electronic Journal of Health Informatics
www.eJHI.net
2012; Vol 7(1):e8

Structured Pathology Reporting for Cancer from Free Text: Lung Cancer Case Study

Anthony Nguyen[1], Michael Lawley[1], David Hansen[1], and Shoni Colquist[2]

[1] The Australian e-Health Research Centre, CSIRO ICT Centre, Brisbane, Queensland, Australia
[2] Queensland Cancer Control Analysis Team, Queensland Health, Brisbane, Queensland, 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. Systematized 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 were 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.*

Keywords: Cancer Staging; Information Extraction; Lung Cancer; Natural Language Processing; Synoptic Reporting; Systematized Nomenclature of Medicine

1 Introduction

Surgical pathology cancer case reporting involves the communication of an extensive amount of scientifically validated clinical information for each tumour and tumour site [1]. 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 [2].

In particular, CAP has produced checklists containing a list of tumour site specific items for structured reporting [3]. 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 [3]. 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. [1]).

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 re-

ports (commonly caused by individual and institutional variations, transcription errors during dictation, and insufficient and omitted clinical data in free text) [3]. 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 Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) [4] 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.

2 Methods

The proposed structured reporting system builds upon the Medical Text Extraction (MEDTEX) system [5], developed using the General Architecture for Text Engineering (GATE) platform [6]. MEDTEX comprises of analysis modules to pre-process the free text such as identifying SNOMED CT concepts and their semantic types, as well as abbreviations/acronyms, shorthand terms, and dimensions. It then uses contextual information such as medical negation phrases to negate relevant disease and finding concepts. Additional modules were developed to identify headings and sections, and apply formal semantics to reason with the clinical concepts for the extraction of items for structured reporting.

The CAP cancer checklist has been encoded with SNOMED CT codes [3] 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 [7], based on the American Joint Committee on Cancer (AJCC) 6th edition staging guidelines [8], was used to illustrate the structured reporting methodology.

Synoptic items from the CAP cancer checklist required for accreditation purposes for the CoC including an optional data element, *Lymphatic (Small Vessel) Invasion*, were implemented. Figure 1 shows the architecture of the lung resection cancer synoptic reporting module. The lung resection cancer checklist is typically used for pathological lung cancer staging.

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 [9].

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 [9]. 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 2 shows an example synoptic item's concept requiring the creation of a SNOMED CT extension. In this case, concept model attributes were used to fully model the concept such that the extension is a sufficient characterisation of the concept in the context of the lung resection cancer checklist. Here, the candidate expression would need to be subsumed by the predicate expression, which is the SNOMED CT extension. In this example, <*disease*>, <*morphology*>, <*margin*>, and <*topology*> are disease, morphology, surgical margin and topology concepts found in the free text, respectively, and <*disease.associated morphology*> is the "associated morphology" attribute value in <*disease*>. Note that the candidate expression is dependent on <*dis-*

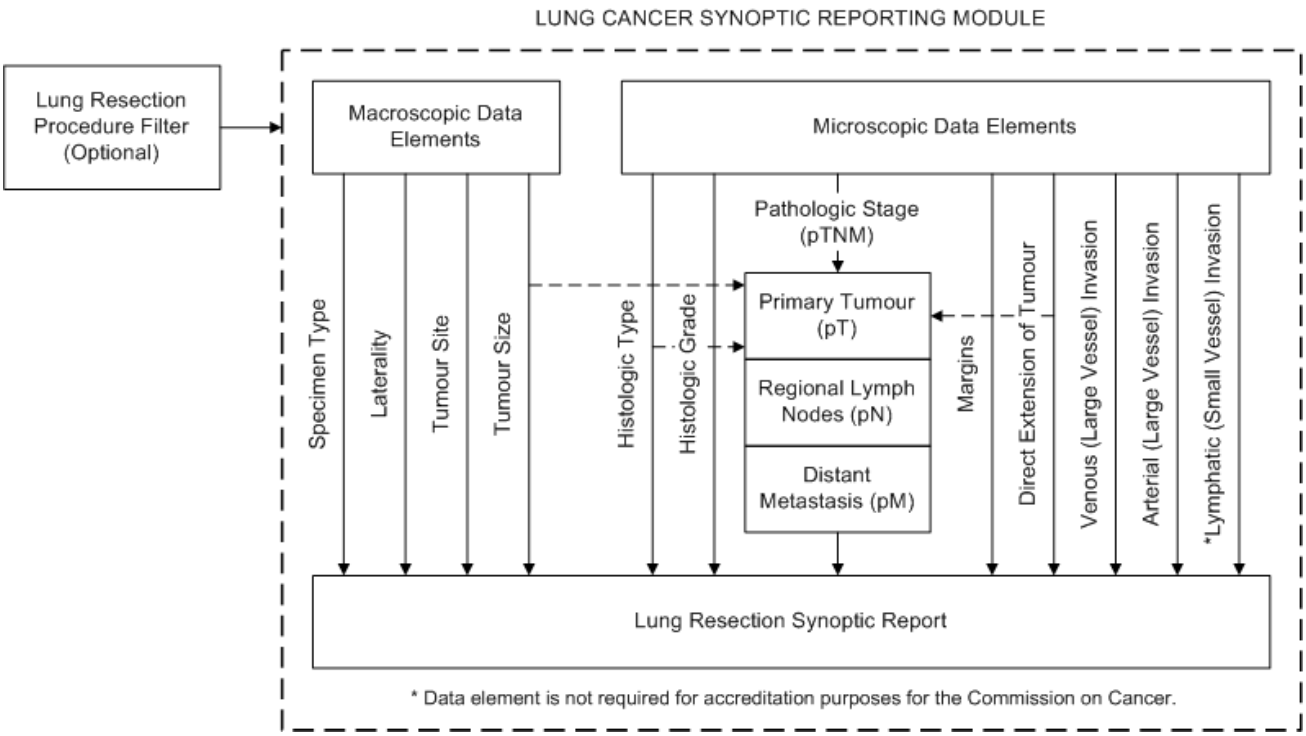

Figure 1: Cancer synoptic reporting module relating to surgical lung resections.

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>*}

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

ease> or <*morphology*>, <*margin*> and <*topology*>. The same methodology in creating SNOMED CT extensions can be applied to other topologies where surgical margins have been involved by malignant neoplasm.

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/or sentence boundaries within the relevant sections of the free text. However, when four or more concepts are identified and used to populate a candidate expression, then the six term scope restriction was removed and the scope of search would be bounded by conjunction phrases and/or the start and end of the sentence.

3 Data Set

The data set was obtained from Queensland Health with research ethics approval from the Queensland Health Research Ethics Committee. 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 remaining 1091 (non-development) reports were used for evaluation purposes.

As a measure of system performance, the synthesised TNM stages (T[X,is,0-4], N[X,0-3], M[X,1] [8]) 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.

4 Results

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 3.

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 2.

5 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 3) 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 synthesised TNM stages was also evaluated against a database of multidisciplinary team staging decisions and a machine learning-based text classification system using support vector machines in [10]. Results from the proposed system against a database of pathological TNM staging decisions were 72%, 78%, and 94% for T, N, and M staging, respectively. The system's performance was also comparable to the support vector machine based classification approach. A more detailed discussion and comparison between the symbolic and machine learning based approaches can be found in [10].

The proposed symbolic rule-based approach using SNOMED CT can be easily adapted to other structured reports for cancer. The methodology is currently used for the extraction and coding of items relevant for the notification of cancers such as basis of diagnosis, histological type and grade, cancer site and laterality from a state-wide pathology repository [11]. The methodology can also be used in health domains beyond cancer, and is currently used for the identification of patients for ad-

Concept	**384955008 | surgical bronchial margin involved by malignant neoplasm |**
Normal Form	384955008 | surgical bronchial margin involved by malignant neoplasm | : 363714003 | interprets | = 15220000 | laboratory test |
Predicate	64572001 | disease | :
Expression (SNOMED CT extension)	116676008 | associated morphology | = 367651003 | malignant Neoplasm (Morphology) | ,363698007 | finding site | = 82868003 | surgical margins | 116676008 | associated morphology | = 367651003 | malignant Neoplasm (Morphology) | ,363698007 | finding site | = 955009 | bronchial structure |
Candidate Expression	64572001 | disease | + *<disease>* : 116676008 | associated morphology | = *<disease.associated morphology>* ,363698007 | finding site | = *<margin>* 116676008 | associated morphology | = *<disease.associated morphology>* ,363698007 | finding site | = *<topology>* 116676008 | associated morphology | = *<morphology>* ,363698007 | finding site | = *<margin>* 116676008 | associated morphology | = *<morphology>* ,363698007 | finding site | = *<topology>*

Table 2: Example SNOMED CT extension used for subsumption querying.

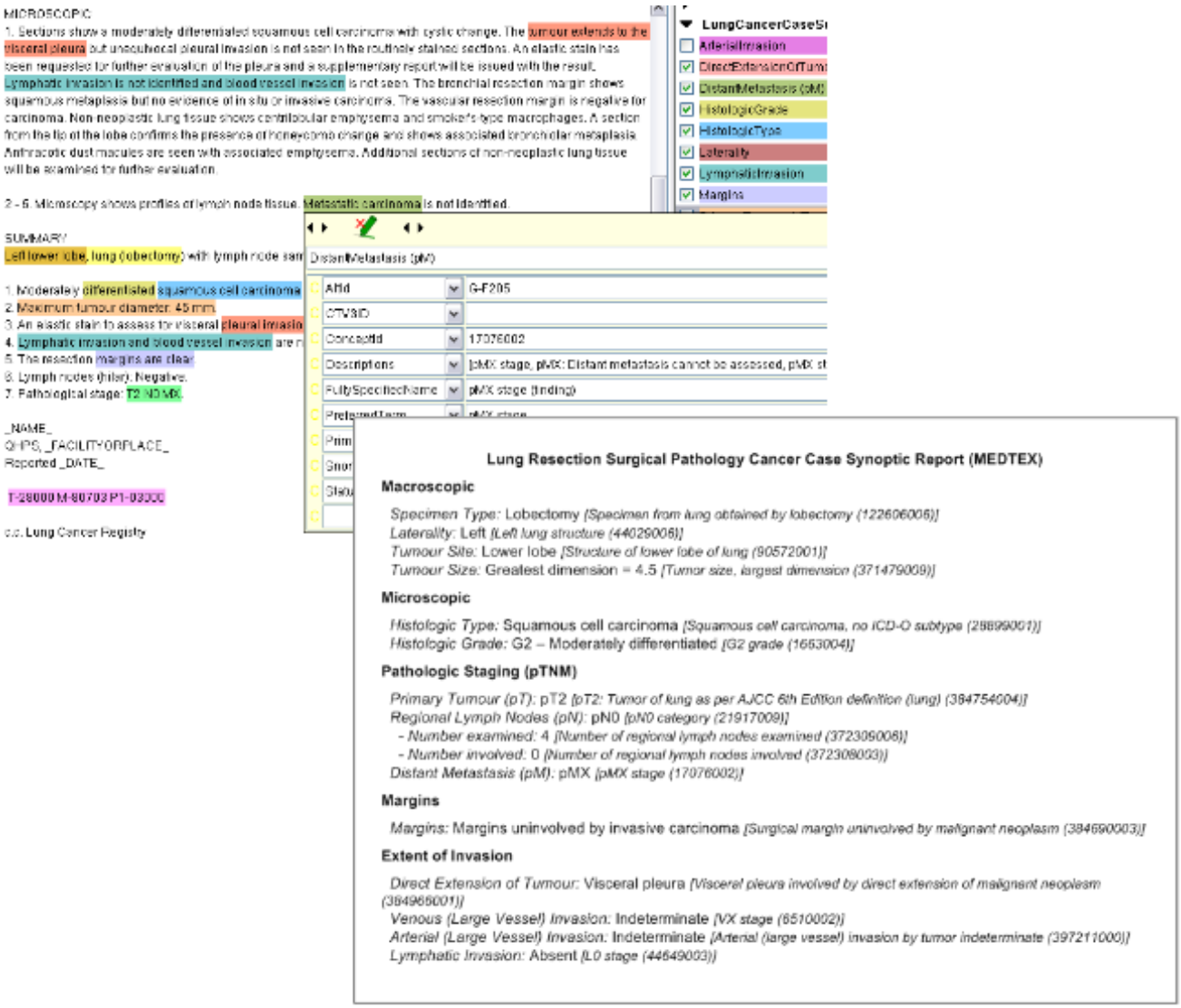

Figure 2: Example system annotated free text and structured pathology report.

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

Table 3: Accuracy of system with respect to TNM stages recorded in reports.

vanced prostate cancer clinical trials [12]. Given the performance levels achieved, the system can be used for retrospective studies for the purposes of population-level research. However, by employing a semi-automated approach to extracting clinical information from free text, the reliance on expert clinical staff can be lessened, thus improving the efficiency and availability of health information.

6 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 clinical information extraction tasks.

Acknowledgements

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

1. Qu Z, Ninan S, Almosa A, Chang K, Kuruvilla S, Nguyen N. Synoptic reporting in tumor pathology – advantages of a web-based system. American Journal of Clinical Pathology. 2007; 127(6): 898-903.

2. Royal College of Pathologists of Australasia. Structured reporting. 2009 Available from: http://www.rcpa.edu.au/Publications/StructuredReporting.htm.

3. College of American Pathologists. An overview of the College of American Pathologists cancer checklists. 2009.

4. International Health Terminology Standards Development Organisation. SNOMED clinical terms user guide. 2008. Available from: http://www.ihtsdo.org.

5. Nguyen AN, Lawley MJ, Hansen DP, Colquist S. A simple pipeline application for identifying and negating SNOMED Clinical Terminology in free text. Health Informatics Conference, Canberra, Australia, 2009. 188-193.

6. Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: A framework and graphical development environment for robust NLP tools and applications. Association for Computational Linguistics (ACL), Philadelphia, USA, 2002.

7. College of American Pathologists. SNOMED CT–encoded cap cancer checklist (v1.5). 2006.

8. Greene F, Page D, Fleming I, Fritz A, Balch C, Haller D, Morrow M, Eds. AJCC cancer staging manual. 6 ed. 2002; Springer: Chicago, IL.

9. Lawley M, Vickers D, Hansen D. Converting ad hoc terminologies to SNOMED CT extensions. Health Informatics Conference, Melbourne, Australia, 2008. 133.

10. Nguyen A, Lawley M, Hansen D, Bowman R, Clarke B, Duhig E, Colquist S. Symbolic Rule-based Classification of Lung Cancer Stages from Free-Text Pathology Reports. Journal of the American Medical Informatics Association, 2010; 17(4): 440-445.

11. Nguyen A, Moore J, Lawley M, Hansen D, Colquist S. Automatic Extraction of Cancer Characteristics from Free-Text Pathology Reports for Cancer Notifications. To be published in Health Informatics Conference, Brisbane, Australia, 2011.

12. Wagholikar A, Nguyen A, Fong M. Patient Identification for Advanced Prostate Cancer Clinical Trials. Australian-Canadian Prostate Cancer Research Alliance Symposium, Gold Coast, Australia, 2010.

Correspondence

Dr. Anthony Nguyen

The Australian e-Health Research Centre
CSIRO ICT Centre
Level 5 – UQ Health Sciences Building 901/16
Royal Brisbane and Women's Hospital
Herston Qld 4029, Australia

Phone: +61 (0)7 3253 3637
http://aehrc.com

Anthony.Nguyen@csiro.au