hhu.

# Task-oriented Conversational Modelling with Subjective Knowledge

Renato Vukovic
Dialogue Systems and Machine Learning

07.12.2023

# Outline



- Task-oriented dialogue with subjective knowledge (SK-TOD)
- DSTC 11 track 5
- Ensemble Methods for SK-TOD
- LLMs for SK-TOD evaluation

# “What do others think?”: Task-Oriented Conversational Modeling with Subjective Knowledge

Chao Zhao[1] Spandana Gella[2] Seokhwan Kim[2] Di Jin[2] Devamanyu Hazarika[2] Alexandros Papangelis[2] Behnam Hedayatnia[2] Mahdi Namazifar[2] Yang Liu[2] Dilek Hakkani-Tur[2]
[1]UNC Chapel Hill [2]Amazon, Alexa

*Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 309–323 September 11–15, 2023.

# Task-oriented Dialogue (TOD)

Heinrich Heine Universität Düsseldorf

- TOD systems aim at fulfilling tasks given by users via natural language
- These tasks range from turning on the light in the living room to booking hotels
- Information is accessed through a data base or API
- Some *aspects* queried by users are not part of structured data bases
  - E.g. whether you are allowed to bring pets to a hotel or how good the WIFI quality is
- This information is present in unstructured knowledge sources such as FAQs or reviews
- Handling **subjective knowledge** in user requests comes with unique challenges

# TOD with subjective knowledge (SK-TOD)

- **Given:** a dialogue context $C = [U_1, S_1, U_2, S_2, \cdots, Ut]$ between a user and a system.
- The dialogue involves one or more **entities**, denoted as $\mathcal{E} = \{e_1, \cdots, e_m\}$.
- Alongside the dialogue, we have a **subjective knowledge source** $\mathcal{B} = \{(e_1, \mathcal{R}_1), (e_2, \mathcal{R}_2), \cdots\}$ containing all the entities and their corresponding customer reviews.
  - Each entity $e$ is associated with multiple **reviews** $\mathcal{R} = \{R_1, R_2, \cdots\}$.
  - Each review can be divided into **segments** $[K_1, K_2, \cdots]$, such as paragraphs, sentences, or sub-sentential units.
    - Here, each review *sentence* is regarded as a knowledge snippet.

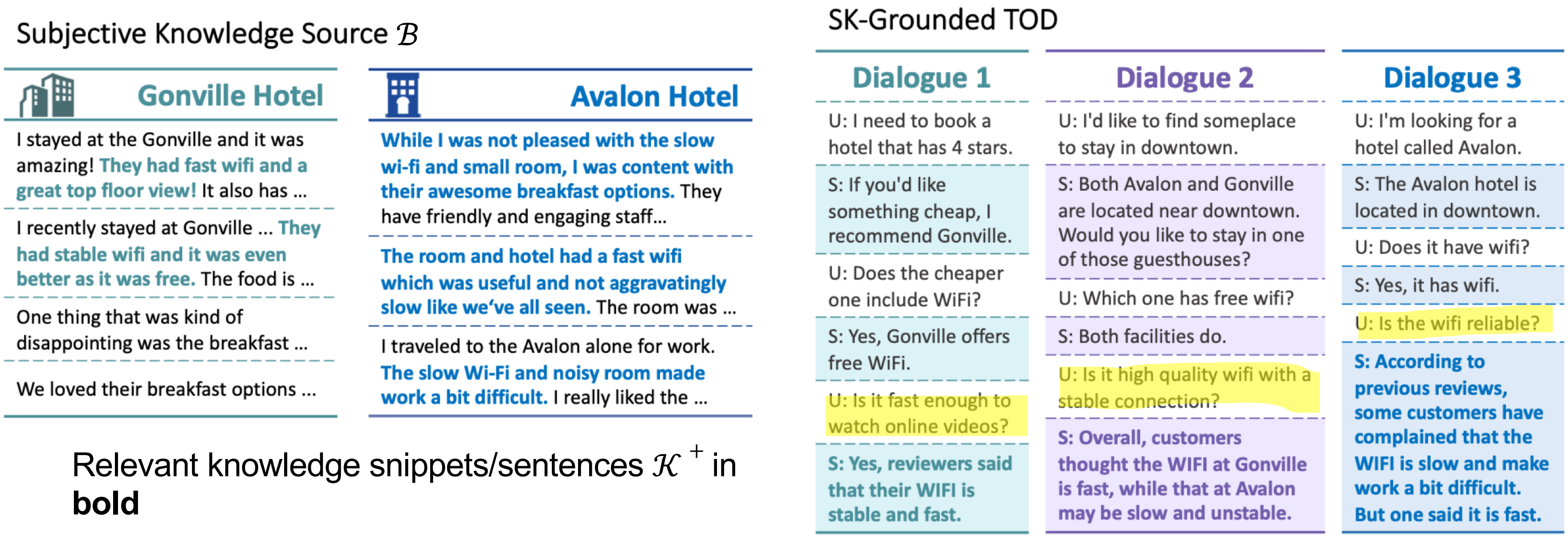
# TOD with subjective knowledge (SK-TOD)

Heinrich Heine Universität Düsseldorf

- **Task:**
  1. Identify whether $U_t$ is a **subjective knowledge-seeking request**
  2. If it is, select **all the relevant knowledge snippets** $\mathcal{K}^{+}$ from the knowledge source $\mathcal{B}$
  3. Generate a response $S_t$ grounded on all the snippets $\mathcal{K}^{+}$.

# TOD with subjective knowledge (SK-TOD)

Heinrich Heine
Universität Düsseldorf




Relevant knowledge snippets/sentences $\mathcal{K}^+$ in **bold**

Relevant entities from dialogue $C$: $\mathcal{E} = \{Gonville\ Hotel, Avalon\ Hotel\}$
**Responses** to the yellow marked subjective knowledge-seeking user requests are **bold**.

# TOD with subjective knowledge (SK-TOD)

Heinrich Heine Universität Düsseldorf

## Subjective Knowledge Source $\mathcal{B}$

**Gonville Hotel**

I stayed at the Gonville and it was amazing! **They had fast wifi and a great top floor view!** It also has ...

I recently stayed at Gonville ... **They had stable wifi and it was even better as it was free.** The food is ...

One thing that was kind of disappointing was the breakfast ...

We loved their breakfast options ...

**Avalon Hotel**

**While I was not pleased with the slow wi-fi and small room, I was content with their awesome breakfast options.** They have friendly and engaging staff...

**The room and hotel had a fast wifi which was useful and not aggravatingly slow like we've all seen.** The room was ...

I traveled to the Avalon alone for work. **The slow Wi-Fi and noisy room made work a bit difficult.** I really liked the ...

## SK-Grounded TOD

**Dialogue 1**

U: I need to book a hotel that has 4 stars.

S: If you'd like something cheap, I recommend Gonville.

U: Does the cheaper one include WiFi?

S: Yes, Gonville offers free WiFi.

U: Is it fast enough to watch online videos?

**S: Yes, reviewers said that their WIFI is stable and fast.**

**Dialogue 2**

U: I'd like to find someplace to stay in downtown.

S: Both Avalon and Gonville are located near downtown. Would you like to stay in one of those guesthouses?

U: Which one has free wifi?

S: Both facilities do.

U: Is it high quality wifi with a stable connection?

**S: Overall, customers thought the WIFI at Gonville is fast, while that at Avalon may be slow and unstable.**

**Dialogue 3**

U: I'm looking for a hotel called Avalon.

S: The Avalon hotel is located in downtown.

U: Does it have wifi?

S: Yes, it has wifi.

U: Is the wifi reliable?

**S: According to previous reviews, some customers have complained that the WIFI is slow and make work a bit difficult. But one said it is fast.**

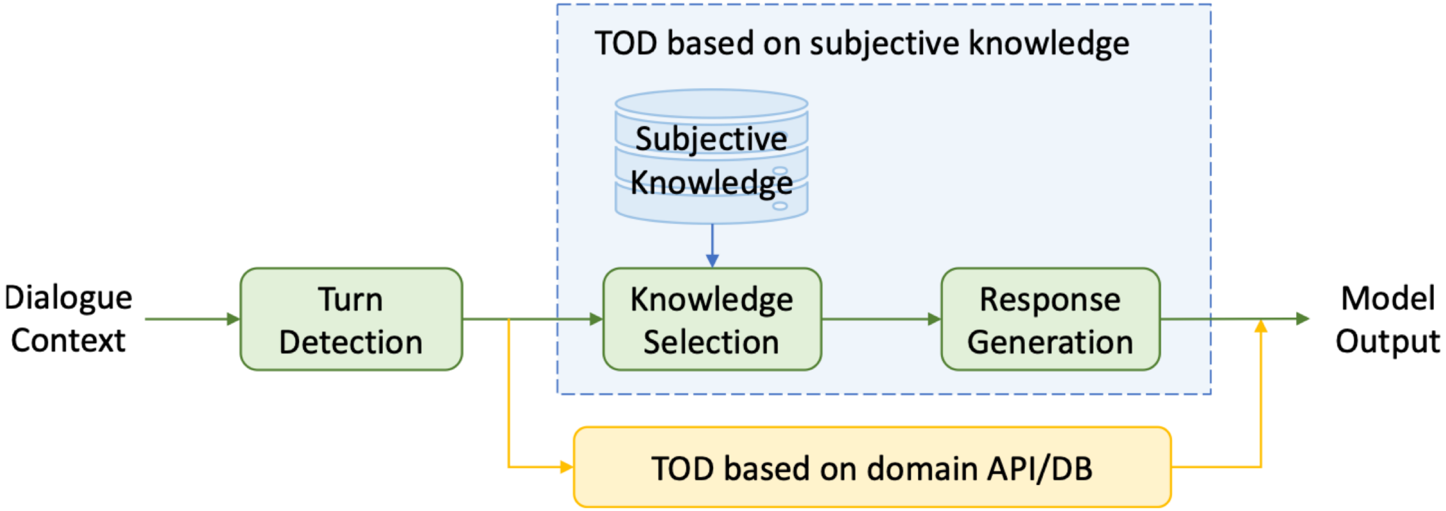
→ The challenge in SK-TOD is including **all the different opinions** found in the knowledge source, as seen in Dialogue 3.

# TOD with subjective knowledge (SK-TOD)



- SK-TOD is divided into four subtasks based on the subtasks for factual unstructured knowledge TOD (Kim et al., 2020):
  1. **Knowledge-seeking turn detection** (KTD): Decide whether the user queried subjective knowledge based on the context
  2. **Entity tracking** (ET): Extract the relevant entities for the user query to minimise the number of possible knowledge snippets to be chosen in the next step
  3. **Knowledge selection** (KS): Based on the context and entities tracked find all the relevant knowledge snippets
  4. **Response generation** (RG): Based on context and the selected knowledge generate a response

# TOD with subjective knowledge (SK-TOD)



# TOD with subjective knowledge (SK-TOD)

Heinrich Heine
Universität Düsseldorf

- Two major differences to using ***subjective* knowledge** compared to *factual* knowledge:
  - The SK-TOD model needs to consider **all relevant knowledge snippets** for the context, i.e. both recall and precision are important
  - The model needs to aggregate these knowledge snippets into a concise response that can faithfully reflect the **diversity and proportion** of the different opinions expressed
- Including both negative and positive responses and their proportions increases trust into the system (Baek et al., 2012)

# SK-TOD Data-set

- Part of the 11th Dialogue Systems Technology Challenge (DSTC) as track 5
- Data is based on MultiWOZ 2.1 data-set (Eric et al., 2020)
  - Assume that it does not contain any subjective knowledge requests
- Reviews, subjective knowledge-seeking requests and corresponding responses were *written by crowd workers* in three steps
- Only entities and dialogues from the **hotel and restaurant** domain are part of the data
  - Overall 33 hotels and 110 restaurants are selected from MultiWOZ and 10 reviews are collected per entity
- In the validation and test set of the SK-TOD data-set there are *seen and unseen* subsets
  - → The unseen subsets contain **aspects** that are **not part of the training data** to test models‘ generalisability

# Data-set Statistics



| | Train | Val | Test |
|---|---|---|---|
| # instances | 14768 | 2129 | 2799 |
| # seen instances | 14768 | 1471 | 1547 |
| # unseen instances | 0 | 658 | 1252 |
| # multi-entity instances | 412 | 199 | 436 |
| Knowledge Snippets | | | |
| Avg. # snippets per instance | 3.80 | 4.07 | 4.21 |
| Avg. # tokens per snippet | 14.68 | 15.49 | 14.5 |
| Dialogue | | | |
| Avg. # uttrances per instance | 9.29 | 9.44 | 9.36 |
| Avg. # tokens per request | 8.65 | 8.94 | 9.12 |
| Avg. # tokens per response | 24.18 | 23.61 | 23.86 |

→ Only few multi-entity instances in the data

Instances consist of subjective user requests and subjective-knowledge-grounded responses

# Data Collection

Heinrich Heine
Universität Düsseldorf

1. **Reviews:**
   - Crowd workers were asked to write reviews for different entities in MultiWOZ, given the **user persona, sentiment and the aspect** that should be covered in the review
2. **Requests:**
   - Write a request given an aspect to write a subjective knowledge request by the user and insert it at an appropriate position in the dialogue
   - Turns after the insertion are removed
3. **Responses:**
   - Given dialogue context and knowledge snippets, worker is asked to give a natural response, which considers the diversity and proportion of the opinions in all relevant knowledge snippets

# SK-TOD Data Collection



**Instruction**

Please assume that you recently visited **MIDSUMMER HOUSE RESTAURANT** alone.
This place serves **British** cuisine and you ordered the following:

- Dishes:
  - Strawberries and Cream
- Drinks:
  - beer

Please write down your review comments based on the following aspects:

- **What you liked:**
  - **Good portion of foods**
  - **High-quality foods**
- **What you disliked:**
  - **Overpriced drinks**

Notes:

- Please do **NOT** copy and paste the aspects as they are.
- Please provide as many details as possible.

**Your review post:**

Write down a review post

Review collection instruction example

# SK-TOD Data Collection

**GOAL: You prefer a place which has a stable internet connection. You must ask a question about this for the hotel.**

NOTE: *Please do NOT insert a question before the hotel name is explicitly mentioned in the conversation.*

---

User: Where can I get some British food that isn't too cheap or expensive?
Agent: There are a total of 5 British restaurants that are moderately priced. Is there any particular location you are wanting such as the centre or west?

○

User: No, no particular location. What is the information for the first listing?
Agent: Restuarant one seven. The address is De Vere University Arms Regent Street City Centre. The phone number is 01223337766.

○

User: Great. Do you know of a high-end hotel I could stay at? It has to have free wifi.
Agent: The University Arms Hotel is an expensive place with free wifi, conveniently located very close to your restaurant. Would you like to book a room there?

◉
User: Do they provide fast and stable wifi to watch online videos?
Agent:

User: I might be interested. First, is that a guesthouse style? I would love to stay in a guesthouse style with free parking!
Agent: Unfortunately no, it is not a guesthouse style lodging. There are none that fall into the expensive category, would you like me to search for something cheaper that is a guesthouse?

○

User: No, that's okay. Are there any hotels available in the moderate price range?
Agent: There are 12 guest houses that have free wifi and free parking. Would you like me to book one for you?

○

Request insertion example (note that subjective DB queries are not possible)

# SK-TOD Data Collection

**Conversation:**

Customer: Hi there. I'm on a really tight budget but need to find me a cheap room to stay.
Agent: We have several cheap htels and guesthouses available. Do you have a particular location in mind?
Customer: Not really. It does need to be a hotel and parking isn't really an issue with us.
Agent: I have one in the cheap range, shall I book it?
Customer: Does it have a star of 0. That is very important to me?
Agent: Unfortunately, it's 4 stars. Would you like to try a different hotel type?
Customer: What about one that is in the moderate price range?
Agent: I found 3 hotels, do you have a preference for the area of centre or north?
Customer: No area preference. Any of them are fine. Can you book one for me starting Thursday?
Agent: Yes, of course. For how many nights and how many people?
Customer: There will be three people and we need it for 3 nights. Can I get the reference number for that once it is booked?
Agent: I've booked you at the Ashley Hotel, and your reference number is 9ZKDNRMV.
Customer: Wait. Is there a good internet connection at the Ashley?

Agent: Write down a response (longer than 100 characters) to the customer to continue the

Submit

**Knowledge:**

**Customer Reviews**

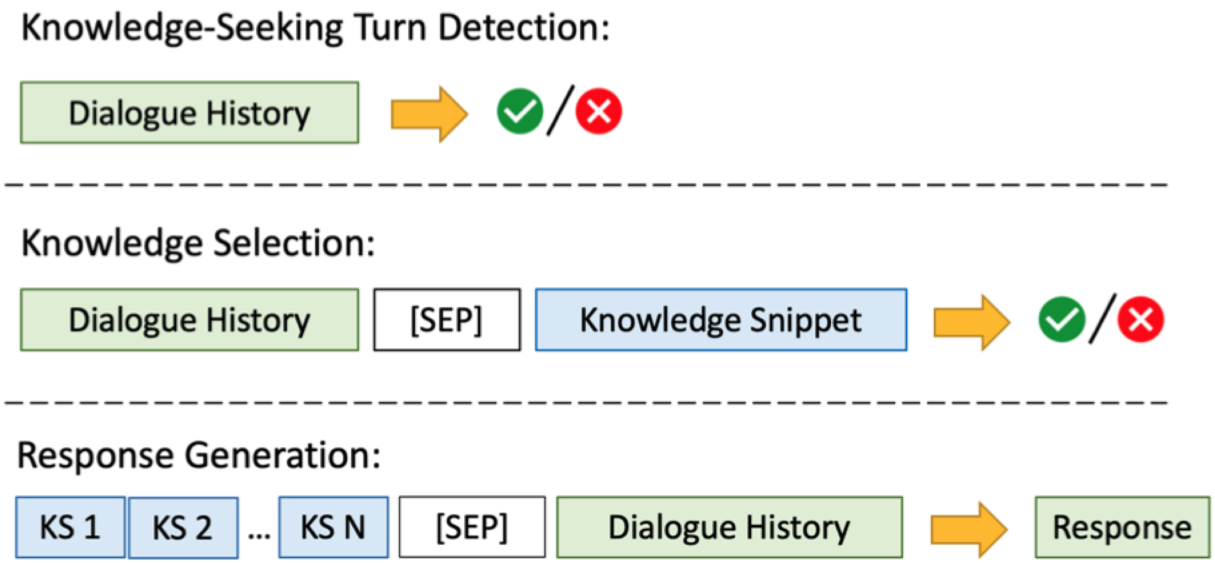
ASHLEY HOTEL
... This is very frustrating when you are trying to work and the delay is huge. ...

... Also, the fast wi-fi was a huge plus for work I had to work to do on my laptop. ...

... Also, the Wi-Fi was extremely slow. ...

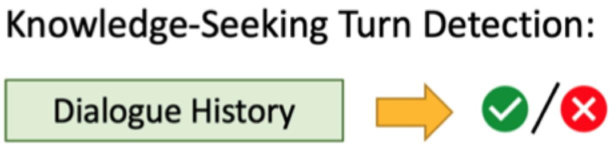
Response collection instruction example

# SK-TOD Baselines



# SK-TOD Baselines

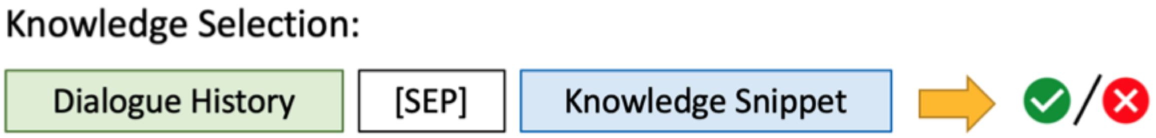
- **KTD**: Train a binary classifier based on BERT embeddings of the dialogue context on predicting whether subjective knowledge is requested
- **ET**:
  - Normalise names in the knowledge source using heuristic rules
  - Find the entities in the dialogue via fuzzy n-gram matching
    - Fuzzy n-gram matching finds the longest contiguous matching sub-sequence
    - Only works if the set of entities is known beforehand
  - Choose the entities present in the last utterance before the request to be relevant for the user query



# Knowledge Selection Baseline

Heinrich Heine Universität Düsseldorf

- Compare to information retrieval baselines such as TF-IDF
- Encode the dialogue context and the set of possible knowledge snippets based on the tracked entities to calculate a pairwise **textual similarity score**
  - Bi-Encoder: encode both inputs on their own and calculate their distance
  - Cross-Encoder: encode the concatenation of both and calculate a score of the embedding via supervised training
  - For training choose all the relevant knowledge snippets as **positive pairs** and randomly sample the same number of **negative pairs**
  - Since recall and precision are important here, a **threshold** is used to choose the knowledge snippets according to their score, rather than a top $k$ approach
    - Adapt the threshold to the validation set



# Response Generation Baseline



- Random extractive baseline:
  - choose a random relevant snippet as response
- Language model baseline: generate a response with GPT-2 (Radford et al., 2019) and BART (Lewis et al., 2020) models based on:
  - **Dialogue context**
  - **Chosen knowledge snippets**
  - **Predicted sentiment** of the reviews concatenated as natural language, e.g. “ambience is great” (Use SOTA aspect-based sentiment analysis (ABSA) model for prediction)

# KTD Results

- Knowledge-seeking turn detection can be solved almost perfectly
- The subjective knowledge-seeking turns are collected differently than the underlying data-set
  - possibly making them easier to distinguish from the Multi-WOZ turns, e.g. regarding style

| | Acc | P | R | F |
|---|---|---|---|---|
| BERT | 99.67 | 99.75 | 99.61 | 99.68 |
| RoBERTa | 99.74 | 99.86 | 99.64 | 99.75 |
| ALBERT | 99.49 | 99.64 | 99.36 | 99.50 |
| DeBERTa | 99.71 | 99.86 | 99.57 | 99.71 |

# ET Results

- Report the instance-level accuracy score
  - An instance is regarded as accurate only if the predicted entities match exactly with all the gold entities.
- The fuzzy n-gram matching method achieves an instance-level accuracy of 92.18%.
- Type of errors:
  - Underprediction: for 1.8% of the instances, there is at least one gold entity missing from the predicted entities.
  - Overprediction: for 7.6% of the instances, the predicted entities contain at least one spurious entity.

# KS Evaluation



- Can be viewed as classification or retrieval task
  - **Classification:** use precision, recall and F1 score at both instance and snippet level, i.e. whether all snippets were found for a request and on snippet level over all context-snippet pairs
  - **Retrieval:** use mean average precision (mAP) which reflects the overall ranking positions of all relevant knowledge snippets according to their score
    - top-k based measures are not used since the number of relevant snippets varies for each instance

# KS Results

| | Instance-level | | | Snippet-level | | | mAP |
|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | |
| *IR Baselines* | | | | | | | |
| TF-IDF | 34.61 | 70.33 | 40.46 | 23.81 | 65.00 | 34.85 | 45.97 |
| BM25 | 31.38 | 40.95 | 32.21 | 31.14 | 32.42 | 31.77 | 45.42 |
| *Bi-encoder* | | | | | | | |
| BERT | 56.66 | 70.06 | 59.31 | 58.87 | 74.69 | 65.84 | 71.59 |
| RoBERTa | 60.98 | 83.06 | 66.47 | 54.40 | **85.38** | 66.46 | 77.25 |
| ALBERT | 70.21 | 78.74 | 70.43 | **63.13** | 78.90 | 70.14 | 81.62 |
| DeBERTa | **71.46** | **83.18** | **72.44** | 62.64 | 83.50 | **71.58** | **83.43** |
| *Cross-encoder* | | | | | | | |
| BERT | 85.18 | 86.01 | 83.33 | 82.40 | 83.82 | 83.11 | 90.06 |
| RoBERTa | 81.59 | 83.62 | 80.53 | 82.20 | 80.77 | 81.48 | 88.98 |
| ALBERT | **86.18** | 87.29 | 84.22 | **83.56** | 84.78 | 84.16 | 90.50 |
| DeBERTa | 86.07 | **87.64** | **84.6** | 82.70 | 85.71 | **84.18** | **91.84** |
| SEEN | 88.80 | 93.45 | 89.93 | 90.83 | 89.90 | 90.37 | 95.70 |
| UNSEEN | 82.68 | 80.47 | 78.03 | 69.98 | 78.29 | 73.90 | 87.07 |

→ The IR baselines perform much worse
→ Cross-encoder works significantly better than the bi-encoder
→ Performance drops significantly on the unseen aspects

# RG Results

- Metrics: BLEU, ROUGE, METEOR, BERTscore and response length

| | BLEU | R-1 | R-2 | R-L | MT | BS | Len |
|---|---|---|---|---|---|---|---|
| EXT | 2.89 | 23.17 | 6.53 | 18.33 | 9.62 | 30.83 | 14.93 |
| GPT2 | 9.04 | 33.9 | 13.52 | 26.73 | 16.27 | 39.73 | 22.66 |
| DialoGPT | 9.19 | 33.6 | 13.62 | 26.81 | 16.15 | 39.72 | 22.05 |
| BART | 10.8 | 36.35 | 15.04 | 28.57 | 17.96 | 41.12 | 24.02 |
| $\text{BART}_{\text{ABSA}}$ | 10.78 | 36.30 | 15.36 | 28.47 | 18.06 | **41.75** | 23.66 |
| T5 | 10.72 | 36.50 | **15.57** | 28.81 | **18.33** | 40.84 | 25.36 |
| $\text{T5}_{\text{ABSA}}$ | **10.97** | **36.66** | 15.51 | **28.88** | 18.15 | 40.94 | 24.75 |

→ Adding the ABSA predictions to RG slightly improves performance and makes responses shorter

# Human Evaluation

- Three metrics evaluated with Likert scale 1-5:
  - **Appropriateness**: whether the response is fluent and naturally connected to the dialogue context.
  - **Aspect Accuracy**: whether the response provides relevant and useful information to the aspect that the user queried.
  - **Sentiment Accuracy**: whether the sentiment proportion provided by the response is consistent with that of the subjective knowledge.
- Including ABSA increases the sentiment accuracy

| | Approp. | Asp-Acc | Senti-Acc |
|---|---|---|---|
| EXT | 2.65 | 3.32 | 3.13 |
| GPT2 | 4.55 | 4.54 | 3.20 |
| BART | 4.55 | 4.68 | 3.56 |
| $\text{BART}_{\text{ABSA}}$ | 4.58 | 4.66 | 3.80 |
| T5 | 4.40 | 4.63 | 3.87 |
| $\text{T5}_{\text{ABSA}}$ | 4.49 | 4.67 | 3.98 |
| Reference | 4.70 | 4.77 | 4.50 |

Reference is the groundtruth annotation

# End-to-End Evaluation

- Start from RG with gold knowledge as input and then gradually add the subtask outputs to the pipeline
- Since most errors happen in KS, it affects the response quality the most
- Entity tracking influences knowledge selection performance

| | KS | | RG | | |
|---|---|---|---|---|---|
| | Macro-F | mAP | BLEU | R-L | BS |
| RG | - | - | 10.80 | 28.52 | 41.12 |
| +KS | 84.60 | 91.84 | 10.20 | 27.78 | 40.64 |
| +ET+KS | 83.47 | 90.45 | 10.29 | 27.80 | 40.56 |
| +KTD+ET+KS | 83.46 | 90.45 | 10.27 | 27.79 | 40.55 |

# Qualitative Analysis

**Dialogue History:**
U: Are there any hotels (not guesthouses) that have 0 stars?
S: Yes, actually. Cityroomz is a moderately priced 0 star hotel.
U: Do they have free wifi?
S: Yes, they do have free wifi! Would you like me to book a stay for you?
U: Does the Cityroomz have strong water pressure in the shower?

**Reviews:**
Review 1: However, I noticed that when taking a shower there was very little water pressure which made the shower disappointing.
Review 2: The water pressure is not good and it comes out really slow.
Review 3: Everything else was a disaster, especially the lack of cleanliness in the room when we checked in, and the incredibly low water pressure in the bathroom when we took our morning showers.
Review 4: I encounter a problem when trying to shower and washing hands.
Review 5: There is no luggage storage service or anything like that, and the water pressure is much to be desired.

The relevant knowledge snippets are all negative

**Model Outputs:**
**GPT-2:** The Cityroomz gets some mixed reviews on their water pressure. Some guests found the water pressure to be very good while others found it to be disappointing. Would you like to know more about them?
**BART:** The Cityroomz hotel has very low water pressure according to most guests. Only one person had a good experience with it.
**$\text{BART}_{\text{ABSA}}$:** The Cityroomz has a lot of complaints about their water pressure being a bit low. Would you like me to find another place with better ratings?
**T5:** The Cityroomz has mixed reviews about their water pressure. Some guests found the water pressure to be very low while others found it to be disappointing. Would you like to know more about them?
**$\text{T5}_{\text{ABSA}}$:** The Cityroomz has a lot of reviews that say the water pressure is very low and disappointing. Do you want to look at some other places?

**Reference:**
No, guests consistently complain about the water pressure, unfortunately. Will that be okay or should I do another search?

GPT-2 and BART falsely include positive opinions in their responses, while including ABSA resolves this problem

# Factual Knowledge Model on SK-TOD

hhu
Heinrich Heine
Universität Düsseldorf

- Train a model on TOD with factual unstructured knowledge and evaluate on SK-TOD to see how transferable models are between the tasks
- The fact-TOD model performs much worse and is not able to find and *include all different opinions* in the response

  → SK-TOD is not transferable from fact-TOD

  → Fact-TOD only learns to predict and include one snippet per response

**Dialogue History:**
U: I am looking for a guesthouse to stay at that has free wifi.
S: Ok. Were you hoping for a particular price range or part of town?
U: I don't care about the part of town or the price range, but it needs to be a three star hotel.
S: The Hamilton Lodge would be perfect! It is a moderately priced 3 star guesthouse on the north side of town. Does that sound okay to you?
U: Does it have comfortable beds?

**Reviews:**
Review 1: The bed really needed a new mattress, too.
Review 2: That being said, the room was spacious and the bed was very comfortable.
Review 3: We both slept extremely well every night we were there.

**Model Outputs:**
**Fact-TOD:** Yes, the Hamilton Lodge has a new mattress. Would you like to make a reservation?
**SK-TOD:** The Hamilton Lodge has really comfortable beds according to most guests, but one guest did say that the bed needed a new mattress.

**Reference:**
There are some mixed reviews on the beds. Some say they're very comfortable while others were not impressed. Would you like me to find another place with better reviews?

| | KTD | KS | | RG | | |
|---|---|---|---|---|---|---|
| | Acc | Macro-F | mAP | BLEU | R-L | BS |
| Fact-TOD | 87.62 | 59.55 | 76.69 | 6.15 | 23.25 | 33.16 |
| SK-TOD | 99.71 | 84.60 | 91.84 | 10.80 | 28.57 | 41.12 |

# Limitations SK-TOD Data-set

- Ideally there should be **more than two domains** covered in the data
  - KTD evaluation should focus on generalisability, as the performance is likely limited to the data and domains it was trained on
- Subjective user requests are collected separately, which might lead to *different data distributions* and unrealistically high KTD performance
- Entity Tracking is easier with **less entities** in the knowledge source

# Limitations SK-TOD Data-set



- The reviews written by crowd workers might be **shorter** than those found in real world scenarios
- In a real world scenario there can be significantly **more reviews** per entity, making more efficient knowledge selection methods important
- It is not possible to do subjective **data base queries** based on the data-set and task design
  - e.g.: “I am looking for a hotel with reliable WIFI and nice atmosphere in the restaurant.”
  - Entity has to be mentioned before

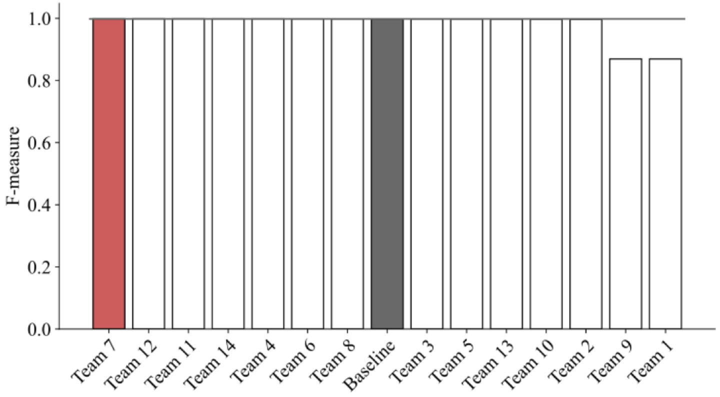
# DSTC 11 Track 5 Results

- 14 teams took part and could submit up to 5 predictions for each sub-task
- For choosing the 7 teams for final human evaluation the best average score over all the sub-tasks decided
- Entity Tracking was not part of the track evaluation
- The team with the highest human evaluation in the end won

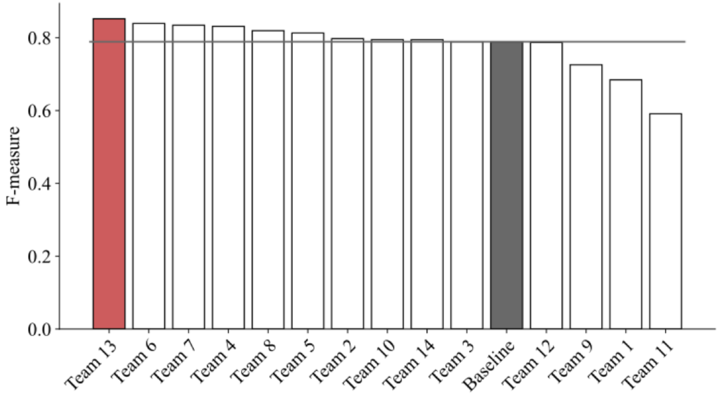
# DSTC 11 Track 5 Results

Heinrich Heine
Universität Düsseldorf

- Most teams show almost perfect performance on **KTD**
- Only half of the teams are able to significantly outperform the baseline in subjective **knowledge selection**



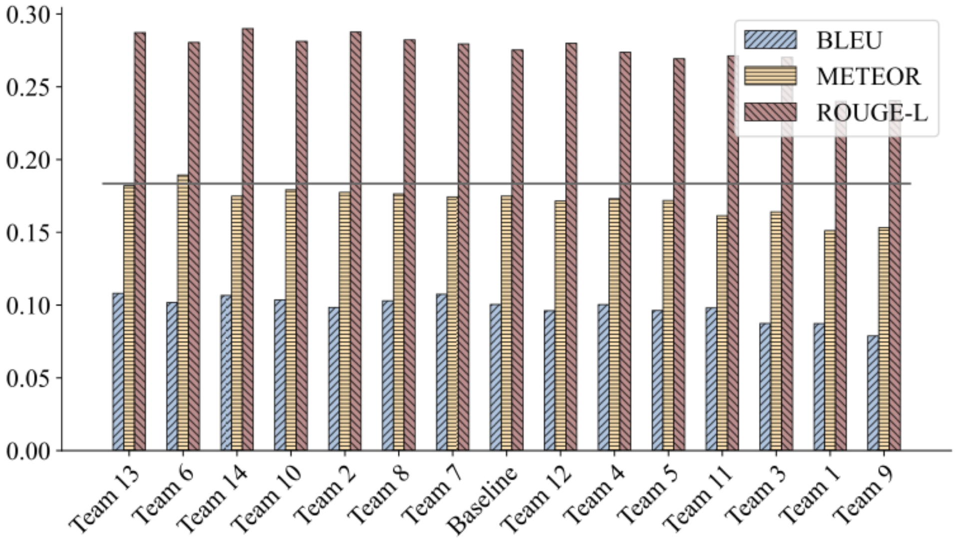

Knowledge-seeking turn detection performance (F-measure) from different entries. The horizontal line indicates the baseline performance.




Knowledge selection performance (F-measure) from different entries. The horizontal line indicates the baseline performance.

# DSTC 11 Track 5 Results

- **Response generation** results, blue line indicates average baseline results
- Only half the teams are able to improve over the baseline



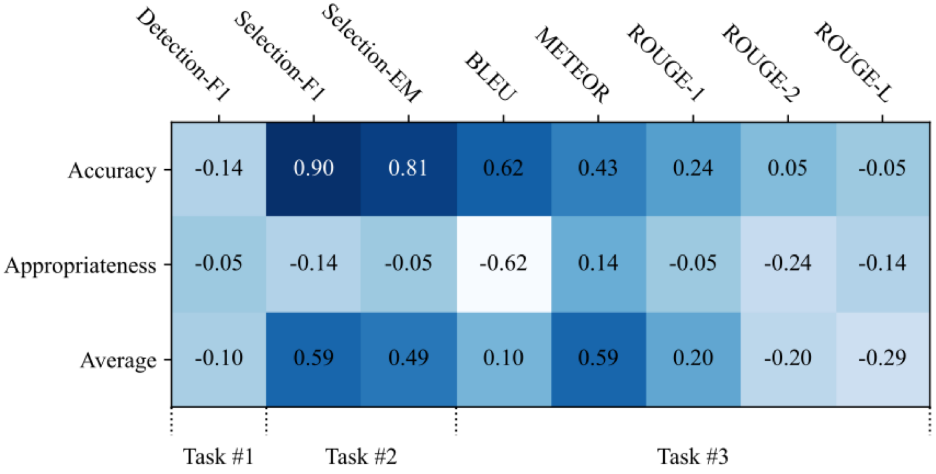
# DSTC 11 Track 5 Results

- Human Evaluation Results, bold indicates best in each metric
- Sentiment and aspect accuracy were combined in one measure
- The performance of the different entries is in a close range

| Rank | Team | Entry | Accuracy | Appropriateness | Average |
|---|---|---|---|---|---|
| Ground-truth | | | 2.9189 | 3.6422 | 3.2806 |
| 1 | 6 | 0 | 2.9095 | **3.6596** | **3.2846** |
| 2 | 8 | 0 | 2.9005 | 3.6535 | 3.2770 |
| 3 | 13 | 3 | **2.9100** | 3.6321 | 3.2710 |
| 4 | 2 | 3 | 2.8908 | 3.6487 | 3.2697 |
| 5 | 7 | 4 | 2.9046 | 3.6348 | 3.2697 |
| 6 | 12 | 2 | 2.8856 | 3.6518 | 3.2687 |
| 7 | 14 | 0 | 2.8912 | 3.6427 | 3.2670 |
| Baseline | | | 2.8715 | 3.6348 | 3.2531 |

# DSTC 11 Track 5 Results

- Spearman‘s $\rho$ **correlation of automatic metrics to human evaluation**




→ **Knowledge selection** metrics are highly **correlated to the accuracy** of the generated response, since this metric is based on the reviews

→ **No metric** is **correlated to appropriateness**, showing the need for more sophisticated metrics for SK-TOD

# DSTC 11 Track 5 Summary

- All teams have almost perfect performance on KTD
- For knowledge selection the majority of entries is better than the baseline
- For response generation only half the teams managed to improve over the baseline
- LLMs were **not the decisive factor** for performance
  - More effective ways of utilising LLMs for SK-TOD should be investigated
  - LLMs cannot perform the subtasks better if they are employed as one-step task solvers (Jung et al., 2023)
- Better automatic evaluation metrics for SK-TOD are also needed since the **correlation** with **human evaluation is low**

# Leveraging Ensemble Techniques and Metadata for Subjective Knowledge-grounded Conversational Systems

Seongho Joo*[1] Kang-il Lee*[1] Kyungmin Min*[1] Joongbo Shin[2] Janghoon Han[2] Seungpil Won[2] Kyomin Jung[1,3]
[1]Seoul National University, [2]LG AI Research, [3]SNU-LG AI Research Center
*Equal contribution.
*Proceedings of The Eleventh Dialog System Technology Challenge*, pages 206–215
September 11, 2023

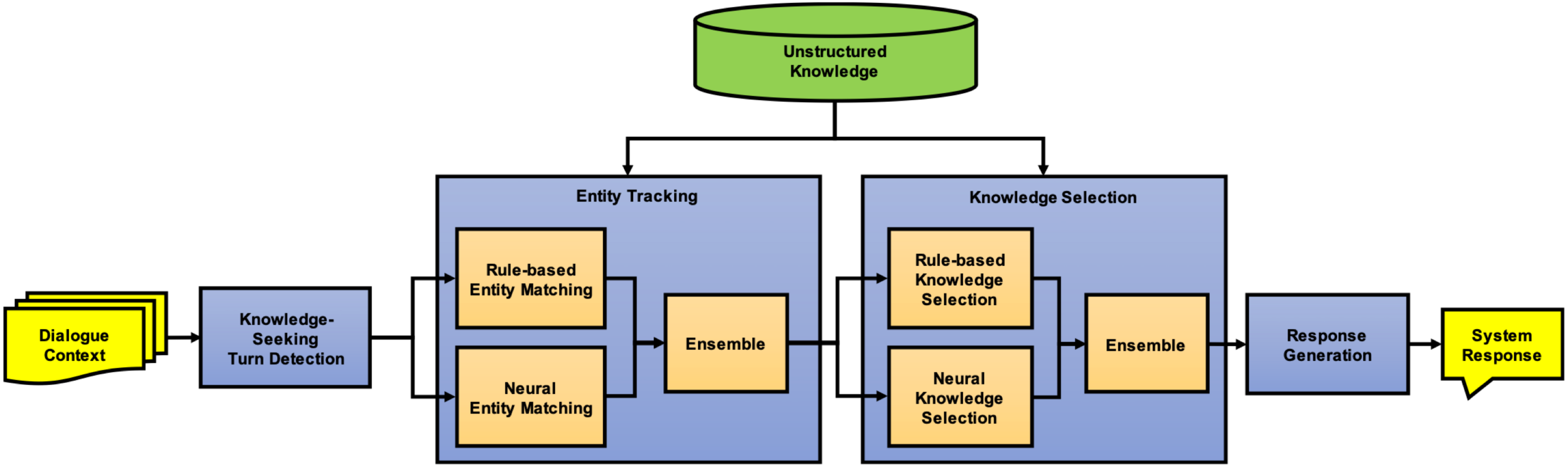
# Overview

Heinrich Heine Universität Düsseldorf

- Best Paper in DSTC11 Track 5
- First place in automatic evaluation and third place in human evaluation
- For knowledge-seeking turn detection the baseline classifier is adopted
- Heuristics to ensemble outputs from **rule-based and neural methods** for entity tracking and knowledge selection
- Utilise available **meta-data** for knowledge selection
  - offers additional information about each review, such as the type of reviewer (e.g. couples, etc.), specific dishes (e.g. beef wellington, etc.) and beverages (e.g. beer, ale, etc.)
- For response generation augment the training data using LLMs

# Architecture

Heinrich Heine
Universität Düsseldorf




For response generation LLM augmented data is utilised.

# Entity Tracking

- More accurate entity tracking leads to
  → Fewer knowledge candidates are considered in the knowledge selection step
  → Improves efficiency and precision
- A heuristic-based ensemble of a rule-based and a neural network-based entity tracker is utilised

# Entity Tracking

- **Rule-based:**
  - In each dialogue turn in context, fuzzy n-gram matching is performed with all 143 entities from the data-set
  - Fuzzy n-gram matching finds the longest contiguous matching sub-sequence between each dialogue turn and all entities and then calculates a matching ratio
  - If the matching ratio exceeds a predefined threshold, the entity and its matching turn from the dialogue history are stored
  - It was observed that *more recently mentioned entities* tend to be more relevant to the user's request

    →To tackle this only track entities from the most recent turn selected by fuzzy matching

# Entity Tracking



- **Neural network based:**
  - Model gets dialogue history and an entity as input in a sentence pair binary classification task
    → classify whether an entity is relevant in context
  - No negative samples are provided by the data-set, so the samples are created in the following ways:
    - *False positives* from the *rule-based entity matching* method are taken as **hard negatives**
    - **Similar-name negatives** are sampled from entities with at least 50% token overlap
    - **In-domain negatives** are sampled from entities from the same domain
    - **Random negatives** are randomly sampled from the whole entity list

# Entity Tracking

- Ensembling the two approaches:
  - **Entity annotation errors** are common in the data-set, especially when there are multiple relevant entities involved in a single turn → deteriorates the neural model performance
  - Rule-based entity tracking is *robust against annotation errors*, but lacks the ability to understand dialogue context and only extracts entities from the most recent dialogue turn
  - Neural entity matching can leverage understanding of the dialogue context for entity tracking
  - → **Heuristic**:
    - If the neural model only **tracks one entity** in the dialogue context, use its predictions, since it has a high level of confidence in the prediction
    - For dialogue contexts where the neural model extracts **several entities**, the output from the rule-based approach is used as prediction

# Entity Tracking

Heinrich Heine Universität Düsseldorf

- Results of entity tracking task on ***custom test set*** with the constructed negative samples
- $P$ is the prediction and $G$ the groundtruth.
- The numbers in the table represent the proportion that each case occupies in the entire test set.

| Method | Exact Match ($P = G$) | Over Prediction ($G \subset P$) | Under Prediction ($P \subset G$) | Incorrect ($(P - G \,!= \emptyset) \cap (G - P \,!= \emptyset)$) |
|---|---|---|---|---|
| Baseline | 0.9076 | 0.0351 | 0.0129 | 0.0444 |
| Rule-based | 0.9316 | 0.0377 | 0.0081 | 0.0226 |
| Neural | 0.8987 | 0.0869 | 0.0002 | 0.0122 |
| Baseline + Neural | 0.9494 | 0.0189 | 0.0129 | 0.0189 |
| Rule-based + Neural | **0.9545** | 0.0233 | 0.0085 | 0.0137 |

→ The proposed ensemble ET approach **outperforms the baseline**
→ Neural model only outperforms if it is ensembled
→ Over prediction and completely incorrect predictions are more common than under prediction

# Knowledge Selection

- Ensemble of rule-based and neural-based approaches
- Instead of only considering single review sentences as knowledge snippets, also concatenate the **previous sentence** of the review to each snippet to include **more context**

  → This is called *“consecutive knowledge snippet”*

# Knowledge Selection

- **Rule-based:**
  - User reviews mention specific dishes or drinks, which can be found in the **meta-data** of reviews
  - Fuzzy n-gram matching is used to compare the user's latest utterance with the **meta-data entity set** to decide whether to leverage metadata for the knowledge selection
  - If the metadata is found to be relevant, perform fuzzy n-gram matching between all candidate knowledge snippets and the corresponding metadata to select relevant knowledge snippets
  - For each document containing relevant metadata, a language model identifies the most suitable knowledge snippets via snippet scoring
  - The output of rule-based knowledge selection is constructed by the **union of knowledge snippets** obtained through *fuzzy n-gram* matching and those selected by the *neural model*

# Knowledge Selection



- **Neural network-based:**
  - Input: the dialogue history and the consecutive knowledge snippet
  - Relevance between a user request and a knowledge snippet is determined via binary classification on the mean pooled last hidden states of the encoding

# Knowledge Selection

Heinrich Heine Universität Düsseldorf

- Ensembling of the two approaches is based on three **heuristic rules**:
  1. If the last user utterance contains metadata, use rule-based knowledge selection approach
  2. If the metadata is mentioned in the user's last utterances but not found in any candidate knowledge snippet, there would be no snippet retrieved via fuzzy n-gram matching
     - Use results from neural knowledge selection in this case
  3. If the user's last utterance does not contain metadata, the neural model is also used

# Knowledge Selection

- Results on validation set:
  - using consecutive knowledge snippets improves performance
  - Meta-data based rules improve performance further

| Method | Precision | Recall | F-1 score | Exact match |
|---|---|---|---|---|
| Baseline-DeBERTa-base | 0.9596 | 0.9416 | 0.9505 | 0.8555 |
| Consecutive-DeBERTa-base | 0.9661 | 0.9533 | 0.9597 | 0.8662 |
| Consecutive-DeBERTa-large | 0.9626 | **0.9638** | 0.9632 | 0.8935 |
| Ensemble | **0.9714** | 0.9553 | **0.9633** | **0.9009** |

# RG Data Augmentation

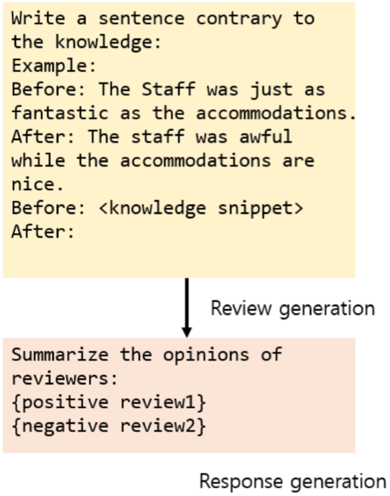


- **Main problem**: reflecting all the sentiments present in the relevant knowledge snippets
  - **augment the available data** with mixed sentiments or **label the presence of mixed sentiment** to train the model on that
- Use GPT-3 with a prompt and few-shot examples to generate *pseudo-labels* that indicate whether the sentiment of selected reviews is mixed
  - Train a model on predicting these pseudo-labels as special tokens during generation

```
Determine whether the following
reviews contain conflicting
options related to the context:
Context: <Question 1>
Reviews: <Knolwedge Snippets1>
Opinions are conflicting: true
.
.
.
Context: <QuestionN>
Revies: <Knolwedge SnippetsN>
Opinions are conflicting: false
Context: <Question>
Reviews: <Test snippets>
Opinions are conflicting:
```

# RG Data Augmentation

- The data-set seems to be biased towards simple cases where there is only one snippet relevant for the response
  - Augment the number of training examples with mixed opinions
    - Prompt GPT-3 to generate a review with a **contrary opinion** to existing ones and **summarising** all the reviews as the corresponding response
  - Using too many GPT-3 generated examples might decrease the BLEU score significantly



# Response Generation

- Evaluation of how many of the mixed opinion generation cases were not handled correctly in the generated response

| Model | BLEU | Mix failed |
|---|---|---|
| Baseline | **0.111** | 52/211 |
| 1600 Aug | 0.106 | 32/211 |
| 3200 Aug | 0.104 | 28/211 |
| 4800 Aug | 0.105 | **22/211** |
| Baseline+Pseudo | 0.103 | 53/211 |
| Baseline+Pseudo+Aug | 0.102 | 45/211 |
| T5-3B+Pseudo | 0.101 | 51/211 |
| T5-3B+Pseudo+Aug | 0.95 | 28/211 |

→ **Augmentation** of mixed sentiment examples **helps** in handling them, while the **pseudo-labels** do **not help**

# DSTC 11 Track 5 Results

Automatic Evaluation, best on average throughout all subtasks and metrics

| Method | | Task1: Turn Detection | | | Task2: Knowledge Selection | | | | Task3: Response Generation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Team ID | Entry ID | Precision | Recall | F1 | Precision | Recall | F1 | Exact Match | BLEU | METEOR | ROUGE-1 | ROUGE-2 | ROUGE-L |
| Baseline | | 0.9661 | 0.9979 | 0.9980 | 0.7901 | 0.7877 | 0.7889 | 0.3906 | 0.1004 | 0.1748 | 0.3520 | 0.1430 | 0.2753 |
| 6 | 0 | 0.9968 | **0.9996** | 0.9982 | 0.8039 | 0.8775 | 0.8391 | 0.5547 | 0.1017 | **0.1894** | 0.3629 | 0.1478 | 0.2804 |
| **13 (Ours)** | 0 | 0.9964 | 0.9982 | 0.9973 | 0.8341 | 0.8716 | 0.8524 | **0.6567** | 0.1024 | 0.1826 | 0.3638 | 0.1524 | 0.2868 |
| | 1 | 0.9964 | 0.9982 | 0.9973 | 0.8511 | 0.8581 | **0.8546** | 0.6474 | 0.1017 | 0.1830 | 0.3630 | 0.1530 | 0.2870 |
| | 2 | 0.9964 | 0.9982 | 0.9973 | **0.8590** | 0.8449 | 0.8519 | 0.6432 | 0.1017 | 0.1819 | 0.3618 | 0.1514 | 0.2865 |
| | 3 | 0.9964 | 0.9982 | 0.9973 | **0.8590** | 0.8449 | 0.8519 | 0.6432 | **0.1081** | 0.1819 | 0.3652 | 0.1528 | 0.2872 |
| | 4 | 0.9964 | 0.9982 | 0.9973 | **0.8590** | 0.8449 | 0.8519 | 0.6432 | 0.0931 | 0.1840 | 0.3591 | 0.1484 | 0.2808 |
| 14 | 0 | 0.9979 | 0.9989 | 0.9984 | 0.7856 | 0.8035 | 0.7944 | 0.4183 | 0.1066 | 0.1748 | 0.3599 | 0.1577 | 0.2899 |

# DSTC 11 Track 5 Results

- Human Evaluation, best in accuracy showing good knowledge selection and mixed opinion handling performance

| Rank | Team ID | Entry ID | Accuracy | Appropriateness | Average |
|---|---|---|---|---|---|
| Ground-truth | | | 2.9189 | 3.6422 | 3.2806 |
| 1 | 6 | 0 | 2.9095 | **3.6596** | **3.2846** |
| 2 | 8 | 0 | 2.9005 | 3.6535 | 3.2770 |
| **3** | **13 (Ours)** | **3** | **2.9100** | 3.6321 | 3.2710 |
| 4 | 2 | 3 | 2.8908 | 3.6487 | 3.2697 |
| 5 | 7 | 4 | 2.9046 | 3.6348 | 3.2697 |
| 6 | 12 | 2 | 2.8856 | 3.6518 | 3.2687 |
| 7 | 14 | 0 | 2.8912 | 3.6427 | 3.2670 |
| Baseline | | | 2.8715 | 3.6348 | 3.2531 |

# Summary

Heinrich Heine Universität Düsseldorf

- Rule-based ET helps with **annotation errors**
- Considering **meta-data** in knowledge selection improves performance
- Adding *preceding context* to knowledge snippets improves performance
- LLM pseudo-labels do not help the model to handle mixed opinions in knowledge snippets in response generation
- LLM data augmentation improves the performance on *mixed sentiment cases*

Heinrich Heine
Universität Düsseldorf

# Ensemble Method via Ranking Model for Conversational Modeling with Subjective Knowledge

Xin Huang, Kye Min Tan, Richeng Duan, Bowei Zou
Institute for Infocomm Research (I2R), A*STAR, Singapore

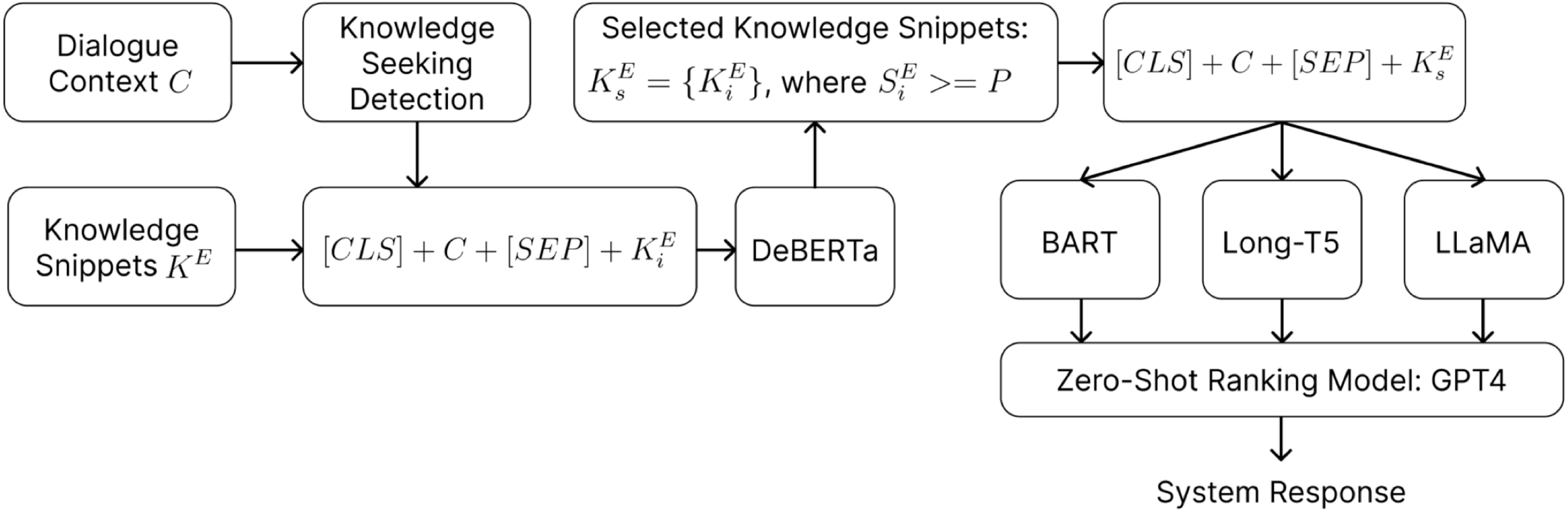
*Proceedings of The Eleventh Dialog System Technology Challenge*, pages 177–184
September 11, 2023

# Overview

- First place in ROUGE_1 score and second in ROUGE_L
- 4th place in human evaluation
- Conduct unseen domain experiments
- Knowledge-seeking turn detection and entity tracking are based on the baselines
- Knowledge selection: **adapt the score threshold** for choosing knowledge snippets based on the validation set and dynamically adapt it during inference
- Response generation: Ensemble of three different LMs for RG:
  - BART
  - Long-T5 (Raffel et al., 2020)
  - LLaMa (Touvron et al., 2023) fine-tuned on RG
- Rank the different model outputs using **scores predicted by GPT-4**

# Approach




Dialogue Context $C$
Knowledge Seeking Detection
Selected Knowledge Snippets:
$K_s^E = \{K_i^E\}$, where $S_i^E >= P$
$[CLS] + C + [SEP] + K_s^E$
Knowledge Snippets $K^E$
$[CLS] + C + [SEP] + K_i^E$
DeBERTa
BART
Long-T5
LLaMA
Zero-Shot Ranking Model: GPT4
System Response

# Unseen domain KTD

- Baseline already shows almost perfect performance
- Test the generalisation capabilities of the baseline by **masking out** one of the two **domains in training**
- Results on KTD when only training on the hotel (H) or restaurant domain (R) in training and masking the other:

| **Data** | | **Metrics** | | |
|---|---|---|---|---|
| Train | Val | P | R | F1 |
| H | H | 99.86 | 99.86 | 99.86 |
| | R | 99.33 | 86.00 | 92.19 |
| R | H | 99.91 | 79.46 | 88.52 |
| | R | 99.86 | 100.0 | 99.93 |
| All | H | 100.0 | 99.86 | 99.93 |
| | R | 99.86 | 99.86 | 99.86 |

→ KTD **recall drops** on then unseen domain, i.e. it is harder to find all subjective requests in that domain

# Knowledge Selection

- Since the number of relevant snippets varies for each instance, a threshold is utilised for knowledge selection similar to the baseline
- This threshold $P$ is chosen based on the performance on the validation set
- The optimal value for $P$ is likely **different on the test set**
- Inference:
  - If there is no knowledge snippet found for a given context, $P$ is lowered by 0.05 until **at least one** knowledge snippet is found for the input

# Knowledge Selection

- Adaptive threshold improves recall and exact match score only slightly
- There are only 31 out of 2796 instances where **no knowledge snippets** could be **found** with the fixed threshold
- Masking the domains in knowledge selection training impacts the performance even more on the unseen domain
- The performance on the **restaurant domain is generally weaker**, possibly due to limited amount of training data or the larger number of knowledge snippets and entities in the restaurant domain (33 hotels vs. 110 restaurants)

| Model | P | R | F1 | EM |
|---|---|---|---|---|
| Fixed $P$ | **77.11** | 82.01 | 79.48 | 42.39 |
| Dynamic $P$ | 77.03 | **82.17** | **79.52** | **42.47** |

Results with dynamic threshold

| Data | | Metrics | | | |
|---|---|---|---|---|---|
| Train | Val | P | R | F1 | EM |
| H | H | 72.94 | 92.99 | 81.75 | 32.34 |
| | R | 33.59 | 83.79 | 47.96 | 18.59 |
| R | H | 72.21 | 80.56 | 76.15 | 28.44 |
| | R | 73.22 | 78.90 | 75.96 | 35.73 |
| All | H | 81.64 | 94.39 | 99.93 | 43.32 |
| | R | 68.59 | 80.70 | 74.15 | 36.46 |

Results with domain masking in training

# Response Generation

Heinrich Heine
Universität Düsseldorf

- Three models used in ensemble:
  - BART is used as in the baseline, increase the maximum input knowledge token size from 256 to 512 to avoid knowledge cutoff for some turns
  - Long-T5 model is used based on hypothesis that a model with strong **summarisation capabilities** will be beneficial for combining information across multiple user reviews
  - The original LLaMa model is directly fine-tuned on the data with low rank adaptation (LoRA) (Hu et al., 2022) to save GPU resources
- For each input the responses of the three models receive quality scores by GPT-4 for ranking them

# Response Generation

Heinrich Heine
Universität Düsseldorf

- Two possibilities of ensembling the outputs of the three models are tested:
  - E1: Select the model response with the **highest GPT-4 quality score**
  - E2: Select the model response with the highest GPT-4 quality score only if the **score of the best reference model is lower than a given threshold** (use threshold $S_t = 3$)
    - The best reference model is **chosen by human evaluation** by sampling $N$ utterances

      → fine-tuned LLaMa is ultimately chosen as the reference model

# Response Generation

Heinrich Heine
Universität Düsseldorf

- Automatic metrics on the left and official DSTC 11 track 5 evaluation on the right
- The worse performance in human evaluation can be mainly attributed to worse KS performance

| **Model** | BLEU | MT | R-1 | R-2 | R-L |
|---|---|---|---|---|---|
| *Baseline* | 10.04 | 17.48 | 35.20 | 14.30 | 27.53 |
| Long T5 | **10.50** | 16.91 | 34.72 | 14.66 | 27.93 |
| BART | 10.20 | 17.50 | 35.28 | 14.34 | 27.60 |
| LLaMA | 9.44 | 17.47 | 36.52 | 14.94 | **28.75** |
| E1. | 9.84 | 17.74 | **36.58** | **15.09** | **28.75** |
| E2. ($S_t = 3$) | 10.24 | **17.76** | 36.33 | 15.08 | 28.65 |

| **Model** | **Approp.** | **Asp-Acc** | **Average** |
|---|---|---|---|
| Baseline | 3.6348 | 2.8715 | 3.2531 |
| Best Team | **3.6596** | **2.9095** | **3.2846** |
| E2. ($S_t = 3$) | 3.6487 | 2.8908 | 3.2697 |

→ E2 is only better in one metric

# Response Generation Example

- In this case, Long-T5‘s response is chosen in the ensemble, the only one to include all sentiments
- GPT-4 as a standalone model generates a response that goes beyond the groundtruth, decreasing the appropriateness

U: Is there any restaurants located in the north side of town?
S: There is one. It is da vinci pizzeria. Shall I book it?
U: Does Da Vinci Pizzeria serve portions that are plentiful?

**BART**: Yes, Da Vinci Pizzeria does serve portions that are plentiful. Past guests have said that the portions are plentiful. Would you like to know more about them?
**Long-T5**: The Da Vinci Pizzeria serves plentiful portions, but some customers have said that the portions are too small.
**LLaMA**: Yes, Da Vinci Pizzeria does serve portion sizes that are plentiful. Would you like to know more about them?
**E2.**: The Da Vinci Pizzeria serves plentiful portions, but some customers have said that the portions are too small.
**GPT-4 (standalone)**: Some customers feel that Da Vinci Pizzeria serves portions that are plentiful and satisfying, while others think the portions are relatively small. It can vary depending on personal preferences and appetites. If you enjoy hearty portions, you might want to explore other options or consider ordering extra items to ensure you’re fully satisfied.

Reference: Some past customers do feel that the portions are plentiful, but there are others who felt that they are too small.

# GPT-4 Scores as Evaluation Metric

- Calculate the Pearson correlation between GPT-4 scores and human evaluation
  → **weak correlation**

| **Model** | **Score** | | **PCC** |
|---|---|---|---|
| | **Human** | **GPT-4** | |
| BART | 3.64 | 4.12 | 0.3626 |
| Long T5 | 3.36 | 4.07 | 0.2856 |
| LLaMA | 3.88 | 4.28 | 0.2884 |
| E1. | **3.90** | 4.52 | 0.1785 |
| E2. ($S_t = 3$) | 3.82 | **4.73** | 0.3026 |

Highest correlation for BART, lowest for E1

# GPT-4 Scores as Evaluation Metric

- In case of low correlation both the human raters and GPT-4 assigned scores are **inconsistent with judging criteria**
- For future evaluations, well-defined judging criteria and experimenting with different prompts may improve accuracy and consistency of scores
- GPT-4's scoring excels in assessing certain aspects such as **grammar**
  - struggles with **understanding context-dependent nuances** that human evaluators are typically adept at capturing
- ROUGE can offer valuable insights into the correctness of text generation
  - It is crucial to **consider multiple evaluation criteria** and perspectives when evaluating response generation models

# Summary

- Further investigation is needed in cases where **unseen domains** are encountered
  - data has to be expanded to cover more than just two domains
- For more realistic set-ups **real reviews** should be incorporated into the data
- LLMs **cannot solve the problem on their own**
- **Worse performance in knowledge selection** has the biggest impact on the following models by inducing noise
- The dependency on GPT-4 as an external model is problematic
  - better have an easily accessible or trainable ranking model instead
- Using three models leads to high **training and inference costs**

# SK-TOD Summary

Heinrich Heine Universität Düsseldorf

- SK-TOD is a challenging task, crucial for developing versatile TOD models capable of using knowledge from a **variety of knowledge sources**
- The main challeng in SK-TOD is including all the **different opinions** in the subjective knowledge sources
- Automatic evaluation of SK-TOD is not ideal, as the automatic metrics only show **weak correlation** to human judgement
- LLMs cannot solve the task on their own
- The proposed data-set comes with several **limitations** regarding the coverage of domains and the number of entities and reviews

  → The data-set and task definition is a first step towards SK-TOD

Heinrich Heine
Universität Düsseldorf

# Thank you for your attention!
# Any questions?

# References




- Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access  Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, Dilek Hakkani-Tur  Amazon Alexa AI, Sunnyvale, CA, USA *Proceedings of the SIGdial 2020 Conference*, pages 278–289 1st virtual meeting, 01-03 July 2020
- LoRA: Low-Rank Adaptation of Large Language Models Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, *Published as a conference paper at ICLR 2022*
- Haein Jung, Heuiyeen Yeen, Jeehyun Lee, Minju Kim, Namo Bang, and Myoung-Wan Koo. 2023. Enhancing Task-Oriented Dialog System with Subjective Knowledge: A Large Language Model-based Data Augmentation Framework. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 150–165, Prague, Czech Republic. Association for Computational Linguistics.
- Hyunmi Baek, JoongHo Ahn, and Youngseok Choi. 2012. Helpfulness of online consumer reviews: Readers’ objectives and review cues. *International Journal of Electronic Commerce*, 17(2):99–126.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019 Language models are unsupervised multitask learners.
- 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 nat- ural language generation, translation, and comprehen- sion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

# References II


- Raffel, Colin, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li and Peter J. Liu. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” *J. Mach. Learn. Res.* 21 (2019): 140:1-140:67.
- Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave and Guillaume Lample. “LLaMA: Open and Efficient Foundation Language Models.” *ArXiv* abs/2302.13971 (2023): n. pag.