

NLP-based Identification of Pneumonia Cases from Free-Text Radiological Reports

Authors: Peter L. Elkin, MD¹, David Froehling, MD¹, Dietlind Wahner-Roedler, MD¹, Brett Trusko, PhD¹, Gail Welsh, MD¹, Haobo Ma, MD², Armen X. Asatryan, MD², Jerome I. Tokars, MD², S. Trent Rosenbloom, MD³, MPH Steven H. Brown, MD^{3,4}

¹Mayo Clinic, Rochester, MN

²Centers for Disease Control and Prevention

³Vanderbilt University

⁴Veterans Administration

Abstract:

Radiological reports are a rich source of clinical data which can be mined to assist with biosurveillance of emerging infectious diseases. In addition to biosurveillance, radiological reports are an important source of clinical data for health service research. Pneumonias and other radiological findings on chest xray or chest computed tomography (CT) are one type of relevant finding to both biosurveillance and health services research. In this study we examined the ability of a Natural Language Processing system to accurately identify pneumonias and other lesions from within free-text radiological reports. The system encoded the reports in the SNOMED CT Ontology and then a set of SNOMED CT based rules were created in our Health Archetype Language aimed at the identification of these radiological findings and diagnoses. The encoded rule was executed against the SNOMED CT encodings of the radiological reports. The accuracy of the reports was compared with a Clinician review of the Radiological Reports. The accuracy of the system in the identification of pneumonias was high with a Sensitivity (recall) of 100%, a specificity of 98%, and a positive predictive value (precision) of 97%. We conclude that SNOMED CT based computable rules are accurate enough for the automated biosurveillance of pneumonias from radiological reports.

Introduction:

This project seeks to determine the accuracy of a Natural Language Processing (NLP) system for the identification of patients with true pneumonias from a corpus of radiological reports. The system is the Multi-threaded Clinical Vocabulary Server (MCVS) developed at the Mayo Clinic.ⁱ NLP techniques have been long used to extract medial knowledge from narrative reports. One study looked at the accuracy of MedLEE in the identification of pneumonias in CXR reports and concluded that manual review was still required to accurately identify true cases from radiological reports.ⁱⁱ Another study by Chapman et al found that expert rules performed better than Bayesian networks or decision trees.ⁱⁱⁱ

Chomsky published, in 1955 in mimeograph form and in press in 1975, his seminal work, *The Logical Structure of Linguistic Theory*.^{iv} This work expressed the view that language was a predictable, systematic, and logical cognitive activity that required a meta-model of language to effectively communicate. He demonstrated that the behaviorists' stimulus-response model could not account for human language. This idea, that language is processed, led to the application of computer science to free text (natural language) processing. Computational linguistics (CL) is the field of computer science that seeks to understand and to represent language in an interoperable set of semantics. CL overlaps with the field of Artificial Intelligence and has been often applied to machine translation from one human language to another.

Researchers have succeeded, to varying degrees, to create CL algorithms for retrieving clinical texts. Sager in 1994 published a paper entitled, "Natural Language Processing and the Representation of Clinical Data."^v Here, Dr. Sager showed that for a set of discharge letters, a recall of 92.5% and a precision of 98.6% could be achieved for a limited set of pre-selected data using the parser produced by the Linguistic String Project at New York University.^{vi,vii,viii}

Researchers have also succeeded, to varying degrees, at representing the concepts underlying clinical texts. In 1999, Wagner, Rogers, Baud and Scherrer reported on the Natural Language generation of urologic procedures.^{ix,x} Here they used a conceptual graph technique to apply translations for 172 rubrics from a common conceptual base between French, German and English. They demonstrated that the GALEN model was capable of technically representing the concepts well; however the language generation was often not presented in a form which native speakers of the target language would find natural. Trombert-Paviot et al. reported the results of the use of GALEN in mapping French procedures to an underlying concept representation.^{xi} Wroe et al in

2001 reported the ability to integrate a separate ontology for drugs into the GALEN model.^{xii} Rector in his exposé “Clinical Terminology: Why is it so hard?” discusses the importance of and ten most challenging impediments to the development of compositional systems capable of representing the vast majority of clinical information in a comparable fashion.^{xiii} In 2001, Professor Rector published one workable method for integrating information models and terminology models.^{xiv}

In 2004, Friedman et al reported a method for encoding concepts from health records using the UMLS.^{xv} In this study, the investigators used their system, MedLEE, to abstract concepts from the record and reported a recall of 77% and a precision of 89%. In 2001, Nadkarni provided a description of the fundamental building blocks needed for NLP.^{xvi} He discussed their method for lexical matching and part of speech tagging in discharge summaries and surgical notes. Lowe developed MicroMeSH an early MUMPS based terminology browser which incorporated robust lexical matching routines.. Lowe, working with Hersh, reported the accuracy of parsing radiology reports using the Sapphire indexing system.^{xvii} Here, they reported good sensitivity and were able to improve performance by limiting the UMLS source vocabularies by section of the report.

Beyond representing clinical concepts, tools are needed to link text provided by clinicians to the concepts in the knowledge representation. Cooper and Miller created a set of NLP tools aimed at linking clinical text to the medical literature using the MeSH vocabulary.^{xviii} Overall, the composite method yielded a recall of 66% and a precision of 20%. Berrios et al reported a vector space model and a statistical method for mapping free text to a controlled health terminology.^{xix} Zou et al, reported a system, IndexFinder, which was principally a phrase representation system.^{xx} Srinivasan et al, indexed Medline citations (titles and abstracts) using the UMLS.^{xxi} Their method took the output of a part-of-speech tagger and feeds the SPECIALIST minimal commitment parser, the lexicon used by the UMLS system. The output of this stage was matched to a set of grammars that yielded a final match.

NLM recently developed MetaMap.^{xxii} It has the capacity to be used to code free text (natural language) to a controlled representation which can be any subset of the UMLS knowledge sources. MetaMap uses a five step process which begins by using the SPECIALIST minimal commitment parser which identifies noun phrases without modifiers. The next step involves the identification of phrase variants. These variants are then used to suggest candidate phrases from within the source material.^{xxiii} Linguistic principals are used to calculate a score for each potential match. Brennon and Aronson used MetaMap to improve consumer health information retrieval for patients.^{xxiv} Brown and Elkin et al published the index article on fully automated electronic

quality monitoring using an earlier version of the MCVS, showing that quality rules can be accurately constructed to differentiate high quality VA disability examinations.^{xxv}

Many emerging infectious diseases can lead to Pneumonia. Surveillance of these conditions may benefit from the ability to rapidly identify pneumonia cases from the reports from free-text radiological examinations. It remained uncertain as to whether NLP systems can perform with high recall and precision with respect to the identification of Pneumonia cases from free-text radiological reports. In this study, the Multi-threaded Clinical Vocabulary Server (MCVS) developed at the Mayo Clinic was evaluated to determine the accuracy rates for the identification of true cases of pneumonia from a corpus of radiological reports. True cases of Pneumonia were identified separately from possible cases of pneumonia and cases where pneumonia was not present.

Methods

Four-Hundred free-text radiological reports were obtained in cooperation with the CDC. All reports were either Chest Xrays or Computed Tomography Examinations. The entire report was parsed using the Multi-threaded Clinical Vocabulary Server (MCVS) coding the output into the SNOMED CT reference terminology. Each section of the report was identified (e.g. History, Findings, Impression). Within each section each concept was coded and tagged as a positive, negative or uncertain assertion.

We developed a SNOMED CT based rule for the identification of Pneumonias, Infiltrates or Consolidations or other types of pulmonary densities. The rule looked first at the impression of the record for the concept and then if not found it looked at the findings. The rules were specified in the Health Expression Archetype Language (HAL-42). This language allows rule designers to specify their coded concepts of interest, what truth value they wish to search for (positive, negative or uncertain assertions and combinations Positive and Uncertain or Negative and Uncertain), whether the concept should be exploded or not and where in the record one should look for the concept (by section of the record (e.g History, HPI, PE, Data, Findings, Impression). Rules can be combined with Boolean operators to make more specific rules (And, Or, Not). The rules can be linked in a series of if – then – else statements to create a chain of rules. Rule sets can be named and

saved for later reuse. Notes regarding rule sets can be saved as annotation with the rules. Concept based rules can be combined with fixed field queries (e.g Metadata like what departments records should be identified) and values can be

associated with concepts using mathematical operators (e.g. Age > 17 years).

The final SNOMED CT encoded rule developed was:

```
if pneumonia 233604007 Positive Assertion Explode Impression Section --> Positive Assertion Pneumonia
else if pneumonitis 205237003 Positive Assertion Explode Impression Section --> Positive Assertion Pneumonia
else if pneumonia 233604007 Uncertain Assertion Explode Impression Section --> Uncertain Assertion of
Pneumonia
else if pneumonitis 205237003 Uncertain Assertion Explode Impression Section --> Uncertain Assertion of
Pneumonia
else if pneumonia 233604007 Positive Assertion Explode Findings --> Positive Assertion Pneumonia
else if pneumonitis 205237003 Positive Assertion Explode Findings --> Positive Assertion Pneumonia
else if pneumonia 233604007 Uncertain Assertion Explode Findings --> Uncertain Assertion of Pneumonia
else if pneumonitis 205237003 Uncertain Assertion Explode Findings --> Uncertain Assertion of Pneumonia
else if (infiltrate 47351003 Positive Assertion Explode Impression Section OR consolidation 9656002 Positive
Assertion Explode Impression Section) --> Positive Assertion of an Infiltrate
else if (infiltrate 47351003 Uncertain Assertion Explode Impression Section OR consolidation 9656002 Uncertain
Assertion Explode Impression Section) --> Uncertain Assertion of an Infiltrate
else if (infiltrate 47351003 Positive Assertion Explode Findings OR consolidation 9656002 Positive Assertion
Explode Findings) --> Positive Assertion of an Infiltrate
else if (infiltrate 47351003 Uncertain Assertion Explode Findings OR consolidation 9656002 Uncertain Assertion
Explode Findings) --> Uncertain Assertion of an Infiltrate
else if ((((((opacity 128305008 Positive Assertion Explode Impression Section OR opacification 263506008
Positive Assertion Explode Impression Section) OR density 125146005 Positive Assertion Explode Impression
Section) OR nodule 27925004 Positive Assertion Explode Impression Section) OR mass 300848003 Positive
Assertion Explode Impression Section) OR granuloma 45647009 Positive Assertion Explode Impression Section)
OR granulomatous disease 128561003 Positive Assertion Explode Impression Section) --> Positive Assertion of an
Opacity
else if ((((((opacity 128305008 Uncertain Assertion Explode Impression Section OR opacification 263506008
Uncertain Assertion Explode Impression Section) OR density 125146005 Uncertain Assertion Explode Impression
Section) OR nodule 27925004 Uncertain Assertion Explode Impression Section) OR mass 300848003 Uncertain
Assertion Explode Impression Section) OR granuloma 45647009 Uncertain Assertion Explode Impression Section)
OR granulomatous disease 128561003 Uncertain Assertion Explode Impression Section) --> Uncertain Assertion of
an Opacity
else if ((((((opacity 128305008 Positive Assertion Explode Findings OR opacification 263506008 Positive
Assertion Explode Findings) OR density 125146005 Positive Assertion Explode Findings) OR nodule 27925004
Positive Assertion Explode Findings) OR mass 300848003 Positive Assertion Explode Findings) OR granuloma
45647009 Positive Assertion Explode Findings) OR granulomatous disease 128561003 Positive Assertion Explode
Findings) --> Positive Assertion of an Opacity
else if ((((((opacity 128305008 Uncertain Assertion Explode Findings OR opacification 263506008 Uncertain
Assertion Explode Findings) OR density 125146005 Uncertain Assertion Explode Findings) OR nodule 27925004
Uncertain Assertion Explode Findings) OR mass 300848003 Uncertain Assertion Explode Findings) OR granuloma
45647009 Uncertain Assertion Explode Findings) OR granulomatous disease 128561003 Uncertain Assertion
Explode Findings) --> Uncertain Assertion of an Opacity
else false.
```

The flow of the rule is to look for positive assertions of pneumonia in the impression section, then look for uncertain assertions of pneumonia from the impression section then look for positive assertions of pneumonia from the findings section and finally look for uncertain assertions of pneumonia from the findings section. This pattern is repeated if an assertion of pneumonia was not

identified for infiltrates or consolidations and then for Opacities of one type or another.

Each of the 400 reports was reviewed independently by two Internal Medicine Physicians (academic physicians at the Mayo Clinic) and the final determination for any disagreements was determined by consensus.

The inter-rater agreement produced a Kappa statistic of 0.936 (95%CI: 0.905, 0.967) which corresponds to

excellent agreement. The accuracy of the MCVS was judged against the gold standard clinician review.

Results:

Performance of the MCVS including Uncertainty							Performance of MCVS without Uncertainty					
Mayo Human Review (Gold Standard)							Mayo Human Review (Gold Standard)					
		PP	PU	DOCI	None	Total		PP	DOCI	None	Total	
Mayo	PP	14	6	0	2	22	MAYO	PP	76	0	2	78
NLP	PU	0	56	0	0	56	NLP	DOCI	0	75	3	78
	DOCI	0	0	75	3	78		NONE	0	0	244	244
	NONE	0	0	0	244	244		Total	76	75	249	400
	Total	14	62	75	249	400						
SENS: 100%							SENS: 100%					
SPEC: 90.3%							SPEC: 97.99%					
PPV: 70%							PPV: 97.44%					

Table 1: Performance of the Hal-42 Pneumonia detection rule with and without the consideration of Uncertainty (Multi-valued Logic). PP = Positive Assertion of Pneumonia; PU is an Uncertain Assertion of Pneumonia and DOCI = Density, Opacity, Consolidation or Infiltrate; None = No Pneumonia or DOCI.

Failure analysis of the MCVS data found that all of the error (6 of 6 missed uncertain assertions of pneumonia) was due to our lack of ability to recognize the positive assertion of pneumonia within a differential diagnosis list as an uncertain assertion. The accuracy for the identification of all lesions correctly was 96.25%.

Conclusions

High levels of accuracy can be obtained using NLP tools to identify cases of Pneumonia from free-text radiological reports. The MCVS performed well with respect to this task of identifying pneumonia cases from free-text radiological

reports. Differentiation of uncertain assertions was very important in producing a highly accurate rule for the automated identification of pneumonia cases. Future work should be aimed at the identification of lists of concepts as positive assertions in a list (e.g. “Findings are compatible with Pneumonia, CHF and Non-Cardiogenic Pulmonary Edema”). We believe that the MCVS has sufficient accuracy to be used as a methodology for the identification of cases of pneumonia from free-text radiological reports. Validating the accuracy of data mining rules is essential to accurate and transparent automated health services research and biosurveillance.

Acknowledgements:

This work has been supported in part by grants from the Centers for Disease Control and Prevention (PH00022 and HK00014, PLE), from the National Library of Medicine (5K22 LM008576-03; STR) and from the Veterans Administration Health Services Research and Development (SAF 03-223 -3, PLE).

References:

- ⁱ Peter L. Elkin, MD, Steven H. Brown, MD, Casey Husser, MD, Brent A. Bauer, MD, Dietlind Wahner-Roedler, MD, S. Trent Rosenbloom, MD, Ted Speroff, PhD, "An Evaluation of the Content Coverage of SNOMED-CT for Clinical Problem Lists", *Mayo Clin Proc.* 2006 Jun;81(6):741-8. June 2006.
- ⁱⁱ Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med.* 1995 May 1;122(9):681-8.
- ⁱⁱⁱ Chapman WW, Fizman M, Chapman BE, Haug PJ, A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. *J Biomed Inform.* 2001 Feb;34(1):4-14.
- ^{iv} Chomsky, N. *The logical structure of linguistic theory.* (New York: Plenum, 1975).
- ^v Sager N, Lyman, M, et al. Natural language processing and the representation of clinical data. *J Am Med Inform Assoc* 1994;1(2):142-160.
- ^{vi} Sager N. Syntactic analysis of natural language. In: *Advances in Computers*, Vol (8) New York: Academic Press, 1967:153-188.
- ^{vii} Grishman R, Sager N, Raze C, Bookchin B. The linguistic string parser. In: *AFIPS Conference Proceedings*. Montvail NJ: AFIPS Press. 1973;42: 427-434.
- ^{viii} Sager N, Gishman R. The restriction language for computer grammars of natural language. *Communications of the ACM.* 1975; 18:390-400.
- ^{ix} Wagner JC, Rogers JE, Baud RH, Scherrer JR. **Natural language generation of surgical procedures.** [Journal Article] *International Journal of Medical Informatics.* 53(2-3):175-92, 1999 Feb-Mar.
- ^x Wagner JC, Rogers JE, Baud RH, Scherrer JR. **Natural language generation of surgical procedures.** [Journal Article] *Medinfo. 9 Pt 1*:591-5, 1998.
- ^{xi} Trombert-Pavio B, Rodrigues JM, Rogers JE, Baud R, van der Haring E, Rassinoux AM, Abrial V, Clavel L, Idir H. **Galen: a third generation terminology tool to support a multipurpose national coding system for surgical procedures.** [Journal Article] *Studies in Health Technology & Informatics.* 68:901-5, 1999.
- ^{xii} Wroe CJ, Cimino JJ, Rector AL. **Integrating existing drug formulation terminologies into an HL7 standard classification using OpenGALEN.** [Journal Article] *Proceedings / AMIA ... Annual Symposium.* :766-70, 2001.
- ^{xiii} Rector AL. **Clinical terminology: why is it so hard?.** [Journal Article] *Methods of Information in Medicine.* 38(4-5):239-52, 1999 Dec.
- ^{xiv} Rector AL. The interface between information, terminology, and inference models. [Journal Article] *Medinfo. 10(Pt 1)*:246-50, 2001.
- ^{xv} Friedman C, Shagina L, Lussier Y, Hripcsak, G. Automated Encoding of Clinical Documents Based on Natural Language Processing. *JAMIA* Sept 2004 Vol 11 (5).
- ^{xvi} Nadkarni P, Chen R, Brandt C. UMLS concept indexing for production databases: a feasibility study. *J Am Med Inform Assoc.* 2001;8:80-91.
- ^{xvii} Huang Y, Lowe H, Hersh W. A pilot study of contextual UMLS indexing to improve the precision of concept based representation in XML-structured clinical radiology reports. *J Am Med Inform Assoc.* 2003;10:580-7.
- ^{xviii} Cooper GF, Miller RA. An experiment comparing lexical and a statistical method for extracting MeSH terms from Clinical free text. *J Am Med Inform Assoc.* 1998;5:62-75.
- ^{xix} Berrios DC. Automated Indexing for full text information retrieval. *Proc AMIA Symp.* 2000:71-5.
- ^{xx} Zou Q, Chu WW, Morioka C, Leazer GH, Kangaroo H. IndexFinder: a method of extracting key concepts from clinical texts for indexing. *Proc AMIA Symp.* 2003:763-7.
- ^{xxi} Srinivasan S, Rindfleisch TC, Hole WT, Aronson AR, Mork JG. Finding UMLS Metathesaurus concepts in MEDLINE. *Proc AMIA Symp.* 2002:727-31.
- ^{xxii} Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, et al. The NLM Indexing Initiative. *Proc AMIA Symp* 2000:17-21.
- ^{xxiii} Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001:17-21.
- ^{xxiv} Brennan PF, Aronson AR. Towards linking patients and clinical information: detecting UMLS concepts in e-mail. *J Biomed Inform.* 2003 Aug-Oct;36(4-5):334-41.
- ^{xxv} Steven H. Brown MS MD, Theodore Speroff PhD, Elliot M. Fielstein PhD, Brent A. Bauer MD, Dietlind L. Wahner-Roedler MD, Robert Greevy PhD, Peter L. Elkin MD, "eQuality: Automatic Assessment from Narrative Clinical Reports", *Mayo Clin Proc.* 2006;81(11):1472-1481.