HIC 2009
Vitali Sintchenko and Peter Croll (Eds)
HISA Ltd, 2009

A Simple Pipeline Application for Identifying and Negating SNOMED Clinical Terminology in Free Text

Anthony N. Nguyen[1], Michael J. Lawley[1], David P. Hansen[1], and 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:* A simple pipeline application is developed to identify and negate relevant SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) concepts in free text medical documents.

Background: SNOMED CT is a large formal ontology of clinical terms that has been identified as the standard set of clinical terms to be used in systems within Australian healthcare. Standardising free-text medical documents using a consistent set of clinical terms will assist in the enhanced aggregation of clinical information for retrieval and analysis.

Method: The system is built using the General Architecture for Text Engineering (GATE) framework and incorporates the MetaMap Transfer (MMTx) application (to identify concepts in the free text which are restricted to the SNOMED CT source) and a negation detection algorithm called NegEx. The SNOMED CT semantics is used to identify concepts related to findings and diseases, and these were used for negation detection. The pipeline application has been tailored to histopathology reports, but has been applied to other medical free text such as radiology, colonoscopy and other pathology reports from lung cancer and inflammatory bowel disease (IBD) patients.

Results: The results from the pipeline application show promise with a high coverage of SNOMED CT concepts and relevant concepts negated as appropriate. The concepts along with their properties can be viewed using GATE's document and annotation viewer. The annotations have been generated so that concepts can also be filtered according to their semantic type (i.e. top-level concept) such as 'Clinical Finding' or 'Body Structure'.

Conclusion: The pipeline application allows the generation of consistent terminology for the aggregation of clinical information for retrieval and analysis of healthcare systems and processes. By generating the SNOMED CT concepts, subsumption relationships in processing the concepts can be taken advantage of to retrieve concepts which relate to the same disease, anatomy or finding, or infer if two descriptions relate to the same concept. Future research can take advantage of the subsumption relationships to generate business rules for extracting key clinical information from medical free text.

Introduction

Medical free texts are non-standardised and make it difficult to aggregate clinical information. This limits the utility of these documents in a broad range of clinical applications. Standardising the text into a reference set of clinical terms can overcome this problem and increase interoperability among different clinical applications, clinical domains, and institutions (Friedman *et al*., 2004).

A number of natural language processing algorithms have been proposed to map medical free text to concepts within a standardised coding system such as the Unified Medical Language System (UMLS) (U.S. National Library of Medicine, 2008). The widely used MetaMap (MMTx) algorithm (Aronson, 2001), developed by the National Library of Medicine (NLM), has been used to map biomedical text to the closest concepts from the UMLS. MetaMap, however, requires an independent negation detection algorithm

Back to Contents

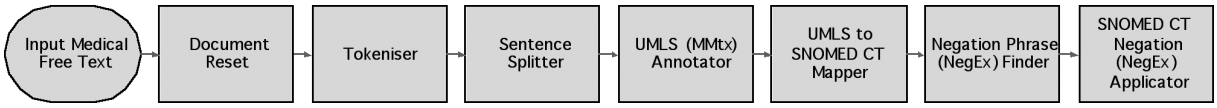

Figure 1: Pipeline application for the annotation of SNOMED CT concepts from free text

such as NegEx (Chapman *et al*., 2001, Chapman, 2009) to detect findings and diseases in free text which are described as absent.

Although much attention has focused on the free text mapping to UMLS, the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) (College of American Pathologist, 2007) has received little attention in this area. SNOMED CT is a large formal ontology of clinical terms that has been identified by the Australian National e-Health Transition Authority (NeHTA) as the standard ontology to be used in systems within Australian healthcare (National E-Health Transition Authority, 2006). The comprehensiveness and inherent structure of SNOMED CT makes it suitable as a reference clinical terminology source. One algorithm (Patrick *et al*., 2007) developed specifically to map free text from clinical notes to SNOMED CT as well as identifying negation terms has proven to perform within acceptable time and accuracy limits.

In this paper, we propose another methodology for the identification of SNOMED CT concepts in medical free text. Since the UMLS encompasses the SNOMED CT ontology, we use the MetaMap algorithm to identify and restrict UMLS concepts to the SNOMED CT source and subsequently apply the NegEx algorithm for the detection of negated concepts. We use the SNOMED CT semantics to identify the subsets of concepts which are related to findings and diseases and use them as terms to be considered for negation. Such use of semantics makes SNOMED CT a powerful tool for understanding clinical information in medical free text. We use General Architecture for Text Engineering (GATE) (Cunningham *et al*., 2002), an open source architecture for natural language processing (NLP), as the development platform for the pipeline application.

Method

The pipeline application, called MEDTEX (Medical Text Extraction), for identifying and negating SNOMED CT concepts was developed using GATE. GATE is a framework and graphical development environment for the development and deployment of language engineering components and resources for various natural language applications such as information extraction (Cunningham *et al*., 2002). It also allows the creation, annotation and evaluation of corpora on the applications generated.

The proposed pipeline application uses seven GATE modules (components) (see Figure 1):

1. *Document Reset:* Restores the document to its original state by deleting existing annotations.
2. *Tokeniser:* Splits the document into tokens (e.g. words, numbers, punctuations and spaces). This module uses regular expressions that define both token delimiters (punctuation and spaces) and special cases where punctuations are a part of certain tokens (e.g. decimal point in floating point numbers and periods in acronyms). Other regular expressions have been created to define length measurements and units (e.g. '1mm', '1 x 3 mm', '1mm x 2 mm x 3mm'), TNM (tumour-node-metastasis) cancer stages (e.g. T0N1MX, T2 N0 MX), legacy SNOMED IDs which define the topology, morphology and procedure in coded form (e.g. T-28000, M-80703, P1-03000), and tokens for de-identified information (e.g. names, facilities, and dates). Although the additional token definitions are typically found in histopathology reports, the Tokensier module is robust to any free text document type and flexible in terms of the specification of additional token definitions, if required.
3. *Sentence Splitter:* Segments the text into sentences using a set of regular expression-based rules that define sentence breaks (e.g. full stop, carriage return). Additional rules to parse question marks which are not sentence breaks (e.g. "Left upper lobe: ? Ca. ? mets. ? nodal involvement") was also incorporated to avoid such occurrences of question marks being annotated as sentence breaks.
4. *UMLS (MMTx) Annotator:* Maps the strings in the free text to the closest concepts in the UMLS medical terminology resource. The UMLS MMTx Annotator for GATE (Roberts, 2008) was used to generate the mappings. The MMTx Annotator provides a wrapper so that MMTx can be used in GATE. MMTx processes phrases in the document to find the best possible mappings. In the event that multiple equally scoring mappings were found, then multiple annotations were generated, one for each of the best mappings. UMLS concepts were restricted to the SNOMED CT source using the MMTx option –R=SNOMEDCT (or --restrict_to_sources=SNOMEDCT).

Back to Contents

5. *UMLS to SNOMED CT Mapper:* Post-processing of the MMTx annotations were performed to remove UMLS concepts that overlap with a dimension, TNM stage, or SNOMED ID token. This was found to improve the UMLS mappings and avoid instances such as "mm" from a dimension token being incorrectly mapped to '*Muscle – MM (C0026845)*' and "T' from a SNOMED ID token being mapped to '*Therapeutic procedure – TX (C0087111)*'.

   The remaining UMLS concepts were then mapped to SNOMED CT using the UMLS Metathesaurus "MRCONSO.RRF" database file, which details the entire UMLS concept structure (U.S. National Library of Medicine, 2008). Since both active (status = 'current' or 'limited') and inactive (e.g. status = 'ambiguous,' 'duplicate,' or 'retired') concepts were mapped, concepts were filtered so that only active concepts were retained for further processing. SNOMED ID tokens annotated by the *Tokeniser* were also mapped to a SNOMED CT concept. Each concept was also examined to determine whether or not the concept is subsumed by 373572006|clinical finding absent| concept; if so, the pre-coordinated concept would be tagged as a negatively asserted concept and be annotated along with the concepts negated by the SNOMED CT Negation (*NegEx*) Applicator module.

   For reference, SNOMED CT properties (e.g. concept ID, fully specified name, and status) were included as part of the concept's annotation. In addition, a new annotation set was created to filter the concepts by their top-level (semantic type) concepts such as 'Clinical Finding,' or 'Body Structure.'

6. *Negation Phrase (NegEx) Finder:* Finds common medical negation phrases (e.g. "no evidence of") in the free text. The negation phrases listed in Chapman (2009) were extended to include those also commonly found in histopathology reports such as "was not present."

7. *SNOMED CT Negation (NegEx) Applicator:* Tags negated concepts by associating negation phrases with neighbouring SNOMED CT findings or disease concepts using the NegEx algorithm. The list of findings and disease concepts considered for negation was derived from the fourteen semantic types (e.g. 'Finding,' 'Disease or Syndrome,' 'Sign or Symptom') related to findings and diseases from the UMLS (Chapman, 2009). These semantic types (and the example concepts given for each) were mapped to the SNOMED CT ontology to derive SNOMED CT concepts which would encompass (or subsume) the fourteen UMLS semantic types. As a result, any concept that is a sub-type (descendant) of any of the following four SNOMED CT concepts were considered for negation:

   - 404684003|clinical finding| (e.g. "decreased capillary fragility," "diabetes mellitus," "dyspnea.")
   - 49755003|morphologically abnormal structure| (e.g. "wallerian degeneration," "carcinoma")
   - 363787002|observable entity| (e.g. "status of invasion by tumour")
   - 272379006|event| (e.g. "exposure to toxin," "accident caused by bench saw")

It is conjectured that the four SNOMED CT concepts selected above encompasses a large proportion of findings and diseases. The terms considered for negation are robust and thus can be modified or updated as more knowledge regarding findings and diseases within SNOMED CT are found.

Results and Discussion

Four datasets containing 1204 histopathology and 6149 radiology reports from lung cancer patients, and 138 colonoscopy and 8 pathology reports from inflammatory bowel disease (IBD) patients were used as test corpora for the proposed pipeline application. Figure 2 shows a screenshot of the pipeline application (MEDTEX) and its output for a sample de-identified histopathology report (HIST000003_REP02). The checked boxes on the right are the annotations highlighted in the 'Document Editor' with SNOMED CT concepts highlighted in yellow, negation phrases in red, and the negated concepts in purple. Also highlighted are dimensions, SNOMED IDs, TNM stages, and de-identified tokens as identified by the Tokeniser module.

Although formal evaluations of the SNOMED CT mappings will be performed in the near future, it was observed that the results from the pipeline application were satisfactory with a high coverage of the free text terms across the corpora. To improve the coverage of the terms, concepts for custom abbreviations (e.g. LLL – Lower lobe of left lung), BAC – Bronchoalveolar carcinoma, and RUL – Upper lobe of right lung) could be identified, if existent; otherwise SNOMED CT reference sets and extensions could be defined for the terms (Lawley *et al*., 2008). Other unmapped free text terms either do not have a clinical concept, excluded because the concept was not part of the best mappings when processed by the MMTx, or are spelling mistakes (which could be resolved by using n-gram statistics to find likely candidates for the correct spelling of misspelled words). Future work will endeavour to

Back to Contents

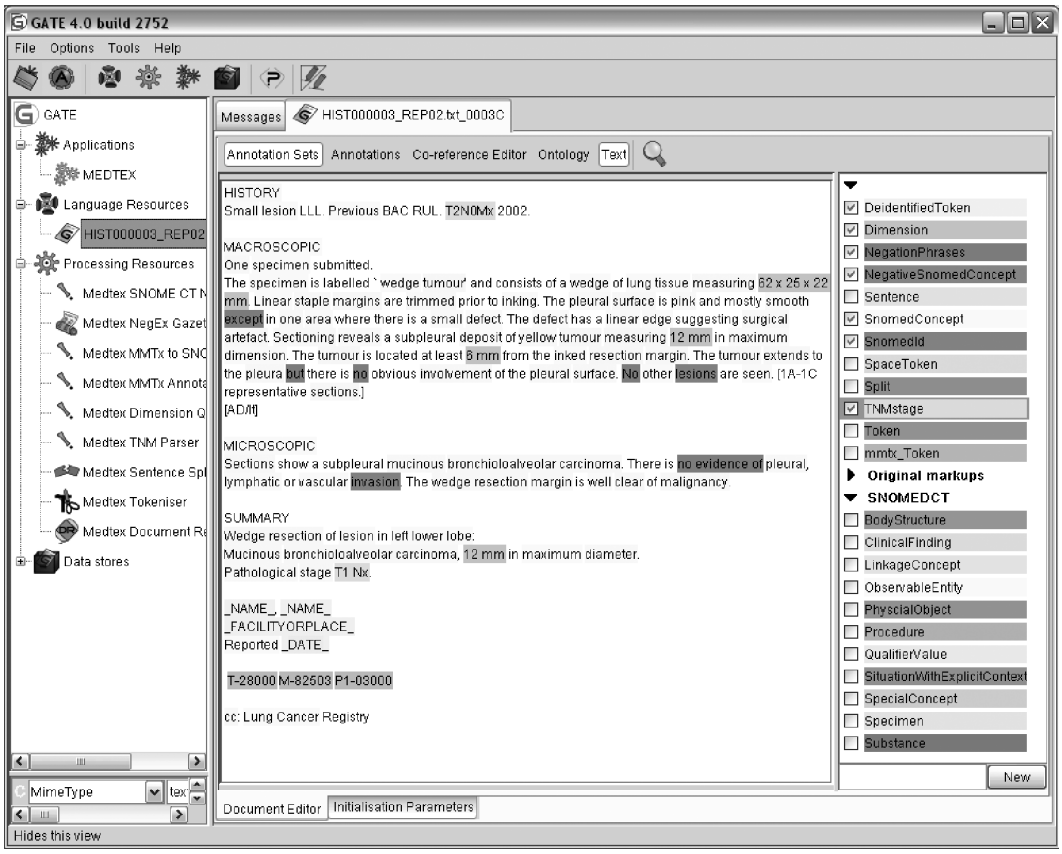

Figure 2: Screenshot of pipeline application output marked up within GATE's annotation view.

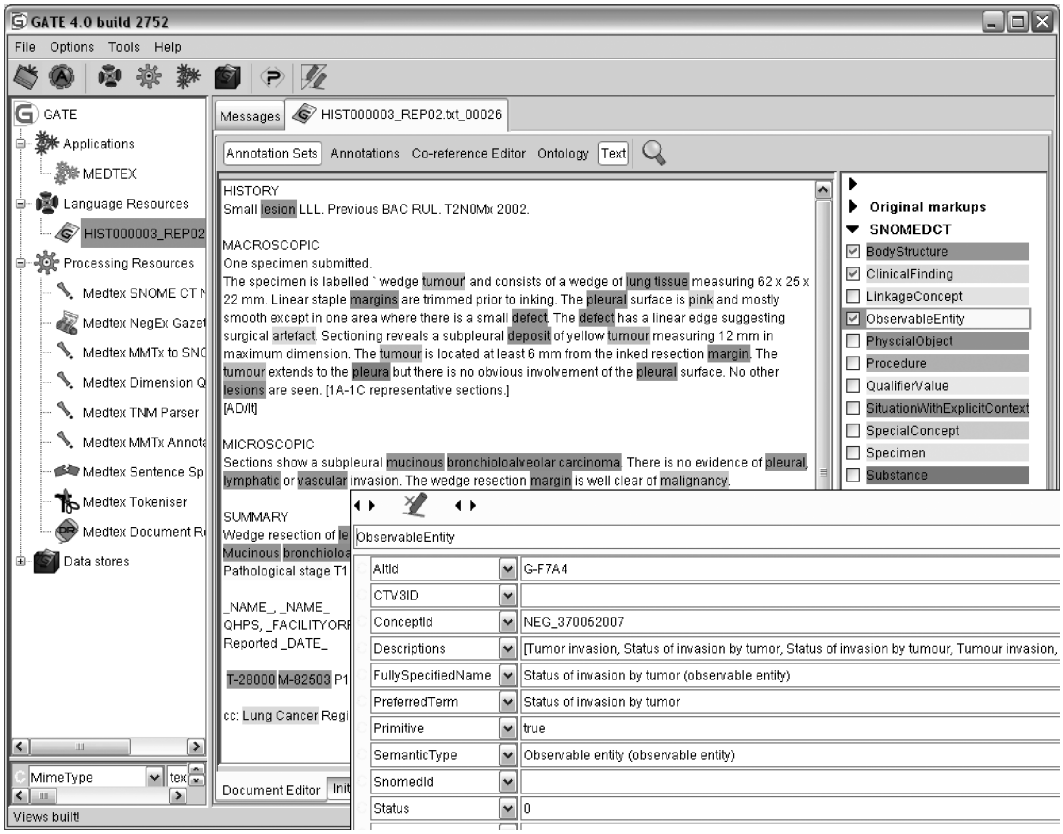

Figure 3: Screenshot of SNOMED CT concepts filtered by the 'Body Structure,' 'Clinical Finding,' and 'Observable Entity' semantic types. Also shown are properties for the negated SNOMED CT concept 370052007|status of invasion by tumour|.

Back to Contents

design a formal framework for the evaluation of the SNOMED CT mappings for the test corpora.

The negation algorithm also performed satisfactorily with the relevant findings and disease concepts negated as required. Note that the negation phrase annotations in Figure 2 also highlights conjunctions (e.g. "except" and "but"), but these terms were used to limit the scope of the search for relevant concepts for negation, as described in the NegEx algorithm.

To view the SNOMED CT properties, a mouse cursor over the desired concept annotation will pop-up a tooltip detailing the properties for the given concept. The "SNOMEDCT" annotation set shown on the right side of Figure 2 can also be used to filter concepts by their semantic type (or top-level concept). Figure 3 shows an example where all 'Body Structure,' 'Clinical Finding,' and 'Observable Entity' related concepts are highlighted. The figure also shows the properties for an example negated concept 370052007|status of invasion by tumour|.

By generating the SNOMED CT concepts, subsumption relationships (such as the one used to determine which concepts were candidates for negation) can be taken advantage of to retrieve concepts which relate to the same disease, anatomy or finding, or infer if two descriptions relate to the same concept. It is conjectured that subsumption relationships can be used to generate business rules for extracting key clinical information from medical free text. For example, by exploiting the substructure in histopathology reports and identifying procedure and topology SNOMED IDs, lung resection related reports can be identified by testing if a post-coordinated expression based on the following template is equivalent to or a subtype of 119746007|lung excision|:

<procedure> :

{260686004|method| = <procedure.method>

,405813007|procedure site - Direct| = <topology> }

where <procedure> and <topology> are the procedure and topology concept IDs, respectively, and <procedure.method> is the concept ID value of the procedure concept's 260686004|method| attribute.

Conclusion

The pipeline application for identifying and negating SNOMED CT concepts from medical free text was developed using GATE. The results from the pipeline application show promise with high coverage of SNOMED CT concepts and relevant concepts negated as appropriate. The pipeline application allows the generation of consistent terminology for the aggregation of clinical information for retrieval and analysis of healthcare systems and processes.

Acknowledgment

This research is a part of the Cancer Stage Interpretation System (CSIS) 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 from lung cancer patients, Ken Manthey from the PACS Support Group at the Princess Alexandra Hospital for his work in collecting the report data from radiology information systems, and Katherine Hanigan from the Queensland Institute of Medical Research (QIMR) for providing the pathology and colonoscopy reports from inflammatory bowel disease patients.

References

Aronson, A.R. (2001), 'Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program,' *American Medical Informatics Association (AMIA) Annual Symposium*, pp. 17–21.

Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., and Buchanan, B.G. (2001), 'A simple algorithm for identifying negated findings and diseases in discharge summaries,' *Journal of Biomedical Informatics*, vol. 34, pp. 301–310.

Chapman, W.W. (2009), 'NegEx version 2: A simple algorithm for identifying pertinent negatives in textual medical records.' Available from: http://www.dbmi.pitt.edu/chapman/NegEx.html; accessed Jan 28, 2009.

College of American Pathologist (2007), 'SNOMED CT Clinical Terms Users Guide,' January 2007

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 40th Anniversary Meeting of the Association for Computational Linguistics (ACL).* Philadelphia, July.

Friedman, C., Shagina, L., Lussier, Y., Hripcsak, G. (2004), 'Automated Encoding of Clinical Documents Based on Natural Language Processing,' *Journal of the American Medical Informatics Association*, 11(5), pp. 392-402, Sep/Oct.

Lawley, M., Vickers, D., Hansen, D. (2008), 'Converting Ad Hoc Terminologies to SNOMED CT Extensions,' *Health Informatics Conference*, pg. 133.

National E-Health Transition Authority (2006), 'Fact sheet: A national approach to clinical terminologies', August; Available from http://www.nehta.gov.au/.

Patrick, J., Wang, Y., Budd, P. (2007), 'An Automated System for Conversion of Clinical Notes into SNOMED Clinical Terminology,' *Australasian Workshop on Health Knowledge management and Discovery (HKMD)*, pp. 219-226.

Back to Contents

Roberts, A. (2008), 'Angus Roberts – GATE Software.' Available from http://www.dcs.shef.ac.uk/~angus/gate-software.html; accessed Jan 22, 2009.

U.S. National Library of Medicine (2008), 'Unified Medical Language System (UMLS).' Available from: http://www.nlm.nih.gov/research/umls/.

Correspondence
Anthony Nguyen, PhD
The Australian e-Health Research Centre
Level 7 UQ CCR building 71/918
Royal Brisbane and Women's Hospital
Herston QLD 4029, Australia
E-mail: anthony.nguyen@csiro.au

Back to Contents