UPSTAGE: Unsupervised Context Augmentation for Utterance Classification in Patient-Provider Communication

Do June Min[1], Veronica Perez-Rosas[1], Shihchen Kuo[2], William H. Herman[2], Rada Mihalcea[1]

DOJMIN@UMICH.EDU, VRNCAPR@UMICH.EDU, SHIHCHK@MED.UMICH.EDU, WHERMAN@MED.UMICH.EDU, MIHALCEA@UMICH.EDU

[1]*Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA*

[2]*Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA*

Abstract

Conversations between patients and providers in clinical settings provide a source of natural language data that may reflect and correlate with the patients' experience and response to the treatment they are receiving. When analyzing utterances in such conversations, it is not sufficient to consider each sentence in isolation, since its context may play a role in determining its semantic meaning. Recently, contextual information in natural language documents has been modeled using various techniques, such as recurrent neural networks with latent variables, or neural networks with attention mechanisms. In this paper, we present UnsuPerviSed conText AuGmEntation (UPSTAGE), a classification framework that relies on both *local* and *global* contextual information from different sources. UPSTAGE uses transformer models with pretrained language models and joint sentence representation to solve the task of classifying health topics in patient-provider conversations. In addition, UPSTAGE leverages unlabeled corpora for pretraining and data augmentation to provide additional context, which leads to improved classification performance.

1. Introduction

Recently, the availability of speech recognition technologies along with advances in Natural Language Processing technologies has made it possible to automatically analyze conversations at large scale. This has been particularly beneficial for conversational analysis in the clinical domain where important information is conveyed in the participant's dialog. Conversations between patient and providers contain rich language information that can provide insights into important aspects of the interaction such as patient experience, response to treatment, time allocation for health issues and quality assurance. Thus, it is important to develop tools to assist researchers during their analysis.

In this work, we propose Unsupervised Context Augmentation (UPSTAGE), a framework that relies on different sources of contextual information for classifying utterances in patient-provider conversations transcripts.Our prediction model is an utterance-level classifier based on the transformer architecture that models local contextual information from surrounding utterances in the input using joint utterance representation. Moreover, since supervised methods rely on manual annotations that are labor intensive and cost prohibitive, we pro-

pose to use unlabeled data from online health forums as a source of additional information to augment utterances with global context from topics extracted from a larger body of semantically related natural language corpora.

Technical Significance. Several studies have focused on analyzing and classifying content in health-related documents such as clinical notes or electronic health records (Dernoncourt et al. (2016); Mork et al. (2010); Roberts and Harabagiu (2011). However, our study focuses on analyzing patient-provider conversations that are not confined to a predefined structure and length, which is the case with the former. An important distinction is that conversations happening between patients and providers are free-form and include not only medical concerns but also other personal issues and small talk. Therefore, it is necessary to develop a model that is robust to the variability of style and length of clinical natural language.

In this context, our technical contributions are as following:

- We formalize the problem of health-related conversation coding as a sequential sentence classification task and propose a transformer-based method to classify utterances in patient-provider conversations into a set of clinically relevant topics. Our model is distinguished from other sentence-level classifiers in that it incorporates local context, by using joint sentence representation to model local context from surrounding sentences.

- We empirically show that pretraining a transformer-based model on an out-of-domain corpus collected from an online diabetes forum leads to increased performance. Moreover, we propose a simple approach to further leverage unlabeled data through Latent Dirichlet Allocation (LDA) topic modeling.

- We demonstrate the importance of contextual information by comparing context-aware and context-unaware models. Moreover, we present a set of comparative experiments for different strategies to model contextual information.

Clinical Relevance The model developed in this paper can provide clinical researchers with an automatic tool to label patient-provider conversation utterances about diabetes-related healthcare topics. This can reduce the manual annotation burden and scale up the number of conversations that can be annotated for further analyses. These allow clinical researchers to quantify and compare aspects of clinical interventions that were previously hard to capture. The model can be deployed to facilitate qualitative and quantitative analyses of the patient-provider interactions in clinical settings. For instance, our model output can be used to analyze whether, how, and how extensively providers discuss specific healthcare topics during their conversations with patients, thus providing a better assessment of the intensity and quality of the medical encounter and their impacts on health outcomes.

More specifically, our work has a direct clinical application for an economic analysis of the Glycemia Reduction Approaches in Diabetes: A Comparative Effectiveness (GRADE) study (Nathan et al. (2013)). The GRADE study is a pragmatic, randomized controlled trial in 36 centers across the U.S. to make head-to-head comparisons of four major glucose-lowering medications added to metformin in people with type 2 diabetes on clinical effectiveness, patient-centered, and health economics outcomes. One of the critical components

for the GRADE health economic analysis is the time spent by GRADE study providers and participants on discussing topics related to diabetes management, counseling, and education. A sample of GRADE study participants has been identified and asked to allow all of their GRADE study clinic visits to be recorded. The digital audio files are transcribed and the tool developed in this paper will be used to automatically code more than 4,900 transcripts of the patient-provider interactions. The resulting coding will be used to analyze how time is allocated during the study visits to address specific diabetes-related healthcare topics by GRADE treatment group, the content of the patient-provider interactions, and their impacts on health outcomes. The information will also be used to estimate the time spent and thus the intervention costs of the four GRADE treatments.

Generalizable Insights about Machine Learning in the Context of Healthcare

- We show that modeling contextual information is beneficial while classifying the topics discussed in utterances that are part of a clinical conversation. We demonstrate that a simple method to model local context based on the joint sentence representation approach by Cohan et al. (2019) shows improved performance when compared to several other methods and baselines.

- We explore strategies to address situations where unlabeled data are readily available whereas labeled data are scarce. In medical research settings, collecting labeled data might be expensive for various reasons, while large amounts of unlabeled text data with similar semantic content can be easily collected from online sources such as online health forums. In addition to pretraining on the collected forum data, we propose that an LDA topic model learned from unlabeled data may be used to augment input with additional information from global context to achieve improved performance.

2. Related Work

Transformer Models. Recently, many natural language processing (NLP) tasks have been successfully tackled using frameworks that make use of the Transformer model as the backbone (Vaswani et al. (2017)). Several NLP problems can be cast as sequence transduction tasks, and the Transformer architecture approaches these using multi-head attention instead of temporal modeling, achieving parallelizable and fast computation during training and inference. Some of the most successful Transformer-based models include BERT, GPT2, and XLNet (Devlin et al. (2018); Radford et al. (2018); Yang et al. (2019)). A key insight behind using these large-scale models effectively is to leverage transfer learning. By pretraining the models using density estimation or masked token prediction, large corpora of unannotated language data can be used to increase the model's performance in a range of downstream tasks. Moreover, practitioners can choose pretraining corpora that are semantically or stylistically similar to the downstream task's domain for more effective transfer learning (Ruder (2019)).

Text Classification. Text classification has been widely studied in NLP and artificial intelligence, not only as a way to analyze natural language data but also as a framework to study various tasks in the domain of natural language understanding (Wang et al. (2019)). Feature engineering approaches and deep learning models are often used for classification

of different units of textual input: document-, paragraph-, sentence (utterance)-, and sub-word-levels (Kowsari et al. (2019)). In our work, the task of classifying utterances in conversations is naturally cast as a sentence-level classification task. However, it is important to note that each session defines a document-level context that individual sentences depend on for meaning disambiguation and thus for correct classification. Context modeling for utterance classification has been tackled using hierarchical models or recurrent models with latent variables in the domain of emotion recognition in conversation (ERC) (Li et al. (2020); Poria et al. (2019)).

Analysis of Health-related Conversations. Medical researchers and clinicians have applied qualitative analysis and NLP to study language use in health-related conversations. For instance, Park et al. studies how machine learning algorithms can be used to detect conversation topics in primary care office visits from transcripts of patient-provider interactions (Park et al. (2019)). In the context of motivational interviewing (MI), a counseling methodology that aims to induce behavior change in patients, researchers have shown that patients' language use during counselling sessions can be used to predict the outcome of the intervention (TR et al. (2014); Pérez-Rosas et al. (2019)). Such high-level analysis is not included in this paper, but we point out that our models could be used for automatic analysis of utterances in such settings.

3. Data

3.1. Clinical Conversations Dataset

The data used in this work is derived from recordings of clinical conversations collected from the GRADE (Glycemia Reduction Approaches in Diabetes: A Comparative Effectiveness) study (Nathan et al. (2013)) – a nationwide clinical trial focused on determining which of the top type 2 diabetes drugs are best for glycemic control. The conversations consist of patient-provider encounters during quarterly follow-up visits that discussed diabetes counseling and management with the study participants. Research protocols, including consent for audio recordings, were approved by the relevant organization' institutional reviewing board.

Data collection for this study is ongoing and our work draws from a set of 4285 recorded conversations from 465 participants collected during 2017-2019. The recordings have an average length of 2.2 hours and mainly portray conversations between patients and providers. In some cases, the recordings also include speech from participants' companions and nurses that conduct medical procedures.

The conversation recordings are first automatically turn-by-turn segmented and transcribed using the Google Cloud API's speech-to-text service [1]. The resulting transcripts consists of 10.8 million words with 593 thousand talk-turns in total across the 4285 conversations. Each conversation transcript has on average 2513 words and 130 turns. A sample transcript excerpt is shown in Table 1.

Since our goal is to identify conversation topics during the clinical encounter, our first step consists of building a dataset of conversation turns labeled with health-related topics. From the available set of transcriptions, we randomly choose a subset of 56 conversations that were annotated by two diabetes experts with 7 diabetes-related topics covering glucose

1. https://cloud.google.com/speech-to-text

Speaker	Utterance	Code
Patient	I like to do yogurt with fresh fruit.	Diet
Provider	Yogurt with fruit. And how about your lunch?	Diet
Patient	Burrito.	Diet
Provider	Your lunch. Said something like burritos.	Diet
Patient	Dinner we usually have some kind of a meat? It's a smaller portion, but we always include some kind of protein or two. Lettuce wraps last night and veggies and greens with my lunch.	Diet
Provider	Okay on your exercise.	Exercise

Table 1: Example Snippet of Patient-Provider Conversation from the annotated transcripts.

management (medication management, self-monitoring of blood glucose and hypoglycemia), diet, exercising, foot care, and other medical issues. These topics were originally designed as part of the clinical study. The annotators labeled each conversation turn with the different topics whenever the speakers discussed issues related to it. Turns that did not discuss any of the 7 topics were annotated as not applicable (NA) as a default code. Table 2 presents a brief description of the different topics along with the percentage of turns assigned to them in our annotated set.

Topic (Abbreviation)	Description	Percentage
Diet (D)	Diet and weight management	8.22%
Exercise (E)	Physical activity	1.17%
Medication management (M)	Medication dosage, side effects, and use	14.19%
Self-monitoring of blood glucose (S)	Measuring and monitoring glucose levels	4.94%
Hypoglycemia (H)	Hypoglycemia management	2.36%
Foot care (F)	Control question about feet condition	2.76%
Other medical issues (O)	Any other medical issue being experienced by the participant	16.38%
Not applicable (NA)	Small talk between provider and patient, or interactions with other care team members	49.98%

Table 2: Diabetes-related topics and percentage of conversation turns in the annotated set

To ensure the reliability of coding, we measure the inter-annotator agreement over different topic codes in a sample of 5 conversations. Raw agreement and Kappa score for each topic are shown in table 3. Overall, we observe very high agreement for all diabetes-related codes (0.86 on average), except for "Foot care"(F), suggesting that our annotators are consistent while assigning the different labels. The low agreement score for the code F "Foot care" can be attributed to both its low frequency in the transcripts and its status as an assigned control code during the conversations.

Code	Raw Agreement	Cohen's Kappa
D	0.96	0.84
E	1.00	1.00
M	0.96	0.89
S	0.98	0.87
H	0.99	0.90
F	0.97	0.15
O	0.96	0.88
NA	0.93	0.86
Overall	0.86	

Table 3: Inter-annotator Agreement using Cohen's Kappa

3.2. Online Diabetes Forum Dataset

Since most current deep learning models are not trained with text in the medical domain and require large quantities of data to provide reliable predictions, we decide to expand our data by collecting additional diabetes-related content from an online diabetes forum. We thus collect forum threads where users create and respond to posts on topics relating to type-2 Diabetes and its treatment from the Diabetes Daily website. We used the Python's `Scrapy` package to extract the data. The final dataset, derived from more than 17.6 thousand threads posted by forum users, consists of 2.6 billion words, taking up to 1.4 Gigabytes of storage space. Note that the forum corpus is different from clinical conversations in several ways. First, the interactions contained in the collected threads are not necessarily dyadic, unlike patient-provider conversations. Also, Daily Diabetes threads are typed by the forum users, not spoken and transcribed like the clinical transcripts. Finally, a patient-provider dynamic is not present in the online threads, although the behavior and language of new and experienced patients might be similar. However, we believe that the large volume of text available in these threads constitute a good source of data for diabetes-related language, which can help to better train the deep learning models used in this work.

4. Supervised Utterance Classification with Unsupervised Context Augmentation

We formulate our task as a supervised utterance classification problem, where each utterance in a patient-provider conversation is assigned with one of the diabetes topics described earlier. Given that utterances are typically short, this is a challenging task even for the most recent neural network architecture. We hypothesize that unsupervised context augmentation can improve the classification accuracy, and we perform this augmentation at two levels: (1) *Local*, by including information drawn from surrounding sentences; and (2) *Global*, by including information drawn by very large in-domain corpora.

We propose Upstage, a system for unsupervised context augmentation for utterance topic classification, which enhances a transformer model with modules that explicitly obtain and represent local and global context. Figure 1 shows the architecture of the system. We discuss below each of these modules.

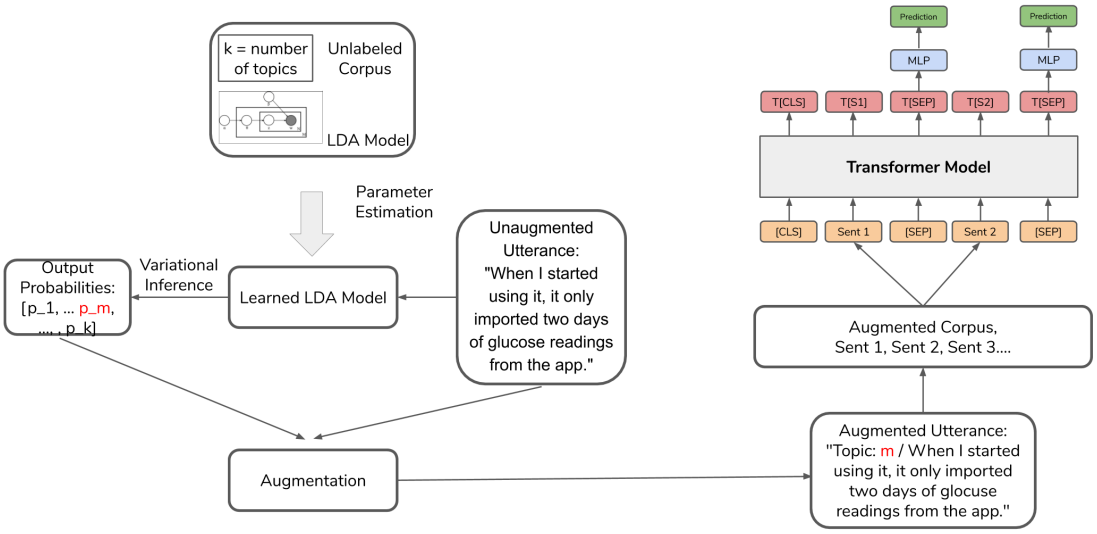

Figure 1: The Upstage Architecture

Transformer-based Model. We use a transformer-based architecture as the backbone of our classifier models. In particular, we use the BERT (Bidirectional Encoder Representations from Transformers) architecture (Devlin et al. (2018)) to compute vector representations of utterances. We then add an additional multilayer perceptron for classification. Specifically, we use the `bert-base-uncased` model which has 12-layers, hidden dimension of 768, and 12 attention heads.

4.1. Local Context

In the medical domain, conversations are free form and frequently have short utterances that can be ambiguous in isolation. Thus, leading to incorrect topic classification when attempting to label them without context. Consider the example shown in Figure 2 which shows two plausible questions by the provider that could have prompted the patient's answer. In Case 1, the provider is asking about a medicine not directly related to diabetes treatment, so both the question and the answer should be classified as "O"(Other medical management). On the contrary, in Case 2, the question is about the patient's exercise regimen, so both utterances should be classified as "E"(Exercise).

To deal with such cases, we augment the context of the target utterance by including utterances from the surrounding local context. We explore two strategies for doing that: (i) concatenation of sentences from before and after; and (ii) joint sentence representation.

Concatenation of Local Context. We implement a simple method that incorporates the surrounding context by using concatenation approaches, following the method proposed by Agrawal et al (Agrawal et al. (2018)). While Agrawal et al. introduced their method for the task of machine translation, we adapt their approach by skipping target side context.

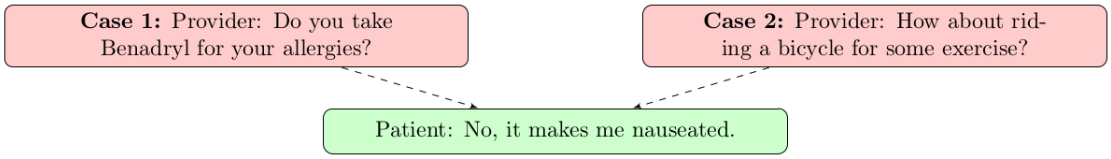

Figure 2: Two constructed patient-provider interactions with a shared response

We concatenate the sentences before and/or after the target sentence to the original input. In the resulting input sequence, the target and context sentences are separated by a `[SEP]` token, but only the `[CLS]` token is used for classification.

Specifically, we implement and evaluate the following alternatives:

- **Before:** k previous utterances are prepended to the original input.
- **After:** k following utterances are appended to the original input.
- **Both:** k utterances from **Before** and **After** are concatenated to the input. If k is not even, the before context will receive $\frac{k+1}{2}$ utterances and the after context $\frac{k-1}{2}$ utterances.

Note that although this approach considers multiple utterances simultaneously, it makes a prediction only for the target utterance, not for its context sentences.

Joint Sentence Representation. We also explore a method that creates a joint representation of neighboring sentences. Our intuition for this strategy is that in order to accurately capture the semantic content of each utterance in the document, it is necessary to use information from local context in the sequence of sentences. Typically, transformer models for classification add a special token (`[CLS]`) token to the sequence to be classified. The hidden representation corresponding to this token is then fed to a feed-forward neural network layer which outputs the prediction. The idea is that through end-to-end training, the transformer model will learn to represent the semantic context with the hidden representation of the `[CLS]` token.

For our task, we implement the joint sentence representation framework by Cohan et al. (Cohan et al. (2019)), as shown in Figure 1. Instead of the `[CLS]` token, the embedding vector for the `[SEP]` separator token is taken to be the hidden embedding vector for each utterance in the input. Since the `[SEP]` token is used to demarcate the boundary between sentences, this allows the model to make multiple predictions using the contextual information from neighboring sentences. After the input sequence is formatted as described, the BERT model computes the contextualized embedding vector for each token in the sequence. Then, the embedding vectors corresponding to the `[SEP]` tokens are passed through a multilayer perceptron (MLP) implemented as a feedforward neural network and a subsequent softmax layer, producing a probability distribution over the topic labels.

4.2. Global Context

Besides adding local context information, we also experiment with ways to provide global context. We consider two strategies for incorporating global context to our model: (i) pretraining for the core transformer model; and (ii) augmenting the input sequence with topic information inferred in an unsupervised way.

Pretraining with Unlabeled Data. Transformer models can be trained in a self-supervised manner, by performing language model (LM) pretraining. Different transformer architectures may use different implementations, such as masked token, permutation token, or next token prediction. This pretraining is typically performed on very large datasets.

In this work, we exploit two sources of language model pretraining: the publicly available pretrained parameters obtained through training on a very large Wikipedia corpus; and domain-specific training using an unlabeled corpus consisting of unannoatated conversations from the diabetes study and scraped data from the Daily Diabetes forum (see Section 3.2). Specifically, we start by initializing our BERT model with `bert-base-uncased` parameters corresponding to Wikipedia pretraining, and then further pretrain the initial model on our unlabeled corpus in a self-supervised manner.

Augmenting Input Utterances with LDA Topics. Given that we have a limited amount of annotated data available for model training, we propose a simple way to further leverage unlabeled data by augmenting each utterance with automatically labeled topics. Specifically, we first train a Latent Dirichlet allocation (LDA) topic model on the entire unlabeled corpus, taking each utterance as a document. We used the Python's NLTK (Loper and Bird (2002)) to process the raw scraped data and Gensim package (Řehůřek and Sojka (2010)) to run the LDA algorithm. LDA is an unsupervised machine learning algorithm that uses a probabilistic topic model to discover a set of latent topics from observed documents (Blei et al. (2003)). For each utterance in the dataset, we choose the topics with the highest probabilities assigned and prepend them to the original sequence. For this step, we choosea threshold value of 0.5 for probabilities to ensure only confident assignments are used. We set the model to automatically discover 8 topics. Table 4 lists the top topics found by the model and sample words in each of them. Interestingly, most of the automatically discovered topics overlap with the topics manually identified by the medical research team in the diabetes study.

5. Experiments and Results

Through our experiments, we aim to determine the effectiveness of our proposed architecture, and to measure the role played by the local and global context. We run all our experiments using five-fold cross validation on the annotated corpus of 56 patient-provider sessions described in Section 3.1. All the models are trained for four epochs for each cross-validation iteration. We implemented our models using the PyTorch (Paszke et al. (2019)) and AllenNLP (Gardner et al. (2017)) packages. For training, we use the BertAdam optimizer with weight decay of 0.01, a constant learning rate of $5e^{-5}$, and a batch size of 8 samples. We also apply a dropout rate of 0.1 to all layers. For evaluation, we use accuracy and per topic F1 score metrics at the utterance level.

Topics	Words in Topic	Description
0	month, better, things, hours, great, control, start, times, first, happen	Treatment Regimen/Schedule
1	carbs, numbers, eating, diabetic, normal, would, range, test, higher, food	Diet
2	exercise, morning, daily, night, welcome, usually, help, dinner, every, point	Lifestyle/Exercise
3	fasting, would, meter, thought, everyone, thanks, testing, anyone, doctor, medical	Fasting/Blood Glucose level
4	using, weight, welcome, around, question, diagnose, different, others, diagnosis, since	Weight/Testing
5	metformin, insulin, taking, doctor, thing, medication, works, liver, unit, nothing	Medication
6	sugar, blood, glucose, level, reading, check, forum, drop, bring, sound	Blood Glucose Level
7	diabetes, insulin, change, could, really, years, maybe, problem, think, would	General Treatment

Table 4: Topics learned from a large unlabeled diabetes corpus using Latent Dirichlet Allocation

We first conducted experiments using the 7 diabetes-related topics plus the NA topic, thus building multi class classifiers that predict whether a given utterance belongs to any of these 8 codes. In addition we conducted experiments at different levels of granularity by aggregating the different codes. More specifically, we conducted five-level classification experiments where we attempt to discriminate between five different topics: the diet and exercise (D/E), the glucose management and hypoglycemia (S/H), and the foot care and other health issues (F/O), and the medication (M), and NA topics. We also conduct binary classification experiments that aim to distinguish between all study related topics and not applicable topics (NA).

Context Models. For all models that include contextual information, we set the context size to 10 (not including the target sentence). We evaluate the following settings:

- **Baseline**: Always predicts the most frequent class
- **Context-blind Model**: Transformer model with single utterance input (No Context)
- **Local Context Augmentation Models**: Look before (Before), Look after (After), Look both sides (Both), Joint Sentence Representation (JSR) transformers.
- **Global Context Augmentation Models:** Pretraining on the unannotated corpus (Pretraining) and Topic Augmented (LDA Topics)

Significance Testing. We conduct significance testing to study the effect of our design choices. Specifically, we report dependent t-test results for the following pairs' performances over 5-fold cross validation:

- Look both sides model (Both) and Joint Sentence Representation model (JSR)
- Joint Sentence Representation model (JSR) (without topic augmentation) and Upstage model, which is a JSR model with topic augmentation

For each of the above pairs and for each evaluation metric, in Tables 5, 6 and 7, we mark the model that was found to have higher performance value in a statistically significant manner with the † symbol (Both vs. JSR) and a ‡ symbol (JSR vs. Upstage), using a significance level of $\alpha = 0.05$.

Results. Tables 5, 6 and 7 summarize the performance of the different classification models for the 8-level, five-level, and binary classification tasks. As results show, both locally and globally context-aware models outperform the No Context models by a large margin, thus indicating that context modeling is beneficial for utterance classification during clinical conversations. Overall, across different coding labels, we observe that the results for 8-level and 5-level classification tasks are similar. However, we notice a couple of cases where the performance of the 8-level classifiers outperform the 5-level classifiers, which seems counter-intuitive as one might expect that fewer classes would lead to an easier classification task. We attribute these differences partially to noise introduced in the automatic transcription process and also to the fact that merging of classes does not necessarily make predicting the combined class easier, since utterances with different semantic values are clustered into a unified category.

Model	Acc	F1							
		D	M	S	H	F	O	E	NA
Baseline	49.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	66.64
No Context	57.54	31.47	44.02	16.06	31.71	38.96	38.95	12.27	71.69
Local Context Augmentation Only									
Before	61.80	49.36	52.94	31.41	29.12	51.60	49.10	20.80	74.82
After	61.59	50.53	49.70	30.21	28.42	50.36	47.18	21.92	74.96
Both	62.93	54.27	51.65	33.62	32.19	50.79	47.92	28.56	76.02
JSR	70.54†	64.98†	63.98†	41.68†	40.73†	67.39†	60.08†	35.37†	80.56†
Global Context Augmentation Only									
Pretraining	61.45	43.82	47.58	22.55	25.47	43.01	44.32	7.28	73.65
LDA Topics	62.86	47.49	49.86	33.82	28.75	46.79	45.81	26.40	74.42
Local + Global Context Augmentation									
Upstage	73.04‡	68.34‡	68.60‡	47.30‡	57.50‡	67.94	62.64‡	46.83‡	81.60

Table 5: Eight-level classification with 5-fold cross-validation with context size=10. Legend: D = Diet, M = Medication, S = Glucose-level monitoring, H = Hypoglycemia, F = Foot care, O = Other medical, E = Exercise, NA = Not Applicable

5.1. Effect of Context Modeling Design Choices

Local Context Across all levels of classification, locally context-aware models generally outperform context-blind models in terms of accuracy and individual F1 scores. Further, we observe a divide between models that use joint sentence representation (JSR) and models that uses context only as an additional input for single target utterance classification.

Model	Acc	F1				
		D/E	M	S/H	F/O	NA
Baseline	49.98	0.00	0.00	0.00	0.00	66.64
No Context	58.81	43.42	40.38	24.87	41.98	73.41
Local Context Augmentation Only						
Before	62.92	50.32	50.79	36.89	50.81	74.87
After	61.07	41.17	46.31	31.87	42.42	72.92
Both	63.00	49.74	52.95	38.84	52.28	74.68
JSR	69.76†	60.62†	63.86†	40.47†	61.16†	78.82†
Global Context Augmentation Only						
Pretraining	62.21	45.11	48.59	30.13	46.10	73.68
LDA Topics	64.42	48.67	51.83	43.12	50.14	74.72
Local + Global Context Augmentation						
Upstage	73.26‡	62.83‡	66.37‡ ‡	56.93‡	63.73‡	80.84‡

Table 6: Five-level classification with 5-fold cross-validation with context size=10. Legend follows that of Table 5, with "/" indicating merged topics

Model	Acc	F1	
		M	NA
Baseline	50.02	66.68	0.00
No Context	68.30	68.14	68.45
Local Context Augmentation Only			
Before	73.53	73.09	73.95
After	71.68	71.32	72.03
Both	74.30	74.02	74.58
JSR	78.68†	78.07†	79.12†
Global Context Augmentation Only			
Pretraining	69.96	68.00	71.62
LDA Topics	70.57	68.31	72.47
Local + Global Context Augmentation			
Upstage	79.94	79.84‡	79.59

Table 7: Binary classification with 5-fold cross-validation with context size=10. NA = Not Applicable, M = All codes except NA

Specifically, we observe that JSR models perform better than Both models in terms of accuracy on all levels of classification in a statistically significant manner.

Also, we note that although both models have access to the same amount of contextual information, there is still a difference in terms of performance. Intuitively, a major difference between the models is that in JSR each utterance in a target sentence's local context is also a target utterance to be classified, whereas in Both models the context utterances are distinct from the target sentence. At the same time, in Both, Before, and After models, the target utterance is not directly identifiable, so, they are additionally required to learn to distinguish between context and target utterances during training time.

On the other hand, joint sentence representation allows for a more direct modeling of the dependency between multiple utterances since the model computes a hidden embedding for each utterance, which is used as an input to a final-layer classifier. In this scheme, there is no distinction between target and context utterances, and the model can focus on learning to produce a contextualized representation for each utterance through `[SEP]` tokens.

Global Context In this study, we hypothesized that although the collected web data might differ in content and style, there is enough similarity in the contained natural language text to provide *global* contextual information about topics discussed in the clinical conversations. The results shown in Tables 5, 6 and 7 (Pretraining, LDA Topics, and Upstage), show that context-aware models tend to have higher performance results compared to their context blind counterparts (i.e. No Context for Pretraining and LDA Topics and JSR for Upstage). Comparing JSR and Upstage, the latter has an additional cost of self-supervised pretraining and estimating LDA topic parameters from unlabeled data and inferring likely topics for each utterance in the dataset. However, we note that these costs are often acceptable as compared to the cost of gathering more labeled data or training very large models.

6. Discussion and Future Work

When developing utterance classifiers to automatically label topics discussed during patient-provider conversations, we identified two important challenges to be addressed. First, patient-provider communication is complex and covers a variety of issues, including patient concerns, medication issues or small talk. Thus, it is important to have a distinction of the different topics discussed during the conversation. In addition, the conversation dialog exchanges are free form, have arbitrary-length and often include short turns, which in some cases, provide little to no direct information regarding the topics or issues being discussed. Second, manually annotating topics during the conversations, which is needed for training supervised classification models, is expensive and time-consuming as it requires very specific human expertise.

To tackle the first problem, we studied how *local* context from neighboring sentences in a conversation can be naturally modeled using joint sentence representation (JSR) as input for Transformer models. Our empirical evaluations showed that incorporating local context in the topic classification for the conversation utterances lead to improved performances. We also provided evidence that JSR-based models outperform alternative strategies that also use BERT-based architectures.

In order to mitigate the lack of labeled data, we focused on the fact that when in-domain annotated data is scarce we can leverage unlabeled and semantically similar text. More specifically, we used related text collected from an online diabetes discussion forum. Although the forum interactions differ from patient-provider interactions in style and content, we note that the semantic similarity and abundance of natural language on diabetes issues can be used to improve the performance of the classifier. In addition to using the unlabeled corpus for self-supervised pretraining of the base Transformer model, we developed an LDA topic model from the unannotated corpus, hence allowing us to include in the representation of input utterances the *global* context information learned from the discussions of online forum users.

Limitations and Future Work. Our work does not fully explore the possible space of document-level context modeling. For example, recurrent models are not considered, as our work focuses on transformer-based models, with emphasis on leveraging transfer learning using pretrained parameters trained from very large corpora. Recurrent architectures like Long Short Term Memory networks (LSTMs) (Hochreiter and Schmidhuber (1997)) are often incorporated into latent variable models or hierarchical models for modeling conversations or long documents. Since joint sentence representation can be applied to any transformer-like architecture, a promising next step is to explore state-of-the-art models that allow very long sequences for each input as a result of improved memory usage.

Furthermore, the LDA module that provides global context from unlabeled corpora has not been systematically studied. Thus, we plan to study how topical information affects the performance of our model under different settings and parameters, such as number of topics to be discovered or size and composition of the unlabeled corpus. Moreover, different designs of incorporating the topic assignment can be considered. Finally, we believe that noisy automatic transcription may be one of the key obstacles to achieving better performance. Thus, we plan on quantifying the effect of noise from automatic transcriptions and developing ways to counter it in the context of utterance classification.

7. Conclusion

In this work, we presented Upstage, a framework for classifying utterances in health-related conversations between providers and patients in a diabetes study. In order to address the challenge of capturing contextual information necessary for correct classification of individual utterances, Upstage uses a transformer model augmented with local and global context. The framework utilizes contextual information obtained from joint sentence representation, as well as pretraining and topic assignments from large semantically similar text data. Experimental results demonstrated the effectiveness of Upstage.

Acknowledgements

This material is based in part upon work supported by the GRADE Study, which is supported by a grant from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) of the National Institutes of Health under Award Number U01DK098246 (Nathan et al. (2013)). The planning of GRADE was supported by a U34 planning grant from the NIDDK (U34-DK-088043). The American Diabetes Association supported the

initial planning meeting for the U34 proposal. The National Heart, Lung, and Blood Institute and the Centers for Disease Control and Prevention also provided funding support. The Department of Veterans Affairs provided resources and facilities. Additional support was provided by grant numbers P30 DK017047, P30 DK020541-44, P30 DK020572 (MDRC Clinical Core), P30 DK072476, P30 DK079626, P30 DK092926 (MCDTR Methods and Measurement Core), U54 GM104940, UL1 TR000439, UL1 TR000445, UL1 TR001108, UL1 TR001409, UL1 TR001449, UL1 TR002243, UL1 TR002345, UL1 TR002378, UL1 TR002489, UL1 TR002529, UL1 TR002535, UL1 TR002537, and UL1 TR002548, and by the Precision Health initiative at the University of Michigan.

Educational materials have been provided by the National Diabetes Education Program. Material support in the form of donated medications and supplies has been provided by BD, Bristol-Myers Squibb, Merck, NovoNordisk, Roche Diagnostics, and Sanofi. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Precision Health Initiative.

References

Ruchit Agrawal, Marco Turchi, and Matteo Negri. Contextual handling in neural machine translation: Look behind, ahead and on both sides. 2018.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S. Weld. Pretrained language models for sequential sentence classification, 2019.

Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. De-identification of patient notes with recurrent neural networks. *CoRR*, abs/1606.03475, 2016. URL http://arxiv.org/abs/1606.03475.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9 (8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes, and Donald E. Brown. Text classification algorithms: A survey. *CoRR*, 2019. URL http://arxiv.org/abs/1904.08067.

QingBiao Li, ChunHua Wu, KangFeng Zheng, and Zhe Wang. Hierarchical transformer network for utterance-level emotion recognition, 2020.

Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002.

James Mork, Olivier Bodenreider, Dina Demner-Fushman, Rezarta Dogan, François-Michel Lang, Zhiyong lu, Aurélie Névéol, Lee Peters, Sonya Shooshan, and Alan Aronson. Extracting rx information from clinical narrative. *Journal of the American Medical Informatics Association : JAMIA*, 17:536–9, 09 2010. doi: 10.1136/jamia.2010.003970.

David M. Nathan, John B. Buse, Steven E. Kahn, Heidi Krause-Steinrauf, Mary E. Larkin, Myrlene Staten, Deborah Wexler, John M. Lachin, and the GRADE research group. Rationale and design of the glycemia reduction approaches in diabetes: A comparative effectiveness study (GRADE). *Diabetes Care*, 36(8):2254–2261, 2013. ISSN 0149-5992. doi: 10.2337/dc13-0356. URL https://care.diabetesjournals.org/content/36/8/2254.

Jihyun Park, Dimitrios Kotzias, Patty Kuo, Robert L Logan IV, Kritzia Merced, Sameer Singh, Michael Tanana, Efi Karra Taniskidou, Jennifer Elston Lafata, David C Atkins, Ming Tai-Seale, Zac E Imel, and Padhraic Smyth. Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. *Journal of the American Medical Informatics Association*, 26(12):1493–1504, 09 2019. ISSN 1527-974X. doi: 10.1093/jamia/ocz140. URL https://doi.org/10.1093/jamia/ocz140.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1088. URL https://www.aclweb.org/anthology/P19-1088.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

Kirk Roberts and Sanda M Harabagiu. A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association*, 18 (5):568–573, 07 2011. ISSN 1067-5027. doi: 10.1136/amiajnl-2011-000152. URL https://doi.org/10.1136/amiajnl-2011-000152.

Sebastian Ruder. *Neural Transfer Learning for Natural Language Processing*. PhD thesis, National University of Ireland, Galway, 2019.

Apodaca TR, Borsari B, Jackson KM, and et al. Sustain talk predicts poorer outcomes among mandated college student drinkers receiving a brief motivational intervention, 2014. URL doi:10.1037/a0037296.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems, 2019.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2019. URL http://arxiv.org/abs/1906.08237. cite arxiv:1906.08237Comment: Pretrained models and code are available at https://github.com/zihangdai/xlnet.