Identifying Significance of Discrepancies in Radiology Reports

Arman Cohan[*] Luca Soldaini[*] Nazli Goharian[*]
Allan Fong[†] Ross Filice[‡] Raj Ratwani[†]

Abstract

At many teaching hospitals, it is common practice for on-call radiology residents to interpret radiology examinations; such reports are later reviewed and revised by an attending physician before being used for any decision making. In case there are substantial problems in the resident's initial report, the resident is called and the problems are reviewed to prevent similar future reporting errors. However, due to the large volume of reports produced, attending physicians rarely discuss the problems side by side with residents, thus missing an educational opportunity. In this work, we introduce a pipeline to discriminate between reports with *significant discrepancies* and those with *non-significant discrepancies.* The former contain severe errors or mis-interpretations, thus representing a great learning opportunity for the resident; the latter presents only minor differences (often stylistic) and have a minor role in the education of a resident. By discriminating between the two, the proposed system could flag those reports that an attending radiology should definitely review with residents under their supervision. We evaluated our approach on 350 manually annotated radiology reports sampled from a collection of tens of thousands. The proposed classifier achieves an Area Under the Curve (AUC) of 0.837, which represent a 14% improvement over the baselines. Furthermore, the classifier reduces the False Negative Rate (FNR) by 52%, a desirable performance metric for any recall-oriented task such as the one studied in this work.

1 Introduction

A key aspect of the education of resident radiologists is the development of the necessary skills to interpret radiology examinations and report their findings. Reports are later examined by an experienced attending physician, who revises eventual interpretation errors or minor mistakes. In case the attending performs substantial edits to the report, we say that *significant discrepancies* exist between the initial and the revised report. These discrepancies are due to potential erroneous image interpretation of the resident. Prevention of such errors is essential to the education of the radiology residents as well as the patient care. On the other hand, if a report has been edited to only address minor errors or style issues, we say that *non-significant discrepancies* exists. In Figure 1, examples of significant and non-significant discrepancies are shown (each example is a small section of a much longer report).

Researchers have studied the frequency of discrepancies in radiology reports [28, 24], as well as their impact on resident learning and patient care [23]. Moreover, recent studies have also determined that residents produce less reports that need to be significantly edited by attending radiologists as their experience increase [9].

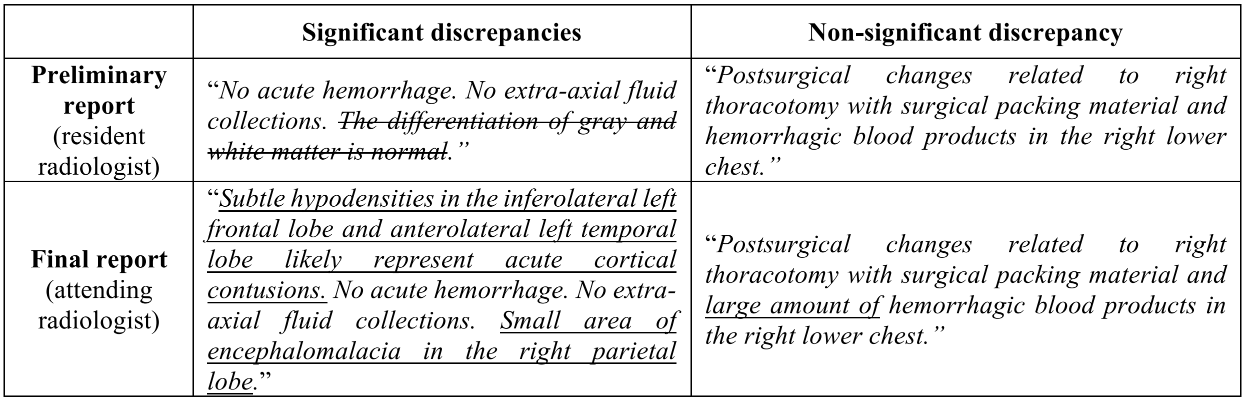
The large volume of radiology reports generated each day makes manual surveillance challenging; thus, in recent years, systems to identify reports that have major discrepancies have been introduced. Sharpe, et al. [25] proposed an interactive dashboard that highlights the differences between reports written by residents alongside the version edited by attending radiologists. Kalaria and Filice [11] used the number of words differing between the preliminary and final report to measure the significance of the discrepancies. However, deviation detected using this measure does not fully capture the difference between reports with significant discrepancies and non-significant ones, as dissimilarities in the writing styles between residents and attending radiologists can also cause differences in word counts.

We propose an accurate and effective two-stage pipeline to distinguish between significant and non-significant discrepancies in radiology reports. In other words, given a set of preliminary radiology reports with the respective final reports, we identify those with significant discrepancies. The first stage of our pipeline

[*] Information Retrieval Lab, Computer Science Department, Georgetown University

[†] National Center for Human Factors in Healthcare, MedStar Health

[‡] Department of Radiology, MedStar Georgetown University Hospital



	Significant discrepancies	Non-significant discrepancy
Preliminary report (resident radiologist)	*"No acute hemorrhage. No extra-axial fluid collections. ~~The differentiation of gray and white matter is normal~~."*	*"Postsurgical changes related to right thoracotomy with surgical packing material and hemorrhagic blood products in the right lower chest."*
Final report (attending radiologist)	*"Subtle hypodensities in the inferolateral left frontal lobe and anterolateral left temporal lobe likely represent acute cortical contusions. No acute hemorrhage. No extra-axial fluid collections. Small area of encephalomalacia in the right parietal lobe."*	*"Postsurgical changes related to right thoracotomy with surgical packing material and large amount of hemorrhagic blood products in the right lower chest."*

Figure 1: Example of significant and non-significant discrepancies between reports. The stroked-through text has been removed from the preliminary report by the attending radiologist, while the underlined sections have been added.

employs an ontology of radiology terms and expressions to identify reports with no significant differences. The remaining reports are then separated by a Support Vector Machine (SVM) classifier. We evaluate the impact of a diverse set of textual, statistical, and assessment score features on the performance of the second-stage classifier. Some of these features have been previously used to assess the quality of the text summarization and machine translation systems. Results illustrate significant improvement over the baseline (up to +14.6% AUC, -52% FNR) and show the effectiveness of the proposed approach. Our focus on false negative rate is motivated by the fact that each missed significant discrepancy is a missed opportunity to educate a resident about a significant error in interpreting an examination.

To summarize, the main contributions of this work are as follows:

- We introduce an approach for automatically classifying the type of discrepancies between preliminary and final radiology reports.
- We explore the use of summarization and machine translation evaluation metrics as features identifying reports with significant discrepancies.
- We provide extensive evaluation of different aspects of the proposed pipeline.

2 Related Works

A related–yet ultimately different–problem to the one studied in this paper is the classification of radiology reports based on their content. In this task, which falls under the text classification domain, the goal is to classify radiology reports into a discrete set of predefined categories. For example, Nguyen and Patrick [19] aimed at grouping radiology reports into cancerous or non-cancerous cases using an SVM. Chapman, et al. [4] presented a system for detecting reports with mediastinal findings associated with inhalational anthrax. Percha, et al. [21] classified reports by breast tissue decomposition using a rule based classification scheme. Johnson, et al. [10] proposed a hybrid approach that combines rules with SVM to classify radiology reports with respect to their findings. Bath, et al. [3] introduced a classifier to determine the appropriate radiology protocol among those available for each disease. Their semi-supervised system takes advantage of the UMLS[1] ontology.

Researchers have also proposed methods for quantifying or comparing the quality of text in various domains. For example, Louis and Nenkova [15] introduced a model for classifying sentences in news articles into general/specific depending on the level of the information carried by each sentence. Their classifier uses word, syntax, and language modeling features. Feng, et al. [7] explored a range of text features such as discourse properties, language modeling features, part-of-speech-based features, and syntactic features to quantify text complexity. Zeng-Treitler, et al. [29] proposed a system to grade the readability of health content; their tool employs lexical, syntactic, semantic and stylistic characteristics to accomplish such goal. Ashok, et al. [2] proposed an SVM classifier based on part of speech and lexical distributions, sentiment features, and grammatical properties to predict the success of novels. Lastly, Louise and Nenkova [16] proposed a model for predicting the appropriate length for a textual content in response to a specific information need.

Another line of related work is detecting plagiarism; systems designed for such task are concerned with determining if a given document was plagiarized from another source. To do so, current approaches in literature attempt to capture the significance of differences between a suspicious text and a source document (e.g., [1, 22, 27]). Most of the previous efforts in plagiarism detection are centered on the retrieval aspect to find the original source of plagiarized content; thus, they focus on information and passage retrieval. Our problem differs from plagiarism detection in that our system takes as input a a candidate-source pair (preliminary and final reports) and attempts at classifying the significance of differences between them; instead, in plagiarism detection, the goal is the retrieval of source document.

3 Methodology

We propose a two stage pipeline for classification of type of discrepancies in radiology reports based on their significance. The overview of our approach is shown in Figure 2. In first stage, we utilize a heuristic based on domain ontology to identify non-significant discrepancies. In next stage, reports that are labeled as significant by the heuristic are processed by a classifier that exploits a variety of textual features. Specifically, we adapt features that are originally used to evaluate text summarization and machine translation systems to our problem. The following sections provide details about each one of these two stages.

3.1 Stage 1: Domain ontology. We first link the significance of the discrepancies to the differences between the domain specific concepts in the reports. To extract domain specific concepts, we use RadLex[1], which is a comprehensive ontology of radiology terms and expressions with about 68K entries.

The domain specific concepts between the preliminary report and the final report are then compared. There might be cases in which there are no difference between the concepts of radiology reports but in one report some concepts are negated. As an example, consider these two sentences: " ... *hypodensities in the inferolateral left frontal lobe* ..." and "... *no hypodensity in the inferolateral left frontal lobe* ...". Although the radiology concepts are identical, the negation might indicate significant discrepancy. Therefore, we also consider the negations in which the RadLex concepts appear to prevent false classification. To detect negations, we use the dependency parse tree of the sentences and a set of seed negation words (*not* and *no*). That is, we mark a radiology concept as negated if these seed words are dependent on the concept. If the RadLex concepts of the reports are identical and the negations are consistent, we classify the type of changes as non-significant. We call this stage, the RadLex heuristic (As indicated in Figure 2). A more comprehensive negation detection algorithm (*NeGex* [5]) was also evaluated; however, its results did not show any significant improvement.

The RadLex heuristic highly correlates with human judgments in identifying non-significant changes, as shown in Section 4.2. However, this simple heuristic is not accurate for detecting the significant discrepancies. In other words, if RadLex terms or their associated negations are not consistent, one can not necessarily classify the report as significant.

3.2 Stage 2: Classification using textual features. In this section, we detail a binary classifier designed to address the shortcoming of the RadLex heuristic, we propose a binary classifier. The classifier uses diverse sets of textual features that aim to capture significance of discrepancies in radiology reports. The features that we use include surface textual features, summarization evaluation metrics, machine translation evaluation metrics, and readability assessment scores. We briefly explain each of these feature sets and provide the intuition behind each one of them.

3.2.1 Surface Textual Features. Previous work used word count discrepancy as a measure for quantifying the differences between preliminary and final radiology reports [11]. We use an improved version of the aforementioned method as one of the baselines. That is, in addition to the word count differences, we also consider the character and sentence differences between the two reports as an indicator of significance of changes.

3.2.2 Summarization evaluation features. Rouge[1] [14], one of the most widely used set of metrics in summarization evaluation, estimates the quality of a system generated summary by comparing it to a set of human generated summaries. Rouge has been proposed as an alternative to manual evaluation of the quality of system generated summaries which can be a long and exhausting process. Rather than using Rouge as evaluation metric, we exploit it as a feature for comparing the quality of the preliminary radiology report with respect to the final report. Higher Rouge scores indicate that the discrepancies between the preliminary and the final reports are less significant.

[1] https://www.nlm.nih.gov/research/umls/

[1] http://www.rsna.org/radlex.aspx

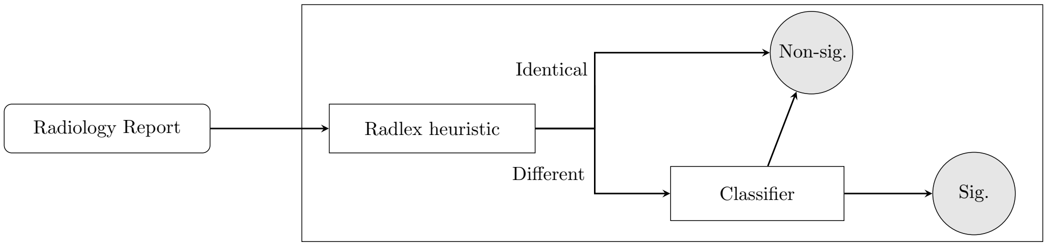


Figure 2: Overview of the proposed approach. The radiology reports are first classified by the Radlex heuristic. If there is no Radlex difference between a preliminary and the associated final report, the case is classified as non-significant discrepancy (*Non-sig* in the figure). Otherwise the case is sent to the a binary classifier for further analysis. The classifier which works based on several textual features, classifies the reports as having either significant (*Sig.* in the figure) or non-significant discrepancies

We utilize the following variants of Rouge:

Rouge-N: In our setting, Rouge-N is the N-gram precision and recall between the preliminary and final report, where N is the gram length (e.g., N=1 indicates a single term, N=2 a word bigram, and so on.) We consider Rouge-1 to Rouge-4.

Rouge-L: This metric compares the two reports based on the Longest Common Subsequence (LCS). Intuitively, longer LCS between the preliminary and the final report shows that the quality of the two reports are closer and therefore differences between the two are less significant.

Rouge-S: Rouge-S computes the skip-bigram co-occurrence statistics between the two reports. It is similar to Rouge-2 except that it allows gaps between the bigrams. Skip-grams are used in different NLP application; they consider additional n-grams by skipping middle tokens. Applying skip-bigrams without any threshold on the distance between tokens often results in incorrect matches (e.g. we do not want to consider all "the the" skip-bigrams in a sentence with multiple "the" expressions). To prevent this, we limit the maximum allowed distance to 10.

3.2.3 Machine translation evaluation features.

The Machine Translation (MT) evaluation metrics quantify the quality of a system-generated translation against a given set of reference or gold translations. We consider the final report as the reference and evaluate the quality of the preliminary report with respect to it. Higher scores indicate a better quality of the preliminary report, showing that the discrepancies between the preliminary and final versions are less significant. In detail, we use the following MT metrics: Bleu [20], Word Error Rate and Meteor [6].

Bleu (Bi-Lingual Evaluation Understudy): In our setting, Bleu is an n-gram based comparison metric for evaluating the quality of a candidate translation with respect to several reference translations. It is conceptually similar to Rouge-N, except being precision-oriented. Specifically, Bleu combines a modified n-gram-based precision and a so-called "Brevity Penalty" (BP), which penalizes short sentences with respect to the reference. Here, we use the Bleu score of the preliminary report with respect to the final report as a feature that indicates the quality of the preliminary report.

Word Error Rate (WER): WER is another commonly used metric for the evaluation of machine translation [26]. It is based on the minimum edit distance between the words of a candidate translation versus reference translations; we consider WER as the following formula:

$$\text{WER} \overset{def}{=} (100 \times (S + I + D)/N)$$

where N is the total number of words in the preliminary report; S, I, and D are the number of Substitutions, Insertions, and Deletions made to the preliminary report to yield the final report.

Metric for Evaluation of Translation with Explicit word Ordering (Meteor): Meteor is a metric for evaluation of machine translation that aligns the translations to the references. Here, we want to find the best alignment between the preliminary report and the final report. In addition to exact matches between terms, Meteor also accounts for synonyms and paraphrase matches between the words and sentences which are not captured by previous features such as

[1] Recall-Oriented Understudy for Gisting Evaluation

		RadLex	A	B
non-significant	RadLex	1.0	0.964	0.942
	A	0.964	1.0	0.906
	B	0.942	0.906	1.0
count=139	Fleiss $\kappa = 0.880$			
significant	RadLex	1.0	0.557	0.492
	A	0.557	1.0	0.934
	B	0.492	0.934	1.0
count=61	Fleiss $\kappa = 0.468$			

Table 1: Agreement rate between the RadLex heuristic and two annotators A and B. Agreement for significant and non-significant reports are separately presented. Both raw agreement rates as well as Fleiss κ between the annotators and the RadLex heuristic are shown.

Rouge. We use both the WordNet [18] synonyms and RadLex ontology synonyms for calculation of the Meteor score.

3.3 Readability assessment features. To quantify complexity of textual content and the style of the reports, we use readability assessment features. Here, "style" refers to reporting style of the radiology reports, such as lexical and syntactic properties. In detail, we use the Automated Readability Index (ARI) [12] and the Simple Measure Of Gobbledygook (SMOG) index [17]. These two metrics are based on distributional features such as the average number of syllables per word, the number of words per sentence, or binned word frequencies. In addition to these statistics, we also consider average phrase counts (noun, verb and propositional phrases) among the features.

4 Empirical Results

4.1 Experimental setup We use a collection of radiology reports with discrepancies obtained from a large urban hospital for evaluation. These reports contain two main textual sections: *findings*, which contains the full interpretation of the radiology examination, and *impression*, which is a concise section that highlights important aspects of the report. We use both sections for evaluation of our proposed pipeline. We use 10 fold cross validation for evaluating the proposed classification scheme.

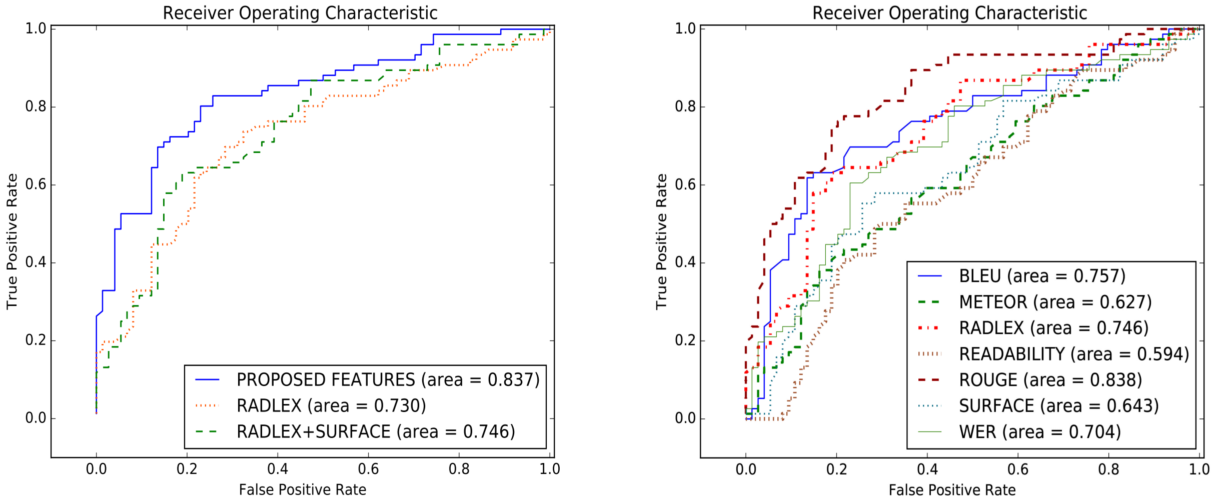
4.2 Classification using RadLex ontology. As explained in Section 3, we first classify the reports using the RadLex ontology and the negation differences between the preliminary and final versions of the report. We ran this method on 200 randomly sampled reports from the dataset; two annotators were asked to label the reports based on significance of discrepancies. The annotators were allowed to label a case as "not-sure" if they could not confidently assign a label for the report. The agreement rates between the annotators and the RadLex heuristic is shown in Table 1. As illustrated, RadLex heuristic is highly correlated with human judgments and the Fleiss κ for non-significant reports is above 0.8, which can be interpreted as perfect agreement [13, 8]. However, the simple RadLex heuristic's performance for the reports that it labels as significant is low. Thus, we conclude that RadLex concept differences between the reports do not necessarily mean that the changes between them is significant. As we show in next section, the proposed classification scheme with the textual features can solve this problem for reports with RadLex differences.

Baselines	F-1	FNR	AUC	ACC
Sf (Improved v. of [11])	0.650	0.329	0.642	0.633
RL	0.690	0.355	**0.746**	**0.707**
Sf+RL	**0.694**	**0.329**	0.730	0.700
Our methods	F-1	FNR	AUC	ACC
Rd	0.568	0.421	0.594	0.553
BL	0.709	0.184*	0.757	0.660
M	0.604	0.368	0.627	0.580
Rg	0.767*	0.197*	0.838*	0.753*
Rg+BL	0.739*	0.237*	0.831*	0.727*
Rg+M	0.775*	0.184*	0.847*	0.760*
Rg+WER	0.702	0.211*	0.746	0.660
Rg+BL+M	0.780*	0.184*	**0.843***	0.767*
Rg+BL+M+RL	0.769*	0.211*	0.841*	0.760*
Rg+BL+M+RL+Rd	**0.797***	**0.171***	0.837*	**0.787***

Table 2: F-1 score (F1) and False Negative Rate (FNR) for significant reports as well as overall Area Under the Curver (AUC) and Accuracy (ACC) based on different set of features. The top part of the table shows the baselines and the bottom part shows our proposed features. Sf: Surface features – character, word and sentence differences; RL: RadLex concepts and their associated negation differences; Rd: Readability features; M: Meteor; BL: Bleu. Rg: Rouge. Asterisk (*) shows statistically significant improvement over all baselines (two-tailed student t-test, $p < 0.05$).

4.3 Classification by textual features. To evaluate our proposed classification approach, a radiologist manually identified types of discrepancies of 150 randomly sampled radiology reports that include RadLex concept differences.

4.3.1 Feature analysis. Table 2 shows the cross validated classification results using the set of features described in Section 3. We use an SVM classifier with linear kernel. We report F-1 score and false negative rates for significant reports, and the overall



(a) Comparison of the proposed pipeline with the baselines

(b) Comparison of individual features.

Figure 3: ROC curves

area under the curve and accuracy. We consider the following baselines: (i) Surface textual features including character, word and sentence differences between the reports (Indicated as "Sf" in the table). (ii) RadLex concepts and associated negation differences (Indicated as "RL"). (iii) Surface textual features along with RadLex concepts and negation differences (RL+Sf). Results based on different sets of features are presented. We experimented with all possible combinations of features; for the sake of brevity, we only report combination of features of significance.

We observe that majority of the proposed features outperform the baseline significantly. One feature set performing worse than the baseline is the readability features. As described in Section 3.3, readability features mostly capture the differences between the reporting styles, as well as the readability of the written text. However, the reporting style and readability of the preliminary and final report might be similar although their content differs. For example, some important radiology concepts relating to a certain interpretation might be contradictory in the preliminary and final report while they both follow the same style. Thus, the readability features on their own are not able to capture significant discrepancies. However, when used with other features such as ROUGE, they are able to capture style differences that are not realized by other features especially in insignificant change category. This causes the performance of combined metrics to increase.

ROUGE features are able to significantly improve over the baseline. When we add METEOR features, we observe a further improvement over ROUGE alone. This is likely due to the fact that METEOR considers synonyms in aligning the sentences as well, which is not captured by ROUGE. However, we note that METEOR by itself underperforms the baseline. We attribute this to the concept drift that may have been caused by consideration of synonyms in METEOR as observed in high FNR of METEOR. The highest scores are achieved when we combine METEOR, ROUGE, BLEU, RadLex and readability features. We attribute the high performance of this setting to different aspects of reporting discrepancies captured by each of the features. ROC curve differences between our best performing features and the baseline (Figure 3a) further shows the effectiveness of our approach. Individual effects of features in terms of ROC curves are also compared in Figure 3b. As shown, ROUGE features are the most informative for identifying significant discrepancies.

4.3.2 Sections of the report.

We evaluated which sections of the radiology report have more influence on the final significance of the discrepancies. As explained in Section 4.1, the reports have two main sections: *findings* and *impression*. As shown in table 3, *impression* section features have higher F-1 scores (+6.68%), lower false negative rates (-31.8%) and higher accuracy (+4.5%) than *findings* section. This is expected, since *impression* contains key points of the report. However, the best results are achieved when both sections are considered, thus indicating that the *findings* section contains valuable information that are not present in the *impression*.

Sections	F-1	FNR	AUC	ACC
Impression	0.772	0.197	0.821	0.760
Findings	0.725	0.289	0.817	0.727
All	0.797	0.171	0.837	0.787

Table 3: Comparison of the results based on features extracted from different sections of the reports.

4.4 Error Analysis. We examined the cases that our approach incorrectly classified. First, many of the false positive cases (i.e., reports that were incorrectly flagged as having significant discrepancies) were due to unnecessarily long length of preliminary reports. We saw that in many cases, the preliminary report, especially in *impression* section, contains extra information that is later removed by the attending editor. In these cases, when almost half of the preliminary report is removed in the final version, our classification scheme fails to classify them as insignificant. According to the domain expert annotator, however, those removed sections do not convey any critical information. Since our features are mostly considering lexical overlaps between the reports, they fail to capture these special cases.

Second, we noticed that some of the false negative cases were due to only slight changes between the two reports. An example is illustrated below which shows a snippet from the preliminary and the final reports:

- **preliminary report**: "*Worsening airspace disease at the left base represents aspiration.*"
- **final report** "*Worsening airspace disease at the left base could represent aspiration.*"

This small change in the report is interpreted as a significant discrepancy between the two reports by the domain expert. Since there is only a slight change between the two reports and the term *could* is not a domain specific term, our features fail to detect this case as significant. In this special case, the term *could* changes a specific interpretation from a definite fact to a possibility, thus can be considered as significant discrepancy.

Although the proposed approach misclassifies these cases, such discrepancies are very rare. In future work, we will focus on designing features that can capture significance of discrepancies in such cases.

5 Conclusions and future work

Identifying significance of discrepancies in radiology reports is essential for education of radiology residents and patient care. We proposed a two-stage pipeline to distinguish between significant and non-significant discrepancies in radiology reports. In the first stage we adopted a heuristic based on the RadLex domain ontology and negations in radiology narratives. In the second stage, we proposed a classifier based on several features including summarization and machine translation evaluation, and text readability features for classification of the reports. We validated our approach using a real world dataset obtained from a large urban hospital. We showed the effectiveness of our proposed pipeline which gains statistically significant improvement (+14.6% AUC, -52% FNR) over the several baselines. A provisional patent based on the proposed approach has been filed at United States Patent and Trademark Office (application number 62280883).

We only focused on the binary classification of changes into two categories: significant and non-significant. Future work will be concerned with exploring the problem of categorizing changes into multiple levels of significance.

Error analysis revealed some rare cases that our features are not designed to capture. Such cases are mostly due to either very small textual differences between the reports that imply significant discrepancy or huge textual differences that do not reflect any significant discrepancies. One natural extension is to design features that can capture such cases. For example, one can consider differences between modality of the reports.

An important goal in detecting significant discrepancies is to prevent future similar problems. One intuitive direction to follow would be clustering discrepancies based on certain textual descriptors. Thus, finding common problems in the collection of initial reports can further promote patient care and resident education.

Acknowledgments

The authors thank Amit Kalaria for helping in annotation and Ophir Frieder for discussions and comments.

References

[1] A. Abdi, N. Idris, R. M. Alguliyev, and R. M. Aliguliyev, *Pdlk: Plagiarism detection using linguistic knowledge*, Expert Systems with Applications, 42 (2015), pp. 8936–8946.

[2] V. G. Ashok, S. Feng, and Y. Choi, *Success with style: Using writing style to predict the success of novels*, Poetry, 580 (2013), p. 70.

[3] A. Bhat, G. Shih, and R. Zabih, *Automatic selection of radiological protocols using machine learning*, in Proceedings of the 2011 workshop on Data mining for medicine and healthcare, ACM, 2011, pp. 52–55.

[4] W. W. Chapman, G. F. Cooper, P. Hanbury, B. E. Chapman, L. H. Harrison, and M. M. Wagner, *Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders*, Journal of the American Medical Informatics Association, 10 (2003), pp. 494–503.

[5] W. W. Chapman, D. Hilert, S. Velupillai, M. Kvist, M. Skeppstedt, B. E. Chapman, M. Conway, M. Tharp, D. L. Mowery, and L. Deleger, *Extending the negex lexicon for multiple languages*, Studies in health technology and informatics, 192 (2013), p. 677.

[6] M. Denkowski and A. Lavie, *Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems*, in Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics, 2011, pp. 85–91.

[7] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad, *A comparison of features for automatic readability assessment*, in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010, pp. 276–284.

[8] A. M. Green, *Kappa statistics for multiple raters using categorical classifications*, in Proceedings of the 22nd annual SAS User Group International conference, vol. 2, 1997, p. 4.

[9] G. Issa, B. Taslakian, M. Itani, E. Hitti, N. Batley, M. Saliba, and F. El-Merhi, *The discrepancy rate between preliminary and official reports of emergency radiology studies: a performance indicator and quality improvement method*, Acta Radiologica, 56 (2015), pp. 598–604.

[10] E. Johnson, W. C. Baughman, and G. Ozsoyoglu, *Mixing domain rules with machine learning for radiology text classification*, (2014).

[11] A. D. Kalaria and R. W. Filice, *Comparison-bot: an automated preliminary-final report comparison system*, Journal of digital imaging, (2015), pp. 1–6.

[12] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom, *Derivation of new readability formulas*, tech. report, Technical report, TN: Naval Technical Training, US Naval Air Station, Memphis, TN, 1975.

[13] J. R. Landis and G. G. Koch, *The measurement of observer agreement for categorical data*, biometrics, (1977), pp. 159–174.

[14] C.-Y. Lin, *Rouge: A package for automatic evaluation of summaries*, in Text summarization branches out: Proceedings of the ACL-04 workshop, vol. 8, 2004.

[15] A. Louis and A. Nenkova, *Automatic identification of general and specific sentences by leveraging discourse annotations.*, in IJCNLP, 2011, pp. 605–613.

[16] A. Louis and A. Nenkova, *Verbose, laconic or just right: A simple computational model of content appropriateness under length constraints*, EACL 2014, (2014), p. 636.

[17] G. H. McLaughlin, *Smog grading: A new readability formula*, Journal of reading, 12 (1969), pp. 639–646.

[18] G. A. Miller, *WordNet: a lexical database for English*, Communications of the ACM, 38 (1995), pp. 39–41.

[19] D. H. Nguyen and J. D. Patrick, *Supervised machine learning and active learning in classification of radiology reports*, Journal of the American Medical Informatics Association, 21 (2014), pp. 893–901.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, *Bleu: a method for automatic evaluation of machine translation*, in Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 311–318.

[21] B. Percha, H. Nassif, J. Lipson, E. Burnside, and D. Rubin, *Automatic classification of mammography reports by bi-rads breast tissue composition class*, Journal of the American Medical Informatics Association, 19 (2012), pp. 913–916.

[22] M. Potthast, M. Hagen, M. Völske, and B. Stein, *Crowdsourcing interaction logs to understand text reuse from the web.*, in ACL (1), 2013, pp. 1212–1221.

[23] A. T. Ruutiainen, D. J. Durand, M. H. Scanlon, and J. N. Itri, *Increased error rates in preliminary reports issued by radiology residents working more than 10 consecutive hours overnight*, Academic radiology, 20 (2013), pp. 305–311.

[24] A. T. Ruutiainen, M. H. Scanlon, and J. N. Itri, *Identifying benchmarks for discrepancy rates in preliminary interpretations provided by radiology trainees at an academic institution*, Journal of the American College of Radiology, 8 (2011), pp. 644–648.

[25] R. E. Sharpe Jr, D. Surrey, R. J. Gorniak, L. Nazarian, V. M. Rao, and A. E. Flanders, *Radiology report comparator: a novel method to augment resident education*, Journal of digital imaging, 25 (2012), pp. 330–336.

[26] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, *A study of translation edit rate with targeted human annotation*, in Proceedings of association for machine translation in the Americas, 2006, pp. 223–231.

[27] E. Stamatatos, M. Potthast, F. Rangel, P. Rosso, and B. Stein, *Overview of the pan/clef 2015 evaluation lab*, in Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer, 2015, pp. 518–538.

[28] J. Walls, N. Hunter, P. M. Brasher, and S. G. Ho, *The depictors study: discrepancies in preliminary interpretation of ct scans between on-call residents and staff*, Emergency radiology, 16 (2009), pp. 303–308.

[29] Q. Zeng-Treitler, L. Ngo, S. Kandula, G. Rosemblat, H.-E. Kim, and B. Hill, *A method to estimate readability of health content*, Association for Computing Machinery, (2012).