Automated Classification of Radiology Reports to Facilitate Retrospective Study in Radiology

Yihua Zhou · Per K. Amundson · Fang Yu ·
Marcus M. Kessler · Tammie L. S. Benzinger ·
Franz J. Wippold

Published online: 30 May 2014
© Society for Imaging Informatics in Medicine 2014

Abstract Retrospective research is an import tool in radiology. Identifying imaging examinations appropriate for a given research question from the unstructured radiology reports is extremely useful, but labor-intensive. Using the machine learning text-mining methods implemented in LingPipe [1], we evaluated the performance of the dynamic language model (DLM) and the Naïve Bayesian (NB) classifiers in classifying radiology reports to facilitate identification of radiological examinations for research projects. The training dataset consisted of 14,325 sentences from 11,432 radiology reports randomly selected from a database of 5,104,594 reports in all disciplines of radiology. The training sentences were categorized manually into six categories (Positive, Differential, Post Treatment, Negative, Normal, and History). A 10-fold cross-validation [2] was used to evaluate the performance of the models, which were tested in classification of radiology reports for cases of sellar or suprasellar masses and colloid cysts. The average accuracies for the DLM and NB classifiers were 88.5 % with 95 % confidence interval (CI) of 1.9 % and 85.9 % with 95 % CI of 2.0 %, respectively. The DLM performed slightly better and was used to classify 1,397 radiology reports containing the keywords “sellar or

suprasellar mass”, or “colloid cyst”. The DLM model produced an accuracy of 88.2 % with 95 % CI of 2.1 % for 959 reports that contain “sellar or suprasellar mass” and an accuracy of 86.3 % with 95 % CI of 2.5 % for 437 reports of “colloid cyst”. We conclude that automated classification of radiology reports using machine learning techniques can effectively facilitate the identification of cases suitable for retrospective research.

Keywords Radiology report classification · Machine learning · Natural language processing · Retrospective studies · Computer analysis · Radiology reporting · Radiology Information Systems (RIS)

Introduction

The field of radiology is ideally suited for relatively inexpensive retrospective research techniques because of the availability of typically large databases consisting of archived radiology examinations and accompanying radiology reports.

A proposed retrospective study in radiology usually begins with a research question that prompts compilation of a list of relevant cases that meet certain inclusion criteria. A common way of developing such a list of cases involves searching a radiology report database using certain keywords. Keyword-based searches are often nonspecific, however, and may yield numerous extraneous reports that are not pertinent to the research project.

In order to enhance efficacy of keyword searches at our institution, a web-based “RIS Search” application has been implemented using the Microsoft SharePoint Search engine to index more than 5 million radiology reports. The typical search result presented to the user resembles the presentation produced by the popular Google search engine and consists of a list of uniform resource locators (URLs) pointing to the

Y. Zhou (✉)
The Department of Radiology, Saint Louis University School of
Medicine, 3635 Vista Blvd at Grand Blvd, Saint Louis, MO 63110,
USA
e-mail: yzhou31@slu.edu

Y. Zhou · P. K. Amundson · F. Yu · M. M. Kessler ·
T. L. S. Benzinger · F. J. Wippold
The Mallinckrodt Institute of Radiology and Siteman Cancer Center,
Washington University School of Medicine, 510 S Kingshighway
Blvd, St Louis, MO 63110, USA

M. M. Kessler
The Department of Radiology, Division of Nuclear Medicine,
University of Arkansas for Medical Sciences, 4301 W. Markham St.
Slot 556, Little Rock, AR 72205, USA

specific reports. Researchers then select each URL to examine the report and determine whether or not the case should be included in the research project. This time consuming manual curation of the results is required in order to refine the list of cases.

For example, in one research project designed to track the surgical results following excisions of third ventricular colloid cysts, the term “colloid cyst” was used to search the radiology report database. The initial result from our institutional Radiology Information System (RIS) search program indicated that approximately 10,240 reports met the criteria of the study, which was clearly an overestimation of the true number of cases based on experience with this population. Following a manual screening of the reports, only 726 reports that contained keywords “colloid cyst” referred to cysts that were in fact within the brain. Of these reports, 50 did not find an evidence of a colloid cyst, 128 described postoperative changes from a prior colloid cyst resection, and 104 mentioned an unsubstantiated history or reported history of a colloid cyst. Clearly the process is quite time consuming and tedious.

In order to facilitate selection of radiology reports for retrospective studies in brain imaging, we implemented radiology report classifiers using the LingPipe [1] package in a case finder application that automatically classifies reports into one of six predefined classes. The purpose of this paper is to report the results and lessons learned from this project.

Materials and Methods

This study was approved by our institutional review board and the requirement for informed consent was waived, as all patient-identifying information was removed before the radiology reports were included in this retrospective study.

Classes Six classes (Table 1) were determined for the purpose of facilitating selection of cases for retrospective studies in the field of neuroradiology, in which the research team may be interested in cases that have findings suggesting a specific diagnosis, a list of differential diagnoses, or post treatment changes. This classification scheme can help the researcher

decide whether to include the specific cases. We also believe that the classes can be applied to other fields of radiology.

Training Dataset To construct the training dataset, a pool of 50,000 (an arbitrary large number) unique sentences were extracted from the impression sections of 26,342 radiology reports randomly selected from a database containing 5,104,594 items in all disciplines of radiology. Four reviewers (three board certified radiologists and one fourth year medical student; all were trained before the start of the work) participated in manually assigning one of the six predefined classes (Table 1) to sentences randomly selected from the report pool. Each sentence in the pool had equal chance to be manually annotated by 0 to 4 reviewers. However, each sentence could only be reviewed once by the same reviewer. Each reviewer was required to review approximately 4,000 sentences independently.

At the end of the reviewing period, a total of 16,724 manual classifications had been performed, representing 11,443 unique sentences derived from 6,756 reports. A total of 3,428 of these sentences were manually classified by more than one expert and used to calculate concordance of manual classification among human classifiers. Majority rule was used to assign classifications to the sentences with discrepancy among the reviewers. One expert reviewer determined the final classifications of the sentences when the majority rule could not be applied. In addition, 2,895 sentences extracted from brain CT and MR reports were also manually classified in the same manner and added to the training set to ensure adequate coverage of brain imaging as one of the immediate goals of the project was to select cases for retrospective studies in the field of neuroimaging. Thirteen sentences that contained no imaging related information were removed manually. Therefore, the final training dataset is composed of 14,325 manually classified sentences extracted from the impression sections of 8,537 reports.

Classification Methods The methods are based on classifiers implemented in LingPipe, a tool kit for processing text using computational linguistics [1], which offers a royalty-free license for noncommercial evaluation purpose. The dynamic

Table 1 Predefined classes

Classes	Description	Example
Positive	The report is positive for a specific search term.	Bilateral small pleural effusions, right greater than left side
DDx	The report indicates differential diagnoses of one or more disease processes.	A supracaloid mass that represents either a meningioma or aneurysm, or less likely a metastasis
Negative	The report is negative for a specific search term.	No radiographic indication of an acute fracture in the cervical spine
Normal	The report indicates a normal examination.	Normal MRI examination of the cervical and thoracic spine
History	The search term is a part of a patient history.	Fifty-three-year-old male, status post left radical orchiectomy
PostTx	The report describes post treatment changes	Postoperative changes of right radical mastoidectomy

language model classifier (DynamicLMClassifier, DLM) and the Naïve Bayesian classifiers (NaiveBayesClassifier, NB) were used in our classification system.

The dynamic language model classifier performs a joint probability-based classification of character sequences into non-overlapping categories based on language models for each category and a multivariate distribution over categories [1]. It is an n-gram classification system based on the frequency distributions of sequences of characters with the length of N [2, 3].

The Naïve Bayesian classifier is based on a uniform whitespace language model and an optional n-gram character language model for smoothing unknown tokens [1]. It is essentially a Bag-of-Words document classification technique where the tokens (words) are assumed to be independent of one another [4]. If words in the text to be classified are not present in the training corpus, an n-gram character language model is applied.

In the output, both methods calculate the probabilities for the text to be analyzed (the representative sentences of each report in this study) and the class with highest probability is assigned to the report.

Validation and Testing First, we tested the performance of the classifiers in classifying the training data itself by running classification of the training dataset using an n-gram of 6, a commonly used n-gram number in text classification. Then, a 10-fold cross-validation approach [5] was used to evaluate the performance of the classification models. Briefly, the training dataset was divided into 10 equal-sized segments. For each fold, nine segments were used to train the classifiers which were then used to classify the remaining segment. After 10 folds, all sentences were classified. The models calculated probabilities of all possible classes for each sentence. The classification assigned to a sentence was the first-best category, the category with the highest probability. The results were compared to the manual classifications, and the performance statistics (precision, accuracy, and recall) were calculated.

To evaluate the effects of the n-gram numbers on the performance, we ran the 10-fold validation process with different n-gram numbers from 2 through 8 for both the dynamic language model classifier and the Naïve Bayesian classifiers.

Implementation of a Case Finder

A case finder computer program was implemented to search the radiology report database and classify the reports into one of the predefined six classifications (Fig. 1). Briefly, the program runs a keyword search against the report dataset and retrieves all reports that contain the keywords of interest. The reports are then parsed into individual sentences using a

language model provided by the natural language processing tool kit, LingPipe [1]. Sentences containing the keywords were extracted and classified using the classification models constructed using the training data described above without further preprocessing. If two or more sentences contain the keyword(s), the sentences are concatenated before being subjected to the classifiers, which treat the combined sentences as a whole. The classifications of the sentences are used to represent the classifications of the reports and used to select cases for retrospective research projects. The case finder provides a simple tool to search the entire database, returning a tabular output of a list of reports with classifications and the representative sentences.

To test the performance of the program against our report database that had over 5 million radiology reports, we used the keywords “sellar mass”, “suprasellar mass”, or “colloid cyst” to search and retrieve the reports that contained the keywords. The reports were classified using the dynamic language model classifier with an n-gram of 4, as well as manually by radiologists who were without knowledge of the classifications by the case finder program. The results were compared to determine the performance of the program.

Results

A total of 14,325 sentences (including 11,430 sentences from 8,537 radiology reports from all disciplines of radiology, and an additional 2,895 sentences from brain CT and MRI reports) were manually classified as one of six predefined classes. The concordance of manual classification by the experts was estimated to be 95.6 %, based on 168 discrepant classifications out of the 3,428 sentences that had been manually classified by at least two experts. The unweighted Cohen’s Kappa was 0.94 with 95 % confidence interval (CI) of 0.01.

When the training dataset was classified using an n-gram of 6, the accuracies for the dynamic language model (DLM) and the Naïve Bayesian (NB) classifiers were 91.6 % with 95 % CI of 0.46 % and 86.0 % with 95 % CI of 0.46 %, respectively. The confidence intervals were estimated using the binomial distribution [1]. The confusion matrices are listed in the Tables 2 and 3.

Ten-fold cross-validations were performed for both the DLM and the NB classifiers using n-gram numbers from 2 to 8. The performance (precision) of the n-gram numbers with each classification method was determined. The quad-grams (n-gram of 4) were found to give the best average performances for the DLM classifier (Fig. 2). For the NB classifier, as expected, the n-gram numbers did not seem to affect the performance, likely due to the large training corpus such that most words are present in the training dataset and only a limited number of words needed n-gram-based smoothing.

Fig. 1 A screenshot of the case finder program

Overall, the DLM classifier performed slightly better than the NB.

When an n-gram of 4 was used in the 10-fold cross-validation analysis, the average accuracies for the DLM and NB classifiers were 88.5 % with 95 % CI of 1.9 % and 85.9 % with 95 % CI of 2.0 %, respectively (Fig. 2).

As the results suggested slightly better performance for the DLM classifier, we then evaluated the performances of DLM classifiers on individual classes using accuracy, recall, and precision as the performance indicators (Fig. 3). The accuracy of all the classifications exceeded 90 % and showed essentially no difference among the groups. However, the recalls and precisions for the class “DDx” were 61.8 % and 71.1 %, respectively, significantly lower than the other categories. A total of 220 sentences manually assigned to the class DDx

were classified incorrectly by the machine learning method, 7 as “Negative”, 2 as “Normal”, 5 as “PostTx”, and 186 as “Positive”. On the other hand, there were 316 sentences manually classified as “Positive” that were incorrectly classified by the machine learning method, among which 3 were assigned to “Normal”, 82 to “Negative”, and 228 to “DDx”.

Next, we tested the performance of the DLM classifier trained using the complete training dataset to classify 1,397 reports containing the keywords “sellar mass or suprasellar mass” or “colloid cyst”. These reports were independently manually classified by radiologists in the same manner as in annotating the training data. When compared to the manual classification, the prediction model produced an overall accuracy of 88.2 % with 95 % CI of 2.1 % for 959 reports that contain “sellar/suprasellar mass”, and an overall accuracy of

Table 2 Confusion matrix of the DLM classifier on the training dataset

	Hx	PostTx	Pos	Neg	Normal	DDx	Total expected
Hx	152	0	0	0	0	0	152
PostTx	73	2,413	9	2	0	5	2,502
Pos	316	77	5,883	56	4	199	6,535
Neg	158	23	23	2,529	45	2	2,780
Normal	104	0	7	2	768	1	882
DDx	84	0	17	1	2	1,370	1,474
Total predication	887	2,513	5,939	2,590	819	1,577	

Table 3 Confusion matrix of the NB classifier on the training dataset

	Hx	PostTx	Pos	Neg	Normal	DDx	Total expected
Hx	152	0	0	0	0	0	152
PostTx	80	2,320	72	14	0	16	2,502
Pos	320	168	5,374	175	110	463	6,610
Neg	160	33	55	2,406	45	16	2,715
Normal	103	1	12	17	745	4	882
DDx	82	5	48	11	2	1,322	1,470
Total predication	897	2,527	5,561	2,623	902	1,821	

86.3 % with 95 % CI of 2.5 % for 437 reports of “colloid cyst”.

Discussion

Computer-assisted, automated classification of radiology reports has been used to analyze radiology reports for a number of applications [6–11], most commonly involving classification of radiology reports into 2–3 categories (i.e., positive, negative, and unknown). The primary goals of most studies have been to help extract clinical information from unstructured reports or assist automated diagnosis of diseases. To our knowledge, our study is the first that uses machine learning-based report classification up to as many as six categories in facilitating case selection for retrospective studies in radiology.

In this study, we demonstrated the effectiveness of automated classification of radiology reports using machine learning-based text classification methods implemented in LingPipe [1] in facilitating case selection for retrospective studies, an important means of radiology research. We used a combination of keyword searches and natural language processing to identify and classify radiology reports from a large database. The keyword search retrieves reports that contain the keywords of interest to reduce the number of reports that the natural language processing step has to

analyze, which makes the approach feasible as an ad-hoc classification tool. The NLP parses the reports into individual sentences. The sentences that contain the keywords are classified into one of the predefined classes. The classifications are used to select cases that can potentially be included in retrospective studies, which can increase efficiency depending on the relative frequency of the specific classification of interest. The evaluation version of the case finder program has received positive feedback from the users. As an example, the case finder correctly identified over 89.3 % cases of patients with cervical myelomalacia based on analysis of the cervical spine MRI reports (data not shown).

The DLM classifier performed better than the NB classifier. In our 10-fold cross-validation, the DLM achieved an accuracy of 88.5 % with 95 % CI of 1.9 % when an n-gram number of 4 was used versus 85.9 % with 95 % CI of 2.0 % for the NB method. We should point out that The DLM and NB classifiers were chosen in this study for no specific reasons. In fact, other frequently used methods such as k-nearest neighbor and logistic regression have also been implemented in the LingPipe [1] tool kit and should be evaluated in the future.

As can be seen from the results of this and other studies [6], classification of unstructured radiology reports using computer analysis continues to be a challenging task. Manual analysis and sorting of reports can also be daunting. The classification concordance among our experts was 95.6 %, indicating some discrepancy among human annotators, which may have also

Fig. 2 Ten-fold cross-validation. The precision represents the overall precision for all classes. The performance for individual classes is not shown

Fig. 3 Performance of different classes using a 4-gram dynamic language model in the 10-fold cross-validation analysis. The performance is measured by accuracy, recall rate, and precision. *All* results of all classes combined, *Hx* history class, *PostTx* post treatment class, *Pos* positive class, *Neg* negative class, *Normal* normal class, *DDx* differential diagnosis class

affected the performance of the machine learning algorithms. Inspection of the discordance among the experts revealed several possible explanations. One of the most important reasons is the ambiguity of word meanings within sentences. For example, the sentence: “Unchanged small likely colloid cyst in the third ventricle near the foramen of Monro with no evidence of ventriculomegaly” was classified as “Positive” by one expert, “Differential” by another expert, and “Negative” by the machine learning method. Each classification could be construed as accurate depending upon the assumed context of the report. In addition, in some circumstances, it may be possible and desirable to classify sentences in multiple categories, although in this article, unique-category classification is assumed. Furthermore, the classes are mutually exclusive in this implementation, which may not be suitable in certain applications.

We also observed that the performance of machine learning for the class DDx was the worst among all classes. This was also true for the experts where the concordance for class DDx was lower than for all other classes (data not shown). Most (85 %) of incorrectly classified DDx sentences were classified as “Positive”. Likewise, most of the incorrectly classified “Positive” sentences were assigned to “DDx”. These results are not surprising because the “DDx” and “Positive” classes are intrinsically similar as a “Positive” classification is essentially a “DDx” with a high certainty while a “DDx” indicates multiple considerations with different degrees of likelihood. Therefore, we believe careful design of the classes and clearly defining classes may improve the performance of automated report classification by reducing ambiguity. In addition, the classifiers were implemented by a third party for general purposes, relying on entirely syntactic features. Using more complex features endowed with domain knowledge might improve performance, particularly with considerations for the style of sentences typical in radiology reports.

Furthermore, we believe that incorporating semantics and ontology into the application as a preprocessing step, which is not a part in this study, may help identify additional reports and facilitate the classification.

One of the initial purposes of the project was to identify cases of brain CT and MR examinations for retrospective research projects in the field of neuroradiology, and therefore included approximately 20 % more reports of brain CT and MRI in the training dataset. However, since most of the training corpus (approximately 80 %) was randomly selected from all disciplines of radiology, we believe the same model can be used in other subfields of radiology. Our cross-validation results support this belief. However, this belief has not been verified with other body parts and other modalities, which is a limitation of this report.

Conclusion

Automated classification of radiology reports using natural language processing and machine learning techniques can effectively facilitate the identification of suitable cases for retrospective research in radiology. Using such a tool could dramatically improve radiologists’ ability to perform retrospective research.

We evaluated an NLP-based report classification system, with an overall accuracy of 88 % and 95 % CI of approximately 2.0 %. The n-gram-based dynamic language model classifier performed slightly better than the bag-of-words Naïve Bayesian classifier. The 4-gram DLM classifier demonstrated the best performance. Performance was the worst for classification of differential diagnosis. Sentence ambiguity and double meaning were among the major factors affecting performance. The effectiveness of the approach was demonstrated in our retrospective research projects.

References

1. Alias-i. 2008. LingPipe 4.0.1. <http://ir.exp.sis.pitt.edu/ne/lingpipe-2.4.0/index.html> Accessed September 25, 2010
2. Cavnar WB, Trenkle JM. N-gram-based text categorization. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994
3. Rahmoun A, Elberrichi Z: Experimenting N-Grams in text categorization. *Int Arab J Inf Technol* 4(4):377–385, 2007
4. Lang K: “Newsweeder: learning to filter news”. Proceedings of the 12th International Conference on Machine Learning 331–339, 1995
5. Kohavi R: A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2:1137–1143, 1995
6. Solti I, Cooke CR, Xia F, Wurfel MM: Automated classification of radiology reports for acute lung injury: comparison of keyword and machine learning based natural language processing approaches. *Proc IEEE Int Conf Bioinformatics Biomed* 2009:314–319, 2009
7. Dreyer KJ, Kalra MK, Maher MM, et al: Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. *Radiology* 234(2): 323–329, 2005
8. Thomas BJ, Ouellette H, Halpern EF, Rosenthal DI: Automated computer-assisted categorization of radiology reports. *AJR Am J Roentgenol* 184(2):687–690, 2005
9. Siström CL, Dreyer KJ, Dang PP, et al: Recommendations for additional imaging in radiology reports: multifactorial analysis of 5.9 million examinations. *Radiology* 253(2):453–61, 2009
10. Waghlikar A, Zuccon G, Nguyen A, et al: Automated classification of limb fractures from free-text radiology reports using a clinician-informed gazetteer methodology. *Australas Med J* 6(5):301–307, 2013
11. Dublin S, Baldwin E, Walker RL, et al: Natural language processing to identify pneumonia from radiology reports. *Pharmacoepidemiol Drug Saf* 22(8):834–841, 2013