Developing an Utterance Classifier for Patient-Nurse Conversations for Comparative Effectiveness Study

F L, F L, F L
NAME@UMICH.EDU, NAME@UMICH.EDU, NAME@UMICH.EDU
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI, USA

Editor: Editor's name

Abstract

When analyzing utterances in long documents, it is not sufficient to consider each sentence in isolation, since its context may play a role in determining its semantic content. Recently, contextual information in natural language documents has been modeled using various techniques, such as recurrent neural networks with latent variables, or through neural networks with attention mechanisms. Automatic classification of sentences in a health-related conversation into a predefined set of topics has a wide range of applications in healthcare, ranging from identifying health problems or complaints to translational research of clinical treatments. In this paper, we adapt Cohan et als approach of using pretrained language models with joint sentence representation [3] to solve the task of analyzing long patient-nurse conversations during regular checkups. Whereas Cohan et al's approach is limited to short documents such as scientific abstracts, our work focuses on utilizing transformer language models for arbitrary-length, free-form dialogue between a patient and a nurse. We assume that in patient-nurse conversations, contextual information is relevant only locally, and use fixed-size context window and explore different ways to aggregate predictions. We compare our approach to models with varying capacity to handle contextual information, and show that our model learns to integrate context as part of its classification procedure.

1. Introduction

Determining the cost of a treatment is a central objective of translational medicine research. To estimate the cost, different measures such as direct medical costs (including time and money of the treatment) and indirect costs are considered. Then, multiple treatments can be compared using the estimated cost in a comparative effectiveness study, allowing researchers or practitioners to make informed decisions.

While some costs of a treatment such as expense can be identified with simple means such as database query, there are other costs associated to a treatment that eludes ease retrieval. For instance, the patient's subjective experience, independent of the outcome of the treatment, is an important aspect that should not be overlooked. At the same time, several treatments require patients to regularly visit and interact with care providers, and during these regular sessions, patients engage in a free-form conversation with medical workers, loosely centered around health-related issues. To utilize information from the natural language data contained such sessions, we propose a transformer-based model to classify

utterances in patient-nurse conversations collected as part of the Glycemia Reduction Approaches in Diabetes (GRADE) study. Our model is a sentence-level classifier based on the transformer architecture, and models contextual information in the input using joint sentence representation.

Technical Significance There have been several studies on analyzing and classifying health-related documents. However, our project is distinguished from them in that patient-nurse conversations are not confined to a predefined structure and length, as is often the case with electronic medical records or clinical reports. The conversations are free-form and each participant is at liberty to talk about any item or issue outside of the relevant medical domain at any time. Therefore, it is necessary to develop a model that is robust to the variability of style, length, and progression of natural language.

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-nurse conversations into a set of clinically relevant topics. Our model is distinguished from other sentence-level classifiers in that it incorporates local context, following the approach of Cohan et al [3].
- We empirically show that pretraining the transformer-based model on an out-of-domain corpus collected from an online diabetes forum leads to increased performance.
- We demonstrate the importance of contextual information by comparing and analyzing classification models which utilize varying amounts of context. Moreover, we compare different ways to represent contextual information.

Clinical Relevance Our work has direct and indirect clinical applications. First, the classifier will be used to provide data for the Glycemia Reduction Approaches in Diabetes (GRADE) study, conducted by the Brehm Center for Diabetes Research at the University of Michigan. The GRADE study is a comparative effectiveness study that seeks to measure and compare the costs of different type-2 diabetes treatment. Given sentence-level annotations of utterances, clinicians and medical researchers can perform statistical analysis using metrics such as fraction of time each topic was discussed, or distribution and change of patients' discussed topic over a period of time. This allows researchers to quantify and compare aspects of clinical treatments that were previously hard to capture, including patients' subjective experience and response.

2. Related Work

Transformer models Recently, many natural language processing tasks have been successfully tackled by frameworks that make use of the Transformer as its backbone [12]. Several NLP problems can be cast as a sequence transduction task, and the Transformer architecture approaches this using multi-head attention instead of temporal sequence, achieving parallelizable and faster computation during training and inference. Some of the most successful Transformer-based models include BERT, GPT2, and XLNet [10, 4, 14] . A key insight behind using these large-scale models effectively is to leverage transfer learning. By

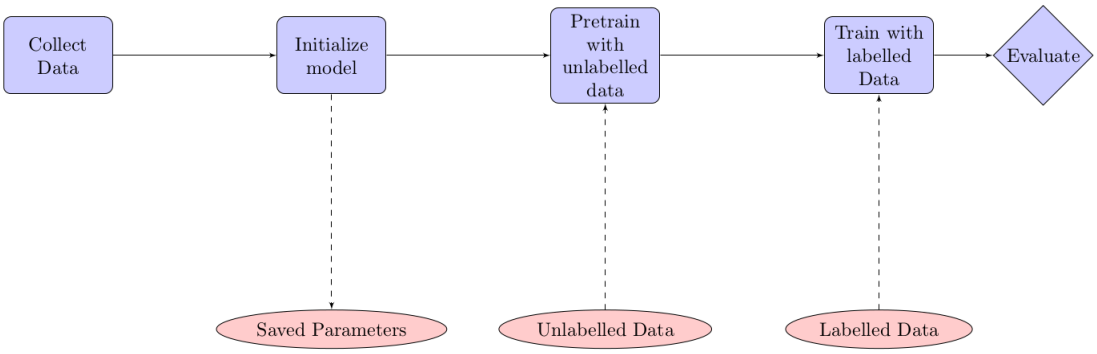

Figure 1: An overview of the framework used in the paper

pretraining the models using density estimation or masked token prediction, large corpuses of unannotated language data can be used to increase the performance of the model in a range of downstream tasks. Moreover, practitioners can choose pretraining corpuses that are semantically or stylistically aligned to the downstream task's domain for more effective transfer learning.

Text classification Text classification has been widely studied in natural language processing and artificial intelligence, both as a way to analyze natural language data but recently also as a framework to study various tasks of natural language understanding [13]. Feature engineering approaches and deep learning models can be used for varying degrees of units are considered for classification: document-, paragraph-, sentence (utterance)-, and sub-word-levels [6]. In this work, the task of classifying is utterances in conversations is naturally cast as a sentence-level classification task. However, it is also important to note that each individual session defines a document-level context that individual sentences might depend on for correct classification. In the context of emotion recognition in conversation (ERC), context modeling has been tackled using hierarchical models or recurrent models with latent variables [7, 9].

Analysis of health-related conversations Medical researchers and clinicians have applied qualitative analaysis and natural language processing to language use in health-related conversations. In the context of motivational interviewing (MI), a counseling methodology that aims to induce behavior change in clients, researchers have shown that clients' language use during counselling sessions can be used to predict the outcome of the intervention [11, 8].

Table 1: 8-, 5-, and 2-level Classification Codes and Fraction of Utterances in each code class in the annotated dataset

8-levels	5-levels	2-levels
Diet/Weight (D) (8.2%)	D/E (13%)	Medical (48.4%)
Exercise (E) (12.7%)		
Medication management (M) (4.8%)	M (12.7%)	
Self-monitoring blood glucose (S) (2.2%)	S/H (7%)	
Hypoglycemia (H) (2.1%)		
Foot care (F) (16.7%)	F/O (17.9%)	
Other medical management (O) (1.2%)		
Not applicable (NA) (51.7%)	NA (51.6%)	NA (51.6%)

3. Task and Dataset

3.1. Task Definition and Notation

Data We denote the set of annotated sessions to be used in supervised training as $D_{s_i} = \{({X_s}^i, {Y_s}^i)\}_{i=1}^{I_s}$. Each annotated sample $({X_s}^i, {Y_s}^i)$ consists of J_{s_i} utterances and topic assignments such that ${X_s}^i = \{{u_s}^j\}_{j=1}^{J_{s_i}}$ and ${Y_s}^i = \{{c_s}^j\}_{j=1}^{J_{s_i}}$. Each ${u_s}^j$ is an English sentence and ${c_s}^j$ is a topic assignment from one of 8-, 5-, and 2-level codes in Table 1. The set of unannotated sessions to be used in self-supervised pretraining is denoted as $D_{u_i} = \{({X_u}^i\}_{i=1}^{I_u}$, with ${X_u}^i = \{{u_u}^j\}_{j=1}^{J_{u_i}}$.

Task Given an unannotated session ${X_u}^i \in D_{u_i}$, our goal is to predict a correct topic prediction ${Y_u}^i = \{{c_u}^j\}_{j=1}^{J_{u_i}}$. Following [3], we identify this task as a sequential sentence classification (SSC), because of the fact 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.

3.2. The GRADE Sessions

Collection The GRADE sessions used in this work are collected as part of the ongoing Glycemia Reduction Approaches in Diabetes (GRADE) study, which is a nation-wide, multi-group clinical trial program that is designed to study and compare different treatments of type-2 diabetes, with the long-term objective of allowing medical practitioners to make better informed decisions in choosing which treatment to administer [2]. Participants are required to make continued, regular follow-up visits to local hospitals, during which a recording device is used to capture the patient-nurse conversation.

Transcription and Annotation After the audio files are prepared, the recordings are then automatically transcribed using Google Cloud API's speech-to-text service. If the session is marked for annotation, human annotators manually annotate each utterance with a single topic code from one of 8-level, 5-level, and 2-level classification categories, as listed

in 1. In this multi-tiered category system, topics are merged based on similar aspects they evaluate, in order to allow for varying levels of granularity in classification.

Dataset (# Labels)	# Documents			# Sentences (Avg #)		
	Train	Val	Test	Train	Val	Test
IEMOCAP (6)	120		31	5810 (48.41)		1623 (52.35)
SEMAINE (8)	63		32	4368 (69.33)		1430 (44.68)
EmotionLines (7)	720	80	200	10561 (14.66)	1178 (14.72)	2764 (13.82)
MELD (7)	1039	114	280	9989 (9.61)	1109 (9.72)	2610 (9.32)
DailyDialogue (7)	11118	1000	100	87832 (7.89)	7912 (7.91)	7863 (7.86)
EmoContext (4)	30159	2754	5508	90477 (3.00)	8262 (3.00)	16524 (3.00)
PubMed (5)	20000			225000 (11.25)		
NICTA (4)	1000			21000 (21.00)		
CSAbstruct (5)	2200			15000 (6.81)		
CSPubSum (2)	21000			601000 (28.61)		
GRADE$_A$ (8/5/2)						
GRADE$_U$ (8/5/2)						

Table 2: Comparison of Multiple Sentence-level Classification Benchmark and Datasets. GRADE$_A$ denotes the annotated GRADE sessions, and GRADE$_U$ the unnoated sessions.

From Table 2, we note that while the number of individual dialogues/documents containing utterances is very sparse for the GRADE dataset compared to other sentence-level text classification benchmarks. On the other hand, each individual session in the GRADE set tends to consist of a larger number of utterances, as indicated by the average number of sentences per session.

3.3. Online Diabetes Forum Data

In addition to the GRADE sessions described above, we utilize language data from an online diabetes forum, Diabetes Daily. We used Python's `Scrapy` package to collect threads on a Type-2 diabetes-themed forum where users start and respond to posts relating to Type-2 Diabetes and its treatment. We note that the forum corpus is out-of-domain for multiple reasons. First, the interaction contained in the collected threads are not necessarily dyadic, unlike the patient-nurse conversations from the GRADE dataset. Also, Daily Diabetes threads are typed, not spoken and automatically transcribed. Finally, a client-caretaker dynamic is not present in the online threads, although the different roles a new patient and an experience patient play in the interaction might be similar. As with the GRADE unannotated set, we pretrain the base transformer model in a self-supervised manner with the forum data.

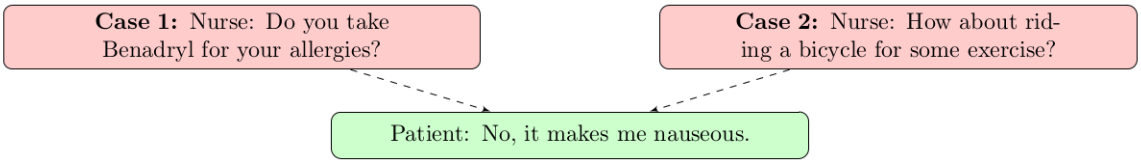

Figure 2: Two constructed patient-nurse interactions with a shared response

4. Methods

4.1. Transformer-based Model

TODO: Brief explanation of the basic transformer model, Multihead attention, Transformer model architecture (BERT, XLNet, GPT2). Explain MLP = Multi-Layer Perceptron, with equation.

4.2. Pretraining with Unannotated Data

TODO: Explain briefly Transfer learning, Explain the process of pretraining for transformer model (masked LM, permutation LM, etc).

4.3. Leveraging Contextual Information with Joint Sentence Representation

In designing the classifier, it is necessary to equip the model with some capacity to model contextual information, as each isolated utterance can often be ambiguous in its meaning and thus its topic classification. Consider the example shown in Figure 2 which shows two plausible questions by the nurse that could have prompted the patient's answer. In Case 1, the nurse is asking about 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, now the question is about the patient's exercise regimen, so both utterances should be classified as "E (Exercise)".

In this work, we follow the joint sentence representation (JSR) approach proposed by Cohan et al in [3]. In common applications of the Transformer models in text classification, the `[CLS]` token is concatenated to the sequence to be classified. Then, the hidden representation corresponding to the token is fed to a feedforward neural network layer which then outputs a 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, as depicted in Figure 3. In the joint sentence representation framework showin Figure 4, the `[SEP]` token replaces the role of the `[CLS]`. 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.

4.4. Alternative Designs of Context Modeling

Beside using the above approach, we also consider alternative ways of including context, mainly following the simple concatenation approaches explored by Agrawal et al in [1]. Al-

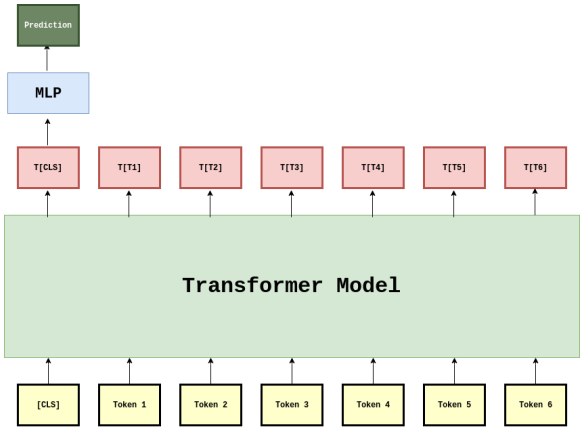

Figure 3: Sentence classification using conventional representation

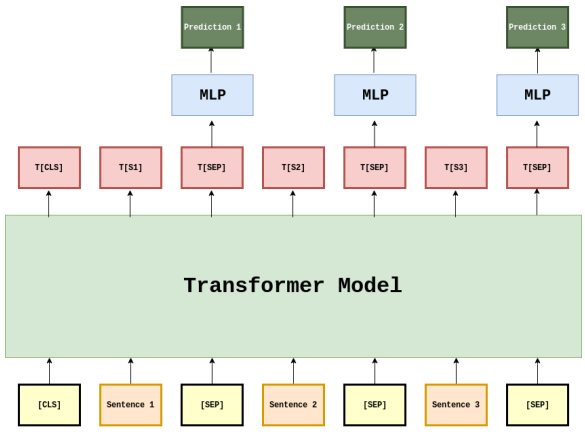

Figure 4: Sentence classification using joint sentence representation

though Agrawal et al employs the design in machine translation, we use a simple adaptation their approach by skipping target side context. In this design, we concatenate sentences before and/or after the sentence to be classified 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.

- **Look-before Context** k previous utterances are prepended to the original input.
- **Look-ahead Context** k following utterances are appended to the original input.
- **Both sides Context** k utterances each from Look-before and Look-ahead are concatenated to the input.

Although this approach considers multiple utterances simultaneously, it makes a prediction only for the target utterance, not its context sentences, as illustrated in Figure 3.

5. Experiments

5.1. Experiment Setup

Configuration TODO: Libraries, Computer Setting, etc.

Context Models TODO: List of models experimented. For all the models that include contextual information, we set the context size = 10 for this experiment.

- Baseline: Majority Class Baseline
- Joint Sentence Representation (JSR), n is the context size
-

Evaluation Metrics TODO: each F1, accuracy

5.2. Results

5.2.1. Analysis of Context Modeling Choice

Table 3 and 4 summarize the performance of the models.

Table 3: 8-level classification with 5-fold cross-validation

Model	Acc	F-score	F_D	F_E	F_M	F_S	F_S	F_F	F_O	F_NA
Baseline										
JSR										
JSR + Pretraining										

Table 4: Binary classification with 5-fold cross-validation

Model	Acc	F-score	F_M	F_NA
Baseline				
JSR				
JSR + Pretraining				

5.2.2. Analysis of Context Size

TODO: focus on JSR (best performing model) and just a few context-including models and plot the performance of the model with varying context size n.

5.3. Qualitative Analysis Through Attention Visualization

TODO: visualize attention and include figure

6. Discussion and Future Work

TODO: limits - Fixed Context Size so Global Context is mixing - Qualitatively analyzing the sessions, it is intuitive that the conversation in each session consists of mini "chat" around certain topics. - limited amount of data, topics were fixed, single label classification

Incorporate larger transformer models like Reformer, [5] Also recurrent models to handle arbitrary length context, hierarchical models

7. Conclusion

In this work, we presented a transformer-based utterance classifier for health-related conversations between nurses and patients participating in a comparative effectiveness study of type-2 diabetes treatments. In order to address the challenge of capturing contextual information necessary for correct classification of individual utterances, we explored and experimented with different design choices for modeling context. We restricted our model search space to transformer-based model in order to leverage transfer learning using pretrained parameters trained from very large corpuses. Our evaluation of these models showed

that a BERT-based classifier that models local contextual information with a joint sentence representation outperforms the alternatives.

References

[1] Ruchit Agrawal, Marco Turchi, and Matteo Negri. "Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides". In: 2018.

[2] GRADE Study Coordinating Center. *The Glycemia Reduction Approaches in Diabetes: A Comparative Effectiveness Study (GRADE Study), Statistical Analysis Plan*. 2017. URL: https://www.clinicaltrials.gov/ProvidedDocs/43/NCT01794143/SAP_001.pdf.

[3] Arman Cohan et al. *Pretrained Language Models for Sequential Sentence Classification*. 2019. arXiv: 1909.04054 [cs.CL].

[4] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[5] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. "Reformer: The Efficient Transformer". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=rkgNKkHtvB.

[6] Kamran Kowsari et al. "Text Classification Algorithms: A Survey". In: *CoRR* (2019). arXiv: 1904.08067. URL: http://arxiv.org/abs/1904.08067.

[7] QingBiao Li et al. *Hierarchical Transformer Network for Utterance-level Emotion Recognition*. 2020. arXiv: 2002.07551 [cs.CL].

[8] Verónica Pérez-Rosas et al. "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*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 926–935. DOI: 10.18653/v1/P19-1088. URL: https://www.aclweb.org/anthology/P19-1088.

[9] Soujanya Poria et al. *Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances*. 2019. arXiv: 1905.02947 [cs.CL].

[10] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2018). URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[11] Apodaca TR et al. *Sustain talk predicts poorer outcomes among mandated college student drinkers receiving a brief motivational intervention*. 2014. URL: doi:10.1037/a0037296.

[12] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[13] Alex Wang et al. *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*. 2019. arXiv: 1905.00537 [cs.CL].

[14] Zhilin Yang et al. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. cite arxiv:1906.08237Comment: Pretrained models and code are available at https://github.com/zihangdai/xlnet. 2019. URL: http://arxiv.org/abs/1906.08237.