Automatic Structured Pathology Reporting of Melanomas

Author: Ying Ou (yiou6374@uni.sydney.edu.au)
Supervisor: Jon Patrick
School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

1. BACKGROUND

Generally, clinical decision making in cancer patients is based on information provided by pathology reports. Pathological reporting of resection specimens provides important information for the clinical management of melanoma patients, allowing accurate diagnosis, staging and determination of prognosis.

There are several issues in traditional narrative pathology reports:

- Essential data are occasionally omitted.
- Difficult to identify the necessary elements to justify a given diagnosis.
- Negative results are not always reported.

2. AIMS

The aim of this project is to use the methods of Natural Language Processing to identify and extract pertinent information from free-text melanoma pathology reports to automatically populate structured reports.

- For pathologists, structured reports can be used as a guide to correct missing and ambiguous contents to ensure significantly higher quality reporting.
- For the referring doctors, structured reports can be used to rapidly identify the essential elements in a report to justify a given diagnosis.

3. MATERIALS

A collection of 380 melanoma pathology reports have been scanned and OCRed to form a training set. After de-identification, all reports were annotated. There are 41 types of concepts for medical entity recognition on the melanoma corpus in total.

4. METHODS

Figure 1 demonstrates the system architecture. From the diagram, raw records are passed through the pre-processing engine, which includes a sentence boundary detector and a tokeniser. A record is split into sentences, and then each sentence is split into tokens. In a separate process, the training corpus is annotated manually to create a gold standard. Subsequently errors in the manual annotations are identified by performing validation on the training data with a 100% train and test strategy (a reflexive validation process).

After pre-processing, four CRF feature sets were prepared to train the CRF to identify the entities embedded in the corpus, and the output from the CRF was sent to the annotation file converter to convert the outputs from the CRF to the format that can be processed by a text annotation tool (the Visual Annotator). The structured output generator populates the outputs to conform to the structure templates.

4.1 Medical entity recognition (MER)

For MER experiments on melanoma reports, the best model was obtained from four feature sets below:

(1) Context features: a) Bag of words with window size of nine words; b) Section heading.
(2) Semantic features: a) Medical category; b) Resources.
(3) Lexical features: a) Lemma, part of speech, chunk; b) Lowercase of words; c) Expansions of abbreviations and acronyms; d) Correction of misspelt words.
(4) Other features: a) Ring-fencing

4.2 Structured output generation

To populate the structured templates, a structured output generator was designed with rule-based methods. Figure 2 displays each module in the structured output generator.

(1) Document classification
At first, the documents need to be divided into multiple specimen documents (documents containing more than a specimen) and single specimen documents (documents only contain one specimen).

(2) Context Detection
Based on the section heading detector, we built a post-processing engine to detect the heading context information for each specimen to facilitate the following procedures. This information is also used for populating the values of “Summary”, “Comment”, or “Description” for each section in the template.

(3) Concept Extraction
Concepts are extracted according to their types from the outputs of the Ann converter and then ranked with particular criteria: a potential candidate is assigned a salience measure based on a series of criteria, and the one with the highest salience measure is selected as the best candidate.

(4) XML Generation
The last step is to generate the XML format outputs from the extracted concepts.

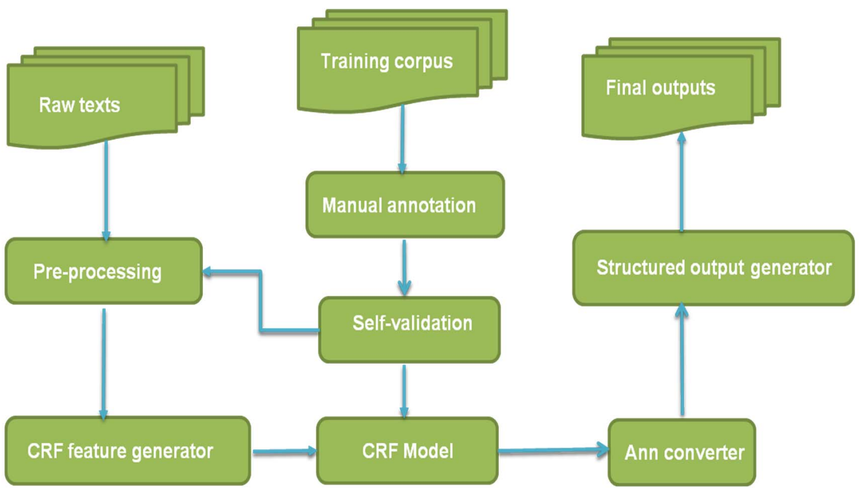

Figure 1. System architecture

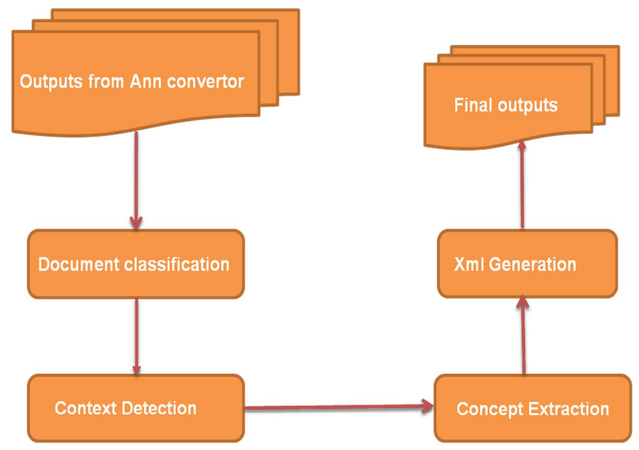

Figure 2. Modules in the structured output generator.

5. RESULTS

The averaged F-scores for medical entity recognition on 100% train and test and 10-fold cross validation were 99.94% and 80.20% respectively. 380 structured reports in xml format were populated by the structured output generator, and prepared for evaluation by pathologists. Figure 3 illustrates an example of the structured output.

Structured Output Report

Section	Title	Description
Diagnostic Summary	Summary	acral lentiginous melanoma, tumour thickness 1.4 mm, Clark level III, excision incomplete at one lateral surgical margin - please consult above report and comments.
	Comment	The slides have been reviewed by Dr X and he agrees with the above findings.
Supporting Information	Clinical Description	See 972284O. Wider excision acral lentiginous malignant melanoma right plantar foot.
	Macroscopic Description	"Excision base of right foot". A roughly oval piece of skin 42 x 32 x 6 mm. Centrally there is an irregular mottled brown lesion 20 x 12 mm. Directly adjacent to this is a linear change 13 mm in length. Towards one end there is a tentative
	Microscopic Description	Sections are of glabrous skin from the foot, which includes the underlying skeletal muscle and subcutaneous tissue. The skin centrally shows scarring and chronic inflammation in the dermis and the subcutaneous tissue from the previous

Items in Supporting Information

Sub-section	Specimen id	Item	Value
CLINICAL	1	Site and laterality	right plantar foot
	1	Specimen type	wider excision
	1	Prev. Rx / Trauma	surgical procedure(previous)
	1	Clinical diagnosis	malignant melanoma
	1	Previous melanoma	N/A
	1	Distant metastasis	N/A
	1	Other medical history	None relevant
MACROSCOPIC	1	Other lesions	N/A
	1	Size of specimen	42mm x 32mm x 6mm
MICROSCOPIC	1	Ulceration(mm diam)	absent
	1	Mitotic rate	N/A
	1	Tumour thickness	1.4mm
	1	Level of invasion (Clark)	III
	1	Lymphovascular invasion	absent
	1	Neurotropism	absent
	1	Microsatellites	absent
	1	Assoc. benign naevus	N/A
	1	Excision margins: In-situ	N/A
	1	Excision margins: Invasive	N/A
	1	Excision margins: Deep	2.2mm
	1	Desmoplasia	N/A
	1	Diagnosis	melanoma
	1	Subtype	acral lentiginous
	1	Cell growth	radial growth phase, single
	1	TILs	absent
	1	TILs: Distribution	N/A
	1	TILs: Density	N/A
	1	Regression	N/A

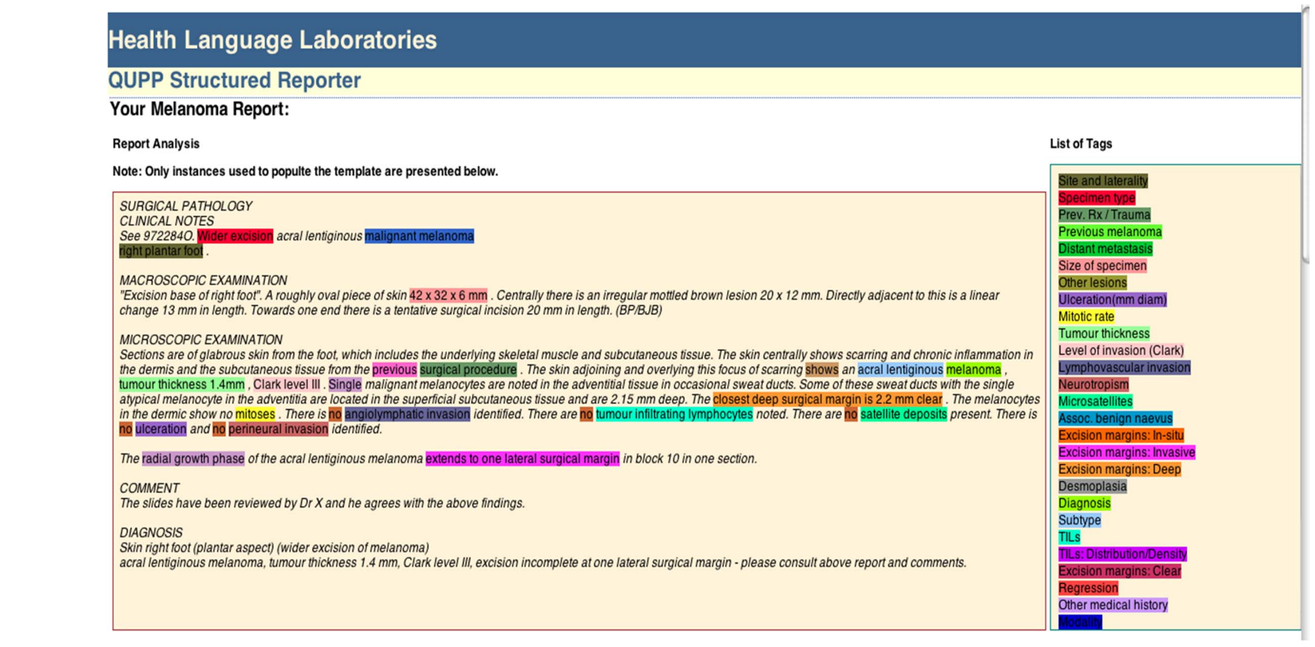

Figure 3. An example of the structured output

6. DISCUSSION AND CONCLUSION

For medical entity recognition on 100% train and test, performances on all entity types are improved by best feature selection and self-validation. Compared to 100% train and test, the F-score for 10-fold cross validation decreased by 19.7%, revealing that there is high variability in language usage among some of the entity types as well as ambiguity in other entity types.

A web page has been built to display the structured outputs for pathologists to evaluate. We also have released the system to the research community as a web page for testing data samples. It is now available at:
http://icims.com.au/QUPPDemo

THIS RESEARCH IS SPONSORED BY
Australian Government
Department of Health and Ageing