Unsupervised Topic Modeling in a Large Free Text Radiology Report Repository

Saeed Hassanpour¹ · Curtis P. Langlotz¹

Published online: 9 September 2015
© Society for Imaging Informatics in Medicine 2015

Abstract Radiology report narrative contains a large amount of information about the patient's health and the radiologist's interpretation of medical findings. Most of this critical information is entered in free text format, even when structured radiology report templates are used. The radiology report narrative varies in use of terminology and language among different radiologists and organizations. The free text format and the subtlety and variations of natural language hinder the extraction of reusable information from radiology reports for decision support, quality improvement, and biomedical research. Therefore, as the first step to organize and extract the information content in a large multi-institutional free text radiology report repository, we have designed and developed an unsupervised machine learning approach to capture the main concepts in a radiology report repository and partition the reports based on their main foci. In this approach, radiology reports are modeled in a vector space and compared to each other through a cosine similarity measure. This similarity is used to cluster radiology reports and identify the repository's underlying topics. We applied our approach on a repository of 1,899,482 radiology reports from three major healthcare organizations. Our method identified 19 major radiology report topics in the repository and clustered the reports accordingly to these topics. Our results are verified by a domain expert radiologist and successfully explain the repository's primary topics and extract the corresponding reports. The results of our system provide a target-based corpus and framework for

information extraction and retrieval systems for radiology reports.

Keywords Topic modeling · Radiology report narrative · Clustering · Text mining · Natural language processing

Introduction

Radiology reports are the major source of information in clinical imaging. These reports contain vital information about many anatomic and pathophysiologic conditions and communicate this information to the referring physicians. Radiology reports, in combination with electronic medical records and genomics data, can explain relations between risk factors and clinical conditions and provide a better understanding of clinical conditions through clinical research.

The majority of report information content, even in presence of structured templates, is captured in free text. The free text format of radiology reports is the major obstacle to identifying similar groups of reports in a large data repository and finding relevant radiology reports in a specific domain for biomedical research. This difficulty is primarily due to the wide range of observations and findings in radiology reports, the complexity of natural language, and variations among different radiologists and organizations [1, 2]. As a result, radiology reports are used only once by the clinician who ordered the study and are rarely used again for clinical research or data mining.

Natural language processing (NLP) techniques have been previously applied on radiology reports to classify and extract their information contents [3–18]. For instance, the Lexicon Mediated Entropy Reduction (LEXIMER) system extracts and classifies phrases with important findings and recommendations from radiology reports through lexicon-based hierarchical decision trees [4]. In another work, Medical Language

✉ Saeed Hassanpour
saeedhp@gmail.com

¹ Department of Radiology, Stanford University, 300 Pasteur Drive, Stanford, CA 94305, USA

Extraction and Encoding System (MEDLEE) was developed to extract information from Columbia-Presbyterian Medical Center's chest radiology report repository [6]. MEDLEE uses a controlled vocabulary and grammatical rules to translate text to a structured database format. Mayo Clinic's Clinical Text Analysis and Knowledge Extraction System (cTAKES) provides a named-entity recognizer to lookup and highlight the Unified Medical Language System (UMLS) Metathesaurus terms in text, in addition to other NLP functionalities, such as tokenizing, part of speech tagging and parsing [11].

These and the majority of other NLP systems that are applicable to radiology reports rely on domain-specific dictionaries and rules for annotation and information extraction [14]. Considering the time and resources required to build NLP systems, identifying the major topics in a radiology report repository and organizing the repository according to these topics can improve NLP system performance [19]. Therefore, our goal is to provide an automatic method to recognize the topic of a radiology report repository and group the reports according to these topics. For example, it may be wise to train separate NLP systems for brain computed tomography and musculoskeletal MRI, due to the differences in modality and anatomic focus.

To achieve this goal, we propose an unsupervised machine learning approach to cluster the radiology reports based on their free text contents. This clustering reveals the major topics in the radiology report repository and detects the reports that belong to each topic's domain. Our unsupervised machine learning approach does not require any manual training data. The number of topics in the repository is decided by a domain expert radiologist based on clustering quality criteria. We applied our approach on a large radiology report repository from three major healthcare organizations. Our evaluation shows the efficacy of our approach in identifying the major topics in the large multi-institutional radiology report repository.

Materials and Methods

Data Set

In this work, we use RadCore radiology report repository to build and evaluate our topic modeling system. RadCore is assembled in 2007 at Stanford as an effort to construct a large multi-institutional radiology report corpus for NLP. RadCore contains 1,899,482 reports from three major healthcare organizations: Mayo Clinic (812 reports), MD Anderson Cancer Center (5000 reports), and Medical College of Wisconsin (1,893,670 reports). The reports belong to 1995 to 2006 time period and were de-identified by their source organizations before the submission to RadCore. All radiology reports used in this work are in free text format and do not contain any metadata about the type and nature of the imaging exams. The

use of RadCore radiology report repository in this project was approved by Stanford Institutional Review Board (IRB).

Computational Platform

Given our goal to build a NLP topic modeling system that is applicable to a large number of radiology reports from various organizations, we use a distributed computational and data storage platform. Our algorithm for topic modeling is based on Apache Mahout open-source distributed machine learning library [20]. Mahout is widely used in academic and commercial NLP applications and uses Apache Hadoop [21] as its underpinning computational and data storage platform. Hadoop is an open-source implementation of MapReduce programming and data storage model [22]. Through MapReduce model, Hadoop provides an efficient, distributed, and fault-tolerant platform for executing algorithms and storing data on a computer cluster.

Data Modeling

We model each radiology report as a vector in Euclidean space. Each dimension of the vector represents a term in the radiology report repository. The term weights in radiology report vectors are calculated using a common weighting scheme in text mining, called term frequency-inverse document frequency (tf-idf) [23]. The tf-idf weight for a term in a radiology report vector is positively correlated with the frequency of the term in the report and is scaled down by the commonality of the term among all radiology reports in the repository. As the result of this modeling process, the radiology reports are represented in a mathematical format in a vector space and can be processed by our NLP algorithm.

Clustering

To automatically identify different domain topics in the radiology report repository and organize the reports according to these topics, we applied k-means clustering algorithm on the vector representations of the radiology reports. The clustering algorithm split the repository to separate partitions based on the similarity among radiology reports [24]. As a result, the radiology reports in each cluster represent a distinct topic in the repository. Among different clustering algorithms, k-means clustering algorithm is one of the most efficient and popular clustering algorithms in data mining [25]. K-means clustering algorithm partitions the vector space to a predetermined number of clusters through iterative refinement steps. To start, a centroid is chosen randomly for each cluster. Then, the algorithm assigns the report vectors to the cluster with the closest centroid, update the centroids based on new clusters, and repeat the process till convergence. The

algorithm converges when clusters do not change in the algorithm's iterative update process.

In order to calculate similarity between radiology report vectors and cluster centroids in k-means clustering algorithm, we use cosine similarity metric [26]. The cosine similarity for two vectors is equivalent to the cosine of the angle between them. The cosine similarity value ranges from 0 for orthogonal vectors to 1 for identical vectors. Because cosine similarity for two vectors does not depend on their size, it provides an accurate and stable similarity measure compared to Euclidean distance in high dimensional vector space. As discussed in the previous section, for scalability and reliability purposes and to manage the computational requirements of k-means clustering algorithm for our large radiology report repository, we used Apache Mahout machine learning library to implement k-means clustering algorithm in this work.

Clustering Evaluation

The number of clusters is required as an input for k-means clustering algorithm. Because the number of clusters in our radiology report repository is unknown, we determined this number through domain expert evaluation. In this evaluation, we ran k-means clustering with different number of clusters from 3 to 35 and one of the authors (CL), a board certified radiologist with more than 20 years of clinical experience, reviewed the clustering results to identify the best clusters as the final partitioning for the repository. In this evaluation, the top 30 terms for each cluster are reviewed by the domain expert. These terms have the highest tf-idf weights in cluster centroid vectors and represent their corresponding clusters. In addition, randomly selected radiology reports from different clusters are reviewed to confirm the validity of the clusters. In this review, the homogeneity and distinction of the clusters are assessed by the domain expert to evaluate the clustering quality.

Results

Nineteen clusters were identified as optimal RadCore radiology report repository clusters in our method. The domain expert confirmed that each of the 19 clusters represents a major radiology report topic. The topic names and the number of reports in each cluster are shown in Table 1.

Discussion

We have described a machine learning method to detect the major radiology report topics based on reports' free text content in a large multi-institutional repository. Our approach successfully identified the underlying topics in radiology reports from three major healthcare organizations, which

Table 1 Radiology report repository clusters

Number	Cluster name	Number of reports
1	XR chest 2 views	383,768
2	Spine imaging	188,933
3	CT brain	171,350
4	US retroperitoneum	142,596
5	XR chest anteroposterior (AP)	131,145
6	XR musculoskeletal	103,827
7	XR lower extremity	101,890
8	XR gastrointestinal	92,708
9	XR musculoskeletal post-operative	89,275
10	MG mammography	72,385
11	IR interventional radiology	71,378
12	CT abdomen	70,916
13	XR upper extremity	67,947
14	NM cardiac stress	57,306
15	MG mammography screening	49,077
16	NM cardiac	45,499
17	US vascular Doppler	26,519
18	DXA bone densitometry	21,134
19	FL upper gastrointestinal	11,829

XR radiography, *CT* computed tomography, *US* ultrasound, *NM* nuclear medicine, *DXA* dual energy x-ray absorptiometry, *IR* interventional radiology, *FL* fluoroscopy

constitute an intuitive set of report topics. The number of associated reports for each topic reflects the expected distribution of patients among these topics. Our topic modeling approach is based on an unsupervised machine learning algorithm and does not require manually labeled data. Therefore, it can be applied easily to other free text data repositories. In addition, our approach uses a distributed machine learning implementation, which is scalable and can handle very large free text corpora.

Our topic modeling method produces a detailed picture of a free text radiology report corpus, which helps biomedical researchers better understand the characteristics of their data. The resulting radiology report clusters provide topic-specific corpora for data mining research, data management, and data pruning for a particular domain. Particularly, these topic-specific corpora could facilitate building domain-specific dictionaries and rule bases in NLP systems. Our work can improve the performance of data driven biomedical information systems such as information retrieval, information extraction, risk prediction, and healthcare surveillance systems through extracting topic-related subsets in a large free text corpus.

In our approach, we rely on domain expert evaluation to decide the optimal number of clusters. As a future work, we are planning to develop and evaluate statistical and information-theoretic techniques to specify the number of clusters automatically. We are also planning to apply our

method on radiology reports from other data sources including the Stanford Translational Research Integrated Database Environment (STRIDE). STRIDE contains more than four million radiology reports from the Stanford Health Care electronic health record system dating back to 1998. In addition, we plan to extend our work by implementing an extrinsic evaluation of our method as a part of an information extraction application.

Conclusion

We have described an unsupervised machine learning approach to group radiology reports by major topics in a large multi-institutional repository of narrative radiology reports. Our approach models radiology reports as vectors and partitions the repository in separate clusters through k-means clustering algorithm based on cosine similarity metrics. A domain expert radiologist evaluated our clusters and verified the efficacy of our results. Our results identified 19 major topics in the radiology report repository and their corresponding reports. Our approach does not require any training data and leverages a MapReduce distributed platform for scalable and stable clustering of large radiology report repositories. Our system can assist researchers in managing large free text radiology report repositories and providing domain-specific radiology report corpora to develop and improve data mining and NLP techniques.

Acknowledgments The authors would like to thank Chuck Kahn, Kevin McNery, and Brad Erickson for their work on compiling RadCore database and Daniel Rubin for his contribution to RadCore and providing access to this database.

References

1. Sobel JL, Pearson ML, Gross K, Desmond KA, Harrison ER, Rubenstein LV, Rogers WH, Kahn KL: Information content and clarity of radiologists' reports for chest radiography. *Acad Radiol* 3(9):709–17, 1996
2. Khorasani R, Bates DW, Teeger S, Rothschild JM, Adams DF, Seltzer SE: Is terminology used effectively to convey diagnostic certainty in radiology reports? *Acad Radiol* 10(6):685–8, 2003
3. 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* 234(2):323–9, 2005
4. Dreyer KJ: Information theory entropy reduction program. U.S. Patent 8,756,234, 2014
5. Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH: A text processing pipeline to extract recommendations from radiology reports. *J Biomed Inform* 46(2):354–62, 2013
6. Friedman C, Hripcsak G, DuMouchel W, Johnson SB, Clayton PD: Natural language processing in an operational clinical information system. *Nat Lang Eng* 1(01):83–108, 1995
7. Hripcsak G, Austin JHM, Alderson PO, Friedman C: Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology* 224(1):157–63, 2002
8. Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF: A reliability study for evaluating information extraction from radiology reports. *J Am Med Informatics Assoc* 6(2):143–50, 1999
9. Elkins JS, Friedman C, Boden-Albala B, Sacco RL, Hripcsak G: Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. *Comput Biomed Res* 33(1):1–10, 2000
10. Johnson DB, Taira RK, Cardenas AF, Aberle DR: Extracting information from free text radiology reports. *Int J Digit Libr* 1(3):297–308, 1997
11. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Informatics Assoc* 17(5):507–13, 2010
12. Goryachev S, Sordo M, Zeng QT: A suite of natural language processing tools developed for the I2B2 project. In: *AMIA Annual Symposium Proceedings*, 2006, p 931
13. Aronson AR: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: *Proceedings of the AMIA Symposium*, 2001, p 17
14. Taira RK, Soderland SG: A statistical natural language processor for medical reports. In: *Proceedings of the AMIA Symposium*, 1999, p 970
15. Taira RK, Soderland SG, Jakobovits RM: Automatic structuring of radiology free-text reports. *Radiographics* 21(1):237–45, 2001
16. Haug P, Koehler S, Lau LM, Wang P, Rocha R, Huff S: A natural language understanding system combining syntactic and semantic techniques. In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 1994, p 247
17. Haug PJ, Koehler S, Lau LM, Wang P, Rocha R, Huff SM: Experience with a mixed semantic/syntactic parser. In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 1995, p 284
18. Christensen LM, Haug PJ, Fiszman M: MPLUS: a probabilistic medical language understanding system. In: *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain*, vol. 3, 2002, pp 29–36
19. Friedman C, Rindfleisch TC, Corn M: Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *J Biomed Inform* 46(5):765–773, 2013
20. Apache Mahout. Available at <http://mahout.apache.org>. Accessed 24 March 2015
21. Apache Hadoop. Available at <https://hadoop.apache.org>. Accessed 24 March 2015
22. Dean J, Ghemawat S: MapReduce: simplified data processing on large clusters. *Commun ACM* 51(1):107–13, 2008
23. Manning CD, Raghavan P, Schütze H: Scoring, term weighting, and the vector space model. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008, p 100
24. Kaufman L, Rousseeuw PJ: Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, New York, 2009
25. Lloyd S: Least squares quantization in PCM. *Inf Theory, IEEE Trans* 28(2):129–37, 1982
26. Singhal A: Modern information retrieval: A brief overview. *IEEE Data Eng Bull* 24(4):35–43, 2001