International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Natural Language-Based Machine Learning Models for Information Extraction from Radiology Reports - Survey

Jewel Sengupta

Visvesvaraya Technological University, Master of Computer Applications, Bangalore, Karnataka 560048, India

Abstract: ***In the current digital world, most of the patients records are stored in the form of electronic health record, where vast amount of digital contents are reported based on the radiological information. The radiological reports are very important source about the patient and helps to researchers to improve the health care departments. The radiological reports are collected and saved for documentation and communication of image diagnosing. Since the radiological information are stored in a free text format, hence it requires appropriate automated information extraction to retrieve the structured data which helps the physician for decision making. Natural language processing (NLP) is the important technique that helps to attain the structured representation of radiological reports. The structured data further processed by the machine learning (ML) algorithm for classification purpose which helps the physician for better the decision making. In this review, the NLP and ML techniques are considered for handling the radiological reports. In this review, the list of approaches for the automatic classification of radiological reports are identified and gathered into four major ways which are rules based approach, machine learning based approach, and hybrid approach. Moreover, the drawbacks and the upcoming challenges, future scopes for enhancing the NLP functionality in radiology is described.***

Keywords: Natural Language Processing, Electrical Health Records, Machine Learning, Classification, Radiology Reports

1. Introduction

Due to the rapid growth of the electronic health records (EHRs), there is an essential requirement to design and develop an automatic information retrieval system to retrieve the information and knowledge from HER for clinical support and translational research. Now a days number of research are increased rapidly based on the adoption of EHRs by huge number of healthcare institutions announced by the Health Information Technology for Economic and Clinical Health Act (HITECH Act) legislation [1]. The utilization of EHRs for review and evaluate the list of methodologies and reporting is conducted in [2] and the systematic review of health care estimation based on the HER data conducted in [3]. However, the most the HER reports are in the form of free-text [4]. When compared with the structured reports, handling of free text radiological reports are more expensive to document clinical events and managing the communication among the health care team. In order solve the above problems, the information extraction (IE) technique helps to simplify the usage of HER information for healthcare decision support and improving the quality of results.

The IE is specialized area in empirical natural language processing (NLP) which enables automatically extraction and encoding of the clinical information from radiological reports. With the assist of NLP technique, the IE not only extracting the information also it identifies the entities, concepts, events and discovers the relations among the attributes [5]. NLP analyse the free-form text based on the linguistics and statistical approaches which discover the rules and patterns form the EHRs data. The NLP initiates its progress by analysing the text to discover the each distinct concept subsequently the useful concepts are extracted based on the feature extraction which will be formulated in a structured format. The next step is to conclude whether the obtained structured data from the EHRs contains or more desired concepts or not. If the obtained results contain more than one concepts which are handled by the set of clinical rules which are constructed by the domain experts or by using statistical or machine learning approaches to automatically infer rules and patterns from a huge list of data [6].

In general, the NLP concentrates on the design and development of computational models for understanding the natural language based on the set of modules such as syntactic processing modules, semantic processing modules [7]. In the health care field, the researchers are interested to utilize the NLP technology to extract the useful information from the EHR for automatic decision support system [8], discharge summaries [9], problem lists [10], nursing documentation [11], and medical education documents [12]. Different NLP techniques was developed and used to retrieve the useful information from the radiology reports such as events and medical concepts. The NLP technique based information retrieval applications such as MedLEE [13], MetaMap [14], KnowledgeMap [15], cTAKES [16], HiTEX [17], and MedTagger [18].

2. Literature Review

Most of the researchers used of NLP technique to extract the information from EHRs specific to radiology. The key objective for this review is to understand the how the NLP techniques are utilized for extracting the useful information from the radiology reports in a systematic, up-to-date overview.

2.1 Syntactic and semantic analysis

MelihaYetisgen-Yildiz et. al. [19] presented a pipeline for text processing which helps to identify the clinical information from the radiology reports automatically. The proposed text processing pipeline has three major modules. The first one is section segmenter, which helps to determine section of the given radiology report. The second module is sentence segmenter which identifies boundaries of each sentence in the identified section. The third module is the binary classifier, that helps to identify the whether the sentences belongs to positive recommendation sentence or negative one. The major contribution of the proposed text processing pipeline is statistical section segmentation method which can adoptable to radiology reports from other institutions. The statistical feature selection methodology is utilized to improve the throughput of classification process. With the intention of enhance the performance of feature selection methodology, various feature sets are utilized for comparison such as N-gram, and UMLS concept features.

Saeed Hassanpour and Curtis P. Langlotz [20] presented information model provides a framework to summarize radiology reports, The proposed information model presented efficient feature selection process such as n-grams, name-entity recognition and part of speech tags. In order to enhance the quality of clinical research and better decision support for physician, the proposed IE system is optimised to extract the useful information and arrange the results in an appropriate manner. In order to classify the radiology reports, the extracted information is subjected to classification algorithm, to achieve that, they utilized two machine learning algorithms such as Conditional Markov Model (CMM) , Conditional Random Field model (CRF). In order to evaluate the outcome of the proposed methodology, they have collected more than 150 radiology documents from various health care organizations. The comparison of the proposed methodology is done with the non-machine learning based classification methodologies.

Paras Lakhani et. al. [21] presented a text mining algorithm methodology to automatically distinguish radiology reports. By considering the set of syntactic features helps to expanded searches and useful to identify the applicable synonyms. In order to classify the critical results of unstructured radiology, they have considered pattern based approaches helps to attain the maximum classification accuracy. In order to represent the critical results, The proposed pattern based approach is design and developed for discovering the common words and phrases. The proposed methodology can be utilized in order to obtain the appropriate results of the radiology reports in a real time monitoring by combining with the existing dashboards.

Yan Xu et. al. [22] presented a hybrid feature extraction methodology to extract the useful information from the radiology. The proposed methodology of hybrid feature extraction integrates two types of feature extraction methodologies such as orthographic and semantic features. Followed by the feature extraction, the classification process made based on the hybridization of both classifiers such as Labelled Sequential Pattern (LSP) classifier with a CRF recognizer was devised. The result of the proposed information extraction methodology represents orthographic features alone achieved performance comparable with that of a semantic features system when the training dataset was large enough. The LSP classifiers used to removes the unnecessary words, on the other hand the CRF classifier extracts the useful information as a concepts from candidate sentences based on the classifier.

Dorothy A. Sippo et. al [23] utilized open source NLP tool BROK to regulate BI-RADS categories for breast image in the form of text reports. The BROK system utilized regular expression with respect to pre-defined medical terminologies such as taxonomy. The proposed NLP algorithm segregates the each radiology reports into body as well as impression. Once the proposed NLP methodology extracted the useful information from the radiology reports, subsequently the proposed methodology utilized the logic to the extracted information for classification. In order to evaluate the proposed methodology, the BROK NLP tool implemented through the utility of randomly discovered training data set which consist of 550 breast MRI reports along with 250 breast ultrasound reports also considered. The proposes NLP algorithm provides a 100.0 % recall and a 96.6 % precision as a classification results.

2.2 Rule-based classifiers

Rule-based information extraction applications are developed based on the domain experts along with an interpreter to execute the rules. Many methods are present to accomplish the classification task. The easiest strategy among them is clinical “logic rule”, if a report contains a combination of findings then that is “true”. Such clinical rules are made on the basis of decent knowledge and it should be readily understood and further extended by others.

Dublin, S et. al. [24] utilized open source NLP system ONYX which integrates knowledge about syntax (the structure of sentences) and semantics (the meaning of words) to understand free text and construct the structured output. In order to extract the useful information from the radiology based on the various radiology reports from the particular domain are collected for the ONYX training purpose. The trained ONYX system generates a concepts set discovered from individual sentences. Finally, the decision rules constructed by the domain expert are applied so that ONYX’s output to categorize the report into three categories. The first one is consistent and the second category is inconsistent with pneumonia and the third category needs manual review. In order to classify the reports based on the rule; this proposed methodology utilized two major classifiers. The first classifier discovers the reports which are require manual review by minimizing the false positives and false negatives. The second classifier identifies the most frequently occurring concept which helps to reduce the manual reviews.

Sohn, Sunghwan, et al. [25] presented NLP methodology to extract the useful information from the radiology reports and design anda rule-based algorithm is developed to discover the abdominal aortic aneurysm (AAA) patients. In order to process the free text radiology reports, the proposed

methodology utilized the NLP dictionary lookups such as MedTagger to discover the events and concepts.Using the reference of MedTagger the text in the radiology reports are annotated into three major categories. The first one is sentence detection parses and the second one token boundaries, and normalization to identify the different kind of morphological variants of AAA. Initially, based on the set of keywords the potential AAA reports are selected. The selected reports are subjected to rule based classifier to identify the AAA-case vs. non-case. Finally the AAA patient cohort classification is done based on extracted information.

Bao H. Do et. al. [26] presented classification methodology for fracture identification. The real time NLP based system helps to extract the fracture knowledge automatically which is very efficient in case of emergency setting. Becauseof uncontrollable vocabularies in health care field, the automated classification performance unable to provide appropriate classification results. In order to enhance the classification accuracy, it is better to design the appropriate rules with the cooperation of domain expert. The proposed NLP algorithm regular expression technique to segregate the unstructured text and utilized the rules to classify the radiology reports. The raw, unstructured texts are accepted by the proposed NLP system and it applies simple a rules-based heuristic to identify fracture concepts. The proposed NLP system identified the involved bone with a accuracy of 97 %.

Nguyen, Anthony N., et al. [27] presented a semantic rule-based MEDTEX system based on the SNOMED CT ontology. In order to classify the retrieved lung TNM, the proposed methodology utilized the SNOMED CT ontology. Based on the ontology, the text from the radiology reports are annotated which helps enhance the classifier performance. With the help of SNOMED CT expression templates, the sub Sumption querying of concepts is utilize to retrieve useful information from the free text. In order to classify the radiology reports, the rules are designed by the domain expert. The rules are further utilized to find out the stages of Lung TNM.

2.3 Machine learning based classifiers

Statistical and machine learning methods infer rules and patterns directly from data. They are intensely mixed into all aspects of NLP.

Masino, Aaron J., et al. [28] developed an automated pipeline for classification of radiology reports. The proposed pipeline labels radiology text reports as normal or abnormal relative. in this research they integrated NLP and standard ML algorithms to recognize the irregular regions described in the radiology reports for the otologic domain. In order to classify the radiological reports various ML algorithms were utilized such as logistic regression, support vector machine (SVM), decision tree, random forest, and naïve Bayes models.

Chen, Po-Hao, et. al. [29] presented a methodology for the automatic detection of radiologist intent in oncologic evaluations. The proposed methodology includes multiple NLP techniques and ML algorithms. In order to classify the unstructured reports, the proposed methodology integrates three NLP techniques such as TF-IDF, TF, and hashing. Along with the three NLP algorithms, five ML classification algorithms also implemented for classification purpose such as logistic regression (LR), random decision forest (RDF), one-vs-all Bayes point machine (BPM), one-vs-all support vector machine (SVM), and fully connected neural network (NN). In the proposed methodology, they found that NLP model consisted of tokenized bigrams and unigrams with TF-IDF performed much better than N-gram. Optimized way with all parameters, SVM had the best performance on the test dataset, with 90.6 average accuracy and F-score of 0.813. the overall performance depends on the combinatorial optimization of both the NLP and ML algorithms.

Sevenster, Merlijn, et. al. [30] presented an automatic methodology retrieving useful information and combination of measurement across successive radiology documents. The automatic methodology is implemented based on the integration of NLP pipeline and ML classifier. Each and every measurement in successive radiology documents is paired with other measurement. In this research, the random forest classifier is utilized for finding the similar measurements. The similarity index is formulated based on the contextual, narrative and volumetric properties of measurements. In order to find the partial uniqueness, they presented post-processing methodology which is applied to the outcome of the random forest classifier which helps to attain the enhancement the accuracy of the classifier. However, the proposed methodology produces small negative effect on area under ROC.

Yadav, Kabir, et al [31] presented a hybrid methodology for automatic classification of emergency department CT imaging documents. The proposed hybrid methodology integrates the NLP based linguistic and statistical machine learning based classifiers. Initially, the linguistic based features are obtained from the CT image documents which are given to the classifier to classification process. The obtained features are subjected to supervised machine learning classification algorithm for classify the reports. Initially, Naïve Bayes classifier is utilized for finding the appropriate set of features based on the events and concepts, relation and modality. Once the appropriate set of features obtained based on the Naïve Bayes classifier, the classification algorithms such SVM and MaxEnt algorithms are applied on the obtained features for classification of radiology reports. The proposed hybrid NLP and machine learning system for automatic classification of emergency department (ED) CT imaging reports. Their proposed approach contains two main approaches are linguistic (natural language processing [NLP]) and statistical (machine learning).

2.4 Hybrid approaches (machine-learning classifier and manually constructed rules)

Garla, Vijay, et al. [32] developed extensions to the clinical Text Analysis and Knowledge Extraction System (cTAKES) which minimize the computation complexity of feature extraction process. The proposed feature extraction

methodology is evaluated by the both rule and machine-learning based classifiers. Initially, the radiology reports are classified based on the rules constructed by the domain experts. In order to represents the radiology documents, the machine-learning based document-classification methodology utilizes the NLP techniques such as 'bag-of-words' or 'term-document matrix'. This research utilized decision trees (C4.5 algorithm), machine-learning analogue of rule-based classifiers; random forests, ensembles of decision trees; and SVMs, for radiology document classification. The proposed YTEX helps to derive the efficient set of features space which enables to construct the feature representation based on the domain knowledge.

Anne-Dominique Pham et al. [33] presented a radiology classification system by integrating the NLP and machine learning to discover the thromboembolic disease. The proposed classification methodology can identify the relevant medical related information are extracted from radiology reports written in French. The set of concepts and events are determined based on the thromboembolic disease. The obtained features are subjected to supervised machine learning classification algorithm for classify the reports. Initially, Naïve Bayes classifier is utilized for finding the appropriate set of features based on the concepts, modality and relations of annotations . Once the appropriate set of features was obtained based on the Naïve Bayes classifier, subsequently the SVM and MaxEnt algorithms were typically used for classification of the radiology reports. The proposed approach achieved an F measure of 0.98 for pulmonary embolism identification.

3. Conclusion

There are various reasons why most of the NLP applications in handling of radiology reports remain in a proof-of-concept stage. Based on the review, it is concluded that three major reasons are discovered. The first on is, unpredictability about least performance requirements may increase the complexity for the implementation of the system, particularly when fully automated classification of radiology reports. However, still there is no appropriate direction on attain the minimum performance. The second reason is Clinicians and researchers may be unwilling to admit obtained output from automatic algorithms due to its difficult or impossible to trace how the output was constructed. The third one is the terminology used in the radiologic documentation needs to be standardized which enables to make the vocabularies can be controllable. The controlled vocabularies help NLP applications to retrieve the useful information from the radiology documents. The enhancement of lexicons in the health care documents and their reporting incorporation software in an appropriate way may enhance the performance of NLP applications.

References

[1] Wang, Yanshan, et al. "Clinical information extraction applications: a literature review." Journal of biomedical informatics 77 (2018): 34-49.
[2] M.A. Ellsworth, M. Dziadzko, J.C. O'Horo, A.M. Farrell, J. Zhang, V. Herasevich,An appraisal of published usability evaluations of electronic health records via systematic review, J. Am. Med. Inform. Assoc. 24 (2017) 218–226.
[3] B.A. Goldstein, A.M. Navar, M.J. Pencina, J. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. 24 (2017) 198–208.
[4] K. Jensen, C. Soguero-Ruiz, K.O. Mikalsen, R.-O. Lindsetmo, I. Kouskoumvekaki, M. Girolami, et al., Analysis of free text in electronic health records for identification of cancer patient trajectories, Sci. Rep. 7 (2017).
[5] S.G. Small, L. Medsker, Review of information extraction technologies and applications, Neural Comput. Appl.. 25 (2014) 533–548.
[6] Cai, Tianrun, et al. "Natural language processing technologies in radiology research and clinical applications." Radiographics 36.1 (2016): 176-191.
[7] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, J.F. Hurdle, Extracting information from textual documents in the electronic health record: a review of recent research, Yearb Med. Inform. 35 (2008) 44.
[8] R.W.V. Flynn, T.M. Macdonald, N. Schembri, G.D. Murray, A.S.F. Doney, Automated data capture from free-text radiology reports to enhance accuracy of hospital inpatient stroke codes, Pharmacoepidemiol. Drug Saf. 19 (2010) 843–847.
[9] H. Yang, I. Spasic, J.A. Keane, G. Nenadic, A text mining approach to the prediction of disease status from clinical discharge summaries, J. Am. Med. Inform. Assoc. 16 (2009) 596–600.
[10] R. Kung, A. Ma, J.B. Dever, J. Vadivelu, E. Cherk, J.D. Koola, et al., A natural language processing alogrithm for identification of patients with cirrhosis from electronic medical records, Gastroenterology 1 (2015) S1071–S1072.
[11] L.L. Popejoy, M.A. Khalilia, M. Popescu, C. Galambos, V. Lyons, M. Rantz, et al., Quantifying care coordination using natural language processing and domainspecific ontology, J. Am. Med. Inform. Assoc. 22 (2015) e93–e103.
[12] C. Di Marco, P. Bray, H.D. Covvey, D.D. Cowan, V. Di Ciccio, E. Hovy, et al., Authoring and generation of individualized patient education materials, in: AMIA Annual Symposium Proceedings: American Medical Informatics Association, 2006, p. 195.
[13] C. Friedman, P.O. Alderson, J.H. Austin, J.J. Cimino, S.B. Johnson, A general natural-language text processor for clinical radiology, J. Am. Med. Inform. Assoc. 1 (1994) 161–174.
[14] A.R. Aronson, F.-M. Lang, An overview of MetaMap: historical perspective and recent advances, J. Am. Med. Inform. Assoc. 17 (2010) 229–236.
[15] J.C. Denny, P.R. Irani, F.H. Wehbe, J.D. Smithers, A. Spickard III, The Knowledge Map Project: Development of a Concept-based Medical School Curriculum Database, Citeseer, AMIA, 2003.
[16] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, et al., Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (2010) 507–513.

[17] S. Goryachev, M. Sordo, Q.T. Zeng, A suite of natural language processing tools developed for the I2B2 project, in: AMIA Annual Symposium Proceedings: American Medical Informatics Association, 2006, p. 931.
[18] H. Liu, S.J. Bielinski, S. Sohn, S. Murphy, K.B. Wagholikar, S.R. Jonnalagadda, et al., An information extraction framework for cohort identification using electronic health records, AMIA Summits Transl. Sci. Proc. 2013 (2013) 149–153.
[19] Yetisgen-Yildiz, Meliha, et al. "A text processing pipeline to extract recommendations from radiology reports." Journal of biomedical informatics 46.2 (2013): 354-362.
[20] Hassanpour, Saeed, and Curtis P. Langlotz. "Information extraction from multi-institutional radiology reports." Artificial intelligence in medicine 66 (2016): 29-39.
[21] Lakhani, Paras, Woojin Kim, and Curtis P. Langlotz. "Automated detection of critical results in radiology reports." Journal of digital imaging 25.1 (2012): 30-36.
[22] Xu, Yan, Junichi Tsujii, and Eric I-Chao Chang. "Named entity recognition of follow-up and time information in 20,000 radiology reports." Journal of the American Medical Informatics Association 19.5 (2012): 792-799.
[23] Sippo, Dorothy A., et al. "Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing." Journal of digital imaging 26.5 (2013): 989-994.
[24] Dublin, S., Baldwin, E., Walker, R. L., Christensen, L. M., Haug, P. J., Jackson, M. L., ... & Chapman, W. W. (2013). Natural Language Processing to identify pneumonia from radiology reports. Pharmacoepidemiology and drug safety, 22(8), 834-841.
[25] Sohn, Sunghwan, et al. "Identifying abdominal aortic aneurysm cases and controls using natural language processing of radiology reports." AMIA summits on translational science proceedings 2013 (2013): 249.
[26] Do, Bao H., et al. "Automatic retrieval of bone fracture knowledge using natural language processing." Journal of digital imaging 26.4 (2013): 709-713.
[27] Nguyen, Anthony N., et al. "Symbolic rule-based classification of lung cancer stages from free-text pathology reports." Journal of the American Medical Informatics Association 17.4 (2010): 440-445.
[28] Masino, Aaron J., et al. "Temporal bone radiology report classification using open source machine learning and natural langue processing libraries." BMC medical informatics and decision making 16.1 (2016): 65.
[29] Chen, Po-Hao, et al. "Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports." Journal of digital imaging (2018): 1-7.
[30] Sevenster, Merlijn, et al. "A natural language processing pipeline for pairing measurements uniquely across free-text CT reports." Journal of biomedical informatics 53 (2015): 36-48.
[31] Yadav, Kabir, et al. "Automated outcome classification of emergency department computed tomography imaging reports." Academic Emergency Medicine 20.8 (2013): 848-854.
[32] Garla, Vijay, et al. "The Yale cTAKES extensions for document classification: architecture and application." Journal of the American Medical Informatics Association 18.5 (2011): 614-620.
[33] Pham, Anne-Dominique, et al. "Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings." BMC bioinformatics 15.1 (2014): 266.