Automatic Retrieval of Bone Fracture Knowledge Using Natural Language Processing

Bao H. Do · Andrew S. Wu · Joan Maley · Sandip Biswal

Published online: 2 October 2012

Abstract Natural language processing (NLP) techniques to extract data from unstructured text into formal computer representations are valuable for creating robust, scalable methods to mine data in medical documents and radiology reports. As voice recognition (VR) becomes more prevalent in radiology practice, there is opportunity for implementing NLP in real time for decision-support applications such as context-aware information retrieval. For example, as the radiologist dictates a report, an NLP algorithm can extract concepts from the text and retrieve relevant classification or diagnosis criteria or calculate disease probability. NLP can work in parallel with VR to potentially facilitate evidence-based reporting (for example, automatically retrieving the Bosniak classification when the radiologist describes a kidney cyst). For these reasons, we developed and validated an NLP system which extracts fracture and anatomy concepts from unstructured text and retrieves relevant bone fracture knowledge. We implement our NLP in an HTML5 web application to demonstrate a proof-of-concept feedback NLP system which retrieves bone fracture knowledge in real time.

Keywords Natural language processing · Decision support · Information retrieval of bone fractures

Introduction

Natural language processing (NLP) techniques to extract data from unstructured text into formal computer representations are valuable for creating robust, scalable methods to mine data in medical documents and radiology reports [1–7]. These implementations, however, are retrospective applications of data mining.

Voice recognition (VR) has become more prevalent in radiology practice in recent years [8–11]. VR generates text data at the time of interpretation and can create an opportunity for implementing NLP in real time to enable context-aware information retrieval and processing. For example, as the radiologist dictates a report, an NLP algorithm can extract concepts from the text and perform decision support tasks such as retrieve relevant classification or diagnosis criteria or calculate disease probability. NLP can work in parallel with VR to potentially facilitate evidence-based reporting (for example, automatically retrieving the Bosniak classification when the radiologist describes a kidney cyst). Figure 1 illustrates a proposed "feedback" NLP system.

For these reasons, the goal of this project is to develop and validate an NLP system which extracts fracture and anatomy concepts from unstructured text and retrieves relevant bone fracture knowledge. We have chosen to study fractures because such a system of automated information retrieval could be useful in the emergency setting, and based on our clinical experience, the syntax and lexicon for describing bone fracture in radiology reporting is limited. We

B. H. Do · S. Biswal
Division of Musculoskeletal Section, Department of Radiology,
Stanford University School of Medicine,
300 Pasteur Drive, S-056,
Stanford, CA 94305, USA

A. S. Wu
Department of Radiology,
Kaiser Permanente Downey Medical Center,
9333 Imperial Hwy,
Downey, CA 90242, USA

J. Maley
Division of Neuroradiology, Department of Radiology,
University of Iowa,
200 Hawkins Drive,
Iowa City, IA 52242, USA

B. H. Do (✉)
Division of Musculoskeletal Imaging, Department of Radiology,
Stanford University Medical Center,
300 Pasteur Dr., S-068B,
Stanford, CA 94305, USA
e-mail: baodo@stanford.edu

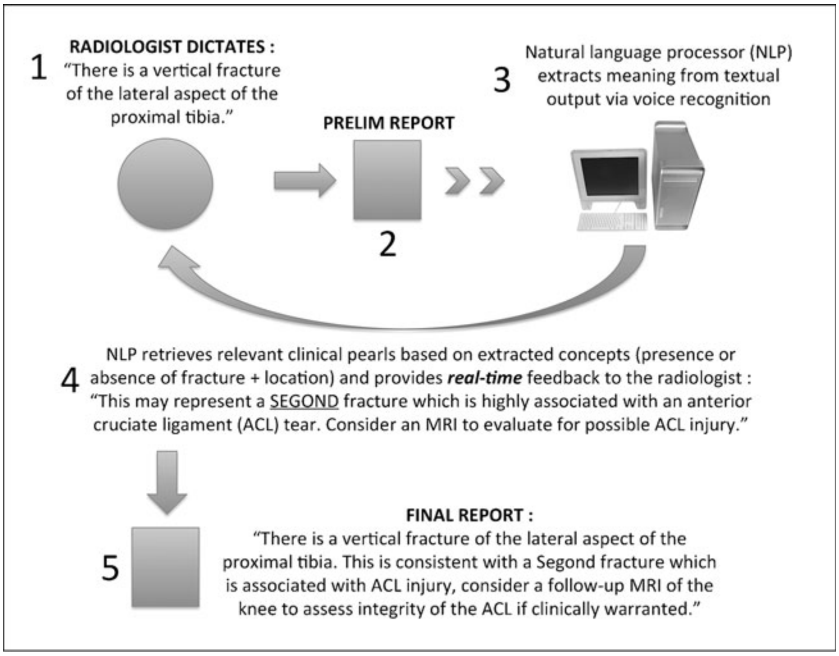

1
RADIOLOGIST DICTATES :
"There is a vertical fracture of the lateral aspect of the proximal tibia."
PRELIM REPORT
2
3
Natural language processor (NLP) extracts meaning from textual output via voice recognition
4
NLP retrieves relevant clinical pearls based on extracted concepts (presence or absence of fracture + location) and provides *real-time* feedback to the radiologist : "This may represent a SEGOND fracture which is highly associated with an anterior cruciate ligament (ACL) tear. Consider an MRI to evaluate for possible ACL injury."
5
FINAL REPORT :
"There is a vertical fracture of the lateral aspect of the proximal tibia. This is consistent with a Segond fracture which is associated with ACL injury, consider a follow-up MRI of the knee to assess integrity of the ACL if clinically warranted."

Fig. 1 Overview of a decision support system driven by speech recognition and NLP. As the radiologist describes a fracture, the natural language processor extracts disease (fracture) and anatomy (tibia) concepts from unstructured text to retrieve relevant fracture knowledge

will implement our NLP in a web application and use speech recognition software to demonstrate a proof-of-concept feedback NLP system which retrieves bone fracture knowledge in real time.

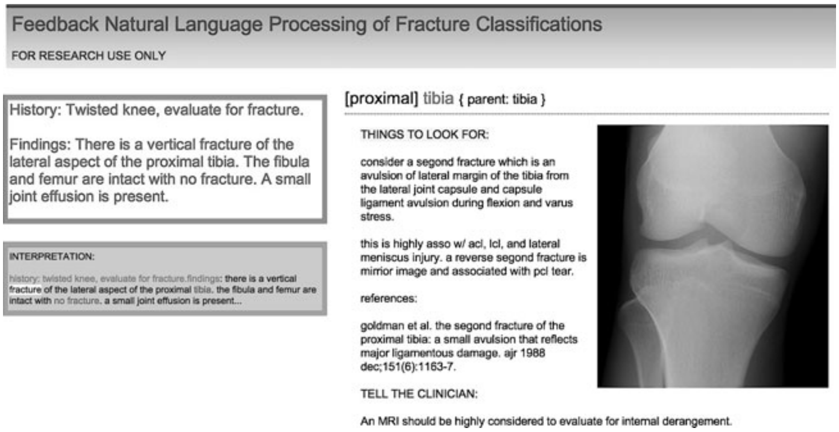
Materials and Methods

Institution review board approval for a retrospective review of 1 year of radiology reports of emergency department studies was obtained, and all reports were de-identified in compliance with the Health Insurance Portability and Accountability Act.

Architecture

A database of 33,090 unstructured radiology reports of emergency department studies over a 1-year period was indexed in a MySql database (Sun Microsystems, Redwood City, CA) running on an Apache Web server (Apache Foundation, Los Angeles, CA).

The feedback NLP system was designed on a web-based architecture using a Hypertext Preprocessor (PHP Group), MySql, and Apache backend. This set-up was chosen for platform versatility, and all software components are available as open-source for the Windows, Mac, and Linux platforms. Figure 2 shows the Web-based interface.



Feedback Natural Language Processing of Fracture Classifications

FOR RESEARCH USE ONLY

History: Twisted knee, evaluate for fracture.

Findings: There is a vertical fracture of the lateral aspect of the proximal tibia. The fibula and femur are intact with no fracture. A small joint effusion is present.

INTERPRETATION:

history: twisted knee, evaluate for fracture.findings: there is a vertical fracture of the lateral aspect of the proximal tibia. the fibula and femur are intact with no fracture. a small joint effusion is present...

[proximal] tibia { parent: tibia }

THINGS TO LOOK FOR:

consider a segond fracture which is an avulsion of lateral margin of the tibia from the lateral joint capsule and capsule ligament avulsion during flexion and varus stress.

this is highly asso w/ acl, lcl, and lateral meniscus injury. a reverse segond fracture is mirrior image and associated with pcl tear.

references:

goldman et al. the segond fracture of the proximal tibia: a small avulsion that reflects major ligamentous damage. ajr 1988 dec;151(6):1163-7.

TELL THE CLINICIAN:

An MRI should be highly considered to evaluate for internal derangement.

Fig. 2 Application interface consists of three components: (1) *upper left* dictation box (DB), (2) *lower left* natural language processor workspace (NLPW), and (3) *right column* NLP output (NLPO). As the radiologist dictates or types in the DB, the system extracts fracture concepts (NLPW) in real-time and displays relevant knowledge (NLPO). In this example, the NLP detects a fracture and reminds the radiologist to consider a Segond fracture and follow-up MRI. Note that the NLP correctly ignores the mention of fracture in the history section of the report and recognizes that the femur and fibula are intact

For input, unstructured report text can be entered by keyboard or speech. Dedicated radiology speech recognition software was not available; therefore, we demonstrated real-time functionality using the generic English language Microsoft Speech Recognition Engine (SRE) in our Windows XP PC as proof of concept of speech triggered real-time information retrieval. The text input triggers the server-side analysis in real-time via Asynchronous JavaScript and XML (AJAX).

NLP System

Figure 2 shows the Web-based NLP system. Text data from the Web browser is passed via AJAX to the back-end natural language processor. The NLP system accepts raw, unstructured text and applies simple a rules-based heuristic to identify fracture concepts. The system evaluates each statement and classifies each for the presence or absence of a fracture, and if there is a fracture, the system extracts the name of the involved bone (mapped to a custom lexicon) and longitudinal location of the fracture (for example, proximal, mid, distal, diaphyseal, metaphyseal, and epiphyseal).

The natural language processor consists of several modules that perform specific functions in series: section parser, negation engine, bone ID module, and anatomy sub-ID module

First, the system uses contextual cues from standard headings of radiology reports to eliminate false-positive matches for fracture. For example, the *section parser* identifies the report sections contextually by evaluating paragraph headings tokens. The purpose of this function is to eliminate statements relating to the history and/or technique section of the report that would otherwise be false-positive mentions of the token fracture. For example, the phrase, “evaluate for femoral neck fracture” is an indication statement rather than a description of an actual fracture and is excluded from further analysis by the section parser.

Next, the *negation engine* uses regular expressions of negative grammatical constructs of to classify statements as containing or not containing an actual fracture concept. The regular expression-based algorithm identifies a limited number of syntactic patterns denoting negations such as “without evidence … fracture” (“…” characters represent wildcard words) which matches to statements such as “without evidence for femoral neck fracture” and “without evidence for femur fracture” without having to define all possible negative phrases for all recognized bone name entities. (A discussion of regular expressions is beyond the scope of this manuscript, but briefly, a regular expression is a sequence of characters which signals the natural language processor to match a pattern of text without requiring all the possible lexical and syntactic variants to be explicitly defined.)

If there is a true positive fracture concept within the statement, the bone ID module identifies the involved bone. The bone ID module maps the statement to a custom bone ID ontology table which defines a simple parent–child hierarchical relationship. For example, in the statement, “there is a fracture of the inferior pubic ramus,” the NLP recognizes that there is a fracture of the inferior pubic ramus which is a child of the parent bone, pelvis, and maps the sentence to the management knowledge associated with pelvis bone. The initial recognition, however, is simple text matching to the phrase, inferior pubic ramus.

Next, the *anatomy sub-ID module* further localizes the fracture site along the longitudinal axis of the bone (for example, proximal, mid, and distal), and if relevant, uses the bone and anatomy sub-ID output to retrieve the appropriate normal comparison bone image and management knowledge associated with the recognized fracture.

NLP Development and Training

A previously validated natural language processor which recognizes single word negation concepts [12] was implemented to recognize fracture negations in our NLP system. From our database of 33,090 unstructured radiology reports of emergency department studies over a 1-year period, we selected a corpus of 91 consecutive X-ray exams from a 1-week period for review. We studied these X-ray reports and developed a database of regular expressions to capture the ways in which fractures were described in reports (lexicon and syntax) and iteratively train the natural language processor. We used background knowledge from clinical practice, RSNA’s RadLex (www.radlex.org), and musculoskeletal textbooks [13, 14] to create the bone ID, anatomy sub-id, and management knowledge tables. The bone ID table consists of 128 records of bones. The management knowledge table consists of 46 records covering 20 unique bones.

Feedback NLP Demonstration

The Microsoft SRE was used to demonstrate the speech-triggered real-time information retrieval capability of the system; however, performance statistics were not recorded.

NLP Validation

To maintain 100 % report fidelity, we developed a function which randomly selected X-ray reports from a pool of 33,090 unstructured emergency department reports and directly sent each report (368 total) to the NLP which processed the unstructured text (thus excluding potential report transmission error relating to speech recognition or keyboard entry). The NLP processed each report and identified the presence or absence of a fracture. The NLP output was then manually graded by investigators AW and BD. The arbitrary one year period was chosen to maximize the unique dictation styles. Exclusionary criteria included reports without a findings or impression section and non-X-ray studies.

The true-positive, true-negative, false-positive, and false-negative counts, using the human grader as ground truth, were tallied, and a 2×2 contingency table was created. Using this table, basic contingency table statistics were calculated (sensitivity, specificity, positive predictive value, negative predictive value, accuracy, diagnostic odds ratio, and Kappa measure of agreement) to summarize NLP performance in identifying fractures. Within the subset of reports describing positive fractures, 43 reports were randomly chosen, and the NLP's accuracy in correctly identifying the involved bone was manually evaluated.

Results

Speech output produced by the Microsoft SRE successfully triggered information retrieval of bone fracture knowledge in real time.

The true-positive, true-negative, false-positive, and false-negative counts were 113, 230, 13, and 12, respectively. These tally results are summarized in Table 1.

Based on this contingency table, the sensitivity, specificity, positive predictive value, negative predictive value, accuracy, diagnostic odds ratio, and Kappa measure of agreement were 90, 95, 90, 95, 93, 166, and 85 %, respectively.

Within the set of positive fractures (n=125), 43 reports were randomly selected, and NLP system accuracy in identifying the bone involved was assessed. The bone involved was identified with an accuracy of 79 % (34 of 43). Of the nine inaccurately localized reports, one did not report the actual bone ("there is a fracture"), one mentioned only "multiple facial fractures", six involved phrases related to vertebral body levels, and one described an undefined structure in the anatomy database ("supracondylar").

Table 1 Contingency table summarizing NLP classifier performance in extracting fracture concepts from unstructured text

NLP classification of unstructured text report	Radiologist classification (gold standard)	
	Fracture	No fracture
Fracture	13 (true positive)	113 (false positive)
No fracture	230 (false negative)	12 (true negative)

The NLP classified each report as containing or not containing a fracture (rows). Human graders reviewed the NLP output (columns) and tallied each NLP call as a true positive, true negative, false positive, or false negative

Discussion

We have developed and validated a natural language processor to retrieve relevant bone fracture knowledge. We implemented our NLP in a web application and used generic speech recognition software to demonstrate a proof-of-concept feedback NLP system which retrieves bone fracture knowledge in real-time.

Automated information retrieval may be advantageous to the traditional search initiated by the radiologist using a text reference or Google. Retrieved information can be presented to the radiologist without need for the radiologist to be aware that the potentially useful knowledge exists. Instead radiologist need is *anticipated*. For example, if the radiologist describes a "fracture of the medial malleolus of the ankle," the system could suggest to the radiologist to examine the tibiofibular joint space and inquire if the patient has pain around the knee to exclude a potential Maisonneuve fracture which is associated with medial malleolar ankle fractures and syndesmotic disruptions. In the traditional paradigm, the radiologist must be aware of (or has time to seek reference to discover) the association. Although beyond the scope of this work, a hypothesis driven study evaluating the effect of a real-time NLP system on radiologist interpretation warrants further investigation.

We chose to study fracture classification because a real-time NLP system automating retrieval of fracture knowledge can be useful in the emergency setting, and based on our clinical experience, there is a limited expression of fracture pathology in radiology reporting. In fact, despite a lack of controlled vocabulary terms for the pathology, fracture, such as buckle, impaction, cortical disruption, or cortical irregularity, the system performance is comparable to described rules based classifiers in medicine (overall accuracy 93 %) [6, 15, 16]. However, we acknowledge that other disease entities may have more lexical and syntactic complexity which may require more sophisticated approaches. For example, a renal tumor can be described as a kidney tumor, kidney mass, kidney lesion, renal tumor, renal mass, and renal lesion. Indeed, the validation process discovered additional controlled vocabulary equivalents for the fracture concept, including compression deformity and cortical irregularity. Despite this, rule-based algorithms have reported high accuracy, exceeding 90 %, for identifying concepts they are specifically designed to study, such as critical findings [16]. In contrast to devising unsupervised machine-learning quantitative models, rules-based classifiers require time- and domain-specific knowledge, along with iterative training, to develop the rules and knowledge database to account for all permutations and styles of expressing a single pathology of interest. Furthermore, institutional differences in reporting should be considered prior to clinical deployment of NLP tools derived from rules based text classification methods at a single institution.

Refining the controlled vocabulary may improve sensitivity, but context is also important. One report counted as a false negative contained the phrase “cast obscures bony details.” For the human reader, it is implied that a cast protects a fracture, and thus a fracture exists although there is no direct mention. An example of a false positive without a negation concept is the mention of fracture in the context of a recommendation, for example, “if patient has persistent pain at this site, recommend repeat radiographs in 7 to 10 days, to evaluate for fracture.” Another reason for the relatively low sensitivity is that true positives were undercounted in reports which contained multiple fractures. This problem became evident after the training period, and instead of adjusting our system, we preferred to report the lowest sensitivity outcome.

The NLP system identified the involved bone with an accuracy of 79 %. Excluding the two statements which did not mention a specific bone/site and six statements related to the spine, the system identified the major bone in 97 % (34 of 35) of statements. In addition to incorporating specific code heuristics to account for spine fractures, incorporating more advanced forms of NLP such as parts of speech (POS) tagging may improve specificity.

Conclusions

We have developed and validated an NLP system which extracts fracture and anatomy concepts from unstructured text and retrieves relevant bone fracture knowledge. We implemented our NLP in a web application and used generic speech recognition software to demonstrate a proof-of-concept feedback NLP system which retrieves bone fracture knowledge in real time.

Acknowledgment of Grants None

Financial Disclosures or Other Assistance None

References

1. Do BH, Wu A, Biswal S, Kamaya A, Rubin DL: Informatics in radiology: RADTF: a semantic search-enabled, natural language processor-generated radiology teaching file. Radiographics 30 (7):2039–2048, 2010
2. Lacson R, Khorasani R: Practical examples of natural language processing in radiology. J Am Coll Radiol 8(12):872–874, 2011
3. Hripcsak G, Austin J, Alderson P, Friedman C: Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology 224(1):157–163, 2002
4. Sistrom CL, Dreyer KJ, Dang PA, Weilburg JB, Boland GW, Rosenthal DI, Thrall J: Recommendations for additional imaging in radiology reports: multifactorial analysis of 5.9 million examinations. Radiology 53(2):453–461, 2009
5. Dang PA, Kalra MK, Blake MA, Schultz TJ, Halpern EF, Dreyer KJ: Original research: extraction of recommendation features in radiology with natural language processing: exploratory study. AJR Am J Roentgenol 191(2):313–320, 2008
6. Thomas BJ, Ouellette H, Halpern EF, Rosenthal DI: Automated computer-assisted categorization of radiology reports. AJR Am J Roentgenol 184(2):687–690, 2005
7. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, Dittus RS, Rosen AK, Elkin PL, Brown SH, Speroff T: Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 306(8):848–855, 2011
8. Mehta A, Dreyer KJ, Schweitzer A, Couris J, Rosenthal D: Voice recognition—an emerging necessity within radiology: experiences of the Massachusetts General Hospital. J Digit Imaging 11(4 Suppl 2):20–23, 1998
9. Quint DJ: Voice recognition: ready for prime time? J Am Coll Radiol 4(10):667–669, 2007
10. Mehta A, McLoud TC: Voice recognition. J Thoracic Imaging 18:178–182, 2003
11. Pezzullo J, Tung GA, Rogg JM, Davis LM, Brody JM, Mayo-Smith WW: Voice recognition dictation: radiologist as transcriptionist. J Digit Imaging 21(4):384–389, 2008
12. Wu AS, Do BH, Kim J, Rubin DL: Evaluation of negation and uncertainty detection and its impact on precision and recall in search. J Digit Imaging 24(2):234–242, 2011
13. Clifford R. Wheeless III, MD. Wheeless’ Textbook of Orthopaedics. http://www.wheelessonline.com
14. Greenspan A. Orthopedic Imaging—A Practical Approach, fourth edition. Baltimore: Lippincott Williams & Wilkins, 2004.
15. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG: A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 34(5):301–310, 2001
16. Lakhani P, Kim W, Langlotz CP: Automated detection of critical results in radiology reports. J Digit Imaging 25(1):30–36, 2012