be extracted from reports by determining whether these items subsume clinical concepts identified in the free text. The extracted items can also be used to build logic to derive synthesised items such as cancer stage. The Systematised Nomenclature of Medicine – Clinical Terms (SNOMED CT) (International Health Terminology Standards Development Organisation, 2008) is an internationally recognised clinical terminology standard and was used as the base ontology for the identification of clinical concepts in free text and subsumption querying. The lung cancer checklist relating to lung resections was used to demonstrate the methodology.

Method

The proposed structured reporting system builds upon the Medical Text Extraction (MEDTEX) system (Nguyen *et al.*, 2009), developed using the General Architecture for Text Engineering (GATE) platform (Cunningham *et al.*, 2002). MEDTEX comprises identification of SNOMED CT concepts in free text, and the detection and application of medical negation phrases to relevant disease and finding concepts. Additional modules include further pre-processing to segment the free text into sections, and the extraction of items for structured reporting.

The CAP cancer checklist has been encoded with SNOMED CT codes (College of American Pathologists, 2009) and was used to identify items to be extracted from free text for the structured report. In particular, concepts identified in the free text were tested for subsumption by the SNOMED CT encoded checklist items. The surgical lung cancer resection checklist (College of American Pathologists, 2006), based on the American Joint Committee on Cancer (AJCC) 6th edition staging guidelines (Greene *et. al.*, 2002), was used to illustrate the structured reporting methodology.

SNOMED CT expressions were used to facilitate retrieval using subsumption querying. Expressions consist of a single concept or a combination of concepts post-coordinated by the user according to SNOMED CT's compositional grammar. To test for the subsumption of a candidate expression by a predicate expression, expressions were transformed to their normal forms and concepts codes from the normal forms were tested for subsumption using rules defined in the SNOMED CT Transforming Expressions to Normal Forms publication (International Health Terminology Standards Development Organisation, 2007).

In the event that concepts were not fully modelled (i.e. a concept's defining relationships do not provide a sufficient characterisation of the concept for subsumption testing), new concepts were created and modelled using post-coordinated expressions conforming to the compositional grammar and thus creating a SNOMED CT extension (Lawley et al., 2008). However, there are cases where the compositional grammar is insufficient to model the required relationships between concepts (and hence it is not possible to create a SNOMED CT extension) in which case, ad-hoc concepts were used to test for subsumption.

An example of a predicate and candidate expression for a fully modelled SNOMED CT concept (i.e. predicate expression is the concept's normal form) is shown in Table 1, where the candidate expression is defined as a template and is filled in by <*procedure*> and <*topology*> concepts identified in the free text, and <*procedure.method*> is the value of the "method" attribute from <*procedure*>. These predicate and candidate expressions allow the identification (and thus filtering) of lung resection reports for the extraction of lung resection checklist items.

Table 1. Example predicate and candidate expression used for subsumption querying

Concept	119746007 | lung excision |
Predicate Expression	71388002 | procedure | : {260686004 | method | = 129304002 | excision – action | ,405813007 | procedure site – Direct | = 39607008 | lung structure | }
Candidate Expression	<*procedure*> : {260686004 | method | = <*procedure.method*> ,405813007 | procedure site – Direct | = <*topology*>}

The search for concepts in the free text for the population of candidate templates has a limited scope of six terms (or concepts) bounded by conjunction phrases and sentence boundaries within the relevant sections of the free text. However, when four or more concept codes were identified in the free text for a single candidate expression, then the six term restriction was removed.

Results

A set of 114 reports pertaining to a random 100 lung cancer patients from a corpus of 1205 de-identified pathology reports for 1054 lung cancer patients was used for system development. The corpus was obtained from Queensland Health with research ethics approval. The remaining 1091 (non-development) reports were used for evaluation purposes.

As a measure of system performance, the synthesised TNM stages were evaluated against reports with explicitly mentioned TNM stages. There were 491 of the 1091 non-development reports that had TNM stages recorded in them. Four of these reports were found to have TNM stages only recorded in the "History" section and therefore were not relevant to the current lung resection examination detailed in the report. Discarding these 4 reports, there were a total of 487 reports which had at least a TNM stage in the non-history sections of the free text. The final TNM stage recorded in the report was used as the ground truth stage for evaluations, and a MX (metastasis cannot be evaluated) stage was assumed if only T (tumour) and N (node) stages were recorded in the free text. Overall TNM stage accuracy with respect to the TNM stages recorded in the reports and those synthesised by the proposed system is shown in Table 2.

Table 2. Accuracy of system with respect to TNM stages recorded in reports

Stage	Reports	Accuracy % (95% CI)
T	487	78 (74–81)
N	487	89 (86–91)
M	487	95 (92–96)

It was also observed that other extracted information from free text show promise with satisfactory results. The extracted checklist items for each report was stored as a XML document and can be associated with a style sheet or parsed for visualisation. An example structured report output by the system is shown in Figure 1.

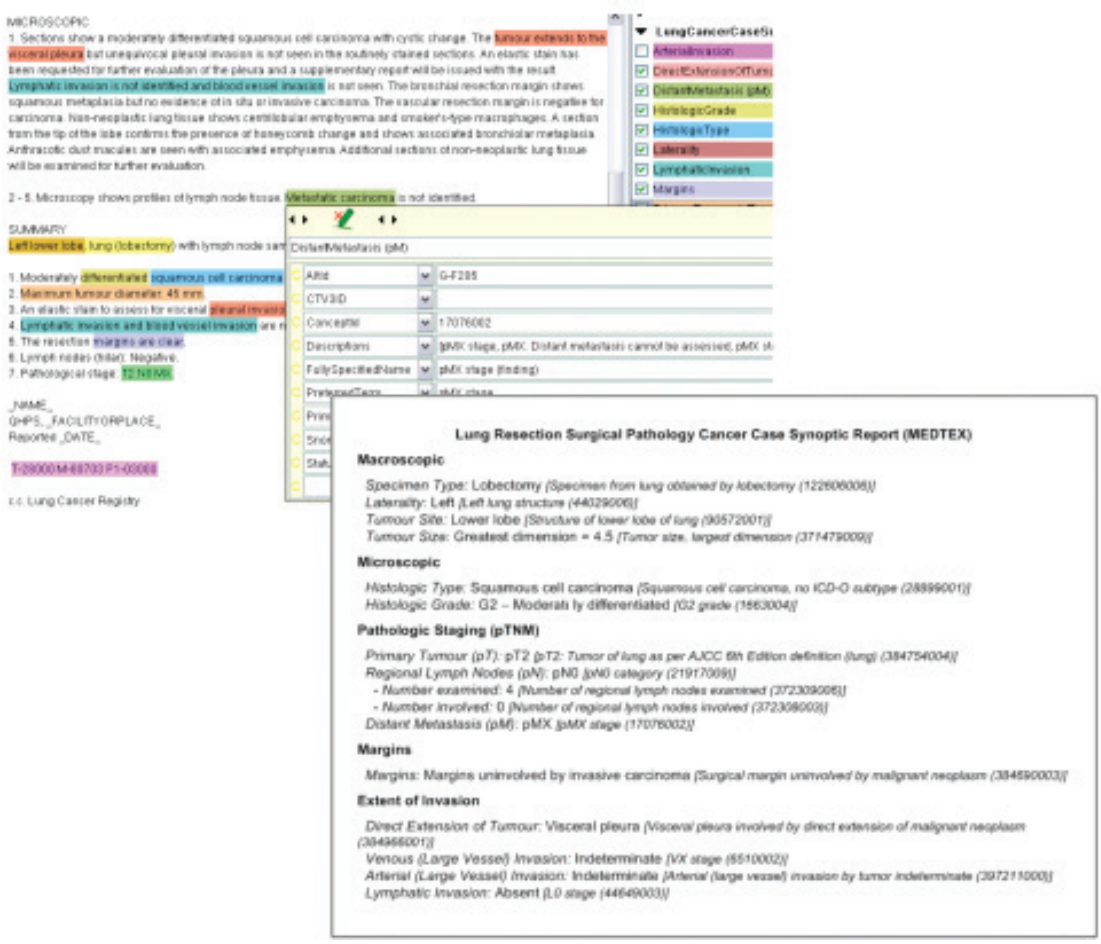

Figure 1. Example system annotated free text and structured pathology report.

Discussion

Examination of the structured reports show promise with satisfactory results for all items extracted. Checklist items other than stage were not evaluated due to the lack of readily available validation data. However, the proposed methodology based on using the SNOMED CT ontology and its semantics is the same for all structured report items. It is proposed that the lung resection synoptic reporting system along with other structured reports for other cancer types will be formally evaluated against independent experts to determine the level of accuracy of the system and the accuracy required for practical deployment.

Overall TNM stage accuracy on the evaluation set with respect to the TNM stages recorded in the reports (Table 2) was very encouraging. Staging errors were found to be a result of the occurrence of proximity and/or possibility terms near relevant findings, and also due to the fact that not all factors relevant to staging were itemised in the checklist to synthesise a cancer stage. These limitations were observed to also cause errors in other structured report items. However, the proposed approach is flexible and extensible in that errors can be fed back into the development process to improve system performance. For example, one solution to address the proximity and possibility terms limitation is to add these terms to the list of "pseudo-negation" terms (i.e. phrases that are not reliable indicators of negatives) in MEDTEX's negation detection module, and use these phrases to neither assert a negative or positive disease or finding concept.

The proposed symbolic rule-based approach using SNOMED CT can be easily adapted to other structured reports for cancer. In fact, the current lung cancer staging component of the system is currently being adapted to perform staging using the recently published 7th edition of the cancer staging guidelines (Sobin *et al.*, 2010).

Conclusion

An automated symbolic rule-based system for generating structured reports from free-text pathology reports was proposed. SNOMED CT concepts identified in the free text were symbolically manipulated to post-coordinate SNOMED CT expressions for subsumption querying against items in the structured report. The method shows promise on lung cancer cases and its utility will be evaluated on other cancer types.

Acknowledgment

This research is a part of the Cancer Information Processing and Reporting (CIPAR) project, a partnership between CSIRO Australian e-Health Research Centre and Queensland Cancer Control Analysis Team (QCCAT) within Queensland Health. The authors would like to acknowledge: QCCAT staff for their help in providing access to histopathology data for lung cancer patients.

References

College of American Pathologists (2006), SNOMED CT–Encoded CAP Cancer Checklist (v1.5).

College of American Pathologists (2009), An Overview of the College of American Pathologists Cancer Checklists. Jan.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. (2002), 'GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications,' Proceedings of the Association for Computational Linguistics (ACL). Philadelphia, July.

Greene FL, Page DL, Fleming ID, Fritz A, Balch CM, Haller DG, et al., editors. (2002), AJCC Cancer Staging Manual. 6th ed. Springer.

International Health Terminology Standards Development Organisation. (2007), SNOMED Clinical Terms® Transforming Expressions to Normal Forms. Jan 31.

International Health Terminology Standards Development Organisation. (2008), SNOMED Clinical Terms® User Guide. Jul.

Lawley M, Vickers D, Hansen D. (2008), Converting Ad Hoc Terminologies to SNOMED CT Extensions. Proceedings of the Health Informatics Conference; Melbourne, Australia. p. 133.

Nguyen AN, Lawley MJ, Hansen DP, Colquist S. (2009), A Simple Pipeline Application for Identifying and Negating SNOMED Clinical Terminology in Free Text. Proceedings of the Health Informatics Conference; Aug: Canberra, Australia. pp. 188-193.

Royal College of Pathologists of Australasia. (2009), 'Structured Reporting,' Available from: http://www.rcpa.edu.au/Publications/StructuredReporting.htm; accessed Feb 18, 2009.

Qu Z, Ninan S, Almosa A, Chang KG, Kuruvilla S, Nguyen N. (2007), Synoptic reporting in tumor pathology – Advantages of a web-based system. Am J Clin Pathol. *Jun;*127(6):898-903.

Sobin LH, Gospodarowicz MK, Wittekind C., editors. (2010), TNM Classification of Malignant Tumours. 7th ed. Blackwell Publishing Ltd.

Correspondence

Anthony Nguyen, PhD
The Australian e-Health Research Centre
Level 5 – UQ Health Sciences Building 901/16
Royal Brisbane and Women's Hospital
Herston QLD 4029, Australia
E-mail: anthony.nguyen@csiro.au