Natural Language Processing Techniques for Extracting and Categorizing Finding Measurements in Narrative Radiology Reports

M. Sevenster[1]; J. Buurman[2]; P. Liu[3]; J.F. Peters[4]; P.J. Chang[3]

[1]Philips Research, Briarcliff Manor, NY, United States; [2] Philips Research, Eindhoven, Netherlands; [3]University of Chicago Hospitals, Chicago, IL, United States; [4]Philips Healthcare, Best, Netherlands

Keywords

Natural language processing, radiology report, measurement, maximum entropy, quantitative imaging

Summary

Background: Accumulating quantitative outcome parameters may contribute to constructing a healthcare organization in which outcomes of clinical procedures are reproducible and predictable. In imaging studies, measurements are the principal category of quantitative parameters.
Objectives: The purpose of this work is to develop and evaluate two natural language processing engines that extract finding and organ measurements from narrative radiology reports and to categorize extracted measurements by their "temporality".
Methods: The measurement extraction engine is developed as a set of regular expressions. The engine was evaluated against a manually created ground truth. Automated categorization of measurement temporality is defined as a machine learning problem. A ground truth was manually developed based on a corpus of radiology reports. A maximum entropy model was created using features that characterize the measurement itself and its narrative context. The model was evaluated in a ten-fold cross validation protocol.
Results: The measurement extraction engine has precision 0.994 and recall 0.991. Accuracy of the measurement classification engine is 0.960.
Conclusions: The work contributes to machine understanding of radiology reports and may find application in software applications that process medical data.

Correspondence to:
Merlijn Sevenster, PhD
345 Scarborough Road
Briarcliff Manor, NY 10510, USA
Email: Merlijn.sevenster@philips.com

Appl Clin Inform 2015; 6: 600–610
http://dx.doi.org/10.4338/ACI-2014-11-RA-0110
received: December 1, 2014
accepted in revised form: July 31, 2015
published: September 30, 2015
Citation: Sevenster M, Buurman J, Liu P, Peters JF, Chang PJ. Natural language processing techniques for extracting and categorizing finding measurements in narrative radiology reports. Appl Clin Inform 2015; 6: 600–610
http://dx.doi.org/10.4338/ACI-2014-11-RA-0110

1. Introduction

There is an increasing emphasis on constructing a healthcare organization in which outcomes of clinical procedures such as interventions, treatments and imaging studies, are reproducible and predictable [1]. This challenge can be addressed by accumulating quantitative data points and providing workflow-relevant views on the data thus accumulated.

The purpose of this work is to develop and evaluate two natural language processing engines that extract measurements of organs and findings from narrative radiology reports and to categorize extracted measurements by their "temporality", i.e. if they quantify an entity observed on the current exam, a prior exam or both. The engines are specified in two technical appendices. The work contributes to machine understanding of radiology reports and may find application in software applications that process medical data [2].

Natural language processing techniques have been applied on medical and radiological content either in the form of general-purpose systems [3, 4] or as targeted systems addressing one particular task [5–16]. To the best of our knowledge, measurement extraction and classification has not been studied before in the literature. MedLEE [3], a landmark general-purpose system for normalizing radiological narrative, does recognize measurements as a separate category of information items, but we are not aware of formal evaluations of its measurement extraction and normalization capabilities.

In imaging studies, measurements are the principal category of quantitative parameters [1]. In ultrasound studies, measurements are routinely made to assess if organs have pathological dimensions. In support of oncology care, measurements are one of the key parameters for tracking treatment response per the World Health Organization [17] and RECIST (Response Evaluation Criteria in Solid Tumors) [18] guidelines. Two complementary surveys investigated the radiologists' [19] and oncologists' [20] views on measurement management, respectively. It was found that "most of the abdominal imagers at NCI-sponsored cancer centers dictate tumor measurements in routine clinical scanning" [19].

2. Methods

2.1 Corpus

Three sets worth respectively 34626, 100 and 200 radiology reports was obtained from a 550-bed university hospital in the Midwest. Informed consent was waived by IRB (13–0379). The reports were authored with dictation technology (PowerScribe, Nuance, Burlington, MA). The corpus was anonymized using an extensive set of regular expressions that had been tuned to the hospital's reporting style. After anonymization, the reports were subjected to a sentence boundary detection (SBD) engine that recognizes sections and sentences in narrative radiology reports. The first sentence of each section was matched against a list of known section headers and normalized with respect to five categories: comparison, technique, clinical history, findings and conclusions. The reports and derived section and sentence objects were persisted in a MySQL (Oracle) database.

2.2 Measurement extraction

2.2.1 Definition

A "measurement" was defined as a textual description of the dimensions of a one, two or three-dimensional entity. Sample measurements include "1.5 centimeters", "1.5 by 2.8 cm" and "1.5 x 2.8 x 2.1 cm". In the context of a sentence, a measurement was the complete textual description of all dimensions discussed in the sentence of one particular entity. For instance, in the sentence "The lesion currently measures 1.5 x 2.8 x 2.1 cm", the strings "2.1 cm" and "2.8 x 2.1 cm" were not considered measurements. The measurement extraction engine essentially extracted a string from a given sentence, e.g. the underlined string in "Submandibular lymph node measures 5 x 5 mm". Ideally, all extracted strings were measurements as defined above.

2.2.2 Software development

A measurement extraction engine was developed based on the measurements in the 100-report set that is driven by a regular expression accounting for typical measurement syntax, ▶Appendix 1. The regular expression was basically a context-free grammar that permitted permutations of sub-expressions that recognize floating numbers (e.g. "1.5"), product markers (e.g. "x" and "by") and units (e.g. "mm" and "centimeters"). In this manner, "1.5 x 2.8 cm" was recognized as an admissible permutation, but "1.5 2.8 x cm" was not. Each subexpression accounts for common variants. For instance, "1.5 by 2.8 cm" and "1.5 x 2.8 centimeters" were both recognized as variants of "1.5 x 2.8 cm" and were therefore accepted as measurements. The extraction engine was configured to find the longest string recognized as a measurement. In this manner, the string "2.8 cm" would not be extracted if "1.5 x 2.8 cm" appears in the sentence.

The extraction engine did not recognize numbers if they were represented by characters (e.g. "four centimeters"), as this style did not appear in our reports. This is presumably a consequence of the fact that all reports were authored with dictation technology.

2.2.3 Evaluation

To evaluate precision of the measurement extraction engine, we validated that the strings that have been extracted by the engine are indeed measurements and are complete (e.g. not only "5 mm" if the complete measurement is "5 x 5 mm"). To this end, we randomly selected 1 000 strings extracted by the measurement extraction engine. Per extracted string, we determined if it is indeed a complete measurement in the sentence. If the extracted substring was incomplete, it was counted as a false positive. Precision was the number of true positives divided by 1 000.

To evaluate recall, we randomly selected 200 reports in such a way that for $0 \leq N \leq 9$, we had 20 reports from which N substrings were extracted by the measurement extraction engine. Per report, we counted the number of measurements and the number of measurements coinciding with an extracted string. Recall was the ratio of correctly extracted measurements.

2.3 Measurement classification

2.3.1 Definition

We introduce a measurement classification scheme in which the top-level concepts identify the nature of the quantified entity:

- Clinical finding: The measurement describes the dimension(s) of a clinical finding.
- Relative position (rel-pos): The measurement characterizes the position of one entity with respect to the position of another.
- Technique specification: The measurement specifies an aspect of the image data or one of its reconstructions.

Clinical finding measurements, or "finding measurements", were subclassified in the following three classes:

- Current: The measurement refers to the current exam and not to the prior exam
- Prior: The measurement refers to the prior exam and not to the current exam.
- Comparison: The measurement refers to both the current and the prior exam within the scope of the sentence in which it appears.

The classification scheme is presented and illustrated by means of sample sentences in ▶Figure 1.

Measurements of class comparison, or "comparison measurements", typically describe the dimensions of its finding on the current exam, but are compared qualitatively to the dimensions of the same finding the prior exam. Using a comparison measurement may be stylistically preferred to mentioning the dimensions of a finding on current and prior, when they are identical. That is, to avoid constructions like: "There is a 1.6 x 0.9 cm lytic focus in the right iliac wing, measuring 1.6 x 0.9 cm on prior exam."

It is important to note that temporality of a measurement is solely determined by contents of the sentence in which it appears, not a wider context. Thus, according to this definition, the measurement in the first sentence would be considered a comparison measurement whereas the measurement in the second sentence would be considered current:

- Submandibular lymph node measures 5 x 5 mm, grossly unchanged since prior examination.
- Submandibular lymph node measures 5 x 5 mm. Grossly unchanged since prior examination.

Obviously, the above two fragments reflect the same radiologic reality, but this is not accounted for in the described approach.

In our corpus, all technique specification measurements appear in the technique sections of the report. Since we shall only deal with measurements from the finding sections, this class will henceforth be ignored. Thus, every measurement has one of the following four classes: relative position (rel-pos), current, prior and comparison.

2.3.2 Development

The measurement classification engine consumed a measurement and converted it into an instance representation based on the measurement's internal structure and narrative context, i.e. the sentence in which it appears called the "containing sentence". This instance representation was essentially a series of binary features. Once converted, the class of a given instance representation was determined based on a maximum entropy model, optimized on the four-class classification task based on training instances [21].

For each of the four classes, a feature encoded if the measurement appeared in a narrative context that is typically associated with that class. For instance, measurements appearing in the context of "there is a …" are generally current, whereas measurements appearing in the context of "previously measuring …" are generally prior. The narrative contexts are detected by means of regular expressions that account for lexical variations, which were constructed using the set of 200 reports disjoint from the ground truth.

For each class but rel-pos, a feature encoded if one or more keywords appeared in the containing sentence. For instance, we had "today", "currently" and "now" as keywords for class current and "unchanged", "larger" and "similar" for comparison. The lists of keywords were composed using a corpus of reports disjoint from the ground truth. ►Appendix 2 lists the narrative contexts and lists of keywords.

Finally, features represented the dimensionality of the measurement (e.g. "5 x 5 mm" is a two-dimensional measurement), and the number of measurements appearing in its containing sentence. For instance, from the sentence "The left kidney measures 13.1 cm with increased cortical echogenicity" the following binary features were extracted: measurement-is-1-dimensional, sentence-contains-1-measurement, measurement-appears-in-present-context ("measures …"), and sentence-contains-stable-keyword ("increased").

2.3.3 Evaluation

A ground truth of 2000 labeled instances was created based on randomly selected sentences with measurements, independently of the sentences and measurements used for evaluating the measurement extraction engine. A subset of 200 instances was randomly drawn in such a way that 100 instances in the sample were classified as current, 40 as prior, 40 as comparison and 20 as rel-pos. A third-year radiology resident annotated this subset. The result is compared to the annotation of the same subset in the ground truth using multi-class κ metric, a chance-adjusted measure for inter-rater agreement [22].

The measurement classification engine was evaluated against the ground truth in a 10-fold cross validation protocol, yielding a four-by-four confusion matrix. From this matrix follows the engine's performance on several binary classification problems that were derived from the original four-class classification problem. Each such binary classification problem was obtained by lumping together one or more classes as the positive class (e.g. rel-pos) and lumping together the remaining classes as the negative class (e.g. current, prior, comparison). We determined the engine's precision, recall, F-measure, accuracy (number of correctly classified instances divided by total number of instances) and κ on each binary decision problem.

The measurement classification engine was previously compared against an (inferior) rule-based algorithm [23].

3. Results

3.1 Measurement extraction

In an initial evaluation, precision of the measurement extraction engine was 0.961 (1922/2000). Three types of errors stood out:

- Velocity or pressure measurements: "3.2 mmHg" or "3.2 mm/sec." The initial recognizer would mark "3.2 mm" in both instances. This occurred 19 times.
- Spaces inside measurements: "2.1 x 3 .2 cm." The initial recognizer would match the string "2.1 x 3," which is incomplete. This occurred 12 times.
- Units written out: "3.2 centimeters." This occurred four times.

In a second evaluation on another 1000 instances, after generalizing the regular expression so as to avoid the aforementioned errors (the result of which is given in Appendix 1), only six instances were incorrectly retrieved. This amounts to a precision of 0.994 (994/1000). In the 200 reports, 891 measurement measurements were found, seven of which were not marked by the measurement recognizer. Thus, recall was 0.991 (884/891).

3.2 Measurement classification

Inter-rater agreement on the 200-instance random subset of the ground truth was κ=0.933 (95% confidence interval: 0.890–0.976). The confusion matrix of the classifier is given in ▶Table 1. Accuracy of the classifier is 0.960 ([1365 + 166 + 259 + 133]/2000). The confusion matrices for discriminating each class from the other classes are given in ▶Tables 2–5, respectively. In a plausible use case scenario, the classification engine will be used to filter out new finding measurements, i.e. current and comparison measurements. The confusion matrix for separating current and comparison measurements from prior and rel-pos measurements is presented in ▶Table 6. The performance of the classification engine on the derived binary classification problems is given in ▶Table 7.

4. Discussion

4.1 Measurement extraction

Performance of the measurement extraction engine is very robust, showing that pattern recognition techniques suffice for recognizing measurements in radiological narrative. Errors were primarily due to uncommon ways of expressing measurements, as are exemplified by the following sentences in which the underlined substrings mark what was (incorrectly) extracted by the engine:

- … previously 4-mm on 5-Nov-2011 by 8mm …
- … now measures 8.6 x 5.9 cm transaxial x 4.4 cm craniocaudal …
- It measures today 5 x 1–2 x less than 2 mm …

4.2 Measurement classification

Inter-rater agreement is satisfactory (κ=0.933), indicating that the measurement classification scheme's concepts were well defined and that the ground truth creation process was reproducible. Disagreement mostly concerned confusion between comparison instances on the one hand and current and prior instances on the other hand.

The measurement classification engine has near-perfect performance on separating relative position measurements from finding measurements (current/prior/comparison). In the one-against-all evaluation of rel-pos versus the finding measurement classes, F-measure was 0.985. Measurements from the third top-level concept in the classification scheme (technique specification measurements) can be detected easily as they only appear in technique sections, which can be recognized by an SBD engine. We conclude that natural language processing techniques can be used reliably to automatically distinguish the top-level concepts in the measurement classification scheme (▶Figure 1).

Comparison measurements are hardest to distinguish for the classification engine. In the one-against-all evaluation of comparison against the other classes, the engine's F-measure is 0.853, which is substantially poorer than its performance on the other two temporal orientation classes: F-measure 0.973 (current) and 0.957 (prior).

Confusion between current and comparison is the main source of error: 34 of the 80 misclassified instances are current mistaken for comparison, whereas 19 are misclassified the other way around. The third error category is constituted by 17 previous instances misclassified as current. These three error categories thus comprise 87.5% (70/80) of all misclassified instances. We discuss each error category below.

Error category 1: current instances misclassified as comparison, accounting for 42.5% (34/80) of all misclassified instances. The majority (82.3% [28/34]) of these misclassified measurements appeared as the sentence's sole measurement, in a present context and together with a comparison keyword. This combination of features is apparently picked up as a typical way of expressing comparative measurements, such as "Submandibular lymph node measures 5 x 5 mm, grossly unchanged since prior examination". Upon close examination, in the majority of these misclassified instances (67.9% [19/28]), the comparison keyword did not concern measurements ("Increased number of mediastinal lymph nodes, measuring up to 8 mm in short axis") or not the particular measurement at hand ("No polyps 6 mm or larger seen anywhere in the colon"). This error source can be addressed by automatically detecting the semantic scope of the comparison keyword and ensuring that the measurement at hand is included therein.

Error category 2: comparison instances misclassified as current, accounting for 23.8% (19/80) of all misclassified instances. The majority (57.9% [11/19]) of these measurements appeared as one of the sentence's two measurements, in a present context and together with a comparison keyword. A typical present measurement is the underlined string in the sentence "Submandibular lymph node measures 5 x 5 mm, previously measuring 7 x 7 mm". It appears in a present context as one of two measurements and the sentence contains a comparison keyword. However, this reasoning pattern is incorrect if the comparison keyword addresses both measurements ("A 7-mm mass within the right upper outer quadrant and a 5-mm mass in the lateral left breast are stable"), which happened 11 times. This error source could be addressed by checking for the appearance of keywords like "both" and conjunctions of measured lesions ("… and …") inside the sentence.

Error category 3: previous instances misclassified as current, accounting for 21.3% (17/80) of all misclassified instances. The vast majority of these instances (88.2% [15/17]) appeared in a previous context that the detector did not picked up because of ungrammaticalities ("These measurements are prior exam of 22.2 x 28.6-mm") or lexical patterns that were not accounted for (e.g. the context pattern "as compared with … previously" in "Previously measured left periaortic lymph node currently measures 0.8 x 0.5 cm, as compared with 1.0 x 1.0 cm previously"). We hypothesize that the classifier assigns present as this is the most prevalent class for instances without previous context evidence.

4.3 Limitations

The ground truth was based on the reports of one single institution, which may have a measurement reporting style that is not necessarily consistent with that of other US-based institutes. The ground truth for the measurement extraction task was not assessed for inter-rater agreement and the ground truth for the measurement classification task was assessed by one reviewer. The work presents the results of a maximum entropy classifier, which is common in natural language processing domains with Boolean variables. However, comparison to classifiers from other paradigms may have been helpful for appreciating its achievements. Finally, to assess recall of the measurement extraction engine, reports with varying numbers of measurements were selected. The error analysis was not broken down along those lines as we considered that this factor was of insufficient interest to be presented in this work.

5. Conclusion

Accumulating quantitative outcome parameters may contribute to constructing a healthcare organization in which outcomes of clinical procedures are reproducible and predictable. In imaging studies, measurements are the principal category of quantitative parameters. We developed two natural language processing engines that extract and classify measurements from narrative radiology reports, respectively. Both engines displayed satisfactory performance in a formal evaluation. The more knowledge-intensive components of the engines have been made available in two technical appendices. The work contributes to machine understanding of radiology reports and may find application in software applications that process medical data.

CLINICAL RELEVANCE STATEMENT

Natural language processing methods were described to extract and classify finding measurements in narrative radiology reports. The work contributes to machine understanding of radiology reports and may find application in software applications that process medical data.

Conflict of interest

The authors declare that they have no conflicts of interest in this project.

PROTECTION OF HUMAN AND ANIMAL STUDIES

Human subjects were not included in this project.

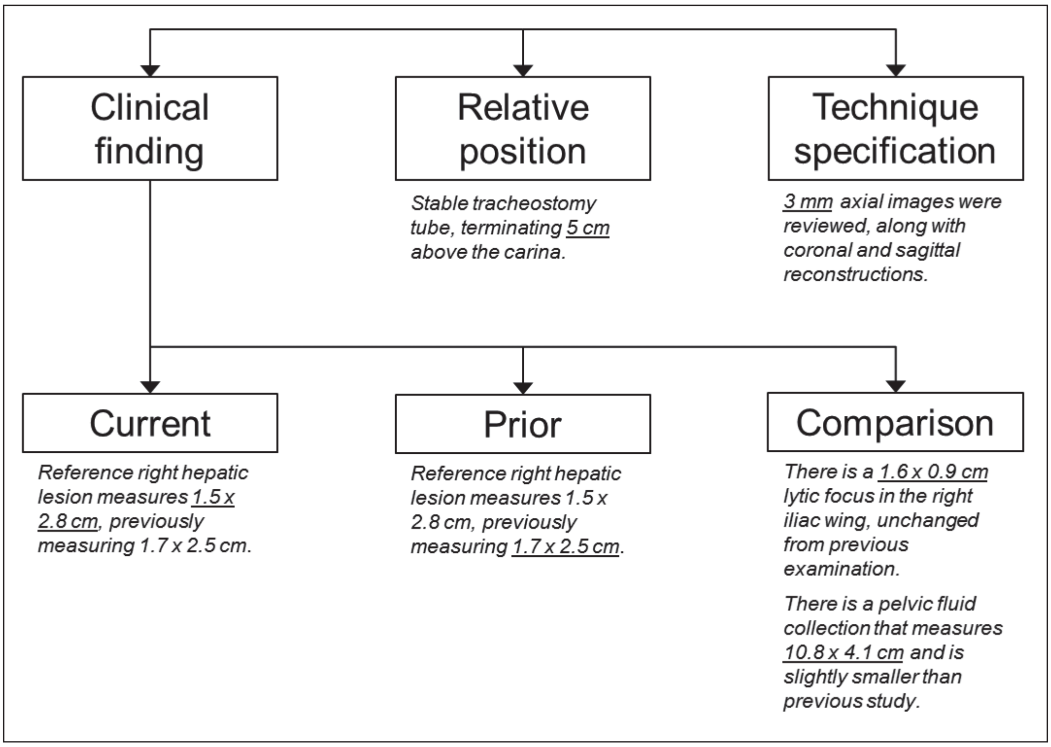
Clinical finding
Relative position
Technique specification
Stable tracheostomy tube, terminating 5 cm above the carina.
3 mm axial images were reviewed, along with coronal and sagittal reconstructions.
Current
Prior
Comparison
Reference right hepatic lesion measures 1.5 x 2.8 cm, previously measuring 1.7 x 2.5 cm.
Reference right hepatic lesion measures 1.5 x 2.8 cm, previously measuring 1.7 x 2.5 cm.
There is a 1.6 x 0.9 cm lytic focus in the right iliac wing, unchanged from previous examination.
There is a pelvic fluid collection that measures 10.8 x 4.1 cm and is slightly smaller than previous study.

Fig. 1 Measurement classification scheme – Each class is illustrated by a sample sentence in which the measurement of that class is underlined.

Table 1 Confusion matrix of the measurement classification engine.

		Ground truth				
		Current	Comparison	Prior	Rel-pos	SUM
Predicted	Current	1 365	19	17	3	1 404
	Comparison	3	166	2	0	202
	Prior	34	1	259	0	263
	Rel-pos	0	1	0	130	131
	SUM	1 402	187	278	133	2 000

Table 2 Derived confusion matrix for recognizing current measurements.

		Ground truth		
		Current	Not current	SUM
Predicted	Current	1 365	39	1 404
	Not current	37	559	596
	SUM	1 402	598	2 000

Table 3 Derived confusion matrix for recognizing prior measurements.

		Ground truth		
		Prior	Not prior	SUM
Predicted	Prior	259	4	263
	Not prior	19	1 718	1 737
	SUM	278	1 722	2 000

Table 4 Derived confusion matrix for recognizing comparison measurements.

		Ground truth		
		Prior	Not prior	SUM
Predicted	Prior	259	4	263
	Not prior	19	1 718	1 737
	SUM	278	1 722	2 000

Table 5 Derived confusion matrix for recognizing rel-pos measurements.

		Ground truth		
		Rel-pos	Not rel-pos	SUM
Predicted	Rel-pos	130	1	131
	Not rel-pos	3	1 866	1 869
	SUM	133	1 867	2 000

Table 6 Derived confusion matrix for separating current and comparison measurements from prior and rel-pos measurements.

		Ground truth		
		Current or comparison	**Prior or rel-pos**	**SUM**
Predicted	**Current or comparison**	1 553	22	1 575
	Prior or rel-pos	36	389	425
	SUM	1 589	411	2 000

Table 7 Performance of the measurement classification engine on the four one-against-all problems. Each column represents one derived problem in which the column´s header defines the positive instances.

	Current	Comparison	Prior	Rel-pos	Current and comparison
Precision	0.972	0.822	0.985	0.992	0.986
Recall	0.974	0.888	0.932	0.977	0.977
F-measure	0.973	0.853	0.957	0.985	0.982
Accuracy	0.962	0.972	0.989	0.998	0.971
κ	0.909	0.838	0.951	0.984	0.909

References

1. Sullivan DC. Imaging as a Quantitative Science. Radiology 2008; 248(2): 328–332.
2. Sevenster M, Bozeman J, Cowhy A, Trost W. A natural language processing pipeline for pairing measurements uniquely across free-text CT reports. J Biomed Inf 2015; 53: 36–48.
3. Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inf. Assoc 1994; 1: 161–174.
4. Dreyer KJ, Kalra MK, Maher MM, Hurier AM, Asfaw BA, Schultz T, Halpern EF, Thrall JH. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. Radiology 2005; 234(2): 323–329.
5. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inf Assoc 2008; 15(1): 87–98.
6. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 2010; 10: 70.
7. Tian Z, Sun S, Eguale T, Rochefort C. Automated Extraction of VTE Events From Narrative Radiology Reports in Electronic Health Records: A Validation Study. Med Care 2015.
8. Elkin PL, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma H, Asatryan AX, Tokars JI, Rosenbloom ST, Brown SH. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annu Symp Proc 2008; 172–176.
9. Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH. A Text Processing Pipeline to Extract Recommendations from Radiology Reports. J Biomed Inf 2013; 46(2): 3543–3562.
10. Asatryan A, Benoit S, Ma H, English R, Elkin P, Tokars J. Detection of pneumonia using free-text radiology reports in the BioSense system. Int J Med Inf 2011; 80(1): 67–73.
11. Mabotuwana T, Qian Y, Sevenster M. Using image references in radiology reports to support enhanced report-to-image navigation. AMIA Annu Symp Proc 2013; 908–916.
12. Sevenster M, van Ommering R, Qian Y. Automatically Correlating Clinical Findings and Body Locations in Radiology Reports Using MedLEE. J Digit Imaging 2012; 25(2): 240–249.
13. Friedman C. A broad-coverage natural language processing system. AMIA Annu Symp Proc 2000; 270–274.
14. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inf Assoc 2005; 12(4): 448–457.
15. He J, de Rijke M, Sevenster M, van Ommering R, Qian Y. Generating Links to Background Knowledge: A Case Study Using Narrative Radiology Reports. Conference on Information and Knowledge Management 2011.
16. Rochefort C, Verma A, Eguale T, Lee T, Buckeridge D. A novel method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data. J Am Med Inf Assoc 2015; 22(1): 155–165.
17. World Health Organization. WHO handbook for reporting results of cancer treatment. World Heal Geneva (Switzerland), 1979.
18. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, van Oosterom AT, Christian MC, Gwyther SG. New guidelines to evaluate the response to treatment in solid tumors. J Natl Cancer Inst 2000; 92: 205–216.
19. Jaffe TA, Wickersham NW, Sullivan DC. Quantitative imaging in oncology patients: Part 1, radiology practice patterns at major U.S. cancer centers. Am J Roentgenol 2010; 195: 101–106.
20. Jaffe TA, Wickersham NW, Sullivan DC. Quantitative imaging in oncology patients: Part 2, oncologists' opinions and expectations at major U.S. cancer centers. Am J Roentgenol 2010; 195(1): 19–30.
21. Nigam K. Using maximum entropy for text classification. IJCAI-99 Workshop on Machine Learning for Information Filtering 1999; 61–67.
22. Carletta J. Assessing agreement on classification tasks: The kappa statistic. Comput Linguist 1996; 22: 249–254.
23. Sevenster M. Classifying measurements in dictated, free-text radiology reports. AIME Symp Proc 2013; 310–314.