Automatic Classification of Free-Text Radiology Reports to Identify Limb Fractures using Machine Learning and the SNOMED CT Ontology

Guido Zuccon, PhD[1], Amol S Wagholikar, PhD[1], Anthony N Nguyen, PhD[1], Luke Butt, Dip IT[1], Kevin Chu, MBBS, MS, FACEM[2], Shane Martin, MBBS[2], Jaimi Greenslade, PhD[2]

[1] The Australian e-Health Research Centre, Brisbane, Queensland, Australia; [2] Royal Brisbane and Womens Hospital, Brisbane, Queensland, Australia.

ABSTRACT

Objective. To develop and evaluate machine learning techniques that identify limb fractures and other abnormalities (e.g. dislocations) from radiology reports.

Materials and Methods. 99 free-text reports of limb radiology examinations were acquired from an Australian public hospital. Two clinicians were employed to identify fractures and abnormalities from the reports; a third senior clinician resolved disagreements. These assessors found that, of the 99 reports, 48 referred to fractures or abnormalities of limb structures. Automated methods were then used to extract features from these reports that could be useful for their automatic classification. The Naive Bayes classification algorithm and two implementations of the support vector machine algorithm were formally evaluated using cross-fold validation over the 99 reports.

Results. Results show that the Naive Bayes classifier accurately identifies fractures and other abnormalities from the radiology reports. These results were achieved when extracting stemmed token bigram and negation features, as well as using these features in combination with SNOMED CT concepts related to abnormalities and disorders. The latter feature has not been used in previous works that attempted classifying free-text radiology reports.

Discussion. Automated classification methods have proven effective at identifying fractures and other abnormalities from radiology reports (F-Measure up to 92.31%). Key to the success of these techniques are features such as stemmed token bigrams, negations, and SNOMED CT concepts associated with morphologic abnormalities and disorders.

Conclusion. This investigation shows early promising results and future work will further validate and strengthen the proposed approaches.

INTRODUCTION

The misdiagnosis of patients true clinical condition due to misinterpretation of radiological evidence by the treating doctor is an occasional problem in hospital emergency departments. There is always a time delay between real-time reporting of the radiologist and clinical treatment by the Emergency Department clinician. The large amount of manual processing of unstructured text is one of the main issues that can be resolved by technology enabled solutions.

A good example of a misdiagnosis issue is the identification of subtle limb fractures. Radiological evidence of limb fractures, when subtle, can be missed by doctors working in the Emergency Department. The reporting of a fracture by a radiologist may not occur in real time and therefore may not be available to the doctor treating a patient. Consequently, patients may be sent home from the Emergency Department without appropriate treatment and follow up. For example, Cameron[1] reported that 2.1% of all fractures were not identified on their initial presentation to the Emergency Department. Furthermore, Sprivulis and Frazer[2] reported that 1.5% of all x-rays have abnormalities not identified in the Emergency Department records. Similarly, Mounts et al.[3] reported that 5% and 2% of the x-rays of the hand/fingers and ankle/foot from a paediatric Emergency Department had fractures missed by the treating doctor. Although small, these percentages are not insignificant.

The need to reduce errors is well recognised[4,5,6]. To ensure a diagnosis is not missed, radiology reports are commonly checked and patient records are reviewed, but this may not happen until days after the initial Emergency Department presentation. The current clinical practice of identifying limb fractures from radiology reports is highly labour intensive and is subject to human error or omissions. There is a need to streamline the process of identifying missed limb-fracture for better patient healthcare outcomes. Technology enabled solutions that can streamline the diagnosis identification would certainly improve efficiency in the existing process.

De Bruijn et al.[7] have focused on automatically detecting free-text radiology reports that identify acute fractures of the wrist. They reported that a support vector machine algorithm (SVM) was able to identify fractures in free-text radiology notes, achieving an overall F-measure of 91.3%. While, Thomas et al.[8] developed a text search algorithm that accurately classified radiology reports into the categories "fracture", "normal" and "neither normal nor fracture"[1].

In this paper we experiment with the automatic classification of free-text radiology reports for identifying fractures and other abnormalities (e.g. dislocations) of limb structures using machine learning algorithms and features such as bigram formed by stemmed tokens, negations, and SNOMED CT concepts extracted from the free-text.

OBJECTIVE

The existing process of determining accurate diagnosis of patients clinical condition from radiology reports is highly manual and labour intensive. The cause of misdiagnosis could be due to delays between diagnosis of treating doctor and response from the specialist radiology doctor whose expert opinion is needed to confirm the existence of certain clinical conditions. Computer solutions that automatically identify fractures and other abnormalities from free-text radiology reports would lessen these problems, providing time savings to doctors and hospital staff, as well as better health care outcomes.

As a first step towards such solutions, in this paper we develop and formally evaluate automated computer algorithms, based on machine learning, that attempt to identify fractures and other abnormalities, such as dislocations, from the narrative of free-text radiology reports. This research makes a contribution in investigating machine learning methods and features to reduce misdiagnosis errors from radiology reports.

MATERIALS AND METHODS

Data

A set of 99 free-text radiology reports of limb structures was acquired from the Emergency Department of a large Australian public hospital. Ethics approval was granted by the Human Research Ethics Committee at Queensland Health to use the non-identifying data. Free-text reports were short in length, containing on average 51.66 words, and an (unstemmed) vocabulary comprising 930 unique words.

Free-text reports were manually annotated by an Emergency Medicine Registrar and a Medical Officer as being either "normal" (e.g. the radiography does not exhibit a fracture of a limb structure) or "abnormal" (e.g. a fracture is identified in the radiography). A software tool was developed to assists clinicians in the recording of their interpretation and to highlight the portion of text in the report that lead to their interpretation.

Initially, assessors agreed on the annotations of 77 reports, with 20 of the remaining 22 reports annotated as normal by one assessor and abnormal by the other. A senior Staff Specialist in Emergency Medicine was then asked to act as third assessor and resolve disagreements. It was found that the 20 reports the first assessor labelled as normal while the second as abnormal referred to situations where patients with known fractures undertook a scheduled or unscheduled review. The third assessor conveyed that these reports should be treated as abnormal cases. The remaining two reports the initial assessors disagreed on were re-assessed by the third clinician; disagreement in these cases were not due to patients with known fractures scheduled for review, but represented true errors made by one of the two first assessors.

While the Fleiss' kappa (κ) calculated on the initial set of annotations provided by the two first assessor is only 0.67 (95%CI 0.51-0.82), thus exhibiting moderate to substantial inter-rater reliability, it is recognised that errors were mainly due to the fact that annotation guidelines did not specify how reports for the re-examination of known fractures should be dealt with. If these errors are to be excluded, then a strong inter-rater reliability is found, achieving almost perfect agreement. If an assessor were to be evaluated against the gold standard as created by the third senior clinician, then the F-Measure averaged over the assessments of the first two clinicians would be 98.03%.

[1]Note that Thomas et al.[8] results are not directly comparable with those of De Bruijn et al.[7] or our results.

Automatic Feature Extraction and Weighting

Machine learning algorithms require documents to be described by features. The text analysis capabilities of the Medtex tool were developed to automatically extract features from the free-text radiology reports[10]. Medtex is a text analysis system that has been previously used for classify cancer-notifiable pathology reports and produce a minimum set of synoptic factors. A wide range of features were initially extracted, including

- token, i.e. a word found in a report
- punctuation
- token stem, i.e. the stemmed version of a word contained in a report
- token negation, i.e. the addition to the token string of a common prefix representing negation if the token was negated in the text of a report; Medtex's implementation of the ConText algorithm[9] was used to identify negations in free-text
- token stem bi-gram, i.e. a pair of adjacent stemmed words as found in a report
- token stem tri-gram, i.e. a 3-tuple of adjacent stemmed words contained in a report
- SNOMED CT concepts
- the fully specified terms of the extracted SNOMED CT concepts
- the fully specified terms of extracted SNOMED CT concepts restricted to morphologic abnormalities and disorders
- SNOMED CT concept bi-gram, i.e. a pair of adjacent SNOMED CT concepts as found in a report.

While a number of these features are commonly used for the classification of free-text documents, SNOMED CT features have not been evaluated by previous works on classification of radiology reports. SNOMED CT provides a clinical terminology which was used to map various descriptions of a clinical concept to a single standard clinical terminology. In this work, the SNOMED CT ontology was used as an underlying mechanism to classify free-text using semantically matching SNOMED CT concepts. Empirical results will show that SNOMED CT concepts, in particular those referring to morphologic abnormalities (e.g. fracture, dislocation, etc.) and disorders (e.g. fracture of bone, traumatic injury, etc.), may provide valuable evidence for representing free-text radiology report data. Table 1 provide an example of feature sets extracted from the free-text of the radiology reports.

	Features																			
	stem						stemBigram					concept			conceptFull				...	
	moder	soft	...	swell	dorsal	disloc	moder_soft	soft_tissue	...	tissue_swell	disloc_present	298349001	...	108367008	Soft tissue swelling	...	Dislocation of joint	Present	...	Abnormal?
Document 1	1	1	0	1	1	0	1	1	0	1	0	1	1	0	1	0	0	0	...	0
Document 2	0	1	1	1	0	1	0	1	1	1	1	1	0	1	1	1	1	1	...	1

Table 1: Features extracted from two example free-text radiology reports; a 1 corresponds to the feature being present.

A number of weighting schemes for capturing the local importance of a feature in a report were tested. Binary coefficients were used to encode the presence or absence of a feature. The weighting schema composed by the feature frequency $f(\mathcal{F})$ of feature $\mathcal{F}$ was used to capture the number of times a specific feature appeared within a document. Variations of this schema were also experimented with. A first variation was to scale the appearance of feature $\mathcal{F}$ in a free-text report by the function $1+\log(f(\mathcal{F}))$ if $f(\mathcal{F}) \geq 1$, and 0 if the feature was absent. This function would capture the fact that little importance is given to subsequent appearances of a feature $\mathcal{F}$ in a radiology report: recall in fact that the logarithm of a number greater than one flats out rapidly. A second variation was to assign increasing weights to features that appear with high frequencies within a free-text radiology report. This was achieved by weighting the appearance of feature $\mathcal{F}$ by the score $e^{f(\mathcal{F})}$, while a zero score was assigned to absent features. It is suggested that, given the short length of the considered free-text radiology reports, the unexpected multiple appearance of a feature would provide strong evidence for determining the presence or absence of abnormalities; using the exponential function to weight the appearance of such feature would assign dominating scores to features that appear frequently in a text. Note that only local weighting functions were used to assign scores to features, i.e. weights were computed by only taking into account the frequencies of appearance of a feature in a text, ignoring thus the distribution of that feature on a global level, i.e. in the dataset. This is because of the small size of the dataset and the prevalence of features that are intuitively of key importance for the identification of abnormalities (e.g. the word "fracture").

Automatic Classification Methodology

Three common classifiers were evaluated. (1) The multinomial Naive Bayes classifier determines a free-text report's class (i.e. normal or abnormal) according to the features that occur in the text and their weights. (2) The SMO classifier is an implementation of a support vector machine (SVM) algorithm where training is performed according to the sequential minimal optimisation algorithm and a polynomial kernel is used. (3) The SPegasos classifier is a variation of the SVM algorithm, where a stochastic gradient descent algorithm and a hinge loss function are used to train a linear SVM. Parameters of all three classifiers are set to the default values (see e.g. Witten et al.[11] for details).

Evaluation Methodology

Given the small size of the dataset, a 10-fold cross validation methodology was used to train and test the classification algorithms. In this methodology, the dataset is randomly divided into 10 stratified folds of equal dimension (in our case nine folds will contain ten reports, while the remaining fold will contain only nine reports). The model for each classifier is then learnt on nine of these folds, leaving one fold out for testing the model. The process is repeated by selecting a new fold for testing, while a new model is learnt from the remaining folds. Classification performances are then averaged across the folds left out in each iteration. F-Measure was used as primary metric to evaluate the efficacy of the implemented classifiers; accuracy, sensitivity (recall) and positive predictive value (precision) were also recorded, along with the confusion matrix for the classifications (i.e. number of true positive, false negative, etc.)

RESULTS AND DISCUSSION

Not all features provided good classification effectiveness. We only report results for features that obtained the highest performance; other features obtained an average F-Measure value across the classifiers lower than 75%. The results obtained by the classifiers when attempting to identify fractures and other abnormalities of limb structures are reported in Table 2 along with error intervals for the considered measures (at 95% confidence level). The reported features are (stem) token stem along with token negations; (bigram) token stem bi-gram along with token negations; (SNOMED) the fully specified terms of extracted SNOMED CT concepts restricted to morphologic abnormalities and disorders along with their negations; (SNOMED+bigram) the combination of bigram and SNOMED), i.e. token stem bigrams with negation information and SNOMED CT concepts (limited to morphologic abnormalities and disorders) with negation information. Bolded results indicate the best values found for a given feature.

$\mathcal{F}$	Classifier	Accuracy	Positive Predicted Value	Sensitivity	F-Measure	TP	FN	FP	TN
stem	NaiveBayes	82.83%(±1.05)	82.69%(±1.05)	84.31%(±1.01)	**83.50%(±1.03)**	43	8	9	39
	SMO	76.77%(±1.17)	75.00%(±1.20)	82.35%(±1.06)	78.50%(±1.14)	42	9	14	34
	SPegasos	80.81%(±1.09)	80.77%(±1.09)	82.35%(±1.06	81.55%(±1.08	42	9	10	38
bigram	NaiveBayes	91.92%(±0.75)	90.57%(±0.81)	94.12%(±0.65)	**92.31%(±0.74)**	48	3	5	43
	SMO	85.86%(±0.96)	82.46%(±1.05)	92.16%(±0.75)	87.04%(±0.93)	47	4	10	38
	SPegasos	82.83%(±1.04)	82.69%(±1.05)	84.31%(±1.01)	83.50%(±1.03)	43	8	9	39
SNOMED	NaiveBayes	71.72%(±1.24)	67.16%(±1.30)	88.24%(±0.89)	76.27%(±1.18)	45	6	22	26
	SMO	74.75%(±1.20)	82.50%(±1.05)	64.71%(±1.32)	72.53%(±1.24)	33	18	7	41
	SPegasos	78.79%(±1.13)	87.50%(±0.92)	68.63%(±1.29)	**76.92%(±1.17)**	35	16	5	43
SNOMED+bigram	NaiveBayes	91.92%(±0.75)	92.16%(±0.75)	92.16%(±0.75)	**92.16%(±0.75)**	47	4	4	44
	SMO	86.87%(±0.93)	83.93%(±1.02)	92.16%(±0.75)	87.85%(±0.91)	47	4	9	39
	SPegasos	83.84%(±1.02)	84.31%(±1.01)	84.31%(±1.01)	84.31%(±1.01)	43	8	8	40

Table 2: Experimental results obtained by three machine learning classifiers when identifying fractures and other abnormalities in free-text radiology reports. Confidence interval scores at 95% are reported in brackets along the values of Accuracy, Positive Predicted Value (Precision) and Sensitivity (Recall) achieved by the classifiers. True positive (TP), false negative (FN), false positive (FP), and true negative (TN) values are also reported for reference; a positive value refers to the absence of a fracture.

The multinomial Naive Bayes classifier is found to achieve the best overall performance when using bigram or SNOMED+bigram features (no statistical significant differences were found between the performance of the two

models when using a paired t-test with $p = .05$). Although the overall performance of this classifier using these two different set of features do not exhibit statistical significant differences, some of the prediction errors do actually differ. This suggests that SNOMED CT features do extract semantic information that, used in combination with textual information in the form of bigrams, may provide enhancements in the performance of a classifier trained using only bigram features. How these two features should be integrated to enhance classification performance is still an open question. The use of different weighing schemas for each of the extracted feature may be a viable solution.

It is also important to note that the negation feature was consistently used across all feature sets reported in Table 2. This affirms the importance of capturing negations from free-text to correctly identifying the presence or absence of fractures and other abnormalities. Feature sets that did not comprise negations perform significantly lower than their counterparts that included negations (not shown in Table 2).

CONCLUSIONS

The automatic identification of fractures or other abnormalities in free-text radiology reports was studied in this paper. Machine learning algorithms and a wide range of features were tested on a set of 99 radiology reports of limb structures obtained from a large Emergency Department. It was found that automatic techniques based on machine learning algorithms and a combination of stemmed token bigram features, negation features, and SNOMED CT concept features related to morphologic abnormalities and disorders, could classify radiology reports with high efficacy (up to 92.31% F-Measure). While these early results are promising, further work is needed to reach similar performance to those of expert clinicians (i.e. F-Measure of 98.03%).

Further work is therefore needed to provide an automated solution able to identify fractures and other abnormalities from free-text with the same accuracy as expert clinicians. To this aim, future investigation will be directed towards testing the described classifiers and features on larger free-text radiology datasets, as well as studing how the promising combination of bigram and SNOMED CT features may be enhanced.

Acknowledgement

This research was supported by the Queensland Emergency Medicine Research Foundation Grant, EMPJ-11-158-Chu-Radiology.

References

1. Cameron MG. Missed fractures in the emergency department. *Emerg Med (Fremantle)*, 6:3, 1994.
2. Sprivulis P. and Frazer A. Same-day x-ray reporting is not needed in well supervised emergency departments. *Emerg Med (Fremantle)*, 13:194–197, 2001.
3. Mounts J., Clingenpeel J., andE. Byers E. McGuire and Kireeva Y. Most frequently missed fractures in the emergency department. *Clin Pediatr (Phila)*, 50:183–186, 2011.
4. James M. R., Bracegirdle A. and Yates D. W. X-ray reporting in accident and emergency departments – an area for improvements in efficiency. *Arch Emerg Med*, 8:266–270, 1991.
5. Saab M., Stuart J., Randall P. and Southworth S. X-ray reporting in accident and emergency departments – reducing errors. *Eur J Emerg Med*, 4:213–216, 1997.
6. Siegel E., Groleau G., Reiner B. and Stair T. Computerized follow-up of discrepancies in image interpretation between emergency and radiology departments. *J Digit Imaging*, 11:18–20, 1998.
7. De Bruijn B., Cranney A., O'Donnell S., Martin J.D. and Forster A.J. Identifying wrist fracture patients with high accuracy by automatic categorization of x-ray reports. *JAMIA*, 13(6):696–698, 2006.
8. Thomas B.J., Ouellette H., Halpern E.F. and Rosenthal D.I. Automated computer-assisted categorization of radiology reports. *American Journal of Roentgenology*, 184(2):687–690, 2005.
9. Chapman W.W., Chu D. and Dowling J.N. Context: An algorithm for identifying contextual features from clinical text. In *Proceedings of the Workshop on BioNLP 2007*, pages 81–88. ACL, 2007.
10. Nguyen A., Moore J., Lawley M., Hansen D. and Colquist S. Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications. In *Health Informatics Conference*, pages 117–124, 2011.
11. Witten I.H., Frank E. and Hall M.A. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.