



Ranking Significant Discrepancies in Clinical Reports

Sean MacAvaney¹(✉), Arman Cohan², Nazli Goharian¹, and Ross Filice³

¹ IR Lab, Georgetown University, Washington DC, USA
{sean,nazli}@ir.cs.georgetown.edu

² Allen Institute for Artificial Intelligence, Seattle, WA, USA
armanc@allenai.org

³ Department of Radiology, MedStar Georgetown University Hospital,
Washington DC, USA
Ross.W.Filice@medstar.net

Abstract. Medical errors are a major public health concern and a leading cause of death worldwide. Many healthcare centers and hospitals use reporting systems where medical practitioners write a preliminary medical report and the report is later reviewed, revised, and finalized by a more experienced physician. The revisions range from stylistic to corrections of critical errors or misinterpretations of the case. Due to the large quantity of reports written daily, it is often difficult to manually and thoroughly review all the finalized reports to find such errors and learn from them. To address this challenge, we propose a novel ranking approach, consisting of textual and ontological overlaps between the preliminary and final versions of reports. The approach learns to rank the reports based on the degree of discrepancy between the versions. This allows medical practitioners to easily identify and learn from the reports in which their interpretation most substantially differed from that of the attending physician (who finalized the report). This is a crucial step towards uncovering potential errors and helping medical practitioners to learn from such errors, thus improving patient-care in the long run. We evaluate our model on a dataset of radiology reports and show that our approach outperforms both previously-proposed approaches and more recent language models by 4.5% to 15.4%.

1 Introduction

Medical errors are a pervasive problem in healthcare that can result in serious patient harm [11]. To identify and reduce the occurrence of preventable errors, many medical centers use reporting systems to document cases. Initial reports are often reviewed and revised by more experienced physicians. The revisions could be due to stylistic reasons or (more importantly) misinterpretations/errors in the initial report. In such cases, to prevent recurrence of the errors, is crucial to identify reports with substantive differences between the original and final report and discuss them with the clinician who wrote the initial report. It is

often challenging to manually identify such cases among the large number of daily written reports in a timely manner. In this work, we propose an approach for ranking revisions of medical reports by the degree of discrepancy between the different versions of the report. This allows medical practitioners to easily find the reports in which they made an error, which helps them learn from their mistakes and prevent future similar errors.

This is a challenging task to automate because the edits that an attending physician makes to a report can range from stylistic differences to significant discrepancies that may have a major effect on the patient (e.g., an unobserved mass). See Fig. 1 for an example of significant and non-significant discrepancies from radiology reports. As we can see, differences between the significant and non-significant discrepancies are often not trivial to identify and requires more than just comparing surface word changes in the versions of the reports. Furthermore, significant discrepancies can occur relatively frequently in practice; in our dataset collected from a large urban hospital, around 7% of reports contained significant errors. With hundreds of reports generated a week at some hospitals, this can amount to a considerable number of errors. We address this problem by proposing a supervised ranking approach for clinician’s revised reports by the degree that there are significant discrepancies between their preliminary report and the final corrected report. I.e., our goal is to rank revisions that are more likely due to errors higher than revisions that are merely due to stylistic changes.

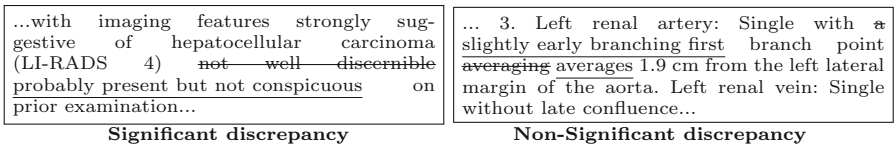


Fig. 1. Example radiology impression revisions (strikeout removed, underline added). We aim to rank report revisions by the significance of the discrepancy.

Prior works have investigated significant discrepancies in medical reports through comparison of surface textual features [14, 19], semantic similarity features [2], and word frequencies [7]. These works often treat the problem as classification and the most successful ones leverage a variety of textual similarity measures. Viewing this problem as ranking is a more suitable and practical form of evaluation; given a doctor’s limited time, it is important for them to be presented with the reports that have the most significant discrepancies.

Document ranking in the broad medical domain have received extensive interest of researchers [8, 13, 15–17, 21]. However, these efforts focus on conventional query-document retrieval. Our goal is to rank significant discrepancies by measuring the semantic overlap between the initial and final report. There have also been efforts to identify semantic similarity between two texts, e.g., for paraphrase identification [5, 9, 12, 18], but these approaches operate on the sentence-level, making them unsuitable for documents (e.g., radiology reports).

To summarize, our contributions are: (i) we propose an end-to-end supervised ranking model for identifying significant discrepancies in medical reports. (ii) We demonstrate that our approach outperforms both previously proposed approaches and more recent language model approaches in a variety of metrics. (iii) We provide an analyses of the importance of different model components.

2 Model

We propose a supervised model that measures the overlap between the preliminary and final report for the purpose of ranking pairs of preliminary and final reports based on their significance over a given period of time. We observe that the central challenge of this task is being permissive of surface-level changes (which may be considerable), while emphasizing changes of substance, which may be subtle (see Fig. 1 for examples of such changes). To address this, we incorporate *importance* and *similarity* scores. The *importance score* weights each term/phrase and is learned during training. This score allows for terms that are not important to have less of an impact on the ranking score of the report (e.g., words like *well* and *but*). Note that this is a special application in which some function words that are often ignored actually have a big impact on the meaning of a report (e.g., *not* is often considered a stop word and removed). We let the model learn which terms are important during training. The *matching score* allows for the model to account for the replacement of similar terms using the cosine distance of word vectors (e.g., *averaging* and *averages* are similar) and synonym information from a domain-specific ontology (*chauffeur fracture* and *Hutchinson fracture* are synonymous). This allows the replacement of semantically-similar terms to have little impact on the ranking score. We calculate three *similarity scores* (addition, deletion, and overlap) using the importance and matching scores, and linearly combine them as a *ranking score*.

Notation and Task Definition. Let R be a set of clinical reports. Each report $r \in R$ consists of a preliminary and final version of the report (p and f , respectively), and a label $l \in \{0, 1, \dots, L\}$ indicating the degree of discrepancy between p and f . Each version of the report consists of a sequence of tokens, denoted by p_i and f_i . The significant discrepancy ranking task produces a ranking score $s \in \mathbb{R}$ for each report $r \in R$ such that the reports with higher degrees of discrepancy are assigned a higher ranking score.

Similarity Scores. Our approach combines several similarity scores to produce a ranking score. Specifically, we measure the weighted soft additions, deletions, and overlap of unigrams, n-grams, and ontological entities. The addition score (S_a , Eq. 1) defines weighted soft similarity as the ratio between the similarity score (weighted by a learned importance score) and the total importance of all terms in the final report. Thus, terms from the final report that do not appear in the preliminary report (i.e., additions) yield a higher score. The deletion score (S_d , Eq. 2) is defined similarly, but in terms of the preliminary report; terms from the preliminary report that do not appear in the final report (deletions)

yield a higher score. The overlap score (S_o , Eq. 3) combines the addition and deletion scores into one succinct measure. We use all three scores to measure term unigram, n-gram, and ontological differences (defined below). We define the similarity functions (where $M_X(y) \in [0, 1]$ is a matching score of term y in X , and $I(y) \in [0, 1]$ is the importance score of term y) as:

$$S_a(p, f) = -\frac{\sum_{f_i \in f} M_p(f_i) I(f_i)}{\sum_{f_i \in f} I(f_i)} \quad (1) \quad S_d(p, f) = -\frac{\sum_{p_i \in p} M_f(p_i) I(p_i)}{\sum_{p_i \in p} I(p_i)} \quad (2)$$

$$S_o(p, f) = -\frac{\sum_{p_i \in p} M_f(p_i) I(p_i) + \sum_{f_i \in f} M_p(f_i) I(f_i)}{\sum_{p_i \in p} I(p_i) + \sum_{f_i \in f} I(f_i)} \quad (3)$$

Unigram and N-gram Matching. Unigram matching can provide valuable signals for significance in radiology reports. For instance, the addition *no* (e.g., *fracture* vs. *no fracture*) could change the meaning of the report considerably. We define the matching function for unigrams as the maximum cosine similarity between the word embeddings ($emb(\cdot)$) of the term and any term in the other report, and a unigram importance function using a simple feed-forward layer with sigmoid activation (W_{imp} and b_{imp} as model parameters):

$$M_X(y) = \max_{x \in X} (\cos(emb(x), emb(y))) \quad (4)$$

$$I(y) = \sigma(emb(y)W_{imp} + b_{imp}) \quad (5)$$

N-gram matching provides another important view of similarity, since there are many multi-word noun phrases in radiological notes. For instance, *right arm* and *left arm* represent completely different parts of the body, and should be treated differently. We handle n-grams by first taking the average of the embeddings over sliding windows. This is a simple and effective way to combine the representations. We use bi-grams and tri-grams in our experiments.

Ontological Matching. Since medical knowledge is broad and extensive, the model may never encounter certain medical entities during training. This knowledge may also not be captured effectively by embeddings. Thus, it is valuable to explicitly encode domain information into the model using an ontology. We use a mapping function that matches any exact ontological name to the corresponding concept, a constant similarity for exact entity matches, and constant weight for all ontology concepts. We use RadLex (v4.0, <http://radlex.org/>), an ontology of radiology concepts (e.g., procedures, diagnoses, etc.).

3 Experiments

Dataset. We train and evaluate using a dataset of 3,368 radiology reports from a large urban hospital. Each sample consists of a preliminary report written by

a resident, and a final report revised by the attending radiologist, who labeled the edit by the degree of discrepancy between the two reports. The labels are 0 (attending doctor fully agrees with assessment of the resident, 81% of reports), 1 (errors exist, but they are insignificant to the overall impression, 12%), 2 (subtle, yet important, error exists, 6%), 3 (an obvious error exists, 1%). We split the dataset into 122 sets based on the combination of resident and week (*ranking sets*, average 27.6 reports per ranking set, min 5, max 148). Since residents often work weekly shifts, this is a valuable setting because it allows residents to review report discrepancies from the past week. We randomly split the ranking sets into 60-20-20 train-dev-test set splits. Each ranking set consists of at least 5 reports, each with at least one report discrepancy. Radiology reports contain several sections; we primarily concern ourselves with the summary section of the reports (called the *impression*) because it contains the main findings.

Baselines. To evaluate the effectiveness of the model, we compare with the variety of methods in the state-of-the-art, including ranking models, domain specific models and textual similarity models, briefly described below:

- **Vector space model (VSM).** We use the traditional TFIDF-weighted vector space similarity score between the preliminary and final report (from lucene).
- **BiPACRR.** We test the PACRR [6] neural IR model because it learns to identify n-gram similarity between two texts. We modify the architecture to learn two scores (one with the preliminary report as the query and the other with the final report as the query), and linearly combine them to produce a final ranking score. We call this variant BiPACRR. We also experimented with other neural rankers (e.g., KNRM [20]), but BiPACRR was the most effective.
- **Textual similarity regression (SimReg) [2].** This approach uses logistic regression to combine several hand-crafted features (mostly consisting of textual similarity measures and lexical features) to identify significant discrepancies in radiology reports. Since this approach performs classification, we use the label score as the ranking score. Our experiments used the authors’ implementation.
- **(Sci)BERT classification.** We use the standard fine-tuned BERT textual similarity method on both the pretrained BERT [4] (**base-uncased**) and SciBERT [1] (**scivocab-uncased**) models. Based on preliminary parameter tuning, we use a learning rate of 10^{-5} for fine-tuning these models.

Evaluation Metrics. Given the time constraints of doctors, we choose evaluation metrics that emphasize placing reports with higher discrepancies at the top. We evaluate using nDCG@1, nDCG@5, nDCG (without cutoff), P@1, P@5, and R-Prec (binary labels test any degree of discrepancy higher than label 0).

Parameters and Training. We train the neural models using pairwise cross-entropy loss [3]. Hyper-parameters are tuned using nCDG@5 on the dev set. We use SciBERT term embeddings [1] in our model and BiPACRR and tune for the

optimal layer’s embeddings akin to [10]. SciBERT is an adaptation of BERT to the biomedical and scientific domains, making it suitable for radiology notes.

Results. Test set performance of our best model configuration are shown in Table 1. Our optimal model consists of unigram, bi-gram, tri-gram, and RadLex scores. When compared to the best prior work (SimReg [2]), our model typically yields a considerable improvement in ranking performance. Our method improves R-Prec by 7.4%, nDCG@1 and nDCG@5 performance by 4.5%, and P@5 performance by 4.7%. In 54% of the test cases, our approach improves the nDCG@5 score over SimReg (decreases performance in only 27% of cases). Our model also outperforms leading language model classification approaches (BERT and SciBERT) and a leading neural ranking approach tuned for this task in most metrics (BiPACRR) by up to 15.4% in nDCG@1. We attribute this improved effectiveness of our approach to the explicit modeling of term importance and overlap, which are critical for the task.

Table 1. Ranking performance of our method and baselines.

Model	nDCG@1	nDCG@5	nDCG	P@1	P@5	R-Prec
VSM	48.1	54.0	70.9	65.4	42.3	49.4
BERT	59.0	69.8	78.7	69.2	53.8	53.9
SciBERT	62.2	68.2	79.3	76.9	51.5	58.3
BiPACRR	64.1	68.6	77.5	69.2	56.2	55.3
SimReg	69.9	70.7	81.1	80.8	51.5	51.8
Our method	74.4	75.2	83.7	80.8	56.2	59.2

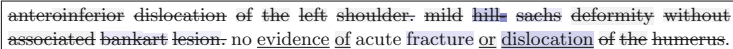
Table 2. Ablation study of our method.

Model	nDCG@5
Full model	75.2
- Replace SciBERT with BERT	64.7
- Replace SciBERT with BioNLP (pubmed-pmc , bio.nlplab.org)	62.2
- Replace SciBERT with FastText (wiki-news-300d-1M , fasttext.cc)	59.1
- Without term importance	68.3
- Without ontology similarity	65.4
- Only overlap score (S_o)	70.8
- Only addition/deletion scores (S_a and S_d)	61.5

Ablations. Table 2 shows the ablation study examining the importance of different components in our system. We observe that both contextualization and domain-specificity of the word embeddings improve the performance of our approach. The term importance mechanism improves nDCG@5 by 6.9% and the

ontology similarity improves performance by 9.8%. All three similarity measures appear to be important, however the overlap score alone can account for most of the performance (last row in table). This may be because it succinctly accounts for both additions and deletions.

Term Importance. To better understand the term importance mechanism of our approach, we present an example report in Fig. 2 (slightly altered for privacy). This report contains highly significant discrepancies and was ranked at position 3 by our approach and position 9 by SimReg (below several non-significant discrepancies). We observe that our model considers many radiological conditions as important, both when unmodified between the reports and when added/deleted (e.g., *fracture*, *dislocation*, *bankart*). Judging by the low textual similarity in this example, we conclude that the SimReg model may be relying too heavily on lexical features. We check the terms that are assigned high importance scores across all reports and find the most common are *no* (12% of reports), *cardiopulmonary* (3%), *process* (3%), and *abnormality* (3%).



anteroinferior dislocation of the left shoulder. mild ~~hills~~ sachs deformity without associated bankart lesion. no evidence of acute fracture or dislocation of the humerus.

Fig. 2. Example unigram importance scores (mean of preliminary and final report). Darker colors indicate higher scores. Underlines: additions. Strikeouts: deletions.

Conclusions. We presented a supervised ranking model based on lexical and ontological overlaps to rank medical reports by their discrepancy significance. On a real-world dataset of medical reports, we demonstrated that our approach outperforms existing approaches by large margins. This direction is a critical step towards addressing the problem of medical errors. By allowing medical practitioners to more easily find and learn from their previous errors, the chance of recurrent errors will be reduced, improving the well-being of patients.

Acknowledgements. This work was supported in part by ARCS Foundation.

References

1. Beltagy, I., Cohan, A., Lo, K.: SciBERT: pretrained contextualized embeddings for scientific text. In: EMNLP (2019)
2. Cohan, A., Soldaini, L., Goharian, N., Fong, A., Ross, F., Raj, R.: Identifying significance of discrepancies in radiology reports. In: SIAM International Conference on Data Mining (SDM) - Workshop on Data Mining for Medicine and Healthcare (DMMH) (2016)
3. Deghani, M., Zamani, H., Severyn, A., Kamps, J., Croft, W.B.: Neural ranking models with weak supervision. In: SIGIR (2017)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)

5. Gan, Z., Pu, Y., Henao, R., Li, C., He, X., Carin, L.: Learning generic sentence representations using convolutional neural networks. In: EMNLP (2016)
6. Hui, K., Yates, A., Berberich, K., de Melo, G.: PACRR: a position-aware neural IR model for relevance matching. In: EMNLP (2017)
7. Kalaria, A.D., Filice, R.W.: Comparison-bot: an automated preliminary-final report comparison system. *J. Digit. Imaging* **29**, 325–330 (2016). <https://doi.org/10.1007/s10278-015-9840-2>
8. Koopman, B., Cripwell, L., Zuccon, G.: Generating clinical queries from patient narratives: a comparison between machines and humans. In: SIGIR 2017 (2017)
9. Liu, L., Yang, W., Rao, J., Tang, R., Lin, J.: Incorporating contextual and syntactic structures improves semantic similarity modeling. In: EMNLP 2019 (2019)
10. MacAvaney, S., Yates, A., Cohan, A., Goharian, N.: CEDR: contextualized embeddings for document ranking. In: SIGIR (2019)
11. Makary, M.A., Daniel, M.: Medical error—the third leading cause of death in the US. *BMJ* **353**, i2139 (2016)
12. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: EMNLP (2019)
13. Roberts, K., et al.: Overview of the TREC 2017 precision medicine track, pp. 500–324. NIST Special Publication (2017)
14. Ruutiainen, A.T., Scanlon, M.H., Itri, J.N.: Identifying benchmarks for discrepancy rates in preliminary interpretations provided by radiology trainees at an academic institution. *J. Am. Coll. Radiol.* **8**(9), 644–648 (2011)
15. Saleh, S., Pecina, P.: Term selection for query expansion in medical cross-lingual information retrieval. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) ECIR 2019. LNCS, vol. 11437, pp. 507–522. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15712-8_33
16. Sankhavara, J.: Biomedical document retrieval for clinical decision support system. In: ACL, pp. 84–90 (2018)
17. Soldaini, L., Yates, A., Goharian, N.: Denoising clinical notes for medical literature retrieval with convolutional neural model. In: CIKM (2017)
18. Tien, H.N., Le, M.N., Tomohiro, Y., Tatsuya, I.: Sentence modeling via multiple word embeddings and multi-level comparison for semantic textual similarity. *Inf. Process. Manage.* **56**, 102090 (2018)
19. Walls, J., Hunter, N., Brasher, P.M., Ho, S.G.: The DePICTORS study: discrepancies in preliminary interpretation of CT scans between on-call residents and staff. *Emerg. Radiol.* **16**(4), 303–308 (2009). <https://doi.org/10.1007/s10140-009-0795-9>
20. Xiong, C., Dai, Z., Callan, J.P., Liu, Z., Power, R.: End-to-end neural ad-hoc ranking with kernel pooling. In: SIGIR (2017)
21. Yates, A., Goharian, N., Frieder, O.: Relevance-ranked domain-specific synonym discovery. In: de Rijke, M., et al. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 124–135. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06028-6_11