5.1 Natural Language Processing

While there is a strong trend towards the direct coding of clinical data, the capture and retention of free text remains essential to record broader narratives about clinical history, physical examinations, clinical procedures and investigation results. Wider deployment of medical transcription technologies featuring speech recognition also add to the volume of free text in electronic format. Medical literature, clinical guidelines and published clinical research also remains largely in free text.

Natural Language Processing (NLP) is a linguistic technique that enables a computer program to analyze and extract meaning from human language. Clinical NLP, using SNOMED CT's concepts, descriptions and relationships, may be applied to repositories of clinical information to search, index, selectively retrieve and analyze free text. These techniques can be used to extract SNOMED CT encoded data from free-text patient records, and also support the retrieval of clinical knowledge documents.

It should be noted that while clinical NLP techniques have increased in sophistication over recent years, it is not possible to guarantee full accuracy or completeness using a computer-based algorithm. Spelling errors, grammatical errors, abbreviations, unexpected synonyms, unusual vernacular (i.e. local) phrases, and hidden contextual information continue to provide challenges that human intelligence is uniquely equipped to handle.

Example

The example shown below in Figure 5.1-1 shows a free text section of a discharge summary that has been processed with clinical NLP to extract a set of potential SNOMED CT clinical findings and procedures. In order to ensure the correctness of this automatic encoding, the application should present this list of extracted codes to the user for confirmation, giving them the opportunity to refine, delete or append codes.

The patient is a frail 88-year-old caucasian male was admitted to our hospital for complaints of nausea and vomiting and suspected urinary tract infection.

He has a past medical history of hypertension, atrial fibrillation and chronic right hip pain after total hip replacement in 2012.

The patient was started on antibiotics. Urine culture confirmed an E. coli urinary tract infection sensitive to trimethoprim.

During admission an episode of possible coffee ground vomiting coupled with his non-steroidal inflammatory drug use prompted an upper GI endoscopy at which no abnormality was detected. Fecal occult blood was negative.

The patient was also provided with physiotherapy and fully remobilised.

Clinical Findings

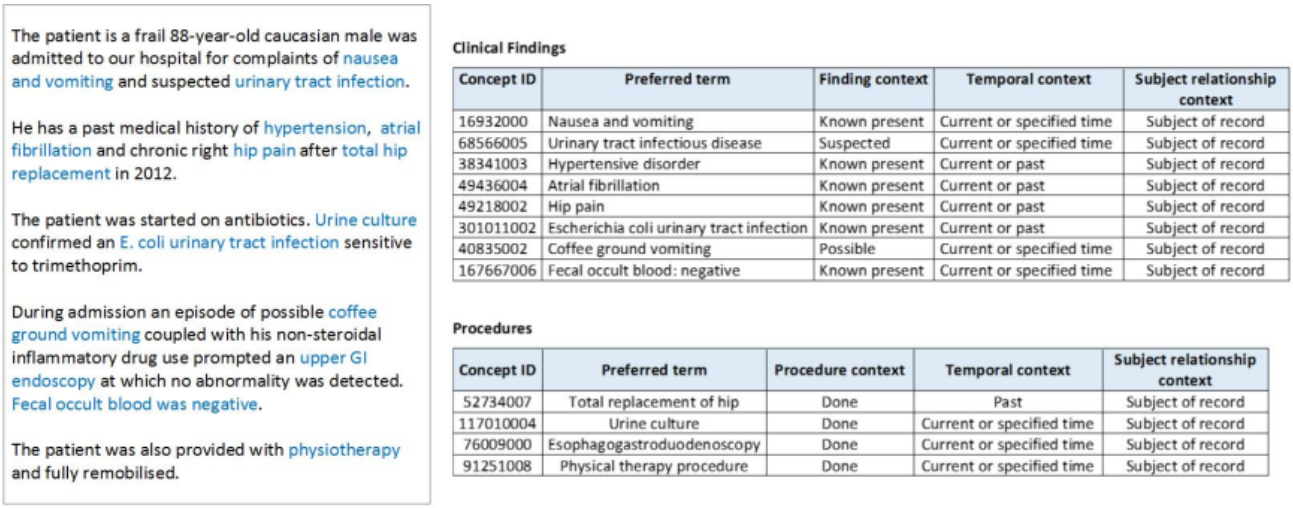
Concept ID	Preferred term
16932000	Nausea and vomiting
68566005	Urinary tract infectious disease
38341003	Hypertensive disorder
49436004	Atrial fibrillation
49218002	Hip pain
301011002	Escherichia coli urinary tract infection
40835002	Coffee ground vomiting
167667006	Fecal occult blood: negative

Procedures

Concept ID	Preferred term
52734007	Total replacement of hip
117010004	Urine culture
76009000	Esophagogastroduodenoscopy
91251008	Physical therapy procedure

Figure 5.1-1: Natural Language Processing encoding SNOMED CT

To improve the accuracy of clinical NLP and the value for analytics processes, it is important that the context of each statement expressed in natural language is clearly identified – for example, past history, suspected and negation/absence. Figure 5.1-2 shows the same discharge summary narrative as in Figure 5.1-1, but this time processed with clinical NLP that also extracts the explicit context of each clinical finding and procedure.



The patient is a frail 88-year-old caucasian male was admitted to our hospital for complaints of nausea and vomiting and suspected urinary tract infection.

He has a past medical history of hypertension, atrial fibrillation and chronic right hip pain after total hip replacement in 2012.

The patient was started on antibiotics. Urine culture confirmed an E. coli urinary tract infection sensitive to trimethoprim.

During admission an episode of possible coffee ground vomiting coupled with his non-steroidal inflammatory drug use prompted an upper GI endoscopy at which no abnormality was detected. Fecal occult blood was negative.

The patient was also provided with physiotherapy and fully remobilised.

Clinical Findings

Concept ID	Preferred term	Finding context	Temporal context	Subject relationship context
16932000	Nausea and vomiting	Known present	Current or specified time	Subject of record
68566005	Urinary tract infectious disease	Suspected	Current or specified time	Subject of record
38341003	Hypertensive disorder	Known present	Current or past	Subject of record
49436004	Atrial fibrillation	Known present	Current or past	Subject of record
49218002	Hip pain	Known present	Current or past	Subject of record
301011002	Escherichia coli urinary tract infection	Known present	Current or past	Subject of record
40835002	Coffee ground vomiting	Possible	Current or specified time	Subject of record
167667006	Fecal occult blood: negative	Known present	Current or specified time	Subject of record

Procedures

Concept ID	Preferred term	Procedure context	Temporal context	Subject relationship context
52734007	Total replacement of hip	Done	Past	Subject of record
117010004	Urine culture	Done	Current or specified time	Subject of record
76009000	Esophagogastroduodenoscopy	Done	Current or specified time	Subject of record
91251008	Physical therapy procedure	Done	Current or specified time	Subject of record

Figure 5.1-2: Natural Language Processing encoding SNOMED CT with context

When SNOMED CT codes with explicit context are extracted from free text narrative, the resulting clinical meanings may be captured using SNOMED CT postcoordinated expressions. For example, the following clinical statement:

Endoscopy revealed an acute gastric ulcer but no evidence of gastric bleeding or perforation of the stomach.

can be encoded using the following SNOMED CT expressions with explicit context (see Clinithink case study):

- 243796009 |situation with explicit context| :

  {408731000 |temporal context| = 410512000 |current or specified time|,

  246090004|associated finding| = 95529005 |acute gastric ulcer|,

  408732007 |subject relationship context| = 410604004 |subject of record|,

  408729009 |finding context| = 410515003 |known present|

- 243796009 |situation with explicit context| :

  {408729009 |finding context| = 410516002 |known absent|,

  246090004 |associated finding| = 61401005 |gastric bleeding|,

  408731000 |temporal context| = 410512000 |current or specified|,

  408732007 |subject relationship context| = 410604004 |subject of record|}

- 243796009 |situation with explicit context| :

  {408729009 |finding context| = 410516002 |known absent|,

  246090004 |associated finding| = 235674005 |perforation of stomach|,

  408731000 |temporal context| = 410512000 |current or specified|,

  408732007 |subject relationship context| = 410604004 |subject of record|}

Implementation

NLP Techniques using SNOMED CT

A clinical NLP engine can use SNOMED CT to encode free text narrative in patient records in a number of ways. Firstly, it can use SNOMED CT descriptions together with techniques such as:

- Stemming: The process of reducing a word to its stem, base or root form – for example "cardiology", "cardiac" and "cardiologist" may be reduced to the stem "cardi".
- Reordering: The process of reordering the words in a phrase – for example, reordering "hip fracture" to "fracture hip".
- Word substitution: The process of substituting a word or word phrase with an equivalent word or word phrase. The SNOMED CT Lexical Resources zip file, available from the SNOMED CT Document Library, includes an English Word Equivalents table that groups together equivalent words and phrases – for example, "Renal stone", "Kidney stone", "kidney calculus", "renal calculus" and "nephrolith" are grouped into the same word block group. This table can be modified or extended with additional word equivalent groups if required.
- Stop word removal: The process of removing words with limited semantic specificity – for example 'a', 'an', 'and', 'as', 'at', 'be', 'by', 'for', 'of', 'the'. The SNOMED CT Lexical Resources zip file , available from the SNOMED CT Document Library, includes an Excluded Words table, which suggests some common English stop words that may be used with SNOMED CT.

The SNOMED CT concept model can also be used to identify potential connections between related concepts – for example, the words "left", "hip" and "fracture" used in close proximity may indicate a |fracture| with finding site |hip| and a laterality of |left|. Similarly, the SNOMED CT concept model may help to identify context that is expressed within the text – for example, past history, certainty and absence.

Another commonly adopted NLP strategy is to use the location of the free text within the structure of a document to restrict the possible SNOMED CT code matches. For example, free text entered into a 'Diagnosis' field may restrict its SNOMED CT encoding to the |disorder| hierarchy, together with other concepts that may be linked to |clinical findings| via the SNOMED CT concept model.

When NLP techniques are applied to non-English (or dialect-specific) text, translations of relevant SNOMED CT descriptions may be required. The NLP methods themselves may also need to be adapted to reflect the structure and style of the language in which the text is written.

Indexing

Another major application for Natural Language Processing technologies is indexing collections of free text transcripts or documents such that topic specific searches may be run on them, or relevant clinical knowledge sources may be identified and linked to a given patient's clinical data. The challenge is to return ranked matches which permit selection of texts with high sensitivity and high specificity (i.e. that relevant documents are rarely overlooked and that irrelevant documents are rarely returned).

SNOMED CT can be used to support these applications by enabling more powerful searching of free text data stores than using a purely lexical keyword-based approach. For example, the clinician may request "all documents which refer to cardiac rhythm disorders". Rather than relying purely on text matching, the search term may be matched with the concept 698247007 |cardiac arrhythmia (disorder)|, based on its synonym |disorder of heart rhythm|. The descendants of this concept (e.g. 276796006 |atrial tachycardia|, 49260003 |idioventricular rhythm|, 233917008 |atrioventricular block|) may then be used to search for any code which is a kind of cardiac arrhythmia. Non-|is a| attribute relationships may also be used in the retrieval process to find associations between the search term and the indexed concepts, and to calculate the relevance of each free text artefact to determine the order in which they should be presented to the user.

Case Studies

Clinical NLP has been implemented for encoding free text narrative in health records by a number of vendors, including Caradigm, Cerner, Clinithink a nd Intelligent Medical Objects).

NLP techniques for indexing and searching have also been implemented by Cerner and Dr Bevan Koopman. Allscript's Sunrise InfoButton™ feature uses encoded patient problem lists and medication data elements, together with SNOMED CT-based indexes provided by third-party medical content providers, to present on-topic information to the clinician without manual searching.