Joint Dialogue Topic Segmentation and Categorization: A Case Study on Clinical Spoken Conversations

Zhengyuan Liu[†], **Siti Umairah Md Salleh**[†],
Hong Choon Oh[‡], **Pavitra Krishnaswamy**[†], **Nancy F. Chen**[†]
[†]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[‡]Health Services Research, Changi General Hospital, Singapore
{liu_zhengyuan,nfychen}@i2r.a-star.edu.sg

Abstract

Utilizing natural language processing techniques in clinical conversations is effective to improve the efficiency of health management workflows for medical staff and patients. Dialogue segmentation and topic categorization are two fundamental steps for processing verbose spoken conversations and highlighting informative spans for downstream tasks. However, in practical use cases, due to the variety of segmentation granularity and topic definition, and the lack of diverse annotated corpora, no generic models are readily applicable for domain-specific applications. In this work, we introduce and adopt a joint model for dialogue segmentation and topic categorization, and conduct a case study on healthcare follow-up calls for diabetes management; we provide insights from both data and model perspectives toward performance and robustness.

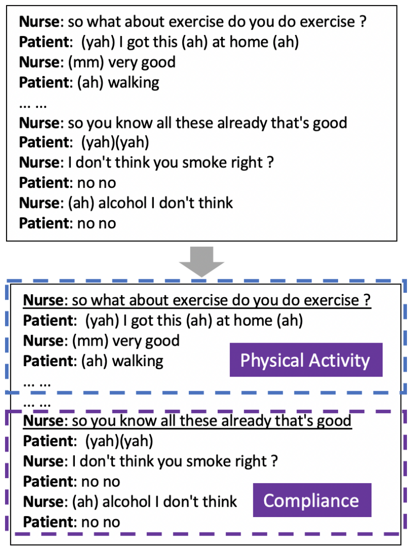

Figure 1: A dialogue example with topic segmentation and categorization. Frames indicate topically-coherent segments, and the corresponding label is highlighted. Utterances at the beginning of segments are underlined.

1 Introduction

The massive records of clinical communication, especially the longitudinal follow-up calls, can be used to scrutinize novel insights into medical history, treatment plans, and customized education (Quiroz et al., 2019); but it is time-consuming and requires domain knowledge for manual operation. Therefore, there has been growing interest in utilizing speech and natural language techniques to analyze and distill information from clinical conversations (Liu et al., 2019b; Krishna et al., 2021; van Buchem et al., 2021). While spoken conversations are often loosely structured, in task-oriented scenarios, interlocutors calibrate the dialogue flow to cover targeted topics and agendas (Sacks et al., 1978). Moreover, when large language models (Brown et al., 2020) are applied, processing the verbose conversations will substantially increase the computational complexity and cost. On the other hand, dialogue segmentation and topic categorization (Arguello and Rosé, 2006; Mei et al., 2007) are useful to handle lengthy inputs, reduce data noise by excluding the task-irrelevant segments, and improve the efficiency of downstream tasks (Liu et al., 2019c; Khosla et al., 2020). More specifically, dialogue segmentation is to extract the structural information by splitting the whole session into topically-coherent segments (Arguello and Rosé, 2006), and topic categorization labels each segment with a particular type, providing features for fine-grained semantic understanding (Mei et al., 2007).

Different from documents, human conversations include ubiquitous verbal and vernacular expressions, along with disfluencies, thinking aloud, and repetition. This leads to lower information density (Sacks et al., 1978) and more topic drifting. The coherence-based methods typically applied to passages cannot perform well on spoken dialogues. Moreover, since there are few corpora constructed with the dedicated annotation, most existing generic (both supervised (Arnold et al., 2019) and unsupervised (Xing and Carenini, 2021)) models cannot meet the requirements of real-world

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 185–193
December 6-10, 2023

applications and provide reliable system outputs. This is because (1) there is no unified segmentation granularity across different data resources, and (2) the variety of topic definitions increases the difficulty of domain adaptation, especially where language resources are limited. In this work, we conduct a case study on a clinical conversation scenario. Because of the chronic nature of diabetes and its associated complications, diabetes requires constant attention and regular follow-up actions (Lawson et al., 2005; Wai Leng et al., 2014). Nurses schedule calls with patients to track their compliance status and health condition, and provide customized coaching and advice (Piette et al., 2001). To facilitate the communication process, dialogues are organized according to a checklist or medical protocol (Kirkman et al., 1994; Taylor et al., 2003). However, due to the characteristics of spoken dialogues such as topic drifting and verbosity, the important information is scattered across the whole conversation, which renders it a representative use case for dialogue segmentation and topic categorization (as the example shown in Figure 1).

Since no existing generic models meet the requirements of our domain-specific application, we investigate a data-driven approach for the clinical conversation processing task, and our contributions of this work are as follows:

- We build our in-domain dataset from follow-up calls for health management with dedicated annotation of dialogue segmentation and topic categorization.
- We conduct quantitative and qualitative analyses on the clinical conversation data, and describe their conversational linguistic features.
- We propose and apply a joint framework for topic segmentation and categorization, by equipping a shared language backbone with functional components.
- We report extensive experimental results, and evaluate the model performance from the accuracy and robustness perspective.

2 Our Clinical Conversation Corpus

2.1 Data Preparation and Annotation

Our data are constructed on recordings of diabetes management follow-up calls. The clinical data were acquired by the Health Management Unit at Changi General Hospital. This research study was approved by SingHealth Centralised Institutional Review Board (Ref: 2019/2803) and A*STAR IRB (Ref: 2019-079). Telephone care programs are a viable strategy for bringing diabetes management services to patients and improving their glycemic control (Wai Leng et al., 2014), and nurses communicate with patients or caregivers following established protocols (Lawson et al., 2005; Taylor et al., 2003). To transform the raw data into a sample set that can be used for developing computational language solutions, we transcribe and annotate the call recordings following two steps: (1) First, speech transcribers are employed for manual speech-to-text conversion to ensure the quality, and transcripts are fully anonymized. Speaker roles (e.g., nurse, patient, caregiver) are added to each utterance. Following previous work (Liu et al., 2019c), the informal and spontaneous styles of spoken interactions such as interlocutor interruption, backchanneling, hesitation, false starts, repetition, and topic drifting are preserved. (2) The annotation of dialogue segmentation and topic categorization is then performed using in-house software. Our linguistic annotators are familiar with clinical conversations, and they have finished a training session on diabetes health management. We formulate the segmentation granularity and topic categories (see Table 1) based on the annotation protocol defined by the healthcare provider. Moreover, there have been three iterations for the corpus construction, where we collect feedback from clinical collaborators, refine the annotation scheme, and update the whole corpus accordingly.

	Segment Number	Averaged Length
1. Introduction	695	97.41
2. Identification	660	65.02
3. General Education	2194	328.4
4. Oral Medication	909	184.5
5. Insulin	468	171.6
6. Self-Monitoring	1276	165.5
7. Programme	766	196.4
8. Vitals	1033	111.8
9. Medical Experience	782	271.4
10. Base Compliance	252	138.8
11. Appointments	711	199.6
12. Social Chatting	296	245.5
13. Physical Activity	455	147.4
14. Diet Management	662	301.3
15. Hyper/Hypo Incident	140	199.5
16. Other	418	244.2

Table 1: Data statistics of topic categorization. We count the number of topically-coherent segments of each topic, and their average word number (length).

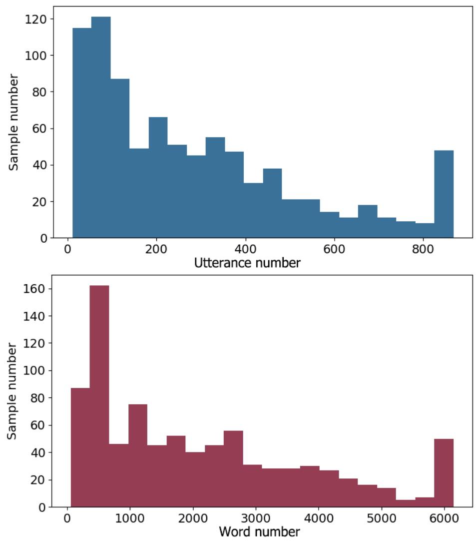

Figure 2: Utterance-level and word-level length distributions of the annotated clinical conversations.

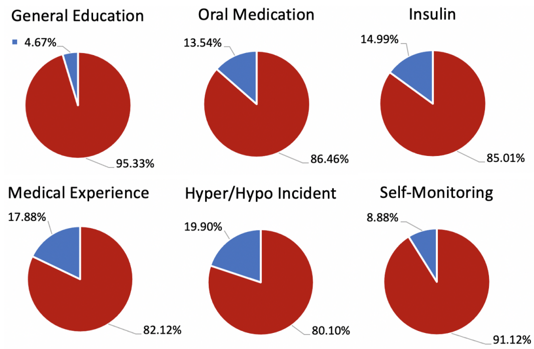

Figure 3: Speaker distribution of selected topic types. The proportion of nurse is in red; others are in blue.

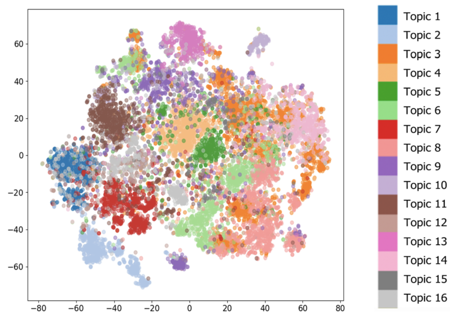

Figure 4: Feature visualization of segment embeddings via t-SNE. The colored points denote topically-coherent segments labeled in different topics.

2.2 Data Statistics

The annotated dataset contains 865 transcripts. As shown in Table 1, for the dialogue topic analysis, there are 16 topic types; the class of *'Other'* includes topics with less medical information such as financial support and caregiver. In our fine-grained annotation, some topics have sub-categories (e.g., *'Customized Coaching'* is one sub-topic of *'Physical Activity'*, *'Insulin'*, and *'Self-Monitoring'*, and we use their base topic type for the labeling task. Figure 1 shows one annotated dialogue example with two topic segments.

(a) Length Distribution With a lower information density, spoken dialogues are often much longer than documents. In our transcribed calls, the maximum, median, and minimum utterance numbers are 1996, 221, and 21, respectively; the maximum, median, and minimum number of words are 16701, 1684, and 70, respectively. The lengthy conversations are usually caused by covering more topics, as well as a detailed discussion. As shown in Figure 2, nearly 5% samples (at the 95% quantile) are comprised of more than 800 utterances (6000 words), which significantly surpasses the input limit of many language backbones (Liu et al., 2019a; Lewis et al., 2020).

(b) Topic Distribution For efficient communication, nurses organize follow-up calls based on patient profiles and health management programmes. As a result, topics present different importance in the form of frequency and length. As shown in Table 1, we calculate the segment number of each topic and their average word number. We observe that some topics are frequent and more well-discussed such as *'General Education'*, *'Medical Experience'*, and *'Diet Management'* (Nazar et al., 2016), while some are more targeted and concise such as *'Identification'*, *'Vitals'*, and *'Insulin'*.

2.3 Conversational Linguistic Features

In order to gain insights into the clinical dialogues, we conduct three quantitative analyses using the annotated topic segments. Here are some findings:

(a) Nurses are the main topic coordinator. We extract the speaker role information from the first utterance of each segment. As shown in Figure 2, the dialogue topic shift is mainly led by the nurses, which is consistent with the purpose of diabetes follow-up calls (Piette et al., 2001), indicating the speaker role can contribute to the topic analysis.

(b) Questions lead the topic shifting. Since punctuation marks are retained in our transcribing, we calculate the number of utterances that end with a question mark, and it shows that 83% of the topic shifting starts with an inquiring utterance.

(c) Different topics show distinct semantics. Aside from the in-topic coherence, different topics will present diverse distribution in a semantic space.

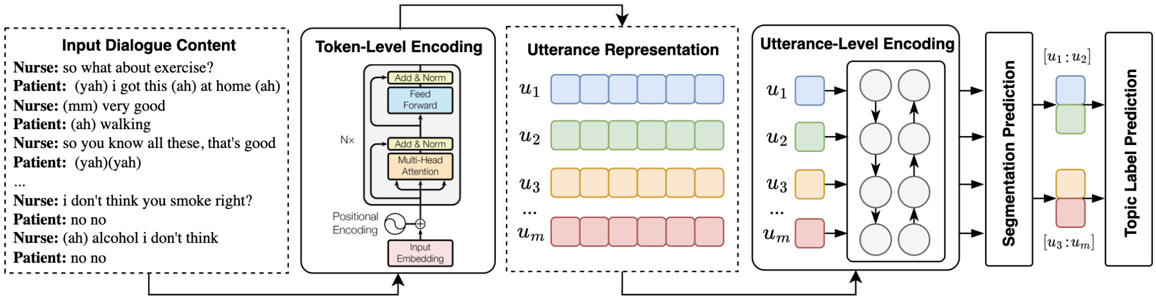

Figure 5: Overview of the joint framework for dialogue topic segmentation and categorization.

Thus, we conduct a semantic feature visualization. We obtain segment representations from an unsupervised sentence embedding model SimCSE (Gao et al., 2021), and use t-SNE (Van der Maaten and Hinton, 2008) to illustrate their distribution in a 2-dimensional space. As shown in Figure 4, the language used in different topics is specific, and varies from one to the other.

3 Joint Model of Dialogue Topic Segmentation and Categorization

3.1 Task Definition

Given a dialogue D which is composed of m utterances $\{u_1, u_2, ..., u_m\}$, (1) a topic segmenter is applied to score each utterance with $y_i^b \in [0, 1]$ that indicates whether it is the first utterance of each segment; (2) a topic classification model is applied to determine the topic label of each segment, where $y_i^t \in [t_1, t_2, ..., t_k]$, and k denotes the categorical dimension (set at 16).

3.2 Framework Description

Following the task-specific fine-tuning paradigm (Liu et al., 2019a; Xing and Carenini, 2021), we build a joint model of dialogue segmentation and topic categorization by equipping a Transformer-based language backbone with functional modules, and its overview is shown in Figure 5.

(a) Token-level Encoder The token-level encoder consists of a stack of Transformer layers; each layer contains a multi-head self-attention and a position-wise feed-forward component. Residual connection and layer normalization are employed. Its input is represented as [<*s*>,u_1,<*s*>,u_2, ...,<*s*>,u_m], where special token '<*s*>' is used as the delimiter. To maximize the receptive field of the token-level encoding, we adopt a sliding-window strategy on full-length input sequences (Wang et al., 2019).

(b) Utterance-level Encoder After token-level context encoding, we obtain the utterance embeddings by extracting hidden states of all delimiters '<*s*>'. Then a Bi-directional LSTM is used for encoding at the utterance level (Liu and Chen, 2021).

(c) Dialogue Segmentation Module The segmentation component F_{seg} (a linear layer) is applied to utterance-level representations, predicting the boundary probability y_i^b. Binary cross-entropy loss is calculated between the model prediction and ground truth. As shown in Figure 5, assuming u_1 and u_3 are the boundary utterances, two topic segments $[u_1{:}u_2]$ and $[u_3{:}u_m]$ are formed.

(d) Topic Categorization Module After dialogue segmentation, for each topically-coherent span, we obtain its segment embedding by aggregating and averaging utterance-level representations. Then the topic categorization module F_{topic} (another linear layer) is applied to predict a categorical probability y_i^t. Cross-entropy is calculated between the model prediction and ground truth as the loss function.

3.3 Enhancement Description

Based on our analysis in Section 2.3, here we investigate three methods to improve the model trained on the limited data.

(a) Conversation Pre-Training Previous work shows that pre-training on dialogic data is beneficial for conversational tasks (Liu et al., 2022), thus we leverage a backbone that is particularly calibrated with utterance-paired contrastive learning (Zhou et al., 2022).

(b) Utterance Dropout One factor that affects segmentation performance is the imbalance ratio of boundary and non-boundary spans, which causes models to overfit on exposure bias. Here we adopt an utterance-level dropout strategy, where each can be excluded before feeding to the encoder by a probability p (set at 0.2).

(c) Windowed Segment Encoding To encourage the segment embedding to capture useful information from a more balanced positional distribution, we adopt a windowed encoding strategy for the

Model Type	Topic Segmentation			Topic Categorization		
	Pk Score ↓	WD Score ↓	F1 Score ↑	Precision ↑	Recall ↑	F1 Score ↑
Roberta-base Model	0.2542	0.1526	0.7486	0.7691	0.7610	0.7581
+ Utterance Dropout	0.2465 [3.0%]	0.1512 [1.0%]	0.7511 [0.3%]	0.7912 [2.9%]	0.7782 [2.3%]	0.7814 [3.1%]
+ Windowed Encoding	0.2401 [5.5%]	0.1325 [13.%]	0.7762 [3.6%]	0.7918 [2.9%]	0.7901 [3.8%]	0.7847 [3.5%]
DSE-base Model	0.2451	0.1394	0.7621	0.7735	0.7703	0.7640
+ Utterance Dropout	0.2375 [3.1%]	0.1341 [3.8%]	0.7756 [1.8%]	0.7937 [2.6%]	0.7915 [2.8%]	0.7883 [3.2%]
+ Windowed Encoding	0.2159 [11.%]	0.1252 [10.%]	0.7853 [3.0%]	0.8093 [4.6%]	0.8139 [5.6%]	0.8110 [6.1%]

Table 2: Experimental results of segmentation and categorization. Values in brackets denote relative improvement.

Model Type	Topic Segmentation			Topic Categorization		
	Pk Score ↓	WD Score ↓	F1 Score ↑	Precision ↑	Recall ↑	F1 Score ↑
Enhanced *DSE-base*	0.2159	0.1252	0.7853	0.8093	0.8139	0.8110
· w/o Punctuation	0.2284	0.1349	0.7751	0.7733	0.7864	0.7855
· w/o Speaker Role	0.2401	0.1405	0.7662	0.7743	0.7838	0.7803
· Typo Injection	0.2238	0.1286	0.7817	0.7723	0.7811	0.7837

Table 3: Robustness analysis of the enhanced model for topic segmentation and categorization.

topic categorization module. More specifically, for each segment, we randomly average utterances within a fixed window size w, which is set at 5.

4 Experimental Results & Analysis

We conduct extensive experiments to assess the model on our domain-specific application.

4.1 Experimental Data

The annotated clinical conversation data (865 dialogue samples) are used for training and evaluation. We retain the original content of dialogue samples, including fillers and punctuation marks, and build model input using sub-word tokenization (Liu et al., 2019a). We randomly select 8% samples for hold-out validation, as well as the test set.

4.2 Model Configuration

We applied and compared two language backbones *Roberta-base* (Liu et al., 2019a) and *DSE-base* (Zhou et al., 2022). *AdamW* optimizer (Loshchilov and Hutter, 2019) was used with learning rate of $1e{-}5$, weight decay of $1e{-}2$, and a linear learning rate scheduler. Model dropout (Srivastava et al., 2014) rate was set at 0.1. Utterance dropout was only applied at the training stage. Batch size and epoch number were set at 8 and 15, respectively. To avoid out-of-memory issues, we split lengthy dialogues into multiple grouped segments by concatenating adjacent topics (set at 5). Best checkpoints were selected based on validation results using averaged F1 scores. Models were implemented with PyTorch[1] and HuggingFace Transformers[2], and all experiments were run on a single Tesla A100 GPU with 40G memory.

[1]https://github.com/pytorch/pytorch
[2]https://github.com/huggingface/transformers

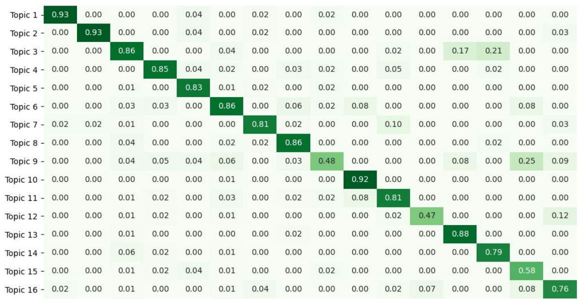

Figure 6: Confusion matrix heatmap of topic categorization predictions. Values are converted to a percentage.

4.3 Evaluation Metrics

For segmentation evaluation, we apply three standard metrics: *Pk* (Beeferman et al., 1999), WinDiff (*WD*) (Pevzner and Hearst, 2002) and macro-average F1. *Pk* and *WD* are penalty metrics (↓ denotes lower scores are better) calculated on the window-based overlap between gold and predicted segmentation. F1 is the standard harmonic mean of precision and recall, where higher scores are better (↑). For topic categorization evaluation, we report F1, precision, and recall scores. At the inference stage, we obtain topic label predictions based on gold segmentation to align with the ground truth.

4.4 Evaluation Results

Table 2 shows quantitative evaluation results on two language backbones and our proposed enhancements. For both segmentation and categorization tasks, *DSE-base* outperforms the *Roberta-base* at all metrics, demonstrating that further pre-training on dialogic data can improve the contextualized modeling of conversations. Moreover, the joint model achieves higher performance by adding utterance dropout and windowed encoding. In particular, dialogue segmentation benefits more from applying windowed encoding. Regarding topic categorization, as the normalized confusion matrix shown in Figure 6, more than half of the topics obtain acceptable topic labeling results (>0.85 accuracy). However, the scores of some topics are much lower, such as *'Medical Experience'* (topic 9) and *'Hyper/Hypo Incident'* (topic 15), we speculate that it is because these two topics are related; speakers discuss some overlapped points, and their utterances are not quite semantically distinct. This observation is also consistent with the embedding distribution shown in 4, where points of *'Medical Experience'* (topic 9) are scattered in the space. While the limited data pose a low-resource training scenario, our methods bring a reasonable performance for bootstrapping the dialogue analysis, and we suppose that the imbalanced categorization scores across topic types can be ameliorated with further corpus extension.

4.5 Robustness Analysis

We further analyze how the conversational linguistic features described in Section 2.3 affect the model's performance, by testing the well-trained and enhanced *DSE-base* model separately on three data perturbation settings: (1) Since questions often lead the topic shifting, the first way is to **remove all punctuation marks** (e.g., question marks, period, comma) at the inference stage. (2) As nurses are the main topic coordinator during the conversation, we **remove the speaker role labels** (e.g., nurse, patient, caregiver) of all utterances to assess model's dependency on such features. (3) Moreover, to simulate the inevitable typos and ASR errors in speech-to-text conversion, we randomly **inject word-level errors**, by randomly replacing or removing words upon a 15% probability of the input text. As shown in Table 3, we observe that these manipulations affect performance, especially removing speaker role labels. However, the model can still provide reasonable results, demonstrating that it utilizes semantic modeling rather than solely relying on lexical features.

5 Related Work

Topic structure analysis plays a pivotal role in dialogue understanding (Arguello and Rosé, 2006; Takanobu et al., 2018). Dialogue segmentation is similar to monologue segmentation, and aims to split a dialogue session into topically-coherent units. Various approaches originally proposed to process documents can also be applied to the dialogue domain. Due to a lack of training data, there are many unsupervised models, that exploit various linguistic features such as the word co-occurrence statistics (Hearst, 1997; Galley et al., 2003), topical distribution (Riedl and Biemann, 2012; Du et al., 2013) to measure the sentence similarity between utterances, so that topical or semantic changes can be detected. More recently, with the availability of large-scale corpora sampled from Wikipedia, by taking the section mark as the ground-truth segment boundary (Koshorek et al., 2018; Arnold et al., 2019), there has been a rapid growth in supervised approaches for monologue topic segmentation, especially neural-based approaches (Somasundaran et al., 2020). In practical use cases, supervised solutions are favored, as they present robust performance and higher learning efficiency.

Language understanding of clinical conversation has attracted a plethora of research work on in-depth analysis regarding clinician-patient communications (Byrne and P.S.Long, 1984; Černỳ, 2007; Wang et al., 2018). More recent work has included the utterances classification according to SOAP sections (Schloss and Konam, 2020), dialogue action detection (Wang et al., 2020), named entity recognition (Jeblee et al., 2019), information extraction (Rajkomar et al., 2019; Du et al., 2019), extractive (Lacson et al., 2006) and abstractive summarization (Liu et al., 2019c; Krishna et al., 2021). Though the downstream language understanding tasks are not explored in this work, dialogue segmentation and topic categorization are beneficial for those tasks by reducing the computational complexity and filtering redundant utterances.

6 Conclusion

The variety of segmentation granularity and topic definition poses challenges to domain-specific dialogue modeling and low-resource training. In this work, we investigated a joint model for dialogue segmentation and topic categorization. From our real-world case study on health management calls, we found that the nurse-to-patient conversations are shown to be topically organized, and modeling conversational features is beneficial for improving performance in practical clinical scenarios.

Limitations

The data and model used in this work are in English, thus to apply the approach to other languages, it will require training data on the specified language or using multilingual language backbones. Moreover, the segmentation granularity and topic definition vary across different domains, while our proposed framework and methods are general, when they are adapted to other conversational data, in-domain annotation is required to obtain reliable results.

Ethics and Impact Statement

We acknowledge that all of the co-authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct. The in-domain samples used in this work are fully anonymized. Participants are enrolled in the health management program with consent for the use of anonymized versions of their data for research. Our proposed framework and methodology in general do not have direct medical implications, and are intended to be used to improve the model's accuracy and robustness for downstream applications.

Acknowledgement

This research is supported by the Agency for Science, Technology and Research (A*STAR), Singapore under its Industry Alignment Pre-Positioning Fund (Grant No. H19/01/a0/023 - Diabetes Clinic of the Future). We thank Ai Ti Aw and Rosa Qi Yue So at the Institute for Infocomm Research (I^2R) for their support and assistance, and thank Siti Maryam Binte Ahmad Subaidi, and Nabilah Binte Md Johan for linguistic resource construction, and Sakinah Binte Yusof and Helen Erdt for data-related discussion. We gratefully acknowledge valuable inputs from Joan Khoo, Anne Teng Ching Ching, Winnie Soo Yi Ling, Lee Yian Chin, Angela Ng Hwee Koon, Sharon Ong Yu Bing, and Bryan Choo Peide at the Changi General Hospital, Singapore. We thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

References

Jaime Arguello and Carolyn Rosé. 2006. Topic-segmentation of dialogue. In *Proceedings of the analyzing conversations in text and speech*, pages 42–49.

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Doug Beeferman, Adam Berger, and John D. Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34(1-3):177–210.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Byrne and P.S.Long. 1984. Doctors talking to patients.

Miroslav Černỳ. 2007. On the function of speech acts in doctor-patient communication. *Linguistica online*, 6(2007):1–15.

Lan Du, Wray Buntine, and Mark Johnson. 2013. Topic segmentation with a structured topic model. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Atlanta, Georgia. Association for Computational Linguistics.

Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. 2019. Extracting symptoms and their status from clinical conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 915–925, Florence, Italy. Association for Computational Linguistics.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Marti A. Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Serena Jeblee, Faiza Khan Khattak, Noah Crampton, Muhammad Mamdani, and Frank Rudzicz. 2019. 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.

M Sue Kirkman, Morris Weinberger, Pamela B Landsman, Gregory P Samsa, E Anne Shortliffe, David L Simel, and John R Feussner. 1994. A telephone-delivered intervention for patients with niddm: effect on coronary risk factors. *Diabetes care*, 17(8):840–846.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. 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.

Ronilda C Lacson, Regina Barzilay, and William J Long. 2006. Automatic analysis of medical dialogue in the home hemodialysis domain: structure induction and summarization. *Journal of biomedical informatics*, 39(5):541–555.

Margaret L Lawson, Nini Cohen, Christine Richardson, Elaine Orrbine, and Ba' Pham. 2005. A randomized trial of regular standardized telephone contact by a diabetes nurse educator in adolescents with poor diabetes control. *Pediatric Diabetes*, 6(1):32–40.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL 2020*, pages 7871–7880. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengyuan Liu and Nancy Chen. 2021. Improving multi-party dialogue discourse parsing via domain integration. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 122–127.

Zhengyuan Liu, Pavitra Krishnaswamy, and Nancy F Chen. 2022. Domain-specific language pre-training for dialogue comprehension on clinical inquiry-answering conversations. In *Multimodal AI in healthcare: A paradigm shift in health intelligence*, pages 29–40. Springer.

Zhengyuan Liu, Hazel Lim, Nur Farah Ain Binte Suhaimi, Shao Chuen Tong, Sharon Ong, Angela Ng, Sheldon Lee, Michael R Macdonald, Savitha Ramasamy, Pavitra Krishnaswamy, et al. 2019b. Fast prototyping a dialogue comprehension system for nurse-patient conversations on symptom monitoring. In *Proceedings of NAACL-HLT*, pages 24–31.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F Chen. 2019c. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821. IEEE.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *The International Conference on Learning Representations (ICLR 2019)*.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499.

Chaudhary Muhammad Junaid Nazar, Micheal Mauton Bojerenu, Muhammad Safdar, and Jibran Marwat. 2016. Effectiveness of diabetes education and awareness of diabetes mellitus in combating diabetes in the united kigdom; a literature review. *Journal of Nephropharmacology*, 5(2):110.

Lev Pevzner and Marti A Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

John D Piette, Morris Weinberger, Frederic B Kraemer, and Stephen J McPhee. 2001. Impact of automated calls with nurse follow-up on diabetes treatment outcomes in a department of veterans affairs health care system: a randomized controlled trial. *Diabetes care*, 24(2):202–208.

Juan C Quiroz, Liliana Laranjo, Ahmet Baki Kocaballi, Shlomo Berkovsky, Dana Rezazadegan, and Enrico Coiera. 2019. Challenges of developing a digital scribe to reduce clinical documentation burden. *NPJ digital medicine*, 2(1):114.

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.

Martin Riedl and Chris Biemann. 2012. TopicTiling: A text segmentation algorithm based on LDA. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42, Jeju Island, Korea. Association for Computational Linguistics.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

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.

Swapna Somasundaran et al. 2020. Two-level transformer and auxiliary coherence modeling for improved text segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7797–7804.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Ryuichi Takanobu, Minlie Huang, Zhongzhou Zhao, Feng-Lin Li, Haiqing Chen, Xiaoyan Zhu, Liqiang Nie, et al. 2018. A weakly supervised method for topic segmentation and labeling in goal-oriented dialogues via reinforcement learning. In *IJCAI*, pages 4403–4410.

C Barr Taylor, Nancy Houston Miller, Kelly R Reilly, George Greenwald, Darby Cunning, Allison Deeter, and Liana Abascal. 2003. Evaluation of a nurse-care management system to improve outcomes in patients with complicated diabetes. *Diabetes care*, 26(4):1058–1063.

Marieke M van Buchem, Hileen Boosman, Martijn P Bauer, Ilse MJ Kant, Simone A Cammel, and Ewout W Steyerberg. 2021. The digital scribe in clinical practice: a scoping review and research agenda. *NPJ digital medicine*, 4(1):57.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Chow Wai Leng, Jiang Jundong, Cho Li Wei, Foo Joo Pin, Fock Kwong Ming, and Richard Chen. 2014. Telehealth for improved glycaemic control in patients with poorly controlled diabetes after acute hospitalization–a preliminary study in singapore. *Journal of Telemedicine and Telecare*, 20(6):317–323.

Nan Wang, Yan Song, and Fei Xia. 2018. Constructing a Chinese medical conversation corpus annotated with conversational structures and actions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nan Wang, Yan Song, and Fei Xia. 2020. Studying challenges in medical conversation with structured annotation. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 12–21, Online. Association for Computational Linguistics.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv preprint arXiv:1908.08167*.

Linzi Xing and Giuseppe Carenini. 2021. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 167–177.

Zhihan Zhou, Dejiao Zhang, Wei Xiao, Nicholas Dingwall, Xiaofei Ma, Andrew Arnold, and Bing Xiang. 2022. Learning dialogue representations from consecutive utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 754–768.