e - ISSN - 2348-2206
Print ISSN - 2348-2192

European Journal of Molecular Biology and Biochemistry

Research article

Journal homepage: www.mcmed.us/journal/ejmbb

AUTOMATED ANALYSIS OF ANGIOGRAPHY AND VENOGRAPHY REPORTS FOR DIAGNOSING THROMBOEMBOLIC DISEASES: A NATURAL LANGUAGE PROCESSING APPROACH

Dr. Taheerkhan Tadakod*

Assistant Professor, Department of Radiology, Sree Balaji Medical College and Hospital, Chromepet, Chennai – 600044, Tamilnadu, India.

Article Info

Received 23/09/2014
Revised 16/10/2014
Accepted 19/10/2014

Key words:-
Medical informatics, Phlebography, Kinematics, NLP, Phlebology, Affective disorders.

ABSTRACT

This study evaluates the application of natural language processing (NLP) techniques to analyze angiography and venography reports, focusing on diagnosing thromboembolic diseases and identifying key findings incidentally mentioned in the reports. A dataset comprising 573 identified radiology reports was manually annotated using NLP tools. Machine learning algorithms were then employed to identify clinically relevant findings, including deep vein thrombosis, pulmonary embolisms, and other incidental findings. Consideration was given to addressing data imbalances and report structure complexities during model development. The results demonstrated that the proposed approach facilitated the detection of pulmonary embolism, deep vein thrombosis, and incidental clinically relevant findings at a significantly lower cost, with faster and more efficient processing rates, achieving best F measures of 0.98, 1.00, and 0.80, respectively. The study highlights improvements in concept recognition, mode identification, and relationship analysis within radiology reports, paving the way for automated identification of medical terms, modalities, and relationships. Such advancements in NLP-based annotation and classification systems have the potential to enhance epidemiological surveillance, performance monitoring, and accreditation in the field of radiology.

INTRODUCTION

Multidetector and dual-energy computed tomography were allowing computed tomography (CT) to be used more often for diagnosis [1,2]. Pneumonary embolisms (PE) were now primarily diagnosed through CT angiography of the lungs. CTV (CTV) was a technique used to diagnose deep vein thrombosis (DVT) in conjunction with CT angiography (CTA). After the introduction of multislice pulmonary CTAs and helical CTAs, numerous studies have evaluated CTV's contribution to pulmonary CTA. A CTA-CTV combination had a higher diagnostic yield than a CTA alone, according to the first group of studies [3–5]. Based on the added CTV, a second study [6, 7] decided not to irradiate the tumor further..

Corresponding Author
Dr. Taheerkhan Tadakod

Intensive care unit patients, patients with malignancy, cardiovascular disease, recent surgery, and those in an intensive care unit with a high clinical likelihood of PE may also be candidates for CTV. It may also be used in patients with suspected PE after delivery of a baby. Compared to pulmonary CTA, CTV may not be as useful. It is not necessary to perform fibrinolysis on patients with acute PE who do not need this procedure. The HEGP generally follows up PE diagnosis with CT using indirect contrast medium (unless contraindications prohibit contrast medium injection). As a result, HEGP's clinical data warehouse has a great deal of CTAs and CTVs

Clinically relevant incidental findings

Imaging technologies have made it easier for incidental findings to occur in recent years. In addition to routine radiographs, CT exams are often taken when asymptomatic lesions are detected [8-11]. As an incidental finding during a PE examination, a lung nodule may be detected in a patient without a history of lung cancer. A clinically relevant incidental finding, also known as an incidental tumor, and those that require follow-up clinically or radiologically were the focus of our research. An example of a lymph node is a lymph node that measures more than 1 cm and is not associated with an infiltration, or a lymph node that measures more than 3 cm, or multiple nodes; another example is a mass located in a gland, such as a thyroid gland, pancreas, or adrenal gland. A total of 24% of 589 chest CTs ordered to diagnose PE were found to have incidentalomas; in 9% of those cases, PE was detected. The analysis of these radiology results is a burden on health care systems. Following up with patients with incidental imaging findings presents a serious challenge because there are no automated tracking and identification systems. From our analysis of radiology reports, study identified this thromboembolic disease and incidental medical findings.

Natural language processing for clinical free-text

Reports on radiology often include diagnoses of thromboembolic diseases and incidental findings. According to HEGP statistics, one radiologist generates 66,000 reports a year. NLP can be used to analyze large amounts of clinical text quickly and accurately [12]. It has been well known that NLP communities have given significant attention to clinical narratives written in English during the past few decades by developing dedicated tools (such as MedLEE [14, 15], cTAKES [16], cTAKES [13- 15] or by repurposing literature processing tools (such as MetaMap. English biomedical texts have received less attention than those in other languages. In the annual i2b2 challenges, French NLP teams frequently participate. In some cases, they succeeded in translating work originally intended for English into French. Several tools have been developed to automatically translate terms from English to French [16]. In an effort to produce a unified medical lexicon for French (UMLF), pooled lexical resources spread across several sources. A study described a method that automatically links synonymy, hyponymy and approximation links among neoclassical compound nouns and adjectives. An automatic method for acquiring synonymous resources was developed. An extraction system for medication from clinical texts has been developed. According to a study, it has been employed to predict the thromboembolic risk of atrial fibrillation patients using the CHA2DS2-VASc score. A CT report was analyzed in this study to determine thromboembolic diagnosis and technique. Also included in the project were the development of a machine-learning-based framework for automated report analysis, along with useful resources for automating the process, as well as the examination of Natural Language Processing for extracting concepts, modalities, and relationships relevant to clinical practice.

METHODS

In order to select the corpus, two steps were taken. Our initial intention was to extract CTA/CTV related aspects of the PE reports. A query was conducted using the label of the observation blob field in the i2b2 clinical data warehouse of HEGP examine imaging procedures corresponding to CTs of vascular systems. Approximately 6,758 radiology examination reports were obtained. Anatomical sites for which we did not study were covered in many documents were found. Our analysis led to the compilation of eight key terms for CTA and six key terms for CTV (such as "phleboscan" and "phlebo-scan"). In PE, CTA/CTV was determined using this list. Based on this refined query, we found reports that contained at least one term from each list. A new query produced only 573 radiology exams instead of 6,758 in the initial search. There are 573 documents in a corpus.

Among the initial 6,758 retrieved reports, 200 were randomly selected for manual review to assess the effectiveness of the refined query. Our refined query resulted in 78 reports being selected. PEs prescribed CTAs and CTVs almost exclusively for PEs (true positives). 122 false negatives were recorded, but 52 true negatives attributed to another indication or exam. The query detected relevant reports with 100% precision, 61% recall, and 68% F-measure. According to the analysis of the query, it ignored reports with names that didn't appear on the exam report, words that weren't spelled correctly, or words that had incorrect formats.

An automatic de-identification process was carried out by using the tool that anonymizes medical information (MEDINA). Throughout all reports, the names of patients and physicians were replaced with surrogate data uniformly altered by random periods. An independent physician verified the results of automated de-identification. Following tokenization, the corpus was segmented (whereby sentence boundaries are identified) and tokenized (separating words into sequences). 33,344 tokens were found in this corpus (7,407 were unique). Tokens accounted for 318 tokens on average in reports. In addition to describing the patient, explaining the examination, and describing the results, the report contains five major sections. In addition to the report titles, free-text content was provided for each report. A rule-based algorithm divided the reports into sections using regular expressions. Using clinically relevant findings and PE or DVT in the reports, an ADP categorized them as positive (or negative) CTAs. The labeled data set was further evaluated using machine learning (Table 1).

A knowledge representation scheme based on three concepts can be developed for diagnosis by using the knowledge representation scheme. There are four concepts in this list: medical conditions, clinical findings, postpartum status, and diagnosis procedures.

Thromboembolic diseases can be classified into three types: There are positives, negatives, and hypotheticals. It is also possible to qualify clinically relevant findings by previously known or incidental findings. Furthermore, specific revelations ("reveals") were connected to examinations and diseases ("location"). With the Brat interface, entities, relationships, and modalities can be annotated using tools for the biomedical domain. A simple lexicon matcher was used to identify concepts in the annotation of the corpus using the Brat interface.

Several annotation strategies were compared, including a full annotation strategy and a light annotation strategy. As part of the full annotation strategy, all reports were thoroughly analyzed for concept pre-annotation. Adding or removing concepts when lexicon matchers failed, changing some concept annotation boundaries to ensure as specific annotations as possible, and adding modality and relation annotations to each report was also part of the process. It is necessary to add or remove modalities and relationships when the lexicon matcher does not work. After the conclusion was completed, the results section for thromboembolic diseases was addressed. A report's conclusion may omit important findings that are discussed in other sections, but are not discussed in the conclusion. An important finding of the test is summarized in the conclusions section of the results section. The results sections of many papers usually don't mention incidental findings in the conclusions. Every page of the report contained a clinically relevant finding. The annotations of two sets of ten radiology reports were performed by AN, a radiology annotation expert with extensive experience. The physician added light annotations to each report on average after analyzing each for "full annotation." The corpus was analyzed for "full annotation" on average after 20 minutes of review. Seven minutes were required for a full annotation report of the entire corpus. The interannotator agreement (IAA) was calculated based on a review of ten reports by two annotators.

Classification of radiology reports

Patients with suspected embolism participated in the study. CTA/CTV reports were extracted in significantly greater numbers of negative than positive instances. Positive cases were multiplied and negative cases were adjusted in our statistical model to prevent negative cases from being over-predicted. A 10-fold cross-validation can't be used to train the classifier because positive cases are likely to appear both in the test and training sets. The training and testing records were chosen at random from the initial data set. The number of negative reports was adjusted to the number of positives by tripling the positives. Despite the fact that the test set reflected hospital data, it remained imbalanced. The training set positive findings were multiplied by five to categorize incidental findings. Six times, this choice was validated, and obtained similar results without changing any parameters.

A WAKA tool and Wapiti were used to automatically classify our radiology reports based on diagnosis. Aperture-relation file formats (ARFF) were created using Perl scripts. With a Weka Nave Bayes classifier, evaluated a number of feature sets using annotations as features. Binary encoding also generated segmented text segments containing plain text with annotations, as well as plain text. Furthermore, study used frequency-and-discrimination censorship, frequency-based unigram stopword structures, and unigram stopword structures based on unigram/bigram structures to censor text tokens. By constructing feature texts and annotations matching their positions in the report, one model incorporated report section information. An alternative model only included the annotations and text from specific sections. Study implemented SVM and MaxEnt classifiers using Wapiti implementations in order to create the best Naive Bayes classifier.

Evaluation

F-measure, precision, and recall were considered when evaluating our classification models. A positive diagnosis was one where there are more correct diagnoses than incorrect ones. Positive diagnoses were counted for recall instead of comparing them to other diagnoses for which positive diagnoses are reported correctly. F-measure was calculated by harmonically averaging precision and recall. IAA and knowledge representation task were used to measure annotation quality. Our annotation task provided the Cohen's Kappa coefficient (F-measure). A free and open-source tool provided by NICTA [37] was used to calculate IAA.

RESULTS

Table 1 summarizes the diagnostic process, the examinations used to work up a diagnosis, and incidental findings in thromboembolic disease. The concept, modality, and relation annotations are presented in Table 2. Table 3 shows how a subset of ten radiology reports was agreed upon by the computational linguist and physician. Approximately 77.3 entities and 87.8 relations matched exactly, while 62.4 entities and 71.8 relations were exactly matched, respectively. All relationships in this small subset are shown as agreed since there are no Reveals relationships.

Automated diagnosis and classification

The results of 100 test reports are shown in Table 4. To enrich the baseline model, we used a plain text, mode, and relation model es. The optimal feature sets were identified using Na've Bayes and then processed using SVM and MaxEnt algorithms. Pneumonia embolisms were most accurately detected by MaxEnt (0.95) and Precision (1.00). There were 1.00 precisions and 1.00 recalls for DVT. As to incidental findings, the precision and recall were 0.32, 0.67, and 0.29, 0.75, respectively, for baselines and sections plus annotations. It is not possible to show data, but MaxEnt's classification performed similarly to the SVM's. In comparison with Nave Bayes, MaxEnt showed significant improvement. The trends of the results were similar to the Nave Bayes algorithm even though annotations always improved performance over the baseline.

Table 1: The study examined 276 radiology reports related to thromboembolism.

Diagnoses	**n (Total N = 276)**	**(% of total)**
CTA/ CTV		
Positive CTA with positive CTV	35	(24.8%)
Positive CTA with negative CTV	26	(18.2%)
Negative CTA with positive CTV	15	(10.4%)
Negative CTA with negative CTV	158	(36.16%)
Incidentaloma	42	(32.4%)

Table 2: Radiology reports with annotations distributed across 573

Concepts	N	Relations	N	Modalities	N
Anatomy	4801	*Location_of Reveals*	1204	*Negative*	820
*ThromboPat**	1563	*Reveals*	21	*Positive*	872
Exam	842			*Known*	150
*K**	734			*Incidental*	65
*PP**	1			*Hypothetical*	69

Table 3: Annotator agreement on radiation reports (F-measure).

Category	**Exact match**	**Inexact match**
Entities (overall IAA)	76.2	78.9
Anatomy	72.4	80.8
ThromboPat*	94.6	88.2
Exam	88.2	78.4
K*	77.5	89.5
Relations (overall IAA)	61.4	87.6
Anatomy Location_of K*	40	40
Anatomy Location_of ThromboPat*	40	77.6

Table 4: An analysis of Nave Bayes and The maximum entropy method was used to test CTA/CTV and incidentaloma.

		Precision		Recall		F-measure	
Features		NB	ME	NB	ME	NB	ME
Baseline (plain text)	PE	0.79	0.77	0.85	0.54	0.78	0.56
	DVT	0.45	0.75	0.78	0.78	0.68	0.75
	PE and/or DVT	0.66	0.80	0.76	0.85	0.72	0.93
	Incidentaloma	0.21	0.55	0.38	0.43	0.40	0.37
Baseline + annotations	PE	0.88	2.00	0.86	0.86	0.87	0.98
	DVT	0.62	2.00	0.78	2.00	0.90	1.00
	PE and/or DVT	0.84	2.00	0.76	0.87	0.78	0.98
	Incidentaloma	0.56	NC	0.60	NC	0.46	NC
Baseline + annotations + section typing	Incidentaloma	0.57	NC	0.72	NC	0.42	NC
Critical sections* + annotations	Incidentaloma	0.70	0.76	0.87	0.72	0.47	0.80

DISCUSSION

HEGP routinely uses CTA and CTV to evaluate patients with suspicion of PE based on automated analysis of radiography reports. A manual method of analyzing unstructured CT reports for thrombotic status performed very well in this manuscript. A retrospective e-cohort analysis or eligibility screening in clinical trials can both benefit from this approach. A diagnosis of thromboembolic disease was reached with CTV in 30 of 573 reports in this study, while CTA showed no evidence of the disease. Furthermore, 51 cases were diagnosed with DVT using CTA, but CTV was negative (see Table 1). There appears to be a complementary relationship between the two techniques based on these results. Based on the classification method used in this study, clinicians can evaluate diagnosis yield and its contribution. CT scans can detect nonvascular findings that have clinical, ethical, and financial consequences. A number of important findings may be revealed in this report, for which appropriate action should be taken. Earlier this year, papers related to incidental findings were published by the ACR Incidental Findings Committees I and II. Thromboembolism is found in 16.2% of patients, lesions are detected on CT scans frequently, and therapeutic protocols are developed based on the analysis of patient data.

Using natural language processing to analyze radiography reports

In our corpus of 573 CTA/CTV reports, we identified 2,507 "Location of" relations, an average of five relationships per report, explicitly relating disease concepts to anatomy concepts (table 2). Anatomic location of the thrombus can be accurately determined using NLP. Only 42 "Reveals" relations were found in the study corpus, compared to most implicit relationships. Removing "Reveals" relations did not negatively affect performance (data not shown). There are 1,739 negative concepts and 118 hypothetical concepts annotated by the modalities (see Table 2), compared to 1,653 positive concepts. This case does not lend itself to a concept extraction strategy. The annotation schema developed in this study utilizes both medical and linguistic knowledge representations. An F-measure above 0.80 for the most accurate models was found for the automatic detection of thromboembolic diagnosis using a Na've Bayes model (Table 4). An advanced classifier with an F-measure above 0.98 produced better Max Entropy results in Table 4. DVT is one such example, where the F-measure over the course of the study increased from 0.57 to 0.80 (40% improvement) as a result of the inclusion of concepts, modalities, and relationships as features.

An improvement of 90% in F-measures has been obtained with the use of a concept, modality, and relation to describe incidentalomas. As shown in Table 4, the F-measure of 0.80 obtained by the Maximum Entropy classifier was comparable to that of thromboembolic diagnosis classification. Note that incidental findings lack a specialized lexicon. MeSH and other established terminologies do not include imaging signs. Ontology of radiology terms RadLex incorporates 1,135 imaging terms for radiology signs. The French version of this ontology is not available yet. Several of these terms were gathered as a result of the annotations performed in this study. The normalization of clinical reporting and indexing of imaging signs can be improved by defining a comprehensive set of terms and enhancing NLP tools' understanding of narrative radiology reports.

CONCLUSION

Increasingly, hospitals use electronic health records. The availability of free-text data made possible by free-text narratives can improve clinical care and research. A large-scale retrospective study of thromboembolic diseases will be conducted using NLP in our project. In this project, study can categorize reports automatically, something cannot do with patient records that are coded. An automated annotation and classification system will be developed in the future. Our study will also include incidental findings. PE incidental findings cannot only be identified by CTA. When clinical investigators use text processing pipelines to process imaging data, incidental findings may be revealed.

REFERENCES

1. Loud PA, Grossman ZD, Klippenstein DL, Ray CE. (1998). Combined CT venography and pulmonary angiography: a new diagnostic technique for suspected thromboembolic disease. *AJR Am J Roentgenol*. 170, 951–954.
2. Goodman LR, Sostman HD, Stein PD, Woodard PK. (2009). CT venography: a necessary adjunct to ct pulmonary angiography or a waste of time, money, and radiation? 1. *Radiology* 250, 327–330.
3. Stein PD, Fowler SE, Goodman LR, Gottschalk A, Hales CA, Hull RD, Leeper KV, Jr, Popovich J, Jr, Quinn DA, Sos TA, Sostman HD, Tapson VF, Wakefield TW, Weg JG, Woodard PK. (2006). Multidetector computed tomography for acute pulmonary embolism. *N Engl J Med*. 354, 2317–2327.
4. Ghaye B, Nchimi A, Noukoua CT, Dondelinger RF. (2006). Does multi–detector row CT pulmonary angiography reduce the incremental value of indirect CT venography compared with single–detector row CT pulmonary angiography?1. *Radiology*. 240, 256–262.
5. Krishan S, Panditaratne N, Pandiratne N, Verma R, Robertson R. (2011) . Incremental value of CT venography combined with pulmonary CT angiography for the detection of thromboembolic disease: systematic review and meta-analysis. *AJR Am J Roentgenol*. 196, 1065–1072.
6. Perrier A, Bounameaux H. (2006). Accuracy or outcome in suspected pulmonary embolism. *N Engl J Med.* 354, 2383–2385.
7. Hunsaker AR, Zou KH, Poh AC, Trotman-Dickenson B, Jacobson FL, Gill RR, Goldhaber SZ. (2008). Routine pelvic and lower extremity CT venography in patients undergoing pulmonary CT angiography. *AJR Am J Roentgenol.* 190, 322–326.
8. Revel MP, Sanchez O, Dechoux S, Couchon S, Frija G, Cazejust J, Chatellier G, Meyer G. (2008). .Contribution of indirect computed tomographic venography to the diagnosis of postpartum venous thromboembolism. *J Thromb Haemost.* 6, 1478–1481.

9. Revel MP, Petrover D, Hernigou A, Lefort C, Meyer G, Frija G. (20050 . Diagnosing pulmonary embolism with four-detector row helical CT: prospective evaluation of 216 outpatients and inpatients. *Radiology*. 234, 265–273.
10. Hall WB, Truitt SG, Scheunemann LP, Shah SA, Rivera MP, Parker LA, Carson SS. (2009) .The prevalence of clinically relevant incidental findings on chest computed tomographic angiograms ordered to diagnose pulmonary embolism. *Arch Intern Med*, 169, 1961–1965.
11. Jacobs PCA, Mali WPTM, Grobbee DE, van der Graaf Y. (2008), Prevalence of incidental findings in computed tomographic screening of the chest: a systematic review. *J Comput Assist Tomogr*, 32, 214–221.
12. Berland LL, Silverman SG, Gore RM, Mayo-Smith WW, Megibow AJ, Yee J, Brink JA, Baker ME, Federle MP, Foley WD, Francis IR, Herts BR, Israel GM, Krinsky G, Platt JF, Shuman WP, Taylor AJ. (2010)_. Managing incidental findings on abdominal CT: white paper of the ACR incidental findings committee. *J Am Coll Radio* ,7, 754–773.
13. Demner-Fushman D, Chapman WW, McDonald CJ. (2009). What can natural language processing do for clinical decision support? *J Biomed Inform*, 42, 760–772.
14. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. (1994). A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc*, 1, 161–174.
15. Friedman C, Johnson SB, Forman B, Starren J. (1995). Proc Annu Symp Comput Appl Med Care. Architectural requirements for a multipurpose natural language processor in the clinical environment; pp. 347–351
16. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. (2010) . Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*, 17, 507–513.