Classifying Electronic Consults for Triage Status and Question Type

Xiyu Ding[1,2] and **Michael L. Barnett**[2,3] and **Ateev Mehrotra**[4] and **Timothy A. Miller**[1,4]
[1]Boston Children's Hospital, Boston, MA
[2]Harvard T.H. Chan School of Public Health, Boston, MA
[3]Brigham and Women's Hospital, Boston, MA
[4]Harvard Medical School, Boston, MA

Abstract

Electronic consult (eConsult) systems allow specialists more flexibility to respond to referrals more efficiently, thereby increasing access in under-resourced healthcare settings like safety net systems. Understanding the usage patterns of eConsult system is an important part of improving specialist efficiency. In this work, we develop and apply classifiers to a dataset of eConsult questions from primary care providers to specialists, classifying the messages for how they were triaged by the specialist office, and the underlying type of clinical question posed by the primary care provider. We show that pre-trained transformer models are strong baselines, with improving performance from domain-specific training and shared representations.

1 Introduction

Electronic consult (eConsult) systems allow primary care providers (PCPs) to send short messages to specialists when they require specialist input. In many cases, a simple exchange of messages precludes the need for a standard in-person referral. eConsult systems decrease wait times for a specialty appointment. (Barnett et al., 2017) An example eConsult question is shown in Figure 1. In general, these questions are much shorter than, say, electronic health record texts. There is a stereotypical structure to these questions, including short histories, descriptions of the current problem, and questions about diagnosis, medication management, procedures, or other issues. When the message is received by a specialist's office, specialist reviewers in that office determine whether the patient needs to be scheduled for a specialist visit or whether the specialist may be able to answer a PCP's question directly without a visit. If a visit needs to be scheduled, the specialists decide whether it is urgent or not (in practice, whether the patient goes to the front of the queue). Because these eConsult messages are unstructured, health systems do not know how they are used. Automatically extracting information about the content and response to these questions can help health systems better understand the specialist needs of their PCPs and guide population health management. Accurately classified eConsults can inform decision-making about how to allocate resources for quality improvement, additional specialist investment and medical education to best serve their patient population.

<age> year old woman with newly diagnosed dermatomyositis who also has significant dysphagia to solids greater than liquids. She has been started on prednisone and methotrexate. She is originally from <country> and has had no prior colon cancer screening. We would appreciate an evaluation for both upper endoscopy and colonoscopy. Upper endoscopy to evaluate her dysphagia and colonoscopy for malignancy screening (dermatomyositis patients are at increased risk for malignancy)

Figure 1: An example eConsult question

In this work, we use standard support vector machine (SVM)-based baselines and transformer-based pre-trained neural networks (i.e., *BERT models) to classify eConsult questions along two dimensions, focusing on referrals to gastroenterology and liver specialists.

First, we build classifiers that attempt to learn to predict triage status (e.g., urgent or non-urgent) assigned to questions by the specialist reviewer. Our goal is to use the ability (or inability) of classifiers to perform this task to understand the consistency of scheduling decisions across individual clinicians. This addresses a concern that specialist reviewers vary too much in their judgment on whether a visit is urgent or non-urgent. To do this, we performed experiments that compare classifiers trained on triage decisions of single specialist reviewers. The magnitude of inconsistency or unexplainable

Proceedings of the 1st Workshop on NLP for Medical Conversations, pages 1–6

https://doi.org/10.18653/v1/P17

decisions among reviewers would inform whether these systems can consistently work as intended to reduce specialist visits safely and effectively.

Second, we build classifiers for the task of understanding the implicit information need that is the cause for the PCP asking the question – we call this *question type*. We developed an annotation scheme and annotated a sample of questions from across eight years for five question types. We then train and evaluate several classifier models, including standard architectures with problem-specific additions. Our results show that triage status is difficult to learn in general, but even more difficult between reviewers, suggesting inconsistent reviewer decisions may be occurring. When classifying question type, the best-performing models are domain-specific pre-trained transformers, and that jointly training to predict different question types is the most effective technique. Our best result occurs when combining domain-specific vocabularies with multi-task learning, suggesting that there is a synergistic effect between these two augmentations.

2 Background

BERT (Bidirectional Encoder Representations from Transformers), along its variants, have been proven to outperform other contextual embedding (e.g. ELMo (Peters et al., 2018)) or traditional word embedding models (e.g. Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), etc.) in a wide variety NLP tasks.

BERT learns contextual embeddings through pre-training on a large unlabeled corpus (including the BooksCorpus (800M words) and English Wikipedia (2,500M words)) via two tasks, a masked language model task and a next sentence prediction task (Devlin et al., 2019).

Domain-specific BERT models have been released, including BioBERT (Lee et al., 2020), which started from a BERT checkpoint and extended pre-training on biomedical journal articles, SciBERT (Beltagy et al., 2019), which is pre-trained from scratch with its own vocabulary, and ClinicalBERT (Alsentzer et al., 2019) which started from BERT checkpoints and extended pre-training using intensive care unit documents from the MIMIC corpus (Johnson et al., 2016). In this work, we use vanilla BERT, SciBERT, and two versions of ClinicalBERT, Bio+Clinical BERT and Bio+Discharge Summary BERT[1].

[1]Bio+Clinical BERT and Bio+Discharge Summary BERT are initialized from BioBERT and respectively trained using MIMIC notes from all types and the notes from discharge summary only.

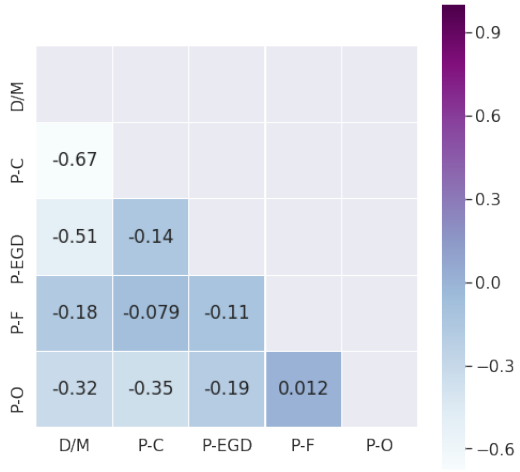

Figure 2: Normalized pointwise mutual information between categories in the question type annotations. Values close to 0 represent variables whose distributions are independent, values above 0 (up to 1) represent pairs of variables that are more likely than chance to occur together, and values below 0 (down to -1) represent pairs of variables that are less likely than chance to occur together.

3 Materials and Methods

3.1 Data

We use de-identified text data from 2008-2017 from the San Francisco Department of Public Health (SFDPH), for which we examined one specialty (gastroenterology and liver) with over 33,000 eConsults.[2] For each eConsult question there are four possible scheduling decisions, *Initially Scheduled* (IS – added to the end of the specialist visit queue), *Not Scheduled* (NS – not added to the queue, typically was resolved via a return message), *Overbook* (OB – added to the front of the queue) and *Scheduled After Review* (SAR – added to the end of the queue after deliberation or additional exchange of messages). Each eConsult also contains meta-data, including a unique identifier referring to the specialist reviewer who first reviewed that question which we use later to train reviewer-specific models. This data was obtained by Boston Children's Hospital under a data use agreement with SFDPH, but unfortunately the terms of that agreement do not allow for public release of the dataset.

For the question type annotation, there are five possible types, *Diagnosis/Management*

[2]This research was approved by our institution's Institutional Review Board as "Not human subjects research."

(D/M), *Procedure-EGD*[3] (P-EGD), *Procedure-Colonoscopy* (P-C), *Procedure-Other* (P-Other), and *Procedure-FlexSig* (P-F). These types are *not* mutually exclusive – a question could, for example, request both a colonoscopy and an EGD. Figure 2 shows the normalized pointwise mutual information between each of the five question type categories. For that reason, it should be modeled as multi-label classification tasks, rather than a multi-class classification task.

This set of categories was created by an iterative process, where clinical experts coded samples of a few dozen eConsult questions at a time, refining after each iteration, with the goal of striking a balance between informativeness to leaders at these health care systems, learnability, and ease of annotation. The annotator who performed the question type annotations is a certified medical coder with several decades of experience in coding clinical documents for various administrative and NLP-related categories. We double annotated a portion of the data and scored them for inter-annotator agreement using Cohen's Kappa. Agreements 0.76 for D/M, 0.94 for P-C, 0.87 for P-EGD, and 0.29 for P-O. P-O is difficult to annotate reliably because it is not clear when it needs to be annotated at all – it is a bit of a default category that probably needs clearer instructions for when it should be annotated.

For the triage classifier, we can use all questions in the data set, because they contain automatic labels for the triage decisions that were made by the specialist reviewers. For the question type classifier, we use a sample of 969 questions annotated by our trained medical coder. We divided the data into training, development, and test splits for training classifiers with an 70%/10%/20% split.

3.2 Machine Learning Algorithms

3.2.1 SVM with Bag Features

The simplest method for classifying text uses an SVM with "bag-of-words" (BoW) features. The text is represented by a "feature vector" $\mathbf{v}$ of size V (i.e., vocabulary size) while the value of i^{th} element, v_i equals to the frequency of the i^{th} word in the vocabulary in the document. A generalization of BoW is "bag-of-n-grams" (BoN). N-grams is a contiguous sequence of n items from a given sample of text. A bag-of-N-grams model has the simplicity of the BoW model, but allows the preservation of more word locality information. In this study, we combine the words and n-grams to create the features. Optimal number of n-grams and the hyper-parameter C of SVM are selected by grid search with 3-fold cross-validation. We performed SVMs with BoN features for both tasks as the baseline reference given that it is surprisingly strong for many tasks in document classification.

One mutation of BoW in the clinical domain is the "bag of CUIs" (BoC). CUIs, or Concept Unique Identifiers map the text spans to medical dictionaries and words with the same medical implications are unified to the same concepts. We use Apache cTAKES (Savova et al., 2010) to extract the medical concepts existing in the text data and apply an SVM on the bag of concepts.

We use the Scikit-Learn implementation of SVMs to implement the training and inference (Pedregosa et al., 2011).

3.2.2 BERT Models

We fine-tune the models (updating the weights of the encoders and classification layer) on our tasks with four different versions of BERT models, BERT (base-uncased), SciBERT (base-uncased), Bio+Clinical BERT (base-cased) and Bio+Discharge Summary BERT (base-cased). For both tasks, we use the last hidden state of the [CLS] token as the aggregate sequence representation. The [CLS] representation is fed into an output layer (softmax for the triage status classifier, sigmoid for the question type classifiers) to get the predicted probability of all labels. All parameters from BERT are fine-tuned by minimizing the overall cross-entropy loss. We use the HuggingFace Transformers library (Wolf et al., 2019) for our BERT implementations.[4] We monitor the training and validation loss for each training epoch and save the model with the highest Macro-F1 score on the validation set before testing on the test split.

3.2.3 Multi-task BERT

For the question task, we also explore a multi-task learning scheme which allows us to jointly fine tune BERT for predicting all the labels with the same model. This forces the fine-tuning process to learn representations that are good for multiple tasks, which can potentially benefit as both regularization and by indirectly sharing information between labels that are known to be correlated. For this model, the same [CLS] representation is fed

[3] Esophagogastroduodenoscopy

[4] https://github.com/huggingface/transformers

	IS	NS	OB	SAR	Ave.
SVM	0.64	0.46	0.54	0.17	0.45
BERT	0.62	0.48	0.54	0.21	0.46
SciBERT	0.64	**0.54**	0.53	0.15	0.47
Bio+Clinical BERT	0.64	0.50	**0.55**	0.22	**0.48**
Bio+DS BERT	**0.65**	0.49	0.54	**0.24**	**0.48**

Table 1: F1 scores of SVM and BERT classifiers for predicting scheduling decisions

	R1	R2	R3	R4
R1	**0.46**	0.28	0.21	0.33
R2	0.35	**0.43**	0.27	0.39
R3	0.18	0.19	**0.37**	0.23
R4	0.32	0.36	0.37	**0.49**

Table 2: Macro F1 scores showing performance of BERT fine tuned on one reviewer's labels and tested on another.

into five separate output nodes with the sigmoid activation function to get the predicted probabilities of five binary outcomes. The BERT parameters are fine tuned by minimizing the aggregated binary cross-entropy loss of all labels.

4 Experiments and Results

4.1 Triage Status

For the triage classifier, we first train several classifiers first on the entire eConsult training split, and test it on the development split. Results of the SVM with linear kernel and a few fine-tuned BERT models show that training across all consults results in poor performance (Table 1). As noted in the introduction, one explanation is that specialist reviewers were not consistent relative to each other. We thus examined was whether reviewers distributions over triage statuses were similar. Figure 3 shows a histogram of each reviewer's distributions of decisions – there are large differences in what fraction are labeled urgent (*Overbook* category). In order to further investigate the consistency of these scheduling decisions among different reviewers, we also trained four reviewer-specific models. Table 2 shows the results of each reviewer-specific model on text from other reviewers. Column headers indicate the reviewer used to train the model and rows indicate test reviewer.

4.2 Question Type Classification

We evaluated several different architectures on this task to explore the value of domain-specific information, as well as the importance of sharing information between the different labels. Tables 3 and 4 shows the results of the experiments for the question type classifiers. We omit results for P-FlexSig because there were only two instances in the split we evaluate on (current work is creating more annotations). The best overall performance was obtained by the SciBERT multi-task learning setup. In the single-task setting, Bio+Discharge Summary BERT alone provides several points of benefit on *Procedure-Colonoscopy* and *Procedure-EGD*. Multi-task learning provides an inconsistent benefit, increasing score in some categories while decreasing in others. However, when these two are combined, multi-task learning and SciBERT provide a large benefit over all other configurations.

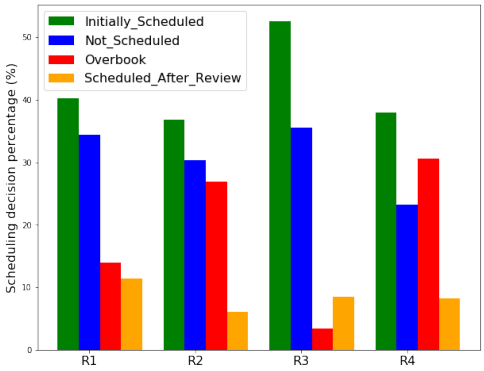

Figure 3: Distribution of scheduling decisions for different reviewers.

Question Type	D/M	P-C	P-EGD	P-Other
Linear SVM+BoN	0.71	0.75	0.81	0.32
Linear SVM+BoC	0.69	0.77	0.83	**0.34**
Kernel SVM+BoC	0.69	0.73	0.85	0.20
BERT	0.71	0.77	0.86	0.32
SciBERT	0.77	0.80	0.84	0.29
Bio+Clinical BERT	**0.78**	0.79	0.85	0.26
Bio+DS BERT	0.77	**0.84**	**0.92**	0.33

Table 3: F1 scores for question type classification with separate classifiers.

Question Type	D/M	P-C	P-EGD	P-Other
BERT	0.71	0.79	0.85	0.21
SciBERT	**0.82**	**0.86**	**0.89**	**0.41**
Bio+Clinical BERT	0.74	0.80	0.85	0.39
Bio+DS BERT	0.77	0.79	0.86	0.39

Table 4: F1 scores for question type classification with multi-task learning with different BERT variants.

5 Discussion & Conclusion

Within-reviewer results (diagonal of Table 2) indicate that predicting scheduling decisions from text alone is difficult, and there are few obvious cues to the urgency of a question. However, we also saw a large decrease in performance across reviewers, suggesting that individual reviewers behave very differently. Improving reviewer consistency may be a viable method for improving efficiency

of specialist referrals in health systems. It still is not totally clear from these results whether the individual reviewers are inconsistent – it is possible that the classifier model we chose is simply the right representation to perform this task. Future work should look deeper at within-reviewer classifier performance to explore the degree to which scheduling decisions are essentially random.

One possible explanation for the improved performance of SciBERT is that it uses domain-specific pre-training as well as a domain-learned vocabulary (ClinicalBERT, in comparison, is pre-trained on clinical data but uses the original BERT vocabulary). Practically speaking, the result is that the SciBERT vocabulary contains more biomedical terms. For example, the term *colonoscopy* occurs as a single token in the SciBERT vocabulary, while the standard BERT vocabulary breaks it into several word pieces. We suspect that this makes it easier for SciBERT to learn domain-specific language, as the meaning is attached directly to the word piece embedding rather than being learned through BERT encoding layers.

Future work should explore further modeling of domain structure, including understanding question text better, but also in modeling relationships between output variables. For example, sometimes the Diagnosis/Management category is clear from expressions like *Please eval*, but in other cases the request is only implicit. In these cases, the best clue is the lack of any specific procedure request. A sequential classification decision process may be able to incorporate this logic. In addition, we are continuing the annotation process, including continuing to revise guidelines to improve agreement, annotating more questions for question type in the gastroenterology specialty, and developing guidelines for additional specialties. Our early results suggest that the question type classifier can still be improved with additional data, despite already-promising performance.

Acknowledgments

Research reported in this publication was supported by the National Institute On Minority Health And Health Disparities of the National Institutes of Health under Award Number R21MD012693. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Michael L. Barnett, Hal F. Yee, Ateev Mehrotra, and Paul Giboney. 2017. Los angeles safety-net program econsult system was rapidly adopted and decreased wait times to see specialists. *Health Affairs*, 36(3):492–499. PMID: 28264951.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke

Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.