Clinical note section classification on doctor-patient conversations in low-resourced settings

Zhuohao Chen[1]*, **Jangwon Kim**[2], **Yang Liu**[2], **Shrikanth Narayanan**[1]
[1]University of Southern California, Los Angeles, USA
[2]Amazon, Seattle, USA
sail.usc.edu, {jangwok, yangliud}@amazon.com

Abstract

In clinical visits, clinical note writing is a time-consuming and cost-prohibitive manual task for clinicians. Although virtual medical scribes have been proposed to generate clinical notes (semi-)automatically, the data sparsity issue is still a challenging problem in practice. Identifying the topic of clinical utterances in doctor-patient conversations is one of the key strategies for automation. In this paper, we propose an utterance-level note section classification method for the situation of the limited amount of in-house data. We leverage an external, unsupervised corpus of medical conversations to transfer knowledge using the framework of Unsupervised Meta-learning with Task Augmentation (UMTA). Our experiments are performed on both manual transcripts and machine transcripts generated by automatic speech recognition (ASR). The results show that our strategies achieve substantial gains in prediction accuracy over several baseline approaches and are robust to ASR errors.

1 Introduction

The information from doctor-patient conversations is typically extracted by electronic health records (EHRs), the digitized clinical notes that summarize the patient's medical history and treatment plan. However, the manual work of generating EHRs increases the burden on physicians and is costly in time, leading to the complaint from medical practitioners (Sinsky et al., 2016; Patel et al., 2018). To mitigate these problems, scientists developed (partly) automated medical scribes to produce structured clinical notes from doctor-patient conversations directly (Finley et al., 2018; Jeblee et al., 2019a; Krishna et al., 2021a). Clinical section classification of speech utterances is one of the key components of automation. Previous studies used various machine learning models, e.g., conditional random field (Wallace et al., 2014)), word sequence embedding (Jeblee et al., 2019b), and recurrent neural network (Rajkomar et al., 2019; Schloss and Konam, 2020; Krishna et al., 2021b) for automatic clinical note writing systems.

The limited amount of supervised data makes it challenging to develop automatic systems. Clinical conversations and medical records are highly sensitive and confidential data with privacy concerns. Also, annotation tasks require medical domain knowledge, thus the process is slow and costly. A couple of studies employed a large number of labeled encounters (Schloss and Konam, 2020; Krishna et al., 2021b), but their datasets are not publicly released. Recent studies leveraged transfer learning and optimization-based meta-learning by using out-of-domain data (Finn et al., 2017; Nichol et al., 2018). They used human transcriptions, not speech audio, which is not available in a real telehealth scenario.

In this paper, we develop an automatic clinical section classification method for clinician-patient speech conversation data. The primary focus of this work is to mitigate the data sparsity issue in our in-house data by using Unsupervised Meta-learning framework with Task Augmentation (UMTA). We also boost model performance by integrating contextual and speaker-role information. Finally, we

Abbr.	Topic description	Nb. of utterances		
		Train	Dev	Test
PS	positive reported symptoms	1017	391	326
NS	negative/denied symptoms	253	76	82
SH	social history	137	53	46
Med	confirmed past medical history	164	37	76
	confirmed allergies			
	confirmed family history			
Plan	what clinician asks patient to do	601	191	203
None	None of above	1866	652	588
Total instances		4069	1407	1358

Table 1: The number of utterances for clinical note sections

*Work done during an internship at Amazon

Proceedings of the Third Workshop on NLP for Medical Conversations, pages 1–12
November 1, 2023.

Speaker	Start time	End time	Text	Section Type
Physician	2.14	7.45	Great. So you said you've had a, a sore throat and a headache. When did those start?	positive reported symptoms
Patient	7.45	8.75	Started a couple of days ago.	positive reported symptoms
Physician	9.02	15.88	OK. Like two days ago. And then, any other symptoms at all? Any fevers coughing, shorts of breath, runny nose, nausea, vomiting, body aches, fatigue? Lots of things.	negative/denied symptoms
Patient	15.88	19.67	No fever.	negative/denied symptoms
Physician	20.22	21.67	Got it. Do you have allergy kinda normally this time of year?	confirmed allergies
Patient	21.67	22.99	Yeah, kinda varies depending on the day.	confirmed allergies
Physician	23.05	25.42	OK. Does this cough feel similar to coughs you had previously with allergies?	positive reported symptoms
Patient	25.42	27.32	No. It's a bit different. Maybe more, more mucus production.	positive reported symptoms
Physician	27.32	27.42	OK.	None

Figure 1: An example episode of doctor-patient conversation.

experiment with manual transcript data and machine transcript data – the 1-best of Automatic Speech Recognition (ASR) output. Results suggest that our approach improves prediction accuracy over several baseline approaches on both types of data.

2 Data

Our (in-house) target task data consists of speech audio recordings of dyad clinician-patient conversations and full clinical documents. Clinician-role participants are real nurses, nurse practitioners, and medical doctors, while patient-role participants are mock patients. In order to minimize concerns over Protected Health Information (PHI), the mock patients were given randomly selected (fake) Reasons for Visit (RFVs), then instructed to mimic specific and realistic situations. Clinicians typically led telehealth sessions as realistically as possible. After the visits, clinicians completed clinical notes according to the SOAP (Subjective, Objective, Assessment, and Plan) coding scheme (Podder et al., 2021).

The in-house speech audio data was manually transcribed at the utterance level by a transcription service provider. Finally, a specialized labeling team manually annotated the clinical note sections: seven sections from (sub)headings of clinical notes and "Other". Fig 1 shows an episode of a labeled snippet. The details of the topic definitions are presented in Appendix A. Table 1 shows statistics of section label distribution. Sections "allergies", "family history" and "past medical history" are merged in our experiments, because the data size of these sections was too small. The in-house data consists of 6,860 utterances in total. We partitioned the data into train/dev/test sets by sessions in a ratio of 28/10/10, without the overlap of telehealth visits.

Another piece of data we used was simulated clinician-patient conversations purchased from external medical data vendors. This data was collected from various specialties and scenarios, including in-patient and out-patient, and telehealth and offline visits. In total, it has 300,000 utterances with role annotations and human transcriptions. However, there is no section label.

For in-house data, machine transcriptions were generated using Amazon Transcribe Medical.

3 Methods

3.1 BERT fine-tuning strategies

This section describes our methods to classify sections by using only in-domain data. A pre-trained bidirectional encoder representations from Transformer (BERT) (Devlin et al., 2019) is fine-tuned with three different strategies: 1) incorporating contextual utterances, 2) incorporating role informa-

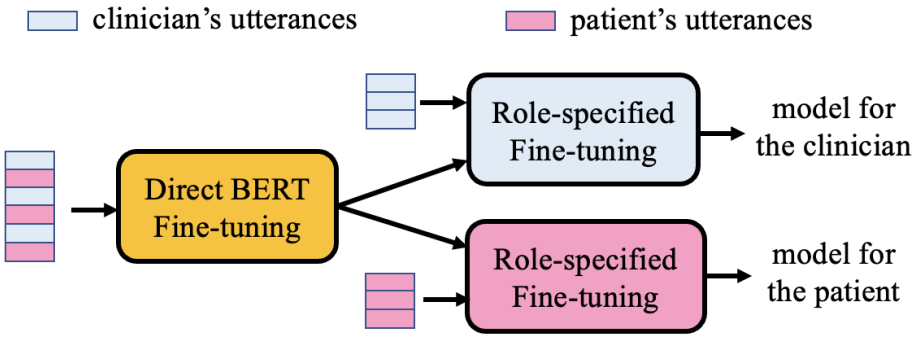

Figure 2: The framework of two-phase fine-tuning.

tion and 3) role-specified fine-tuning.

First, we added preceding and following utterances instead of feeding a single utterance as the input to the model and examined their impact on accuracy. Our hypothesis is that contextual information can benefit prediction accuracy. Second, we learned that the speaker-role information could improve topic categorization (Khosla et al., 2020) and incorporated it into the input by using role-specific tokens: "[PAT]" for patient's utterances and "[CLI]" for clinician's utterances. They were placed in front of the utterances. For example, $\{[PAT], U_{i-2}, [CLI], U_{i-1}, [CLI], U_i, [PAT], U_{i+1}, [CLI], U_{i+2}\}$ were used as the input for predicting the section of clinician's utterance U_i with the context size of 2 (ranging from U_{i-2} to U_{i+2}).

Finally, we examined the benefit of two-phase fine-tuning of Fig 2 to learn role-specific language patterns. This is motivated by our observations of different language patterns used by patients and clinicians for the same clinical note section. For example, clinicians use questions a lot for section "History of Present Illness", while patients use answer statements. In the first phase fine-tuning, we performed the regular BERT fine-tuning on all utterances. Then, in the second phase, we fine-tuned the model on role-specific utterance data. In the end, we trained two BERT models, one for each role.

3.2 Leveraging external data

To overcome the data sparsity issue of the in-house data, our strategy is to leverage external datasets and transfer knowledge to our task domain. We propose an algorithm of Unsupervised Meta-learning with Task Augmentation (UMTA). Fig 3 shows the framework of UMTA. This meta-transfer framework learns from external data first before fine-tuning the model on in-domain data. The challenges on meta-transfer learning from external data are (1) that the external data is unsupervised (no section label), and (2) that there is a shift between the external data and the in-house data, because of their differences in clinical visit scenarios and specialties, and (3) that technically, meta-learning calls a large number of source tasks. To address these challenges, we incorporate three more steps before normal meta-learning.

Step 1 We perform utterance clustering on the source corpus (the external dataset) and produce latent reasoning labels. Specifically, we use BERT to extract features from the data and then use k-mean clustering to group them into M clusters. The extracted features are the pooled output (embedding of the initial [CLS] token) so that the instances within the same cluster are semantically similar. To align the clustering results more closely with the target classes, we utilize the BERT model (referred to as F) following the initial fine-tuning on our in-house dataset. Next, we label utterances by their cluster indices. These clusters are used to construct simulated source tasks, and we hypothesize that they benefit meta-learning performance by increasing task variability.

Step 2 Let $z \in \mathcal{Z} = \{1, 2, ..., N\}$ be the target label variables. We define the distance between a cluster label and a target label variable as follows:

$$D(y = l, z = k) = \frac{\sum_{i=1}^{m_l} \mathbb{1}\{F(x_i^l) = k\}}{m_l} \quad (1)$$

where $D(y = l, z = k)$ denotes the proportion of the utterances in the l-th cluster for which the model F does not assign the label k. A lower value of $D(\cdot)$ indicates that the utterances with the cluster label $y = l$ and target label $z = k$ have

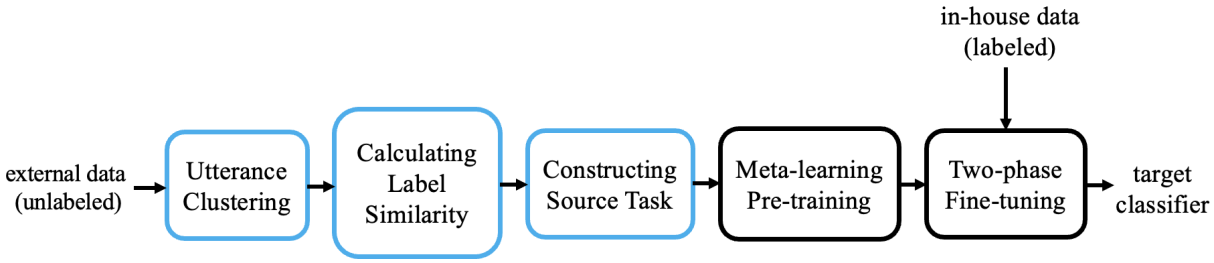

Figure 3: The framework of unsupervised meta-learning with task augmentation (UMTA)

greater similarity.

Algorithm 1 Unsupervised Meta-learning with Task Augmentation

Producing Source Tasks:
Initialize with a pre-trained BERT; $K, M, N \in \mathbb{N}$
Fine-tune BERT with in-domain data
Perform K-mean clustering on in-house data using embeddings from the fine-tuned BERT to get the cluster labels set $Y = \{1, 2, ..., M\}$
Create empty cluster groups $G_1, G_2, ..., G_N$.
for i = 1 to K **do**
    for j = 1 to N **do**
        Compute $D(y, z = j), y \in Y$ using Equation(1), $y^* = \arg\min_{y \in \mathcal{Y}} D(y, z = j)$
        Add y^* to G_j; remove y^* from $\mathcal{Y}$
Pick one label from each group in $G_1, G_2, ..., G_N$ to produce N^K different source tasks $\{T_i\}_{i=1}^{N^K}$

Meta-Transfer Learning:
Initialize the model parameters Φ with a pre-trained BERT; $n_s \in \mathbb{N}$, $\lambda, \delta > 0$
while not done **do**
    Select a batch of tasks $\{T_i\}$ with the probability proportional to the task size
    for all T_i **do**
        Perform by n_s steps of gradient descent with the learning rate λ to obtain $\Phi_i^{n_s}$.
        Update: $\Phi = \Phi + \delta \frac{1}{|\{T_i\}|} \sum_i (\Phi_i^{n_s} - \Phi)$

Step 3: After calculating the distance between each pair of source clusters and target classes, we take turns picking the most similar cluster label for each target label. Algorithm 1 describes this process. Then, for each target class, we create cluster groups G_1, G_2,... G_N, each of which contains K clusters that represent the K source clusters with the smallest $D(\cdot)$ value. We select one cluster from each group to produce K^N source tasks for meta pre-training. The labels of these source tasks are in one-to-one correspondence with the target classes, which we hypothesize that it is beneficial to knowledge transfer.

Fig 3 shows an illustration of the UMTA framework. After the first three steps above (colored in blue in Fig 3), we feed the generated source tasks into the meta-learning pre-training to learn information from the in-house data. We adopt the Reptile algorithm for training because it achieves the best performance among the optimization-based meta-learning algorithms on the benchmark dataset (Dou et al., 2019). As described in Algorithm 1, we select a task with the probability proportional to the size of its dataset, following the work of (Dou et al., 2019) and (Chen et al., 2022). The parameters λ and δ denote the learning rate of the inner loop and that of the outer loop, respectively. And n_s denotes the inner update step. After performing the intermediate task with meta-transfer learning, we continue training BERT with the two-phase fine-tuning as described in Section 3.1 to obtain the final prediction model.

4 Experiments and Discussions

4.1 Experimental Setup

All of our models were implemented in PyTorch (version 1.12.0) (Paszke et al., 2019) with CUDA 10.2. We ran each task 10 times and report the average performance of accuracy. The language model we adopt is *BERT-base*[1]. We set the max sequence length depending on how many contextual utterances we incorporated, which covered more

[1]https://github.com/huggingface/pytorch-pretrained-BERT

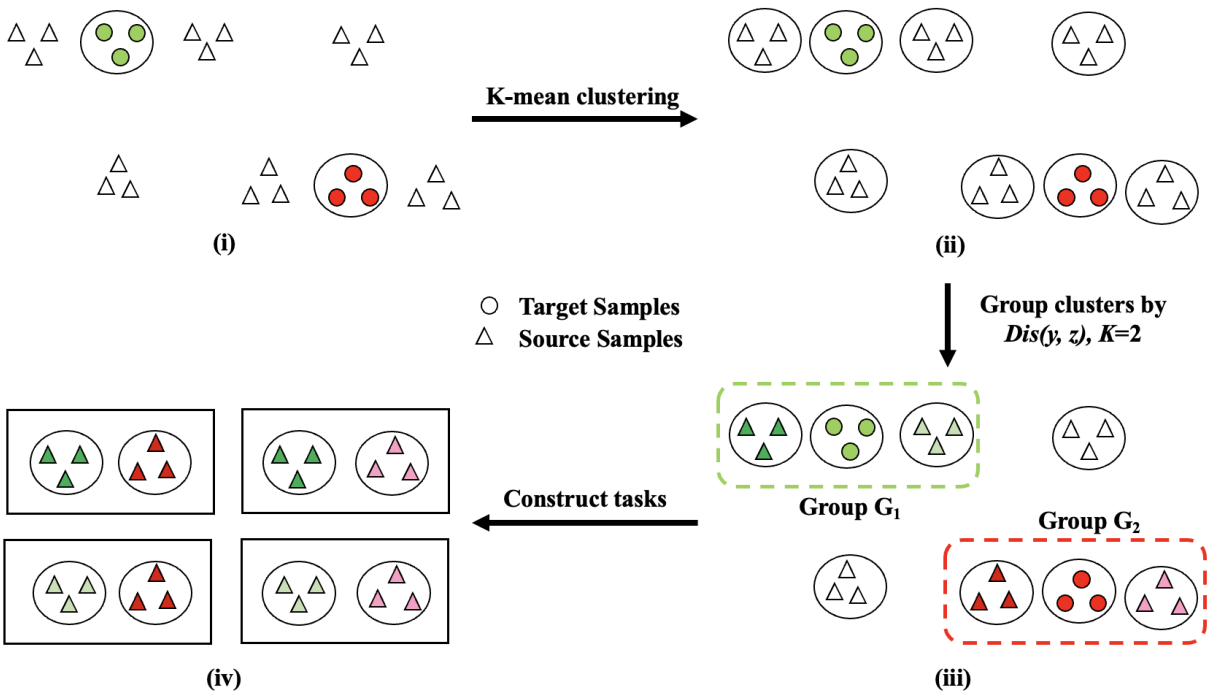

Figure 4: An example of producing source tasks for a two-way classification target task, K: size of cluster groups

than 99% of the sentences. For the BERT fine-tuning processes, we selected the best learning rate among {1e-5, 2e-5, 3e-5} on the validation set. We employed a decoupled weight decay regularizer and a linear learning rate scheduler for optimization (Loshchilov and Hutter, 2018). The model was trained with a batch size of 64 and 5 epochs, selected by their lowest validation loss. For the intermediate task of meta-learning, we set $\lambda = 5e-5$ and $\delta = 2e-5$. We pre-trained the model for 3 epochs, sampled 8 tasks per step, and fixed the inner update step n_s to be 5.

4.2 Language Model Adaptation

To achieve a better pre-trained language model for our task, we adapted BERT to the healthcare conversation domain via domain-adaptive pre-training (Gururangan et al., 2020) using masked word prediction and next sentence prediction using external data. To learn the roles and contextual information, we prefixed the role tokens "[CLI]" and "[PAT]" to each utterance and splice the corpus every three utterances. We trained BERT for 20,000 steps with the external data, setting the learning rate to 2e-5, the batch size to 32, and the maximum sequence length to 128. The adapted language model is denoted as *careBERT*.

4.3 Experimental Results

The experimental results of accuracy we present are all relative values compared to the baseline (the results are masked by dash symbols) in each table.

4.3.1 Results with In-domain Data Only

Table 2 shows experimental performance of incorporating contextual utterances and role information with in-house data only. Both the average and standard deviation of the prediction accuracy are reported. If we do not use the role tokens and set the context size to zero, the method is identical to normal BERT fine-tuning. We also compared with the performance of support vector machine-based (SVM) with bag-of-word (BOW)

Approach	role tokens	context size	Acc(%)
BOW+SVM	N/A	0	-
TF-IDF+SVM	N/A	0	+2.2
BERT-based Methods	no	0	+5.1±1.1
	no	1	+2.0±1.6
	no	2	+0.2±1.7
	yes	0	+4.2±1.2
	yes	1	**+10.7±1.2**
	yes	2	+8.9±1.4

Table 2: Effect of incorporating contextual utterances and role information in clinical topic classification.

	Acc(%) of clinician's data	Acc(%) of patient's data	overall Acc(%)
BERT with clinician's data	-	N/A	N/A
BERT with patient's data	N/A	-3.3	N/A
BERT with all data	+1.8^{+}	-1.3^{+}	+0.1*
two-phase fine-tuning	**+4.7**$^{+}$	**+2.1***	**+3.7***

* is significantly higher than $^{+}$ at $p < 0.05$.

Table 3: The comparison between direct BERT fine-tuning and two-phase fine-tuning approaches. N/A: not applicable.

	BERT-base	blueBERT	careBERT
Two-phase Fine-tuning	-	-2.9	**+2.3**

Table 4: The results of clinical topic classification tasks with different pre-trained language models.

Language Model	Two-Phase Fine-tuning	UMTA K=2	UMTA K=4	UMTA K=6	UMTA K=8
BERT-base	-	+1.4	+2.7	**+3.5**	+3.0
careBERT	+2.3	+3.7	+4.4	**+5.0**	+4.5

Table 5: Comparison between UMTA and two-phased fine-tuning approaches. K: size of cluster groups for UMTA.

and term frequency-inverse document frequency (TF-IDF) transformations, denoted by *BOW+SVM* and *tfidf+SVM*, respectively. In Table 2 shows that adding contextual utterances improves accuracy only when incorporating role tokens for speakers, presumably because otherwise, the model can hardly detect the target utterance. We found that the best context size for this task was one, and increasing the value degraded the performance. We hypothesize that a bigger context size makes the input more complicated, which hurts performance in low-resource situations. Hence, for direct BERT fine-tuning with in-domain data, the best performance is achieved by specifying role tokens and employing one utterance before and after the target utterance as an input sequence. We used this input configuration for all other experiments. Table 3 demonstrates that the two-phase fine-tuning outperforms the direct BERT fine-tuning, indicating the benefit of adding role tokens. Finally, training with data of both roles shows better performance than training with data of a single role, suggesting that additional information from a different role data offers useful information for better prediction accuracy in the low-resource setting in this study.

4.3.2 Results of Leveraging Out-Of-Domain Data

Table 4 shows results of various pre-trained BERT models. *blueBERT* is a publicly available model that was trained on written clinical notes and other medical data (Peng et al., 2019). *careBERT* model achieves the best accuracy, suggesting that adapting the language model to the medical conversational domain boosts accuracy (2.3% normalized accuracy boost in two-phase fine-tuning, compared to BERT-base). *blueBERT* performed worst, which suggests that the impact of different data type (clinical note v.s. dialogue) is significant.

Table 5 compares the performance of UMTA and the two-phase fine-tuning algorithm. UMTA leverages the external corpus by using our proposed algorithm, while two-phase fine-tuning uses in-domain data only. For UMTA, the number of clusters was 50. K is the size of cluster groups. Results show that UMTA makes good use of out-of-domain data and improves accuracy. Finally, K = 6 was the optimal. We speculate that it's related to the impact of K to model: It leverages too little data and generates too few source tasks when K is too small, while it includes more irrelevant source clusters, reducing the task similarity, when K is too big.

To better understand how the UMTA framework improves the classification tasks, we performed an ablation study by grouping the clusters randomly at the stage of Fig 4(iii), instead of using any similarity metric. We denote the modified framework as UMTA-random. Fig 5 presents the comparison between the UMTA and UMTA-random for different values of K. Unlike UMTA, which has an optimal value of K, the result of UMTA-random improves monotonically as K increases, presumably because the random selection procedure does not affect the similarity between the produced source and target tasks. The UMTA-random still improves accuracy with all K values, compared to two-phase fine-tuning (baseline in Table 5). However, it degrades the performance of the UMTA. It indicates that our proposed framework benefits the predicting tasks by implementing meta-learning for the intermediate task and the cluster grouping strategy to increase task similarity.

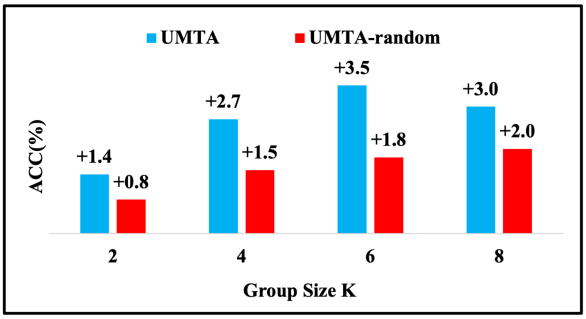

Figure 5: The comparison between standard UMTA and the random version of UMTA in which we form the cluster groups by selecting clusters randomly. We report the relative values of accuracy compared to two-phase fine-tuning. Language model: BERT-base

Finally, we compared UMTA with self-training (Vu et al., 2021) which is another popular way of leveraging external data. Also, we compare the impact of using ASR transcripts, compared to using human transcripts. *careBERT* was used as the base pre-trained model. Table 6 shows the results. First, using ASR transcripts instead of manual transcripts decreases performance of all models, which is expected. It is noteworthy that self-training led to enhanced prediction accuracy compared to a two-phase fine-tuning approach using human transcripts. Conversely, employing ASR transcripts resulted in a decline in accuracy. However, UMTA improved accuracy on both human transcripts and ASR transcripts over two-phase fine tuning, suggesting the robustness of UMTA to the ASR error. In addition, the best accuracy was achieved by UMTA for both manual transcripts and ASR transcripts.

Data Format	Two-Phase Fine-tuning	Self-training	UMTA
Manual	-	+0.7±0.3	**+2.7±0.1**
ASR	-1.3±0.7	-1.9±0.5	**+1.6±0.1**

Table 6: The comparison between the results on the manual transcripts and ASR derived transcripts using careBERT.

5 Conclusion and Future Work

In this study, we performed clinical utterance classification on manual transcripts and ASR transcripts with several strategies for low-resource scenarios. We incorporated contextual and role information into the model and showed their accuracy boost. To handle the data sparsity issue, we leveraged a larger size of unsupervised external dataset by adapting BERT to the medical conversational domain and using our proposed UMTA to improve the knowledge transfer toward the target task. Our experiment results show that UMTA showed the best performance and robust against ASR errors.

We believe there is still room for improvement. Contextual information from longer time window (e.g., at the beginning, in the middle, or the final in a clinical visit) may benefit section classification performance, because clinicians typically lead conversations with patients, following their protocols.. For instance, clinicians typically ask History of Present Illness (HPI) early, discuss assessment and plan at the end. Another dimension of improvement is to extend its application to offline visits by speaker diarization, where the role of an utterance cannot be captured by audio channels.

Limitations

Apart from the one mentioned in Sec 5 that we do not incorporate longer temporal information of the clinical visits, this work has two more limitations. Firstly, we purchased the external data from a medical data vendor, which is still costly. Consequently, it is imperative to conduct further investigations to explore whether the UMTA can efficiently transfer knowledge from the publicly available dataset. While the clinical section coding scheme of SOAP is designed for written notes, adapting it to clinical visit transcripts necessitates the development of annotation rules to map utterances to specific clinical section topics. Furthermore, the labeling task was carried out by non-experts who underwent training and calibration demonstrations prior to the coding process, which might introduce some level of noise into the labeled data.

References

Zhuohao Chen, Nikolaos Flemotomos, Zac Imel, David Atkins, and Shrikanth Narayanan. 2022. Leveraging open data and task augmentation to automated behavioral coding of psychotherapy conversations in low-resource scenarios. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5787–5795, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1192–1197, Hong Kong, China. Association for Computational Linguistics.

Gregory Finley, Erik Edwards, Amanda Robinson, Michael Brenndoerfer, Najmeh Sadoughi, James Fone, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. An automated medical scribe for documenting clinical encounters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–15, New Orleans, Louisiana. Association for Computational Linguistics.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Serena Jeblee, Faiza Khan Khattak, Noah Crampton, Muhammad Mamdani, and Frank Rudzicz. 2019a. Extracting relevant information from physician-patient dialogues for automated clinical note taking. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 65–74, Hong Kong. Association for Computational Linguistics.

Serena Jeblee, Faiza Khan Khattak, Noah Crampton, Muhammad Mamdani, and Frank Rudzicz. 2019b. Extracting relevant information from physician-patient dialogues for automated clinical note taking. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 65–74, Hong Kong. Association for Computational Linguistics.

Sopan Khosla, Shikhar Vashishth, Jill Fain Lehman, and Carolyn Rose. 2020. MedFilter: Improving Extraction of Task-relevant Utterances through Integration of Discourse Structure and Ontological Knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7781–7797, Online. Association for Computational Linguistics.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021a. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Kundan Krishna, Amy Pavel, Benjamin Schloss, Jeffrey P Bigham, and Zachary C Lipton. 2021b. Extracting structured data from physician-patient conversations by predicting noteworthy utterances. In *Explainable AI in Healthcare and Medicine*, pages 155–169. Springer.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Rikinkumar S Patel, Ramya Bachu, Archana Adikey, Meryem Malik, and Mansi Shah. 2018. Factors related to physician burnout and its consequences: a review. *Behavioral sciences*, 8(11):98.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2021. Soap notes. In *StatPearls [Internet]*. StatPearls Publishing.

Alvin Rajkomar, Anjuli Kannan, Kai Chen, Laura Vardoulakis, Katherine Chou, Claire Cui, and Jeffrey Dean. 2019. Automatically charting symptoms from patient-physician conversations using machine learning. *JAMA internal medicine*, 179(6):836–838.

Benjamin Schloss and Sandeep Konam. 2020. Towards an automated soap note: classifying utterances from medical conversations. In *Machine Learning for Healthcare Conference*, pages 610–631. PMLR.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of internal medicine*, 165(11):753–760.

Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. 2021. STraTA: Self-training with task augmentation for better few-shot learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5715–5731, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Byron C Wallace, M Barton Laws, Kevin Small, Ira B Wilson, and Thomas A Trikalinos. 2014. Automatically annotating topics in transcripts of patient-provider interactions via machine learning. *Medical Decision Making*, 34(4):503–512.

A Appendix

Our in-house data is a mock customer dataset simulating the nursing situation from a specific business institution. For every conversation, both transcriptions and clinical notes are provided by the in-house experts. Since the dialogue is different from the written notes, we propose a coding scheme to map the utterances to the materials from the clinical notes. A specific labeling team performed the annotation work who demonstrated calibration prior to coding process, which resulted in high inter-rater reliability (ICC > 0.8 (Cohen, 1960)). Details and examples of each topic label are presented as follows:

PS: positive reported symptoms, patient's assessment, vaccines states (in many cases it is related to the symptoms).

PS is short for positive reported symptoms. As shown in the Fig 6, we not only code for clinician's words but also annotate patient's feedback.

Note: Onset/Timing/Duration (OTD) and Severity/Pain (SP) are included in PS.

NS: negative/denied symptoms.

If the patient's answer to the clinician is a denial, we annotate the utterances of both clinician and patient to be negative/denied symptoms and 'NS' for short.

Med: confirmed past medication history.

Med means "confirmed medication history". Following the clinical note convention of in-house data, we only code for confirmed medications. For the medication related to the patient's current symptoms, we code it as "PS". For the denied medication, we code it by "N".

A: confirmed allergies.

For the allergies related to the patient's current symptoms, we code it as "PS" following the clinical note convention of in-house data. For the denied family allergies, we code it by "N".

SH: social history.

The social history section records the substances in the patient's personal life that have the potential to be clinically significant, such as alcohol and drugs. Unlike medication history, both confirmed and denied items are recorded. In this example of Fig 10, the content of the social history section is "no smoke".

FH: confirmed family history.

For the denied family history, we code it by "N".

Plan: what doctor asks patient to do.

The doctor usually explains why the plan is good before or after he tells the patient about the plan/suggestion. We only code for the plan in this case and ignore the explanation.

N: none of above.

If an utterance belongs to none of previous topics, we label it by "N".

Physician	Got it. Do you have allergy kinda normally this time of year?	
Patient	Yeah, kinda varies depending on the day.	
Physician	OK. Does this cough feel similar to coughs you had previously with allergies?	PS
Patient	No. It's a bit different. Maybe more, more mucus production.	PS
Physician	OK.	

Figure 6: An example of the label *positive reported symptoms*.

Patient	7.45	8.75	Started a couple of days ago.	
Physician	9.02	15.88	OK. Like two days ago. And then, any other symptoms at all? Any fevers coughing, shorts of breath, runny nose, nausea, vomiting, body aches, fatigue? Lots of things.	NS
Patient	15.88	19.67	No fever.	NS
Physician	20.22	21.67	Got it. Do you have allergy kinda normally this time of year?	

Figure 7: An example of the label *negative/denied symptoms*.

Patient	OK. OK. Yes, that sounds good.	N
Physician	mm OK. All right. And then do you take any medications including over-the-counter supplements?	Med
Patient	Just some vitamins, you know, and if my stomach isn't feeling well, I'll do some enzymes, you know, digestive enzymes.	Med
Physician	OK. And what type of vitamins do you take?	Med
Patient	Vitamin D, B twelve, and vitamin C.	Med
Physician	OK.	N

Figure 8: An example of the label *confirmed past medical history*.

Patient	No fever.	
Physician	Got it. Do you have allergy kinda normally this time of year?	A
Patient	Yeah, kinda varies depending on the day.	A
Physician	OK. Does this cough feel similar to coughs you had previously with allergies?	

Figure 9: An example of the label *confirmed allergies*.

Patient	Vitamin D, B twelve, and vitamin C.	
Physician	OK.	
Physician	All right. And then you had, do you smoke cigarette?	SH
Patient	No.	SH

Figure 10: An example of the label *social history*.

Patient	So, looking out at least five to six times a week whether walking, biking, running, yoga.	
Physician	And maybe, do you have any family history or personal history of any heart problems, or lung problems, anything like that?	FH
Patient	Let's see. My grandmother had, I know she had hypertension.	FH
Patient	She had a heart attack and she's been she's passed away. But it's been a while. But my mom also has hypertension. My dad also has hypertension and he also has diabetes.	FH
Physician	OK. And is that your maternal grandmother that passed away? OK. And I'm sorry about that.	

Figure 11: An example of the label *confirmed family history*.

Patient	That's good.	N
Physician	All right. And the other thing you can do is warm soaks also. In that sense, it helps soothe the foot as well, and then just make sure you keep it up.	Plan
Patient	Yeah, water.	Plan
Physician	Are there any other questions you have?	N

Figure 12: An example of the label *plan*.