Combining Structured and Unstructured Knowledge in an Interactive Search Dialogue System

Svetlana Stoyanchev[1] **Suraj Pandey**[2] **Simon Keizer**[1]
Norbert Braunschweiler[1] **Rama Doddipatla**[1]
[1] Speech Technology Group - Toshiba Europe Ltd. Cambridge, UK
[2] The Open University, Milton Keynes, UK
[1]{firstname.lastname}@crl.toshiba.co.uk
[2] surajjung@gmail.com

Abstract

Users of interactive search dialogue systems specify their preferences with natural language utterances. However, a schema-driven system is limited to handling the preferences that correspond to the predefined database content. In this work, we present a methodology for extending a schema-driven interactive search dialogue system with the ability to handle unconstrained user preferences. Using unsupervised semantic similarity metrics and text snippets associated with the search items, the system identifies suitable items for the user's unconstrained natural language query. In a crowd-sourced evaluation, the users were asked to chat with our extended restaurant search system. Based on objective metrics and subjective user ratings, we demonstrate the feasibility of using this unsupervised low latency approach to extend a schema-driven search dialogue system to handle unconstrained user preferences.

1 Introduction

We extend a schema-driven dialogue system with the ability to handle unconstrained user queries and to allow users to specify preferences flexibly as they would when using a search engine.

Interactive search dialogue systems, such as search for restaurants, hotels, trains, books, shows, venues, are task-oriented systems that provide a natural language interface for interactive search and information extraction. In these systems, a user typically starts by typing (or speaking) a search query. Next, the system's policy chooses an optimal action, which may be either asking the user to provide additional information or presenting one or more search result options. Once the system presents an option, the user may provide another query, narrowing down or changing the search criteria, or ask a question about the presented option(s).

In a schema-guided approach to designing a dialogue interface (Rastogi et al., 2020a), a set of 'informable' and 'requestable' slots derived from the fields of the underlying database table (or schema), define the natural language interface capability. A user can specify the values of 'informable' fields as search criteria and ask questions to retrieve information stored in the 'requestable' fields. The schema-guided method simplifies authoring dialogue systems for new domains. With this approach, a dialogue interface for a new database may be bootstrapped from the schema/ontology and database content of the domain.[1]

One of the drawbacks of the schema-guided approach is that the criteria by which the user can search for an item and the types of questions that the user may ask are limited by the database schema. For example, a restaurant search query '*Find a romantic place that serves great wines*' cannot be handled by a schema-driven system unless the schema includes the relevant properties of the restaurant atmosphere and wine quality. It is possible to design a system to notify the user of its limitations using help messages (Komatani et al., 2005), but the constraint on the interaction remains, as the user is unable to retrieve items using criteria other than those defined by the 'informable' slots. Given that in many domains, additional unstructured information beyond the database fields may be available, it is natural to extend schema-guided systems to use this unstructured information. Kim et al. (2020) describe a system that extends the schema-guided functionality with the ability to ask follow-up questions. In this work, we propose to extend the information search dialogue interface functionality to retrieve items for unconstrained user queries.

To handle user queries that cannot be grounded in terms of a domain schema and ontology, we propose to use semantic similarity metrics to retrieve search results from unstructured data. We evaluate the proposed approach with crowd-sourced

---

[1] The schema/ontology refers to the definition of the database tables.

Proceedings of the SIGdial 2022 Conference, pages 531–540
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022 ©2022 Association for Computational Linguistics

users interacting with the restaurant search system through a chat interface. In previous work, restaurant search systems were evaluated by giving users predefined 'goals' which primed users and lead to rigid interactions. In our evaluation, the users are given a general instruction to find an ideal restaurant and are free to specify any search query. The results show the users' preference for the proposed flexible system that allows the use of unconstrained search queries. We release the dialogues with the automatically annotated intents and the subjective user judgements collected during the evaluation to the research community.[2]

2 Related Work

Interactive search can be modelled as task-oriented dialogue using structured knowledge, symbolic dialogue state representation, and a statistical policy that addresses both task and conversational phenomena, such as clarifications and social dialogue acts (Budzianowski et al., 2018; Yan et al., 2017). However, users of dialogue systems that are based only on structured knowledge are limited in expressing their preferences by the underlying database schema. In response to an out-of-schema user request, a task-oriented dialogue system may produce an informative help message guiding the user to adapt to its limitations (Komatani et al., 2005; Tomko and Rosenfeld, 2004). Alternatively, system capabilities may be extended beyond a domain API. For example, Kim et al. (2020) proposes a method for handling user's follow-up questions in task-oriented dialogue systems. To support pragmatic interpretation, Louis et al. (2020) explores users' indirect responses to questions. To extend a task-oriented system to handle natural preferences, a corpus of natural requests for movie preferences was collected using preference elicitation (Radlinski et al., 2019).

End-to-end approaches to dialogue, where the system generates a response without explicitly modelling intent or storing a dialogue state, have been successfully applied to open-domain chitchat (Serban et al., 2016). The use of unstructured knowledge was shown to improve open domain chitchat systems (Dinan et al., 2018; Ma et al., 2022; Zhou et al., 2018). In recent work, interactive search has been modelled as end-to-end generation task using text and images as the knowledge source (Varshney and Anushkha Singh, 2021).

Search tasks in natural human-human dialogues can be complex and are often resolved interactively (Trippas et al., 2017, 2018), motivating the need for methods capable of handling natural conversational phenomena as well as extracting information and generating knowledge-grounded responses. In this work, we evaluate a task-oriented dialogue system with a semantic-level policy extended with the use of unstructured knowledge.

Task-oriented dialogue systems require accurate models to extract information from unstructured text. Pretrained transformer models, such as BERT (Devlin et al., 2019), have shown to be effective in extracting information from text, leading to significant improvements on many NLP tasks, including open-domain question answering, FAQ retrieval, and dialogue generation (Wang et al., 2019; Sakata et al., 2019; Kim et al., 2020). Following previous work, we use BERT in an unsupervised setting to extract relevant information from text (Izacard and Grave, 2021; Zhan et al., 2020).

3 Method

3.1 System Overview

We implement a schema-driven restaurant search dialogue system that uses a database with 422 restaurants in Cambridge, UK.[3] Following the database schema used in previous work (Henderson et al., 2014), each restaurant is described in terms of the following attributes: *name*, *cuisine*, *price range*, *area*, *phone number*, and *address*. As in previous systems, the price range is mapped to *cheap, moderate*, or *expensive* and location to *north, south, east, west,* or *city centre*. In contrast to the schema in the DSTC2 domain (Henderson et al., 2014), *cuisine* in our database is mapped to a list of values rather than a single value for each entity. In addition to the specific restaurant attributes, each restaurant is associated with a set of text snippets including *meals* (breakfast, lunch, dinner), *special diets* (e.g., vegan, gluten free) and *reviews*. Only positive reviews (rating 4 or 5 stars) are used, as we expect user queries to mention desirable properties of the restaurant. 99% of the snippets are reviews and the average number of text snippets per restaurant is 147, varying between 2 and 1.6K.

---

[2] `https://github.com/sstoyanchev/Unstructured_restaurant_search_dialogues_Sigdial2022`

[3] The database is compiled by crawling the Web in January 2021.

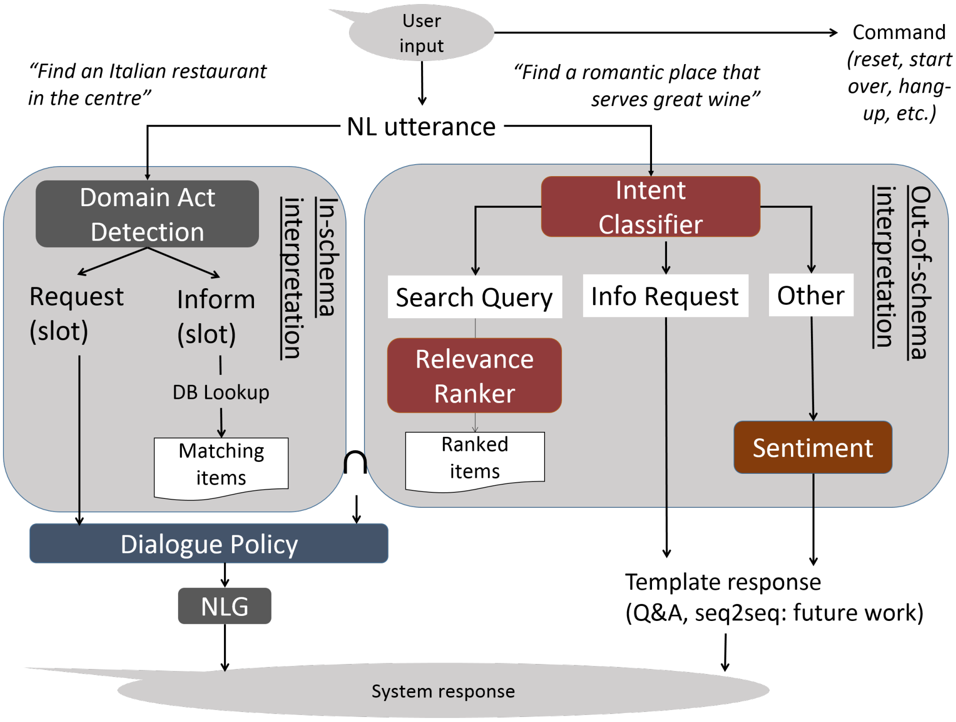


Figure 1: System diagram showing processing of in-schema and out-of-schema user input.

Figure 1 outlines the system architecture. The left-hand side shows the components handling in-schema user acts, such as requesting and providing information for one of the restaurant-specific attributes. The Domain Act Detection module interprets in-schema user acts (Stoyanchev et al., 2021) and a database lookup results in a new list of matching restaurants. A statistical dialogue policy component trained in simulation in the purely schema-driven DSTC2 domain generates the system response for in-schema user utterances (Keizer et al., 2021). The right-hand side of Figure 1 shows the components for handling out-of-schema utterances that do not mention any of the schema-specific attributes, e.g. *'Find a romantic place that serves great wine'*. In (Kim et al., 2020), the authors build a binary model that determines whether to access unstructured data for follow-up question answering. In the proposed system, a prediction of the intent classifier triggers access to the unstructured dataset of restaurant reviews.

3.2 Intent Classification

A task-oriented dialogue system is designed to handle generic dialogue acts and domain-specific intents. Dialogue act taxonomies (Core and Allen, 1997; Bunt et al., 2010) distinguish general purpose acts based on the surface form of the utterance, such as inform or question. However, in interactive search dialogue, the user can formulate a query either with a question, *'Can you find me...?'* or a statement *'I would like...'*. Hence, a distinction between the surface forms is not sufficient and instead, we define the intents specific to the search task *Search Query* (SQ), *Info Request* (IRq), and *Other*. Utterances labeled as SQ include initial and follow-up queries triggering information extraction and resulting in a retrieved set of items. Utterances labeled as IRq include information seeking requests related to one of the restaurants in context, e.g. *'Are dogs allowed?'*, which may trigger a question answering module.[4] The *Other* class includes utterances that are neither SQ nor IRq, for example, an exclamation *'Great!'*. While these utterances do not trigger data access, it is important to detect them and respond appropriately to, in order to maintain fluent conversation.

We tune the pre-trained uncased BERT transformer model (Devlin et al., 2019) on this 3-way classification task. Table 1 shows the statistics for the training dataset. We obtain the initial training dataset from two publicly available task-oriented dialogue corpora: Schema Guided Dialogue (SGD) and Frequently Asked Questions (FAQ) (Rastogi et al., 2020b; Kim et al., 2021). SGD contains semi-

---

[4] Question answering from unstructured data for this system remains future work.

Intent	**Initial Dataset**		**Collected with the system**	**Overall**	**Average #words±stdev**
	SGD (1,698)	**FAQ** (2,198)	(554)	(4,450)	
Search Query (SQ)	72%	-	47%	33%	11.124±5.794
Info Request (IRq)	-	100%	41%	54%	7.186±2.21
Other	28%	-	12%	12%	4.458±2.026

Table 1: Statistics of the dataset used to train the intent classifier showing the numbers of utterances extracted from the schema-guided dialogue corpus (SGD), DSTC9 Beyond Domain APIs track (FAQ), and collected with the system.

automatically generated task-oriented dialogues in 26 domains, including restaurant search, annotated with dialogue acts. We confirm that the initial utterances in the restaurant search domain are search queries and use them as the training examples for the SQ class[5]. Since the utterances in the SGD dataset are authored by people, they include a wide variety of queries outside of our system's domain schema. We use the utterances from the restaurant search domain in SGD annotated as 'AFFIRM', 'NEGATE', or 'SELECT' for the *Other* class. The FAQ dataset includes 2.2k manually authored questions in the restaurant search domain for the Beyond Domain APIs track of The Ninth Dialog System Technology Challenge (Kim et al., 2021; Budzianowski et al., 2018). We use the questions as examples for the IRq class.

Next, we train an intent classifier using the data from the SGD and FAQ datasets and evaluate the system internally. We collect an additional 554 utterances where the authors and colleagues interacted with the system using a web-based chat interface. As the initial set of SQ did not contain any follow-up queries, the intent classifier tended to fail on such utterances. We manually annotate the collected utterances with the dialogue act label and include them in the training set.

3.3 Relevance Ranking

The Relevance Ranker accesses unstructured data producing a ranked list of candidate items (restaurants in Cambridge) for a user's search query. The unstructured data includes reviews and restaurant details extracted from the Web, stored as text snippets associated with each item. Restaurants with snippets that have higher semantic similarity to the user query are more likely to be relevant for the user (see Table 2).

Query: I am looking for a place that serves vegan food and also allows dogs inside.

Relevant snippets	
Special diets	Vegan friendly
Review	It was such a happy surprise that they allowed dogs inside their premises.

Table 2: Query and relevant snippet examples

First, we score each snippet with the semantic similarity according to the user's query. In previous work, we have shown that a supervised model based on BERT encoding and trained on in-domain data achieves F1=.86 on the binary task of identifying relevant query-snippet pairs (Pandey et al., 2021). However, using such a model is computationally expensive as it requires 60k snippets to be classified during run-time, making it intractable for a real-time system. Instead, we use a less accurate low latency approach which achieves F1=.66 on this task.

To measure semantic similarity between the query and the snippet, the user query (Q_i) and each snippet (S_j) are mapped into a fixed-sized vector using an encoding function E. The cosine similarity score between the user request and each snippet is used to measure the semantic similarity:

$$Score(Q_i, S_j) = cos(E(Q_i), E(S_j)) \quad (1)$$

As the encoder, we use pretrained Sentence-BERT (SBERT) optimized on the semantic similarity task and further tuned on the SNLI corpus of semantic entailment which was previously shown to improve sentence classification performance (Reimers and Gurevych, 2019; Bowman et al., 2015). The intuition is that the tuned model's capability will extend to capture not only semantic similarity between the encodings but also the entailment, which may be a relation between search

---

[5] 30% of the initial SGD utterances were manually examined to confirm that they correspond to search queries.

query and a relevant snippet. SNLI is a collection of 570,000 sets of premises and hypotheses sentences annotated with the labels *contradiction*, *entailment*, and *neutral* as in the example in Table 3. We use the pairs of *Premise&Entailment* as positive examples and *Premise&Contradiction/Neutral* as the negative examples to further tune the SBERT model.

Premise:	A boy is jumping on skateboard in the middle of a red bridge.
Entailment:	The boy does a skateboarding trick.
Contradiction:	The boy skates down the sidewalk.
Neutral:	The boy is wearing safety equipment.

Table 3: Example from SNLI dataset

Next, the items (restaurants) are ranked based on the average score of the top M snippets for each item. The top N items are returned.[6]

A user's search query may specify a schema-specific attribute as well as additional information. For example, a query *'Italian restaurant with great desserts'* specifies a food type (Italian) as well as an out-of-schema preference (great desserts). Such queries are processed both by in-schema and out-of-schema modules. The in-schema processing narrows down the set of results to *Italian* and the out-of-schema processing ranks the restaurants based on the snippets' similarity with the query. The result is the ranked list of Italian restaurants where the restaurants with the snippets mentioning the high quality of desserts (if there are any) are at the top.

3.4 System Response

If at least one domain-specific user action (inform-slot or request-slot) is detected in the user's utterance, the utterance is considered in-schema. The system's response to an in-schema utterance may be an *Offer* (e.g., *'Zizzi is an Italian restaurant in the centre'*), a clarification (e.g., *'Did you mean in the centre?'*), or a request for additional information (e.g., *'What price range do you prefer?'*). The response act is selected using a statistical policy which maximizes the expected reward, and the surface form of the response is generated using templates. The out-of-schema user utterances do not mention any of the slots and can not be directly handled with the policy trained on a schema-driven dataset. The system uses the prediction of the intent classifier to determine the method of response selection.

When an out-of-schema utterance is classified as search query, the system updates the state with an inform action and the dialogue policy selects the response act. If the *Offer* act is selected, the system presents the top-ranked restaurant and its top-matching review is appended to the template-generated description. See Figure 2(b) for examples of system responses.

While searching for an item, a user may ask questions about the previously discussed items (restaurants). The intent classifier labels such questions as *Info Request*, differentiating them from the search queries produced in a question form. If the user requests information about one of the schema attributes (phone number, address, price range, etc.), the statistical policy determines the system's response. However, a user may also ask for information outside of the database schema, such as *'Are dogs allowed inside?'*. Currently, the system informs the user that the question cannot be answered signalling understanding of the user's intent. In future work, we plan to handle such questions with a question answering model, e.g. following the approach proposed in (Kim et al., 2020).

According to the initial data collection with the dialogue system, 12% of user utterances are neither search queries nor information requests (see Table 1). These utterances are typically exclamations like *'Sounds great!'* or *'Too bad!'* and are labeled as *Other* by the intent classifier. The system responds to these utterances by generating a sentiment-appropriate template-based response. A user utterance labeled as *Other* is processed with an off-the-shelf RoBERTa model trained on $\sim$ 58M tweets and fine-tuned for sentiment analysis with the TweetEval benchmark (Barbieri et al., 2020). The model outputs either a positive, negative, or neutral sentiment class of the input text. Based on the predicted sentiment, the system selects one of the appropriate template responses, e.g. *Brilliant! How else can I help you?* for a positive sentiment or *'OK, calm down now.'* for a negative sentiment, maintaining the dialogue flow and adding a bit of template-based humour.

---

[6] We use empirically chosen M=5 and N=5 in this work.

4 Evaluation

Our goal is to evaluate the use of unstructured data in an interactive task-oriented dialogue system. The proposed approach involves two statistical models: 1) intent classifier and 2) relevance ranker, which accesses unstructured data, depending on the prediction of the intent classifier. We first show the performance of the intent classifier on the collected dataset and then describe the human evaluation of the overall system.

4.1 Intent Classification

Model	Train/Test data	Accuracy
SVM	All	88.2%
BERT	All	**99.8%**
BERT	Initial/Collected	70.9%

Table 4: Intent classification accuracy.

We evaluate the intent classification performance using the dataset described in Table 1. We compare the performance of the BERT model with bag-of-words SVM baseline using stratified 10-fold cross validation on the full dataset of 4.5K utterances. The BERT model and the SVM model achieve 99.8% and 88.2% overall accuracy (see Table 4).[7] This shows that the pre-trained transformer model is able to effectively capture the distinction between the three intent classes in the restaurant search domain. The intent classifier trained on the initial data subset (the utterances from SGD and FAQ) achieved accuracy of only 70.9% on the utterances collected with the dialogue system, which is not sufficient for the interaction with the real users.

4.2 Human Evaluation

4.2.1 Experimental Setup

System	Intent classifier	Relevance Ranker	Snippet in Offer
SCHEMA	-	-	-
RAND-RANK	+	-	+
EXP-RANK	+	+	+

Table 5: Experimental Conditions.

When evaluating schema-driven dialogue systems with recruited experiment participants, the users are typically given concrete goals, e.g. *'You are looking for a cheap Italian restaurant and don't care about the price range'*. Since the proposed system handles unconstrained queries, we instruct the users to imagine a situation like going out with a family, a romantic date, a business lunch, or use any other preferences. Given more general instructions, the user can come up with authentic in-schema or out-of-schema search preferences. The users are instructed to retrieve at least three restaurant options and ask for the address of their preferred restaurant. In all experimental conditions, the system initiated the dialogue with a general prompt: *'Please describe the restaurant you would like'*. The users are instructed to use the command 'start over' if they feel that the system is stuck.

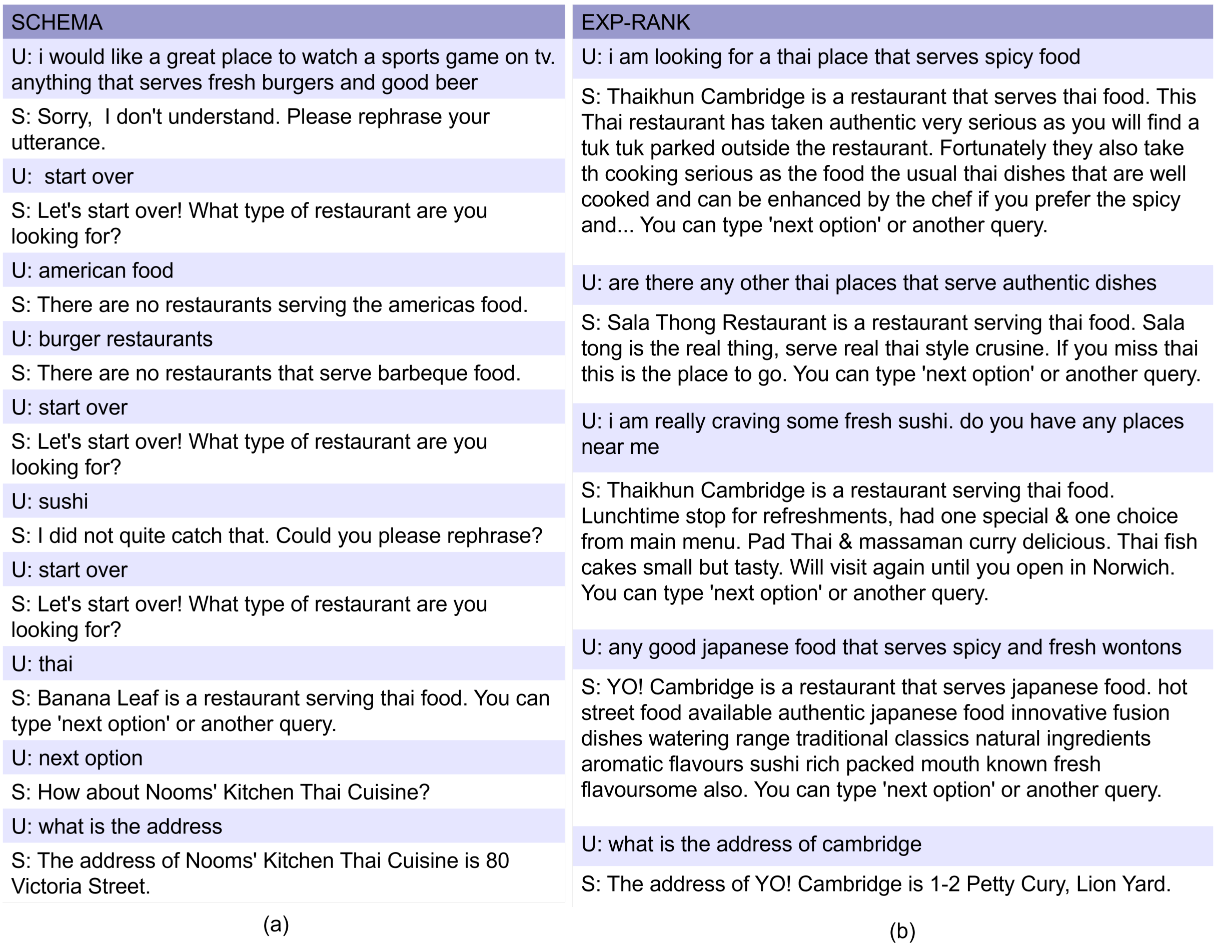
We evaluate three system variants:

- SCHEMA: the baseline condition where the system handles only in-schema user queries. In the initial prompt, the system informs the user of its limitations: *'You can search by food type, price range, or area (centre, north, etc)'*. For the out-of-schema user queries the system responds with the indication of non-understanding, e.g. *'Sorry, I don't understand. Please rephrase'*.
- RAND-RANK: the control condition where the system responds to the user's out-of-schema requests with a randomly selected search result.
- EXP-RANK: the experimental condition where the system that handles unconstrained requests using the method described in Section 3.

Table 5 summarizes the differences between the three systems. The SCHEMA baseline does not use the intent classifier and the relevance ranker. Its *Offer* act does not include the text snippet associated with the restaurant. RAND-RANK is used as the control condition to isolate the effect of the relevance ranker. The RAND-RANK system uses the intent classifier and includes a snippet in the system's *Offer* dialogue act. However, it assigns random relevance scores to the text snippets resulting in a random selection of the proposed restaurant. If RAND-RANK and EXP-RANK receive similar user ratings, the preference over SCHEMA may be due to the intent classifier and snippets in the offer output. We hypothesize that the users

---

[7] This result was achieved after 5 epochs.



SCHEMA

U: i would like a great place to watch a sports game on tv. anything that serves fresh burgers and good beer

S: Sorry, I don't understand. Please rephrase your utterance.

U: start over

S: Let's start over! What type of restaurant are you looking for?

U: american food

S: There are no restaurants serving the americas food.

U: burger restaurants

S: There are no restaurants that serve barbeque food.

U: start over

S: Let's start over! What type of restaurant are you looking for?

U: sushi

S: I did not quite catch that. Could you please rephrase?

U: start over

S: Let's start over! What type of restaurant are you looking for?

U: thai

S: Banana Leaf is a restaurant serving thai food. You can type 'next option' or another query.

U: next option

S: How about Nooms' Kitchen Thai Cuisine?

U: what is the address

S: The address of Nooms' Kitchen Thai Cuisine is 80 Victoria Street.

(a)

EXP-RANK

U: i am looking for a thai place that serves spicy food

S: Thaikhun Cambridge is a restaurant that serves thai food. This Thai restaurant has taken authentic very serious as you will find a tuk tuk parked outside the restaurant. Fortunately they also take th cooking serious as the food the usual thai dishes that are well cooked and can be enhanced by the chef if you prefer the spicy and... You can type 'next option' or another query.

U: are there any other thai places that serve authentic dishes

S: Sala Thong Restaurant is a restaurant serving thai food. Sala tong is the real thing, serve real thai style crusine. If you miss thai this is the place to go. You can type 'next option' or another query.

U: i am really craving some fresh sushi. do you have any places near me

S: Thaikhun Cambridge is a restaurant serving thai food. Lunchtime stop for refreshments, had one special & one choice from main menu. Pad Thai & massaman curry delicious. Thai fish cakes small but tasty. Will visit again until you open in Norwich. You can type 'next option' or another query.

U: any good japanese food that serves spicy and fresh wontons

S: YO! Cambridge is a restaurant that serves japanese food. hot street food available authentic japanese food innovative fusion dishes watering range traditional classics natural ingredients aromatic flavours sushi rich packed mouth known fresh flavoursome also. You can type 'next option' or another query.

U: what is the address of cambridge

S: The address of YO! Cambridge is 1-2 Petty Cury, Lion Yard.

(b)

Figure 2: Example dialogue for the SCHEMA and EXP-RANK systems.

prefer EXP-RANK over both RAND-RANK and SCHEMA.

	SCHEMA	RAND-RANK	EXP-RANK	Overall
#Dlg	81	81	81	243
#Utts	636	610	557	1803
SQ	47.3%	32.5%	34.8%	38.4%
IRq	25.3%	36.1%	32.9%	31.3%
% Out-of-schema				
SQ	44.5%	47.5%	29.4%	41.1%
IRq	11.8%	28.2%	11.5%	14.2%

Table 6: Statistics for the user utterances based on the automatic predictions. % of Search Query (SQ) and Info Request (IRq); % of in-schema SQ and IRq utterances.

4.2.2 Data

We collected 243 dialogues (81 for each system variant), with a total of 1,803 user utterances summarized in Table 6. 38.4% are classified as search queries and 31.3% are classified as information requests. The SCHEMA system is unable to process out-of-schema queries, leading to the longer dialogues where users have to change their initial query.

41.1% of the search queries are out-of-schema (no slots are detected) indicating that the users' preferences constructed without specific instructions are likely to mention information other than price range, area, and food type. Surprisingly, we find that in EXP-RANK system, only 29.4% of search queries are out-of-schema. We notice that the users tend to provide queries with both in-schema and out-of-schema info, e.g. *'I'd like to find a mexican restaurant that has excellent customer service'* which are considered in-schema yet they can benefit from the use of relevance ranking.

Despite the instructions given to the users to find out the address of their preferred restaurant, 14.2% of information requests ask about the information outside of the schema. This shows the need for the more flexible question answering capability. Example dialogues with the SCHEMA and the

Question	SCHEMA	RAND-RANK	EXP-RANK
Self-Reported Subjective Ratings			
I was able to find a satisfactory restaurant option.	4.086±1.769	4.358±1.559	**4.888±1.423**†
The restaurant descriptions matched my preferences.	4.074±1.787	4.308±1.578	**4.876±1.354**†
The system understood me well.	3.592±1.909	3.753±1.684	**4.518±1.696**†
The conversation felt natural.	3.666±1.936	3.679±1.723	**4.209±1.633**
I would recommend the system to my friends.	3.358±2.020	3.543±1.837	**4.320±1.808**†
Objective Metrics			
Average dialogue length (# exchanges)	7.9	7.5	**6.9**
Success rate	80.2%	91.4%	**95.1%**
‘Start over’ rate	3.8%	**1.5%**	**1.5%**

Table 7: Average scores and standard deviation for the subjective user judgements and objective metrics. † indicates a statistical significance with the SCHEMA condition ($p < .05$). Success rate is the % of the dialogues where a user made a choice of a restaurant in the questionnaire.

EXP-RANK systems are shown in Table 2.

4.2.3 Results

We asked the users to score each dialogue on a scale from 1 (strongly disagree) to 6 (strongly agree) for the five subjective statements shown in Table 7. For all statements, the users prefer the EXP-RANK over SCHEMA and over RAND-RANK. The difference in the scores between SCHEMA and EXP-RANK is statistically significant based on the two-tailed t-test with p<0.05 for all statements (except for *‘The conversation felt natural’*). The biggest difference (nearly 1 point) between the scores of EXP-RANK and SCHEMA systems was observed for the questions *‘The system understood me well’* and *‘I would recommend the system to my friends’*. We did not observe a significant difference in subjective ratings between the RAND-RANK and SCHEMA systems. These results suggest that relevance ranking together with the intent classification and additional information in the system response lead to the higher user rating.

We also report the objective scores: average dialogue length, success rate, and ‘start over’ rate. The dialogues with the EXP-RANK system are the shortest, while the dialogues with the SCHEMA system are the longest, on average 6.9 and 7.9 exchanges respectively. This result shows that users were able to complete the task quicker using the EXP-RANK system than the baseline systems.

In the questionnaire, the users were asked to record the name of their preferred restaurant or ‘None’ if no restaurants matched their preference. Success rate is the % of the dialogues where the user indicated a preferred restaurant name in the questionnaire. EXP-RANK system achieves the highest success rate of 95.1% in comparison with 91.4% and 80.2% for the RAND-RANK and SCHEMA conditions.

The users had an option to use *‘start over’* command when they felt stuck in the dialogue. We observe a higher proportion of *‘start over’*s in the SCHEMA system than in the other two systems which use intent classifier and return a suggestion for out-of-schema response leading to fewer non-understanding system responses. This result indicates a functional improvement of the RAND-RANK over the SCHEMA system, which, however, was not reflected in the users’ subjective ratings.

5 Conclusions

In this work, we propose a hybrid design for information navigation dialogue systems combining structured and unstructured knowledge. We present a restaurant search dialogue system where the users specify preferences flexibly as they would to a search engine. The proposed system uses the structured knowledge in a database to extract matching restaurants when a user’s natural language search query mentions one of the database fields and unstructured text when it does not. The system is evaluated in the interactive experiments with crowdsourced users. The results show a preference for the proposed approach. In future work, we will extend the system to answer follow-up questions and introduce a response generation model.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium.

Harry Bunt et al. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Mark G. Core and James F. Allen. 1997. Coding dialogs with the damsl annotation scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Cambridge, MA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *In NAACL-HLT*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *CoRR*, abs/1811.01241.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual SIGdial Meeting on Discourse and Dialogue*, pages 263–272.

Gautier Izacard and Édouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880.

Simon Keizer, Norbert Braunschweiler, Svetlana Stoyanchev, and Rama Doddipatla. 2021. Dialogue strategy adaptation to new action sets using multi-dimensional modelling. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*, pages 977–983. IEEE.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.

Seokhwan Kim, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, and Dilek Hakkani-Tur. 2021. Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access track in dstc9. *arXiv preprint arXiv:2101.09276*.

Kazunori Komatani, Shinichi Ueno, Tatsuya Kawahara, and Hiroshi G. Okuno. 2005. User modeling in spoken dialogue systems to generate flexible guidance. *User Model. User Adapt. Interact.*, 15(1-2):169–183.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. “I'd rather just go to bed”: Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, Online. Association for Computational Linguistics.

Longxuan Ma, Mingda Li, Wei-Nan Zhang, Jiapeng Li, and Ting Liu. 2022. Unstructured text enhanced open-domain dialogue system: A systematic survey. *ACM Trans. Inf. Syst.*, 40(1):9:1–9:44.

Suraj Pandey, Svetlana Stoyanchev, and Rama Doddipatla. 2021. Towards handling unconstrained user preferences in dialogue. In *Proceedings of The 12th International Workshop on Spoken Dialog System Technology (IWSDS)*.

Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020a. Schema-guided dialogue state tracking task at DSTC8. In *Proceedings of the AAAI Dialog System Technology Challenges Workshop*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020b. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. 2019. Faq retrieval using query-question similarity and bert-based query-answer relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1113–1116.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*.

Svetlana Stoyanchev, Simon Keizer, and Rama Doddipatla. 2021. Action state update approach to dialogue management. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7398–7402.

Stefanie Tomko and Roni Rosenfeld. 2004. Speech graffiti vs. natural language: Assessing the user experience. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 73–76, Boston, Massachusetts, USA. Association for Computational Linguistics.

Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the design of spoken conversational search: Perspective paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 32–41, New York, NY, USA. ACM.

Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 325–328, New York, NY, USA. ACM.

Deeksha Varshney and Asif Ekbal Anushkha Singh. 2021. Knowledge grounded multimodal dialog generation in task-oriented settings. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 421–431, Shanghai, China. Association for Computational Lingustics.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage bert: A globally normalized bert model for open-domain question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5881–5885.

Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4618–4625, San Francisco, California, USA. AAAI Press.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. *An Analysis of BERT in Document Ranking*, page 1941–1944. Association for Computing Machinery, New York, NY, USA.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A dataset for document grounded conversations. *CoRR*, abs/1809.07358.