Natural Language Processing of Radiology Reports

Andrew Brooks, EPCC (Edinburgh University); Honghan Wu (Edinburgh University); eDRIS (Public Health Scotland)

Introduction

How can we make use of all the unstructured free-text written by doctors and radiologists which accompanies images such as X-Rays, CT scans, etc.?

```
Description: Anterior cervical discectomy and fusions C4-5, C5-6, C6-7 using Bengal cages and Slimlock plate C4 to C7; intraoperative
x-ray. Herniated nucleuses pulposus, C5-6 greater than C6-7, left greater than C4-5 right with left radiculopathy and moderate stenosis
C5-6.
(Medical Transcription Sample Report)
FINAL DIAGNOSES:
1. Herniated nucleuses pulposus, C5-6 greater than C6-7, left greater than C4-5 right with left radiculopathy.
2. Moderate stenosis C5-6.

OPERATION: On 06/25/07, anterior cervical discectomy and fusions C4-5, C5-6, C6-7 using Bengal cages and Slimlock plate C4 to C7;
intraoperative x-ray.
```

The Scottish Medical Images (SMI) Service, run by Public Health Scotland (PHS) and EPCC (University of Edinburgh) has an archive of radiology images taken from all Scottish health boards. These images come with the text of the radiologists' findings. The text can be useful in several ways:

- As a resource for researchers to consult when interpreting the images they are associated with
- As a standalone resource for researchers using their own Natural Language Processing (NLP) tools
- As a way to find images about a specific topic (disease, drug or medical intervention), i.e. a searchable database for building cohorts
- As a way to create new metadata

We describe the system we have developed and deployed inside the Scottish National Safe Haven for de-identifying, cataloguing and releasing to researchers the archive of clinical reports.

1. De-identification

In all use cases, before being allowed to use or release the text, all documents must be de-identified, so they must be examined and have Personally Identifiable Information (PII) removed. This includes names, addresses, postcodes, dates of birth, telephone numbers, GMC registration numbers, and so on.

The first problem to solve was the file format. Reports are stored in DICOM which is designed for storing images. The text is stored as an item of metadata in a file which has no image pixels. The metadata includes specific fields for Patient Name, Date of Birth, and others which obviously contain PII, plus a set of fields relating to the study and associated images, which may or may not contain PII.

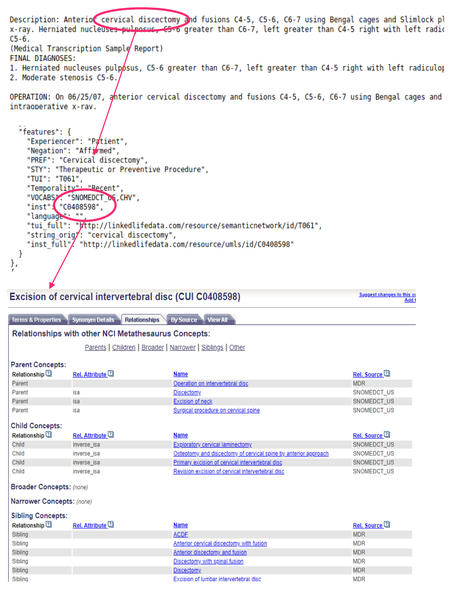
A tool called the Clinical Trial Processor (CTP) is used to remove the fields containing PII, which will include removal of the actual clinical report text. In parallel our own de-identification tool parses the DICOM file to extract the text, de-identifies it, and places this text into the output from CTP.

The second problem to solve was the text structure. The reports use a sub-type of DICOM called Structured Reports, but the text itself is not structured into useful sections (about the patient, the condition, the treatment, the outcome) so it needs to be decoded before it can be used.

After the text has been extracted it is de-identified using a mixture of techniques including dictionaries, rules-based and contextual information. NLP may be used but care must be taken not to remove names of body parts or diseases such as Monro, Parkinson, etc.

2. Annotation

The free-text reports will contain mentions of body parts, diseases, conditions, drugs, treatments, and so on. It is essential to identify these concepts so that we can search for them and also use them for medical analytics. We use NLP to parse the de-identified text and extract concepts from phrases, including context such as who experienced it, whether it is in the past or ongoing, and whether it is negated (e.g. "scan shows patient does *not* have lesions"). The NLP handles concepts which can be described using different words or phrases, and gives them a unique identifier (CUI or inst in this diagram):

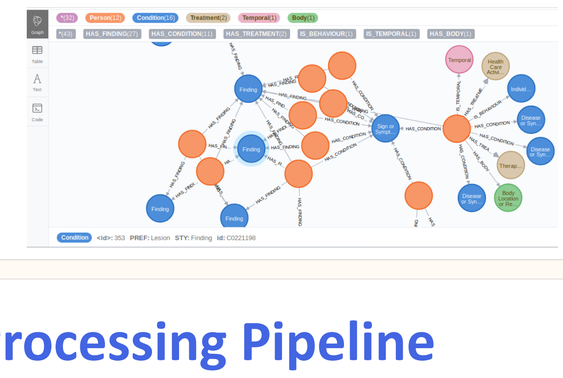


The phrase has been recognised as concept C0408598; the document is now annotated with this "feature" data structure. The concept comes from a hierarchical "metathesaurus" ontology shown here:

The database of concepts has been built using the JSON features of PostgreSQL, with indexes on the array of concepts in each document and on the free text itself. An example shown below highlights all of the concepts found in the above text fragment, showing the word with all of the concept attributes:

Description: Anterior **DISEASE OR SYNDROME (PATIENT, AFFIRMED, RECENT)** cervical discectomy **THERAPEUTIC OR PREVENTIVE PROCEDURE (PATIENT, AFFIRMED, RECENT)** and fusions **THERAPEUTIC OR PREVENTIVE PROCEDURE (PATIENT, AFFIRMED, RECENT)** C4-5 **CARDINAL** C4 **BODY LOCATION OR REGION (PATIENT, AFFIRMED, RECENT)** -5, C5-6 **PRODUCT** C5 **BODY LOCATION OR REGION (PATIENT, AFFIRMED, RECENT)** 5-6 **FINDING (PATIENT, AFFIRMED, RECENT)** , C6-7 **PRODUCT** C6 **BODY LOCATION OR REGION (PATIENT, AFFIRMED, RECENT)** -7 using Bengal **ORG** cages and Slimlock **PRODUCT** plate C4 **BODY LOCATION OR REGION (PATIENT, AFFIRMED, RECENT)** to C7 **PRODUCT** C7 **BODY LOCATION OR REGION (PATIENT, AFFIRMED, RECENT)** ; intraoperative **TEMPORAL CONCEPT (PATIENT, AFFIRMED, RECENT)** x-ray **DIAGNOSTIC PROCEDURE (PATIENT, AFFIRMED, RECENT)** . Herniated nucleuses pulposus, C5-6 **PRODUCT** C5 **BODY LOCATION OR REGION (PATIENT, AFFIRMED, RECENT)** 5-6 **FINDING (PATIENT, AFFIRMED, RECENT)** greater than C6-7 **PRODUCT** C6 **BODY LOCATION OR REGION (PATIENT, AFFIRMED, RECENT)** -7, left greater than C4-5 **CARDINAL** C4 **BODY LOCATION OR REGION (PATIENT, AFFIRMED, RECENT)** -5 right with left radiculopathy **DISEASE OR SYNDROME (PATIENT, AFFIRMED, RECENT)** and moderate stenosis **PATHOLOGIC FUNCTION (PATIENT, AFFIRMED, RECENT)** C5-6 **PRODUCT** C5 **BODY LOCATION OR REGION (PATIENT, AFFIRMED, RECENT)** 5-6 **FINDING (PATIENT, AFFIRMED, RECENT)** .

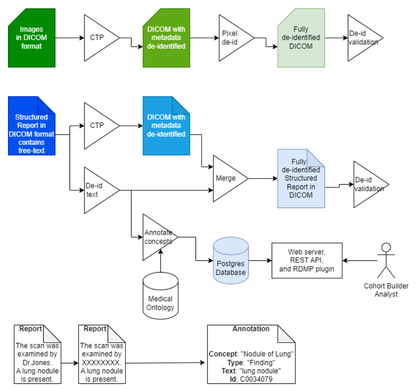
The concepts could be loaded into a graph database for further analysis, as in the Neo4J example:



3. Processing Pipeline

The top row (green) shows image de-identification. In blue is the free-text de-identification and annotation of medical concepts into a database.

In both cases a separate tool is used to validate the resulting DICOM file to ensure that PII does not appear in the text or in any of the metadata fields.

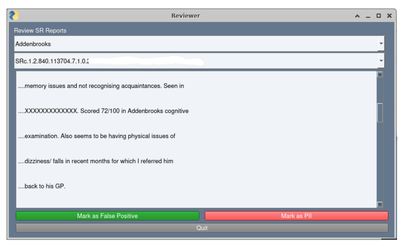


4. Validation

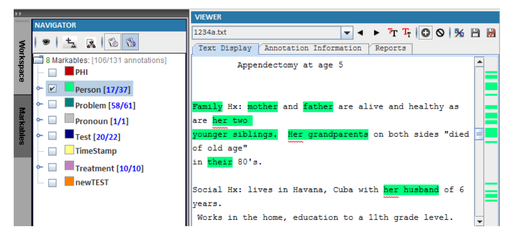
Validation has been performed in both aspects of the project:

- To ensure that the de-identification removes as much PII as possible, and does not remove non-PII, thus reducing the risk of releasing documents to researchers that could identify any person
- To ensure that the medical ontology classifies concepts in a way which is useful for searching through the archive and building cohorts

Two tools have been used for this. One is a very specific tool for checking reports that gives a very easy Yes/No facility to classify words as PII:



The other tool is a customised version of eHost which allows multiple annotators to work on a set of reports, checking for PII, marking elements which were missed or were incorrectly flagged. It has also been used on the concept annotations, for example to classify body parts.



The system has the ability to learn from researcher corrections to the annotations, to make future search queries return more appropriate results for the study.

Conclusions

We have developed a robust method for de-identifying the free-text clinical reports which accompany radiology images. It achieves accuracy high enough to be used by Public Health Scotland for supplying reports to researchers.

We have developed methods and deployed a pipeline for using NLP to extract meaningful information from free-text reports which can subsequently be used for search queries, improved cohort creation and for research analytics.

We have developed and deployed GUI tools streamlining the validation workflow and the annotation workflow.

Acknowledgements

University of Edinburgh, University of Dundee,

Public Health Scotland (eDRIS),

PICTURES project - This work was supported by the Medical Research Council (MRC) grant No. MR/M501633/1 and the Wellcome Trust grant No. WT086113 through the Scottish Health Informatics Programme (SHIP). This project has also been supported by MRC and EPSRC (grant No. MR/S010351/1) and by the Scottish Government through the "Imaging AI" grant award.

CogStack-SemEHR; GATE; BioYodie; UMLS.