DSTC12 Track Proposal: Controllable Conversational Theme Detection

Igor Shalyminov, Hang Su, Jason Cai, James Gung, Raphael Shu, Jake Vincent and **Saab Mansour**
Amazon
{shalymin,shawnsu,cjinglun,gungj,zhongzhu,jakevinc,saabm}@amazon.com

1 Motivation

We propose *Theme detection* as a key task in conversational analytics: themes are supposed to highlight the topics discussed in the conversation that are useful for categorizing and further analyzing them according to the nature of the conversation, e.g. customer support, sales or marketing calls. Automatically discovering and categorizing themes can potentially save hours to days of work on manual analysis of lengthy conversations.

We see the theme detection task being closely related to dialogue intent detection — however, the subtle difference lies in the purpose of the detection result. Dialogue intents are supposed to be used in the downstream dialogue system where there might be some business logic dependent on the exact intent value. Therefore, intents are expected to be of a fixed set of values or mapped to this fixed set via a separate alignment model. In turn, themes themselves are supposed to be the final result directly presented to the user, e.g. a call-center analyst. Therefore, they are expected to provide a gist of the dialogue from the customer's inquiry perspective, which leaves room for a variety of surface forms and user preference based customizations.

The problem of open conversational intent induction was initially explored in a DSTC11 track by Gung et al. (2023b) focusing on utterance clustering in two setups of a varied challengingness. In our theme detection task, we pose the problem as *joint clustering and theme labeling* for the input utterances. We do not restrict the surface form of the resulting theme labels — the language style and quality of the themes will be evaluated based on the guideline which we also provide as the key resource for model development (see Appendix A).

Furthermore, in the proposed task, we put an additional requirement on the custom granularity of the clusters and the corresponding theme labels — it is supposed to be inferred from the user preference data that will be provided as part of the model inputs. The motivation behind customizing the granularity is the fact that each specific customer may have their own business-motivated preferences on which themes should be looked into more closely, and which ones could be observed at a birds-eye view.

We assume the task to be completed in the zero-shot setup on a domain unseen during the training/development process — the extra input data for guiding theme labeling and aligning cluster granularity (described in Section 3) will facilitate that. The extensions to the clustering task we propose here are particularly interesting to explore in the context of Large Language Models (LLMs) but do not necessitate them.

2 Proposed Task

The task we are focusing on in this track is *Controllable Theme Detection*. Given a dataset of unlabeled utterances, the goal of the task would be to cluster them into themes and give each theme a short and concise natural language label. We highlight that for the purpose of theme detection, a variety of possible surface forms can serve as perfectly fine labels — therefore, we provide the Theme Label Writing Guideline (Appendix A) which serves both as the additional input data for the label generation logic and the primary reference for the human and LLM-based label evaluation (more detail on evaluation in Section 4).

Moreover, the theme granularity is supposed to be controlled via additional input data, users' preferences on whether a pair of utterances should belong to the same theme or not (loosely following the Stage 2 approach of Zhang et al. 2023). In this way, if the users' preferences suggest that the utterances *"I want to purchase pet insurance"* and *"I want to purchase travel insurance"* should belong to the same theme, all the utterances like these two would be associated to the single theme whose la-

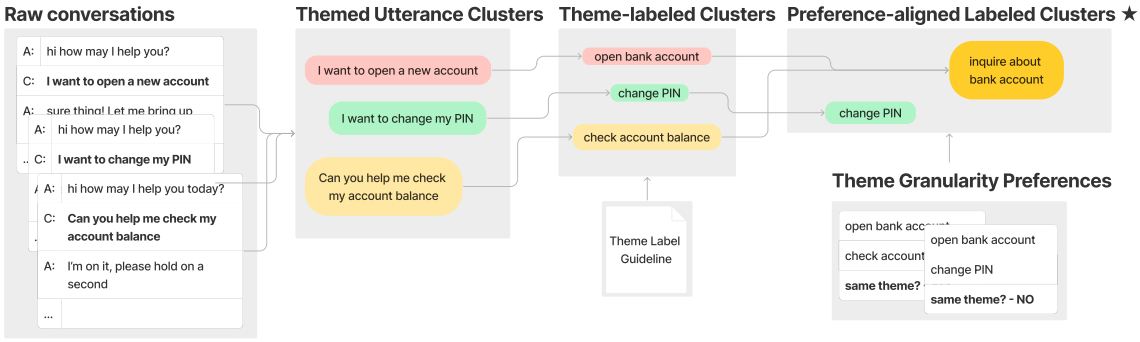

Figure 1: Diagram of the proposed task in the form of an example processing pipeline. The inputs to the "system" are raw conversations, user preferences on the theme granularity and theme label guidelines; the output is preference-aligned utterance clusters with the corresponding theme labels (marked with ★)

bel semantically unifies both of the two utterances' meanings e.g. *"purchase insurance"* or some close paraphrase of it. On the other hand, if the preferences elicit that *"I want to find the closest branch"* and *"Give me the directions to the closest ATM"* should not belong to the same theme, the corresponding themes *"find branch"* and *"find ATM"* as well as the clusters of utterances belonging to them should be kept as separate. This information can be used to enable contrastive fine-tuning of utterance representation as done by e.g. Chu et al. (2023) and Zhang et al. (2021) or to adjust the initial clusters/themes, as depicted in Figure 1.

A visualization of the overall task is presented in Figure 1 where we gave a potential sequential pipeline as an example. The actual submissions can vary in architecture and the types of models used. We encourage the participants to use techniques from both LLM-based and traditional Machine Learning paradigms that adequately correspond to the problem being tackled.

A successful completion of the task would assume assigning each utterance a theme label so that:

- theme labels are concise, exhaustively cover all the examples and are mutually exclusive,
- label wording conforms to the Theme label writing guideline (Appendix A, will be provided to the participants),
- theme granularity matches the 'gold' held out assignment which is supposed to be inferred from the provided user preference samples.

Table 1: Data Statistics

Domain	# Dialogues	# Themed utterances
Banking (train)	933	2504
Finance (dev)	1154	2449
Undisclosed (test)	528	1081

3 Datasets

We build our task on top of the NatCS (Gung et al., 2023a,b), a multi-domain dataset of human-human customer support conversations — the dataset statistics per domain are provided in Table 1.

We intend for the participants' submissions to work in a zero-shot setup naturally supported within the LLM-centered framework. As such, we will provide the data in two domains, **Banking** and **Finance**, for the participants to use for the training/development purposes and assess the domain generalization of their approaches; the test domain would have little to no overlap with the train/dev data.

For the train and dev domains, we will provide the following data:

1. Utterances to cluster/label with dialogue contexts — since some of the valid themed utterances might be elliptical/have otherwise incomplete explicit information, we will provide full dialogue contexts along with them. The utterances marked as themed would be the datapoints to run prediction on. See also Appendix B for an example input.
2. User preferences for clustering — a set of utterance pairs (with the corresponding dialogue contexts) along with binary decisions whether

they should belong to the same cluster/theme or not. This works as the main input for inferring the desired theme granularity. See also Appendix B for a user preference example.

3. The 'gold' theme labeling for all the themed utterances.

For the final evaluation phase, we will release the themed utterances with dialogue contexts and the user preferences set for the Travel domain; the gold theme labeling will be held hidden.

In addition, all the train/dev/test domains will share the same theme label writing guideline (Appendix A) that will be available for the participants from the challenge's start.

We are conducting research on the optimal number of user preference samples to sufficiently describe the desired granularity while maintaining the challengingness of the task.

4 Evaluation

Our task is composed of two subtasks: clustering of the utterances into themes, and assigning natural language labels for those themes. Therefore, our evaluation metrics will focus on both clustering quality and that of label generation. The evaluation metrics described above are automatic and will be provided to the participants together with a simple baseline solution in the starter code.

4.1 Clustering metrics

- **NMI** score (Vinh et al., 2010) — *Normalized Mutual Information* is a function that measures the agreement of the two cluster assignments, reference and predicted, ignoring permutations. Normalization is performed over the mean of the entropies of the two assignments

- **ACC** score (Huang et al., 2014) evaluates the optimal alignment between the reference cluster assignment and the predicted one, with the alignment obtained using the Hungarian algorithm.

4.2 Label generation metrics

We will evaluate the labels generated on top of the predicted clustering: for each cluster, the reference labels of its utterances will be compared to the predicted label:

$$Score_i(Y_i, \hat{y_i}) = \frac{\sum_j sim(Y_{i,j}, \hat{y_i})}{|Y_i|}$$

where Y_i are all the reference labels for the utterances of the ith predicted cluster, $\hat{y_i}$ is the predicted label for the cluster, and sim is one of the similarity metrics described below.

The gold labeling will also include multiple reference labels per each datapoint — the resulting similarity values would then be max over the references.

We will use the following label metrics:

- **Cosine similarity** — the semantic similarity measure over Sentence-BERT embeddings of the reference and predicted labels,

- **ROUGE** score (Lin, 2004) — an N-gram overlap metric useful for comparing short and concise word sequences,

- an **LLM-based score** for evaluating theme labels against the guideline. For the sake of preventing evaluation metric hacking, the actual prompt that will be used in the final submission evaluation would be kept private to the organizers. It will be fully consistent with the guideline in Appendix A in terms of the requirements to the labels.

Based on the number of the participant teams, there is a possibility for the final evaluation phase to include human evaluation covering the top performing submissions which are in turn determined by the aggregated automatic scores.

5 Track Organizers

Igor Shalyminov, Hang Su, Jason Cai, James Gung, Raphael Shu, Jake Vincent, Saab Mansour.

The organizers of this track are machine learning scientists and linguists affiliated with Amazon, focused on research & development in Conversational NLP.

6 Ethics Discussion

The data to be used in this track is collected specifically for research & development purposes and contains no personally identifiable information. The annotators were paid a competitive wage as estimated across the US market. The collection and annotation of datasets used in this track, along with

the baseline & evaluation code, will be released to the public for future academic research.

References

Caiyuan Chu, Ya Li, Yifan Liu, Jia-Chen Gu, Quan Liu, Yongxin Ge, and Guoping Hu. 2023. Multi-stage coarse-to-fine contrastive learning for conversation intent induction. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 31–39, Prague, Czech Republic. Association for Computational Linguistics.

James Gung, Emily Moeng, Wesley Rose, Arshit Gupta, Yi Zhang, and Saab Mansour. 2023a. NatCS: Eliciting natural customer support dialogues. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9652–9677, Toronto, Canada. Association for Computational Linguistics.

James Gung, Raphael Shu, Emily Moeng, Wesley Rose, Salvatore Romeo, Arshit Gupta, Yassine Benajiba, Saab Mansour, and Yi Zhang. 2023b. Intent induction from conversations for task-oriented dialogue track at DSTC 11. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 242–259, Prague, Czech Republic. Association for Computational Linguistics.

Peihao Huang, Yan Huang, Wei Wang, and Liang Wang. 2014. Deep embedding network for clustering. In *2014 22nd International Conference on Pattern Recognition*, pages 1532–1537.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5419–5430, Online. Association for Computational Linguistics.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. ClusterLLM: Large language models as a guide for text clustering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13903–13920, Singapore. Association for Computational Linguistics.

A Theme Label Writing Guideline

An acceptable theme label is structurally and semantically well-formed according to the rules outlined in this appendix. *Structurally well-formed* means that the words and their arrangement in the theme label are acceptable. *Semantically well-formed* means that the meaning and usability of the theme label are acceptable.

A.1 Theme labels exclude unneeded and undesirable words.

Theme labels should be concise (2–5 words long). They should only include essential words (see A.1.1 and A.1.2 below). Essential words will primarily include content (open-class) words. Function (closed-class) words should be excluded. Prepositions may be included as needed but should be avoided when there is a synonymous alternative label without a preposition.

Theme labels should also exclude context-sensitive words like pronouns (*him*, *her*, *them*, *it*, *us*, etc.) and demonstratives (*this*, *that*, *those*, etc.).

A.1.1 Word types

- Content/open-class words:
  - nouns (*items*, *insurance*, *information*, *order*, etc.)
  - main verbs (*check*, *inquire*, *add*, *explore*, etc.)
  - adjectives (*new* patient, *missing* item, etc.)
  - other modifying words (*shipping* information, *product* options, etc.)
- Function/closed-class words:
  - articles/determiners (*the*, *a*, etc.)
  - auxiliary verbs (*have* or *be*, as in *I have eaten* or *I am eating*)
  - copulas
  - negation (*not* or *-n't*, as in *not on time* or *didn't arrive*)
  - conjunctions (*and*, *or*, *but*, etc.)
  - complementizers (clause-embedding uses of *that*, *for*, *if*, *whether*, *because*, etc.)
  - modals (*can*, *could*, *will*, *would*, *may*, *might*, *must*, *shall*)
  - question words (*who*, *what*, *where*, *when*, *how*, *why*)

- Context-sensitive words:
  - pronouns (*she*, *he*, *they*, *it*, *her*, *his*, etc.)
  - demonstratives (*this*, *these*, *that*, *those*, etc.)
  - temporal adverbs (*yesterday*, *tomorrow*, *next week*, etc.)
  - other context-sensitive language
    - *one*, as in *I'm looking for a nearby branch. Can you find* one*?*
    - deleted nouns (noun ellipsis), as in *I found his order, but not yours* __.

A.1.2 Examples

For a theme covering order tracking:

- Good: track order
- Good: track shipment
- Bad: track an order (includes an article)
- Bad: track their order (includes a pronoun)

For a theme covering finding the nearest branch of a chain:

- Good: find nearest branch
- Good: find closest branch
- Bad: find nearest one (includes context-sensitive *one*)
- Bad: check if there's a nearby branch (includes a complementizer *if*; includes a form of *be*)

A.2 Theme labels are verb phrases that classify events.

A verb phrase begins with a verb and may include arguments or modifiers of the verb (such as a direct object). The verb should be in its citation form, lacking any complex morphology such as tense or agreement suffixes. The citation form of a verb is what would normally follow the infinitive *to*, such as *sign up* in *I'd like to* sign up. Theme labels should not be other phrase types, such as noun phrases.

The verb phrase should describe a class of events. Events are things that can be said to **happen**, unlike states (e.g. *learn* [event] vs. *know* [state]), entities (e.g. *redeem* [event] vs. *redemption* [entity]), properties (e.g. *complain* [event] vs. *angry* [property]), and claims (*report defect* [event] vs. *product is defective* [claim]).

A.2.1 Examples

For a theme covering membership sign-ups:

- Good: sign up for membership (verb phrase; describes a kind of *signing up* event)
- Bad: signing up for membership (verb phrase, but verb is not in citation form)
- Bad: membership sign-up (noun phrase; describes a kind of entity)
- Bad: memberships (noun phrase; describes a kind of entity)

For a theme covering requests to check in early at a hotel:

- Good: request early check-in (verb phrase; describes a kind of requesting event)
- Bad: requested early check-in (verb phrase, but verb is not in citation form)
- Bad: request for early check-in (noun phrase; describes a kind of entity)
- Bad: customer wants early check-in (this is a claim)

For a theme covering reporting a defective product:

- Good: report defective product (verb phrase; describes events)
- Bad: reporting defective product (verb phrase, but verb is not in citation form)
- Bad: believe product is defective (verb phrase, but describes a state rather than an event)
- Bad: defective product (noun phrase; describes a kind of entity)

A.3 Theme labels are informative and actionable yet sufficiently general.

Theme labels should be informative enough to substantially narrow down the set of possible customer issue resolution steps (the steps to resolve the problem/need that drove the customer to make contact). For example, *check balance* is probably associated with a standard procedure for checking the balance of a range of customer account types, but *perform check* is so broad that it could be associated with an extremely diverse group of issue resolutions. Non-actionable theme labels may be excessively vague or uninformative, and hence not very useful.

A.3.1 Examples

For a theme covering appointment-scheduling themes:

- Good: schedule appointments
- Bad: ask about appointments (probably too general)
- Bad: schedule appointment for next week (too specific)
- Bad: schedule appointment for elderly parent (too specific)

For a theme covering adding a recognized user to an existing account or policy:

- Good: add user
- Bad: add one (too general)
- Bad: add oldest child (too specific)

For a theme covering user password issues:

- Good: reset password
- Good: troubleshoot password
- Bad: secure account (too general)
- Bad: reset password again (too specific)

For a theme covering credit or debit card charge disputes:

- Good: dispute charge
- Bad: complain about charge (too general)
- Bad: file card complaint (too general)
- Bad: dispute charge for defective blender (too specific)

B Input/Output Data Examples

Below is an input datapoint for a dialogue with one utterance marked as themed. For the train/dev domains, the theme labels will be available as in the example below. For the test domain, only the flag that an utterance is themed will be provided.


```
{
  "conversation_id": "Banking_123",
  "turns": [
      {
        "speaker": "Agent",
        "utterance": "Thank you for
            calling Intellibank. This is
             Melanie. How can I help you
            ?"
      },
      {
        "speaker": "Customer",
        "utterance": "Yeah, hey. This is
             John Smith. I've got a
            quick question."
      },
      {
        "speaker": "Agent",
        "utterance": "OK, John. What can
             I help you with?"
      },
      {
        "speaker": "Customer",
        "utterance": "Yeah I need to
            know what your ATM
            withdrawal limits are for
            the day.",
        "theme_label": "get daily
            withdrawal limit",
      },
      {
        "speaker": "Agent",
        "utterance": "Certainly. Our ATM
             withdrawal limit is on a
            per day basis and it is up
            to two hundred dollars."
      },
      {
        "speaker": "Customer",
        "utterance": "Oh perfect,
            perfect. Yeah, I think I'll
            just see if I can head down
            to the ATM now. Thank you."
      },
      {
        "speaker": "Agent",
        "utterance": "OK, thank you. You
             have a great day."
      },
      {
        "speaker": "Customer",
        "utterance": "You too."
      }
    ]
}
```


Below is an input datapoint with the example user preference on clustering granularity:


```
{
  "utterance_a": {
    "utterance": "Yeah, so I need to
        change the account number thing
        that I put in whenever I go to
        the ATM."
    "conversation_id": "Banking_123",
    "turn_id": 4
  },
  "utterance_b": {
    "utterance": "OK. Excellent. Thank
        you Ms. Crystal. And while I got
         you on the phone I see it's
        been a little bit since you've
        authenticated your account here.
         Would you like to add a PIN
```

```
        number to your account for
        security reasons?"
    "conversation_id": "Banking_345",
    "turn_id": 10
  },
  "belong_to_same_theme": "yes"
}
```