DASH: Dialogue-Aware Similarity and Handshake Recognition for Topic Segmentation in Public-Channel Conversations

Sun Sijin[1,3], **Liangbin Zhao**[1,*], **Ming Deng**[2], **Xiuju Fu**[1]

[1]Institute of High Performance Computing, Agency for Science Technology and Research (A*STAR IHPC)
[2]Shanghai University
[3]National University of Singapore

Abstract

Dialogue Topic Segmentation (DTS) is crucial for understanding task-oriented public-channel communications, such as maritime VHF dialogues, which feature informal speech and implicit transitions. To address the limitations of traditional methods, we propose DASH-DTS, a novel LLM-based framework. Its core contributions are: (1) topic shift detection via dialogue handshake recognition; (2) contextual enhancement through similarity-guided example selection; and (3) the generation of selective positive and negative samples to improve model discrimination and robustness. Additionally, we release VHF-Dial, the first public dataset of real-world maritime VHF communications, to advance research in this domain. DASH-DTS provides interpretable reasoning and confidence scores for each segment. Experimental results demonstrate that our framework achieves several sota segmentation trusted accuracy on both VHF-Dial and standard benchmarks, establishing a strong foundation for stable monitoring and decision support in operational dialogues.

Code — https://github.com/StanleySun233/dash-dts

Datasets — https://github.com/StanleySun233/dash-dts/dataset/vhf.json

Demos — https://github.com/StanleySun233/dash-dts/blob/main/demo/demo.gif

Introduction

Understanding topic structure in dialogue is essential for a wide range of downstream tasks, including summarization (Han et al. 2024), event detection (Davies et al. 2009), dialogue planning, and regulatory decision support. At the core of such understanding lies the task of Dialogue Topic Segmentation (DTS)—the process of identifying boundaries where the conversation transitions from one topic to another (Zhang and Zhou 2019). While considerable progress has been made in open-domain and customer-service dialogue segmentation (Artemiev et al. 2024), public-channel conversations—such as those in maritime Very High Frequency (VHF) radio, air traffic control, or emergency dispatch networks—remain underexplored despite their societal and operational importance.

Copyright © 2026, Trustworthy Agentic AI Workshop@ Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Public-channel dialogues differ fundamentally from traditional open-domain (Feng, Feng, and Qin 2021) or daily conversations. They are characterized by extremely short, fragmented utterances, dynamically shifting speaker roles, and high-stakes operational intent. In such settings, topic transitions are often implicit, driven by speaker intent and interactional coordination rather than explicit lexical cues or discourse markers. These structural and pragmatic characteristics challenge the assumptions of existing DTS models, which typically rely on surface-level continuity, utterance embedding similarity, or turn-level encoders—while largely overlooking speaker-driven signals that indicate topical shifts.

These challenges are not merely academic. In industrial maritime regulation, for instance, analysts are often tasked with reviewing hours of VHF voice communications to identify key events such as near-miss collisions or miscommunications. These public-channel interactions, while rich in safety-critical information, are rarely structured or archived due to their volume and fragmented nature. In practice, identifying task shifts, escalation cues, or behavioral patterns still heavily relies on manual transcription and expert judgment. This bottleneck not only limits the coverage of maritime oversight systems but also results in missed opportunities to learn from near-miss cases—which are far more frequent and informative than actual accidents. A robust DTS system tailored for public-channel dialogues (Gao et al. 2023) would drastically reduce this cost by offering structure-aware segmentation that supports downstream analytics, compliance auditing, and real-time alerting.

Existing DTS approaches primarily rely on surface-level lexical transitions or embedding similarity between utterances, often using cosine similarity to retrieve relevant examples. While such methods work reasonably well in structured or open-domain settings, they struggle in task-oriented, public-channel dialogues—where utterances are terse, speaker roles shift dynamically, and topic shifts are often implicit. In particular, these conversations frequently contain short, functional "dialogue handshakes"—such as "Star Alpha calling port control"—which act as subtle signals of upcoming topical change. However, such interactional patterns are rarely modeled in existing systems. Meanwhile, in-context learning (ICL) with large language models (Rubin, Herzig, and Berant 2021) offers a promising alterna-

tive for few-shot topic segmentation, but selecting semantically appropriate exemplars remains challenging in domain-specific, sparse-data environments like VHF communication.

To address these challenges, we propose DASH-DTS (Dialogue-Aware Similarity and Handshake recognition for Dialogue Topic Segmentation), a structure-aware framework for segmenting topics in public-channel conversations. DASH-DTS incorporates three core components: (1) a handshake recognition module that identifies short interactional cues marking the onset of new topical segments; (2) a dialogue similarity-guided in-context learning strategy, which retrieves semantically relevant exemplars to enhance segmentation in sparse-data conditions; and (3) a context-aware labeling mechanism that utilizes surrounding discourse to produce more coherent and accurate topic annotations. Besides, to support trustworthy deployment in the follow-up applications, our framework additionally generates segment-level justifications and confidence scores, allowing users to assess the reliability of predicted topic boundaries.

In addition, to support research and benchmarking in this domain, we construct and release the first publicly available DTS dataset for maritime VHF communications. This dataset captures the unique characteristics of public-channel dialogues—such as brief utterances, implicit transitions, and dynamic speaker roles—and serves as a valuable resource for evaluating topic segmentation methods in real-world, safety-critical communication environments.

The main contributions of this work are as follows:

- Propose a handshake recognition mechanism that captures speaker interaction cues to identify topic boundaries in public-channel dialogue, addressing structural challenges in dialogue topic segmentation.
- Introduce a similarity-guided in-context learning strategy that selects semantically relevant exemplars to enhance segmentation performance for DTS in sparse and domain-specific settings.
- Introduce an interpretable and trustworthy output mechanism that generates segment-level justifications and confidence scores, enabling downstream applications to assess the reliability of topic boundaries.
- Construct and release the first publicly available dataset for dialogue topic segmentation in the VHF public-channel communication domain, termed **VHF-Dial**, providing a benchmark for real-world application.

Related Work

Early Methods Traditional approaches to Dialogue Topic Segmentation (DTS) primarily relied on lexical cohesion and surface-level continuity. Techniques like TextTiling (Hearst 1997) segmented text by identifying lexical valleys in cosine similarity between adjacent blocks. While effective for monologic texts, these methods struggle with dialogues due to their fragmented utterances, speaker shifts, and informal language (Misra and Jose 2009). For task-oriented dialogues, where topics shift implicitly without explicit lexical cues, such approaches exhibit significant limitations in robustness (Song et al. 2016).

Sequence Modeling Methods To capture contextual dependencies, later works adopted sequence modeling architectures. SimCSE (Gao, Yao, and Chen 2021) models leveraged utterance representations to predict topic boundaries sequentially, improving coherence modeling. However, these methods often require extensive labeled data and fail to generalize across diverse conversational domains. Their reliance on local context also overlooks global dialogue structure.

Pre-trained Models The advent of PLMs like BERT revolutionized DTS by enabling deeper semantic understanding. Baseline evaluate of DTS (Feng et al. 2021) fine-tuned BERT for utterance-pair coherence scoring, using topical relevance as a segmentation signal. Similarly, unsupervised frameworks like Topic-aware Utterance Representation (Artemiev et al. 2024) leveraged pseudo-segmentation tasks on unlabeled dialogues. While PLMs improved accuracy, they remain constrained by domain transferability issues and high computational costs (Lee et al. 2025), particularly in low-resource, noisy environments (e.g., emergency dispatch systems).

Large Language Models Recent efforts integrate LLMs for few-shot segmentation. DEF-DTS (Lee et al. 2025) employs multi-step deductive reasoning via structured prompting. S3-DST (Das et al. 2024) uses self-supervised learning to mitigate data scarcity. Despite promising results, LLM-based methods still face challenges on limited interpretability of topic transitions, and neglect of structural dynamics (Fan and Jiang 2023).

Unlike prior work, DASH-DTS uniquely integrates dialogue handshake recognition to detect structural cues and semantic pivots for topic shift detection. We avoid LLMs' prompt engineering limitations by leveraging similarity-based in-context learning with dynamically retrieved examples. Our framework is specifically designed for noisy, high-stakes dialogue, where implicit transitions and dynamic speaker roles invalidate assumptions of existing methods (Lee et al. 2025). The release of a curated maritime DTS dataset further addresses chronic data scarcity in this domain.

Methodology

Dialogue-level Similarity

To effectively leverage in-context learning (ICL) in the domain of public-channel dialogue topic segmentation, it is crucial to select semantically relevant examples that can guide the model towards more accurate segmentation. Traditional ICL approaches often rely on cosine similarity or surface-level lexical matching, which may not be sufficient in scenarios with sparse and highly specialized data like VHF maritime communications. Therefore, we introduce a novel dialogue similarity-guided in-context learning strategy to enhance the performance called Dialogue-Level Similarity ICL Samples in Figure 1.

Semantic Similarity Calculation The first step in our approach is to compute the semantic similarity (Lavi et al.

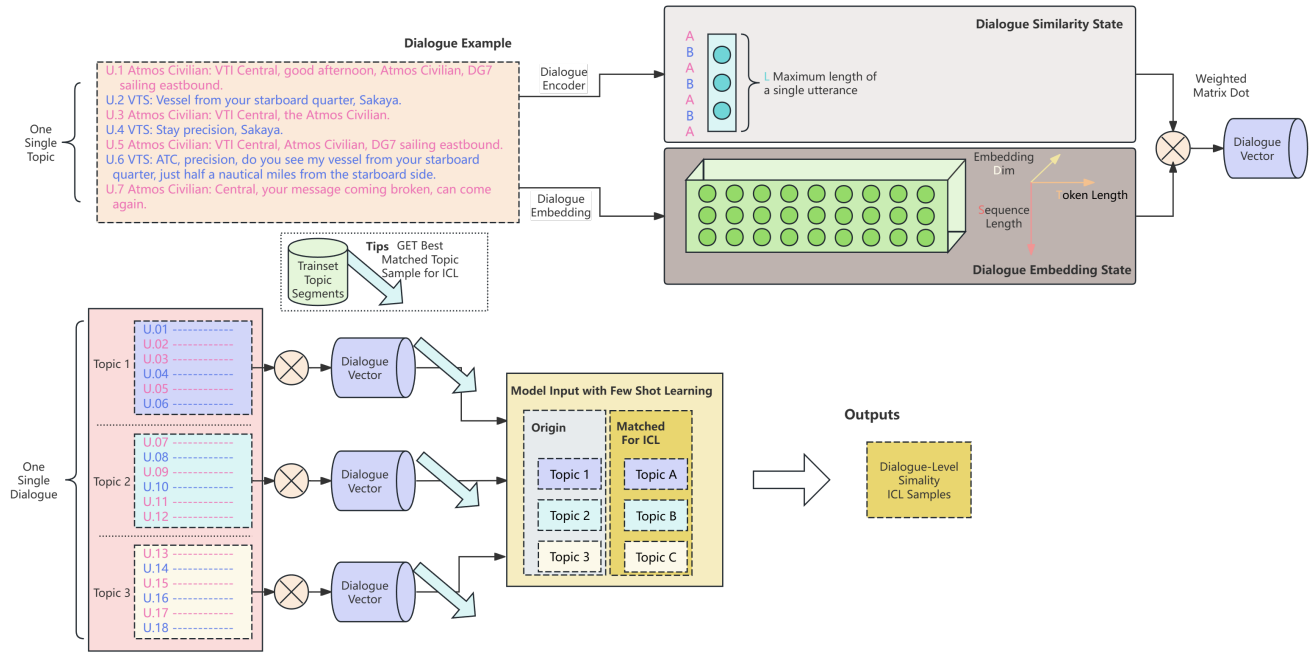

Figure 1: Dynamically select the most semantically relevant exemplars for each input conversation, enhancing the accuracy of topic segmentation in contextual learning for large models.

2021) between the query dialogue and the available exemplars. We use a pre-trained language model, such as BERT or RoBERTa, fine-tuned on a large dialogue dataset, to generate embeddings for each utterance. These embeddings capture the contextual and semantic information of the utterances (Abro et al. 2022), making them more suitable for comparing the underlying meanings rather than just the surface text.

Given a query dialogue D_q and a set of exemplars $E = \{E_1, E_2, \ldots, E_n\}$, we compute the embedding for each utterance in D_q and E_i. Let $U_{q,j}$ be the j-th utterance in D_q and $U_{i,k}$ be the k-th utterance in E_i. The embeddings are denoted as $Emb(U_{q,j})$ and $Emb(U_{i,k})$.

We then calculate the pairwise cosine similarity between the utterances in the query dialogue and the exemplars:

$$\text{sim}(U_{q,j}, U_{i,k}) = \frac{Emb(U_{q,j}) \cdot Emb(U_{i,k})}{\|Emb(U_{q,j})\| \|Emb(U_{i,k})\|} \quad (1)$$

Next, we aggregate the similarities at the dialogue level. One way to do this is by averaging the cosine similarities of all utterance pairs:

$$\text{sim}(D_q, E_i) = \frac{1}{|D_q| \times |E_i|} \sum_{j=1}^{|D_q|} \sum_{k=1}^{|E_i|} \text{sim}(U_{q,j}, U_{i,k}) \quad (2)$$

Alternatively, we can use a weighted aggregation method, where the weights are determined by the importance of each utterance in the dialogue. For example, we can assign higher weights to utterances that are more likely to indicate a topic change, such as those containing dialogue handshakes.

Exemplar Selection Once we have computed the semantic similarity between the query dialogue and the exemplars, we select the most relevant exemplars to include in the context. We rank the exemplars based on their similarity scores and choose the top m exemplars, where m is a hyperparameter that can be tuned based on the specific application and available computational resources.

$$E_{\text{selected}} = \arg\max_{E_i \in E} \text{sim}(D_q, E_i) \quad (3)$$

These selected exemplars are then concatenated with the query dialogue to form the input for the in-context learning process. The model uses these exemplars to learn the patterns and structures that are indicative of topic transitions in the given dialogue.

In-Context Learning with Selected Exemplars With the selected exemplars, we perform in-context learning using a large language model. The concatenated input consists of the selected exemplars followed by the query dialogue. The model is prompted to predict the topic boundaries in the query dialogue, guided by the examples provided in the context.

$$\text{Input} = [E_{\text{selected}}; D_q] \quad (4)$$

The model outputs a sequence of labels indicating the topic boundaries in D_q. By providing semantically relevant exemplars, the model can better generalize to the query dialogue, even in the presence of sparse and specialized data.

Handshake Statement Tag

In public-channel dialogues, particularly maritime VHF communications, "handshake" statements function as critical interactional markers that demarcate topical boundaries. These statements typically comprise concise, functionally-oriented utterances signaling conversational focus shifts or topic transitions (Konigari et al. 2021). Representative examples include phrases such as "Star Alpha calling port control" or "Delta Echo, this is Bravo Hotel," which serve as subtle cues facilitating topic segmentation. Accurate identification of these statements is essential for robust topic segmentation, especially in high-stakes communication environments where implicit topic transitions and dynamic speaker roles introduce additional complexity.

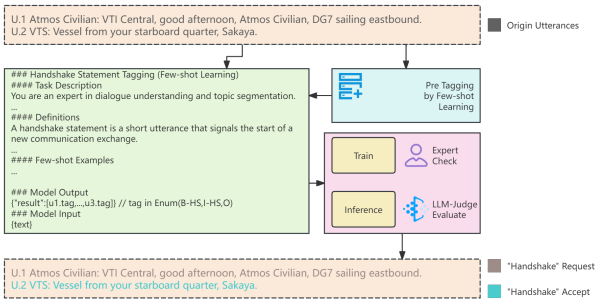

Figure 2: Workflow of the Handshake Statement Tagging Component Based on Few-shot LLM Learning.

Given a dialogue sequence $D = \{U_1, U_2, \ldots, U_n\}$ where each U_i represents an utterance decomposed into tokens, the handshake identification task is formulated as a token-level sequence labeling problem. Each token is assigned one of three labels: (1)HS-BEG: Beginning of a handshake statement, (2) HS-END: End of a handshake statement, (3) O: Inside or indeterminate portions of dialogues.

To systematically identify handshake statements, the Handshake (HS) Agent is introduced as a reusable component within the DASH-DTS framework, as illustrated in Figure 2. The HS agent leverages the capabilities of large language models to recognize these interactional cues (Brysbaert and Lahousse 2022). Rather than relying solely on deterministic pattern matching, the HS agent is designed to produce structured, interpretable outputs that facilitate both automated decision-making and human verification.

The HS agent processes dialogue utterances and generates structured predictions for each token. Each prediction is formulated as a principled triplet in Equation (5).

$$P_i = (l_i, s_i, r_i) \tag{5}$$

where $l_i \in \{\text{HS-BEG, HS-END, O}\}$ denotes the predicted label, $s_i \in [0, 1]$ represents the trustworthiness score indicating prediction confidence, and r_i encapsulates the reasoning justification underlying the classification decision.

This structured restriction ensures that predictions are not only accurate but also transparent and auditable. The reasoning component r_i documents the linguistic and contextual evidence supporting each classification, thereby enabling domain experts to assess prediction reliability and identify potential failure modes.

For each dialogue, the agent is prompted with contextual specifications and exemplars of typical handshake statement patterns. The LLM component generates label predictions for token sequences, accompanied by explicit trustworthiness assessments and reasoning justifications. Post-processing constraints are applied to ensure label coherence: each HS-BEG label must be paired with a corresponding HS-END label, maintaining proper demarcation of handshake statement boundaries. This step enforces structural consistency and eliminates malformed predictions. The final output comprises:

- Token-level label sequence: Predictions of handshake statement boundaries (onset and termination points)
- Trustworthiness scores: Confidence metrics reflecting prediction reliability for each token
- Reasoning chains: Interpretable justifications documenting the evidence supporting each classification decision

Proposed handshake statement predictions, enriched with trustworthiness scores and reasoning chains, are integrated into the DASH-DTS framework to identify structural cues signaling the onset of new topical segments. This principled integration enhances both the robustness and interpretability of the overall topic segmentation process, particularly in scenarios characterized by implicit transitions and dynamic speaker configurations.

Content-Aware Topic Generation

For enhancing the in-context learning (ICL) process and improve the accuracy of topic segmentation, we propose a content-aware topic generation mechanism. This mechanism leverages the full context of the dialogue to generate positive and negative samples, which are then used to guide the model in identifying potential topic boundaries. The key idea is to create synthetic examples that help the model understand the characteristics of both topic transitions and non-transitions.

Given a dialogue segment with contextual window $W = \{U_{i-m}, \ldots, U_i | U_{i+1}, \ldots, U_{i+n}\}$ where U_i and U_{i+1} form a potential segmentation point, we formulate the sample generation task as a structured prediction problem. Let $\mathcal{S} = (D_p, D_n, \mathcal{E})$ denote the generated sample triplet, where $D_p = \{s_1^p, s_2^p, \ldots, s_7^p\}$ is a positive dialogue sample with ground truth label $y_p = 1$, $D_n = \{s_1^n, s_2^n, \ldots, s_7^n\}$ is a negative dialogue sample with ground truth label $y_n = 0$, and $\mathcal{E} = \{\xi_p, \xi_n, \mathcal{C}\}$ represents explainability artifacts including reasoning chains and confidence scores. The core challenge is to automatically generate balanced, diverse samples while maintaining semantic coherence and providing verifiable reasoning traces that enable human inspection.

Contextual Input Given the contextual window W, the L first performs deep semantic analysis of the dialogue through a structured analysis. This analysis stage instructs the model to extract thematic elements and discourse topics

Table 1: Experimental Results on Our Conducted Datasets of public VHF channel dialog

Model	Reference	DialSeg711		Doc2Dial		VHF-Dial	
		P_k	W_d	P_k	W_d	P_k	W_d
Text Tiling	(Hearst 1997)	40.4	44.6	52.0	57.4	54.3	61.7
LLM	gemini-2.5-flash	36.5	67.9	46.4	53.5	38.6	78.6
DyDTS	(Lv et al. 2025)	24.7	**27.6**	39.9	44.0	38.2	39.7
UPS	(Yang et al. 2025)	–	–	35.1	36.5	27.4	34.5
SumSeg	(Artemiev et al. 2024)	47.7	48.3	–	–	32.7	35.1
CSM	(Gao et al. 2023)	26.8	28.2	45.2	47.3	27.7	31.6
BERT	(Devlin et al. 2019)	39.3	41.2	53.7	55.3	44.9	49.1
ours		**20.7**	34.3	**33.9**	36.6	**21.9**	**33.9**

from each utterance, identify lexical and pragmatic markers such as discourse particles and topic shift indicators, characterize speaker roles and dialogue coherence patterns, and detect domain-specific terminologies and contextual dependencies. The output $\mathcal{A}$ is a structured analysis document that serves as grounding for subsequent stages, ensuring that sample generation is semantically faithful to the original dialogue and reflects the actual discourse structure rather than superficial patterns.

$$\mathcal{A} = \mathrm{L}_\theta(\text{analyze}(W) \mid \tau_{\text{analysis}}) \tag{6}$$

Then, building on the analysis $\mathcal{A}$, we perform dual-mode generation structure ($\mathcal{GC}$) to create positive and negative samples through contrastive synthesis.

$$(D_p, D_n) = \mathrm{LLM}_\theta(\mathcal{GC}(W, \mathcal{A}) \mid \tau_{\text{synthesis}}) \tag{7}$$

Posi/Nega-tive Generation For positive samples, the synthesis prompt instructs the LLM to generate utterances that maintain thematic continuity in the previous segment (positions 1-3), create a pivot utterance at position 4 that exhibits explicit topic shift markers including transitional phrases such as "By the way" or "Speaking of", sudden perspective changes, or domain switches, and extend with utterances in the next segment (positions 5-7) that cohere around the new topic. The generated dialogue should present a clear and unambiguous boundary that enables the model to learn definitive topic transition signals.

For negative samples, the synthesis prompt requires the LLM to maintain strong thematic and pragmatic continuity across all seven positions, ensure that the pivot utterance advances, elaborates on, or clarifies the previous topic rather than shifting away, and employ within-topic discourse patterns such as agreement, clarification, and detail addition rather than boundary markers. These negative examples are particularly important as they teach the model to distinguish between genuine topic boundaries and discourse continuations that might superficially resemble transitions.

Crucially, both synthesis operations preserve the stylistic, register, and speaker role patterns observed in $\mathcal{A}$, ensuring that D_p and D_n remain authentic exemplars of real dialogue phenomena rather than synthetic artifacts that might introduce spurious patterns.

The complete output is structured to provide comprehensive traceability:

$$\mathcal{S} = \begin{cases} \text{positive} : \{D_p, y_p = 1, \mathcal{C}_p, \xi_p\} \\ \text{negative} : \{D_n, y_n = 0, \mathcal{C}_n, \xi_n\} \end{cases} \tag{8}$$

Trustworthy CoT This multi-layered output format offers several critical advantages for trustworthy AI systems. The reasoning traces ξ_p, ξ_n provide human-readable justifications that enable auditors to verify that samples reflect intended classification logic rather than spurious correlations learned from training data. The confidence scores $\mathcal{C}_p, \mathcal{C}_n$ enable downstream systems to weight samples appropriately during training, downweighting low-confidence exemplars and preventing overfitting to unreliable signals. When the topic segmentation model makes errors during deployment, practitioners can inspect the explanations to diagnose whether errors stem from inadequate sample generation or model learning failures, supporting systematic debugging and improvement. By explicitly recording the reasoning chains and confidence scores, our method creates an auditable paper trail suitable for regulatory compliance and scientific reproducibility, addressing key concerns in trustworthy machine learning.

Segment Reliability and Explanation

To enhance the trustworthiness and usability of topic segmentation in high-stakes scenarios, DASH-DTS outputs a confidence score for each predicted segment, reflecting the model's certainty in its boundary decisions. Additionally, a brief natural language explanation is generated to justify each segmentation point, allowing users to better interpret and assess the system's output. This design supports human-in-the-loop workflows and enables more informed downstream decision-making.

Prompt example. To support interpretability and trust calibration, we prompt the LLM to provide an explanatory rationale and a confidence score for each predicted topic segment:

```
For each predicted topic segment,
provide:
```
(1) A brief explanation of the topical focus or transition, and whether it constitutes a complete dialogue task.

(2) A confidence score between 0 and 1 indicating the model's certainty.

`Example:`

Explanation: This segment involves a vessel initiating contact with port control to request entry clearance. The conversation forms a self-contained exchange where the intent, response, and acknowledgment are completed, indicating a coherent dialogue unit with a clear operational focus.

Confidence: 0.91

Experiment

Comparison Study

The metrics used for evaluation are P_k(Hearst 1997) in Equation (9) and W_d (Windo$W_d$iff)(Pevzner and Hearst 2002) in Equation (10), which are standard measures for assessing the quality of topic segmentation.

$$P_k = \frac{\sum_{i=1}^{N-k} \delta_P(i, i+k) \oplus \delta_R(i, i+k)}{N-k} \tag{9}$$

where k is the window size, δ_R and δ_P are indicator functions for a boundary in the reference and prediction, respectively, and $\oplus$ denotes the XOR operation that detects a disagreement.

$$W_d(ref, hyp) = \frac{1}{N-k} \sum_{i=1}^{N-k} (|b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})| > 0) \tag{10}$$

where $b(i, j)$ represents the number of boundaries between positions i and j in the text, N represents the number of sentences in the text, and k is the window size parameter. This metric compares the number of reference segmentation boundaries with the number of hypothesized boundaries within each sliding window, penalizing the algorithm when these counts differ.

The experimental results, as presented in Table 1, provide a comprehensive comparison of our proposed DASH-DTS method with several state-of-the-art baselines on three public datasets: **DialSeg711**, **Doc2Dial**, and proposed **VHF-Dial**.

Our proposed VHF-Dial dataset demonstrates distinct challenges and opportunities for dialogue segmentation models. As evidenced in Table 1, our method significantly outperforms all baselines on this dataset, achieving a P_k of 21.9 and W_d of 33.9, which represent notable improvements over the best-performing baseline, UPS, accounting for P_k at 27.4. This suggests that our approach is particularly well-suited to the unique characteristics of VHF (Very High Frequency) channel dialogues, which may involve shorter turns, domain-specific jargon, or structured communication protocols.

On the DialSeg711 dataset, our method achieves a P_k score of 20.7, which is 4 percent lower than most baselines of DyDTS. This indicates that while our method captures some topic segments correctly, it may have a higher rate of false positives or over-segmentation. However, our method exhibits room for improvement on the Doc2Dial dataset, which may be attributed to the inherent characteristics of the dataset. This discrepancy suggests that dataset-specific features, such as document structure or dialogue complexity, may influence segmentation effectiveness. Future work could explore adaptive mechanisms to better handle such variations.

Ablation Study

To evaluate the contribution of each component to the performance of DASH-DTS, we conducted ablation experiments on the **VHF-Dial** in Table 2. In No.1, we included both the Handshake and Dialogue Similarity components but excluded the Topic Generation component. This configuration resulted in a P_k score of 26.5 and a W_d score of 39.6, indicating that while these two components alone can achieve reasonable performance, the absence of Topic Generation slightly degrades the overall segmentation quality. In No.2, we included only the Dialogue Similarity and Topic Generation components, excluding the Handshake component, which yielded a P_k score of 27.1 and a W_d score of 39.7. This suggests that Dialogue Similarity and Topic Generation are crucial for segmenting the dialogue, but the lack of Handshake recognition leads to a slight decrease in performance.

Table 2: Ablation Experiment on VHF-Dial

No.	Component			P_k	W_d
	Handshake	Dialogue Similarity	Topic Generation		
1	√	√		26.5	39.6
2		√	√	27.1	39.7
3	√		√	24.3	34.7
Ours	√	√	√	**21.9**	**33.9**

In No.3, we included the Handshake and Topic Generation components but excluded the Dialogue Similarity component, resulting in a P_k score of 24.3 and a W_d score of 34.7. The removal of Dialogue Similarity led to a higher W_d score, highlighting its importance in maintaining the alignment of segment boundaries. Finally, the full model (Ours) with all three components achieved the best performance, with a P_k score of 21.9 and a W_d score of 33.9. This indicates that the combination of all three components is necessary to achieve optimal topic segmentation, as they collectively contribute to detecting structural cues, enhancing in-context learning, and ensuring semantic coherence.

Discussion

Experimental results provide valuable insights into the performance and contributions of our proposed DASH-DTS method. Here, we discuss three key points based on the findings from the comparison and ablation studies.

1) The ablation study highlights the importance of the handshake recognition mechanism in detecting structural cues and identifying topic boundaries. When the Handshake component is included, the model achieves a lower W_d score, indicating better alignment with the true segment boundaries. This is particularly evident in the DialSeg711

and VHF-Dial, where the absence of Handshake recognition leads to higher W_d scores. The Handshake mechanism effectively captures speaker interaction cues, which are crucial for accurately segmenting topics in public-channel dialogues. This contribution addresses the structural challenges in dialogue topic segmentation, making the model more robust in real-world communication settings.

2) The inclusion of the Dialogue Similarity component significantly enhances the model's performance by selecting semantically relevant exemplars. This is demonstrated in the ablation experiment (No. 3), where the removal of Dialogue Similarity results in a higher W_d score, indicating a decrease in the alignment of segment boundaries. The similarity-guided in-context learning strategy is particularly beneficial in sparse and domain-specific settings, where the availability of relevant training data is limited. By leveraging semantically similar examples, the model can generalize better and achieve more accurate segmentation, thereby addressing the challenges of data sparsity and domain specificity.

3) The context-aware topic labeling module, which incorporates surrounding discourse, plays a critical role in generating more accurate and coherent topic annotations. Ours model achieves the best performance, with a P_k score of 21.9 and a W_d score of 33.9, indicating that all three components (Handshake, Dialogue Similarity, and Topic Generation) are necessary for optimal performance. The Topic Generation component ensures that the segments are semantically coherent, enhancing the overall quality of the topic segmentation. This is particularly important in public-channel conversations, where the context and discourse structure are complex and dynamic. The context-aware approach not only improves the accuracy of the topic labels but also makes the model more applicable to real-world scenarios, such as maritime radio and air traffic control communications.

Conclusion

This paper presents DASH-DTS, a structure-aware framework for Dialogue Topic Segmentation in public-channel conversations. It integrates handshake recognition, similarity-guided in-context learning, and context-aware topic labeling to effectively capture both structural and semantic cues in fragmented, high-stakes dialogues. In addition, to facilitate human-in-the-loop oversight in safety-critical settings, DASH-DTS outputs each predicted topic segment with an accompanying explanation and a confidence score, providing interpretability and trustworthiness estimation to support downstream decision-making.

Experiments on DialSeg711, Doc2Dial, and our newly collected public VHF channel dialog dataset **VHF-Dial** demonstrate that DASH-DTS achieves consistent improvements over state-of-the-art baselines, particularly in sparse-data and operational settings such as maritime VHF communication. Ablation studies validate the contribution of each module to the overall segmentation quality.

To support future research and real-world deployment, we release **VHF-Dial**, the first publicly available DTS dataset tailored to public-channel communication. Looking forward, we envision expanding this framework to other domains, such as air traffic control and emergency dispatch systems, where accurate, automated understanding of dialogue structure can directly improve situational awareness, reduce manual workload, and enable intelligent decision support in mission-critical environments.

Limitation

While DASH-DTS demonstrates significant improvements in topic segmentation for public-channel dialogues, several limitations remain. The framework's performance can be constrained by the sparsity and domain-specific nature of the data, particularly in specialized domains like maritime VHF communications. Additionally, the handshake recognition mechanism, though effective, may struggle with more subtle or complex interactional cues. The computational demands of the in-context learning component also pose a challenge for real-time and resource-constrained environments. Finally, enhancing the interpretability and explainability of the model's decisions is crucial for building trust and facilitating better integration with human operators in high-stakes communication settings. Future work should focus on addressing these limitations to further improve the robustness, efficiency, and practical applicability of DASH-DTS.

Acknowledgments

This research was funded by the Maritime Artificial Intelligence (AI) Research Project, supported under funding grant number SMI-2025-MTP-02 by the Singapore Maritime Institute (SMI).

References

Abro, W. A.; Aicher, A.; Rach, N.; Ultes, S.; Minker, W.; and Qi, G. 2022. Natural language understanding for argumentative dialogue systems in the opinion building domain. *Knowledge-Based Systems*, 242: 108318.

Artemiev, A.; Parinov, D.; Grishanov, A.; Borisov, I.; Vasilev, A.; Muravetskii, D.; Rezvykh, A.; Goncharov, A.; and Savchenko, A. 2024. Leveraging summarization for unsupervised dialogue topic segmentation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 4697–4704.

Brysbaert, J.; and Lahousse, K. 2022. Marking contrastive topics in a topic shift context: Contrastive adverbs versus emphatic pronouns. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (31).

Das, S. S. S.; Shah, C.; Wan, M.; Neville, J.; Yang, L.; Andersen, R.; Buscher, G.; and Safavi, T. 2024. S3-DST: Structured Open-Domain Dialogue Segmentation and State Tracking in the Era of LLMs. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 14996–15014. Bangkok, Thailand: Association for Computational Linguistics.

Davies, S.; McCallie, E.; Simonsson, E.; Lehr, J. L.; and Duensing, S. 2009. Discussing dialogue: Perspectives on the

value of science dialogue events that do not inform policy. *Public understanding of science*, 18(3): 338–353.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

Fan, Y.; and Jiang, F. 2023. Uncovering the Potential of ChatGPT for Discourse Analysis in Dialogue: An Empirical Study. In *International Conference on Language Resources and Evaluation*.

Feng, S.; Patel, S. S.; Wan, H.; and Joshi, S. 2021. Multi-Doc2Dial: Modeling Dialogues Grounded in Multiple Documents. *ArXiv*, abs/2109.12595.

Feng, S.; Wan, H.; Gunasekara, C.; Patel, S.; Joshi, S.; and Lastras, L. 2020. doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8118–8128. Online: Association for Computational Linguistics.

Feng, X.; Feng, X.; and Qin, B. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

Gao, H.; Wang, R.; Lin, T.-E.; Wu, Y.; Yang, M.; Huang, F.; and Li, Y. 2023. Unsupervised dialogue topic segmentation with topic-aware utterance representation. *arXiv preprint arXiv:2305.02747*.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894–6910. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Han, Q.; Yang, Z.; Lin, H.; and Qin, T. 2024. Let topic flow: A unified topic-guided segment-wise dialogue summarization framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 2021–2032.

Hearst, M. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Association for Computational Linguistics*.

Konigari, R.; Ramola, S.; Alluri, V. V.; and Shrivastava, M. 2021. Topic shift detection for mixed initiative response. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 161–166.

Lavi, O.; Rabinovich, E.; Shlomov, S.; Boaz, D.; Ronen, I.; and Anaby-Tavor, A. 2021. We've had this conversation before: A Novel Approach to Measuring Dialog Similarity. *arXiv preprint arXiv:2110.05780*.

Lee, S.; Yoo, Y.; Jung, M.; and Song, M. 2025. Def-DTS: Deductive Reasoning for Open-domain Dialogue Topic Segmentation. *ArXiv*, abs/2505.21033.

Lv, Y.; Tao, W.; Dai, Q.; Chen, Z.; Lu, Q.; and Jiang, N. 2025. Dynamic Topic Segmentation in Dialogues: Enhancing Boundaries with Topic-Aware Propagation. *Companion Proceedings of the ACM on Web Conference 2025*.

Misra, H.; and Jose, J. M. 2009. Text Segmentation via Topic Modeling: An Analytical Study ABSTRACT. *DBLP*.

Pevzner, L.; and Hearst, M. A. 2002. A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 28: 19–36.

Rubin, O.; Herzig, J.; and Berant, J. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Song, Y.; Mou, L.; Yan, R.; Yi, L.; Zhu, Z.; Hu, X.; and Zhang, M. 2016. Dialogue Session Segmentation by Embedding-Enhanced TextTiling. *Annual Conference of the International Speech Communication Association*.

Xie, H.; Liu, Z.; Xiong, C.; Liu, Z.; and Copestake, A. 2021. TIAGE: A Benchmark for Topic-Shift Aware Dialog Modeling. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1684–1690. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Xu, Y.; Zhao, H.; and Zhang, Z. 2020. Topic-Aware Multi-turn Dialogue Modeling.

Yang, S.; Zhang, Z.; Jiang, Y.; Qin, C.; and Liu, S. 2025. A Unified Supervised and Unsupervised Dialogue Topic Segmentation Framework Based on Utterance Pair Modeling. In *North American Chapter of the Association for Computational Linguistics*.

Zhang, L.; and Zhou, Q. 2019. Topic segmentation for dialogue stream. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1036–1043. IEEE.

Appendix

DTS Problem Formulation and Public-Channel Dataset

Problem Statement Formally, Dialogue Topic Segmentation (DTS) can be framed as a structured prediction task that aims to detect latent topic boundaries within a dialogue $D = \{u_1, u_2, ..., u_n\}$, where each utterance u_i is assigned to a contiguous topic segment T_j. The output is a segmentation $S = \{s_1, s_2, ..., s_k\}$, in which each $s_j \subseteq D$ denotes a consecutive span of utterances with high topical coherence and minimal semantic drift. Unlike conventional utterance-level classification, DTS inherently requires modeling discourse-level dependencies, temporal progression, and transition phenomena that may span multiple turns. This is particularly challenging in task-oriented, public-channel environments, where lexical cohesion is weak, explicit discourse markers are absent, and topic changes often arise from pragmatic coordination rather than lexical shifts. Consequently, effective DTS systems must go beyond local similarity to incorporate structural, speaker-centric, and context-aware cues indicative of topical boundaries.

Dataset Overview and Application scenarios A number of established datasets (Xie et al. 2021) have advanced research in Dialogue Topic Segmentation, including **DialSeg711** (Xu, Zhao, and Zhang 2020) and **Doc2Dial** (Feng

et al. 2020). These datasets are primarily drawn from structured interviews, multi-party meetings, or online discussions, where utterances tend to be longer, well-formed, and contextually rich. While effective for modeling discourse in open-domain or cooperative settings, such datasets fall short in representing the terse, fragmented, and pragmatically complex nature of public-channel communications.

We are motivated by pressing real-world needs observed in industrial maritime regulatory practice. In particular, a new dataset was purposefully collected and constructed to support DTS under public-channel communication scenarios—specifically, maritime VHF radio dialogues, namely **VHF-Dial**. In real-world coastal monitoring zones, massive volumes of open-channel voice communications take place between ships and maritime regulators (and, to a lesser extent, between ships themselves), following international conventions on shared frequency use. These conversations are mission-critical, highly dynamic, and span a wide array of tasks ranging from port entry coordination to emergent risk mitigation.

However, structuring such data at scale is almost impossible through manual means. Of even greater concern, near-miss events—non-accidental but high-risk interactions—are frequently omitted from formal reporting pipelines due to the absence of physical consequences. These events are valuable for proactive safety assessment, yet their analysis demands hours of manual transcription and annotation. This imposes high labor costs and limits the ability of current digital oversight systems to learn from such rich but latent sources.

We envision DTS as a fundamental building block to alleviate this bottleneck. If we can accurately and automatically identify topic transitions within public-channel VHF conversations, it would unlock structured behavioral insights at scale and provide regulatory authorities with improved situational awareness and data-driven decision support.

Association for the Advancement of Artificial Intelligence
1101 Pennsylvania Ave, NW Suite 300
Washington, DC 20004

AAAI COPYRIGHT FORM

Title of Article/Paper: DASH: Dialogue-Aware Similarity and Handshake Recognition for Topic Segmentation in Public-Channel Conversations

Publication in Which Article/Paper Is to Appear: AAAI2026 Workshop

Author's Name(s): Sun Sijin, Liangbin Zhao, Ming Deng, Xiuju Fu

Please type or print your name(s) as you wish it (them) to appear in print

PART A – COPYRIGHT TRANSFER FORM

The undersigned, desiring to publish the above article/paper in a publication of the Association for the Advancement of Artificial Intelligence, (AAAI), hereby transfer their copyrights in the above article/paper to the Association for the Advancement of Artificial Intelligence (AAAI), in order to deal with future requests for reprints, translations, anthologies, reproductions, excerpts, and other publications.

This grant will include, without limitation, the entire copyright in the article/paper in all countries of the world, including all renewals, extensions, and reversions thereof, whether such rights current exist or hereafter come into effect, and also the exclusive right to create electronic versions of the article/paper, to the extent that such right is not subsumed under copyright.

The undersigned warrants that they are the sole author and owner of the copyright in the above article/paper, except for those portions shown to be in quotations; that the article/paper is original throughout; and that the undersigned right to make the grants set forth above is complete and unencumbered.

If anyone brings any claim or action alleging facts that, if true, constitute a breach of any of the foregoing warranties, the undersigned will hold harmless and indemnify AAAI, their grantees, their licensees, and their distributors against any liability, whether under judgment, decree, or compromise, and any legal fees and expenses arising out of that claim or actions, and the undersigned will cooperate fully in any defense AAAI may make to such claim or action. Moreover, the undersigned agrees to cooperate in any claim or other action seeking to protect or enforce any right the undersigned has granted to AAAI in the article/paper. If any such claim or action fails because of facts that constitute a breach of any of the foregoing warranties, the undersigned agrees to reimburse whomever brings such claim or action for expenses and attorneys' fees incurred therein.

Returned Rights

In return for these rights, AAAI hereby grants to the above author(s), and the employer(s) for whom the work was performed, royalty-free permission to:

1. Retain all proprietary rights other than copyright (such as patent rights).
2. Personal reuse of all or portions of the above article/paper in other works of their own authorship. This does not include granting third-party requests for reprinting, republishing, or other types of reuse. AAAI must handle all such third-party requests.
3. Reproduce, or have reproduced, the above article/paper for the author's personal use, or for company use provided that AAAI copyright and the source are indicated, and that the copies are not used in a way that implies AAAI endorsement of a product or service of an employer, and that the copies per se are not offered for sale. The foregoing right shall not permit the posting of the article/paper in electronic or digital form on any computer network, except by the author or the author's employer, and then only on the author's or the employer's own web page or ftp site. Such web page or ftp site, in addition to the aforementioned requirements of this Paragraph, shall not post other AAAI copyrighted materials not of the author's or the employer's creation (including tables of contents with links to other papers) without AAAI's written permission.
4. Make limited distribution of all or portions of the above article/paper prior to publication.
5. In the case of work performed under a U.S. Government contract or grant, AAAI recognized that the U.S. Government has royalty-free permission to reproduce all or portions of the above Work, and to authorize others to do so, for official U.S. Government purposes only, if the contract or grant so requires.

In the event the above article/paper is not accepted and published by AAAI, or is withdrawn by the author(s) before acceptance by AAAI, this agreement becomes null and void.

(1) Sun Sijin

Author/Authorized Agent for Joint Author's Signature

30/11/2025

Date (MM/DD/YYYY)

A*STAR IHPC

Employer for whom work was performed

Mr

Title (if not author)

(For jointly authored Works, all joint authors should sign unless one of the authors has been duly authorized to act as agent for the others.)

Association for the Advancement of Artificial Intelligence

1101 Pennsylvania Ave, NW Suite 300
Washington, DC 20004

PART B – U.S. GOVERNMENT EMPLOYEE CERTIFICATION

This will certify that all authors of the above article/paper are employees of the U.S. Government and performed this work as part of their employment, and that the article/paper is therefore not subject to U.S. copyright protection. The undersigned warrants that they are the sole author/translator of the above article/paper, and that the article/paper is original throughout, except for those portions shown to be in quotations.

(2)

U.S. Government Employee Authorized Signature	Date (MM/DD/YYYY)
Name of Government Organization	Title (if not author)

*(Please read and sign and return Part B **only** if you are a government employee and created your article/paper as part of your employment. If your work was performed under Government contract, but you are not a Government employee, sign only at signature line (1) in Part A and see item 5 under returned rights. Authors who are U.S. government employees should also sign signature line (1) in Part A above to enable AAAI to claim and protect its copyright in international jurisdictions.)*

PART C–CROWN COPYRIGHT CERTIFICATION

This will certify that all authors of the above article/paper are employees of the British or British Commonwealth Government and prepared the Work in connection with their official duties , and that the article/paper is therefore subject to Crown Copyright and is not assigned to AAAI as set forth in the first sentence of the Copyright Transfer Section in Part A. The undersigned warrants that they are the sole author/translator of the above article/paper, and that the article/paper is original throughout, except for those portions shown to be in quotations, and acknowledges that AAAI has the right to publish, distribute, and reprint the Work in all forms and all media.

(3)

British or British Commonwealth Government Employee Authorized Signature	Date (MM/DD/YYYY)
Name of Government Organization	Title (if not author)

*(Please read and sign and return Part C **only** if you are a British or British Commonwealth Government employee and the Work is subject to Crown Copyright. Authors who are British or British Commonwealth government employees should also sign signature line (1) in Part A above to indicate their acceptance of all terms other than the copyright transfer.)*