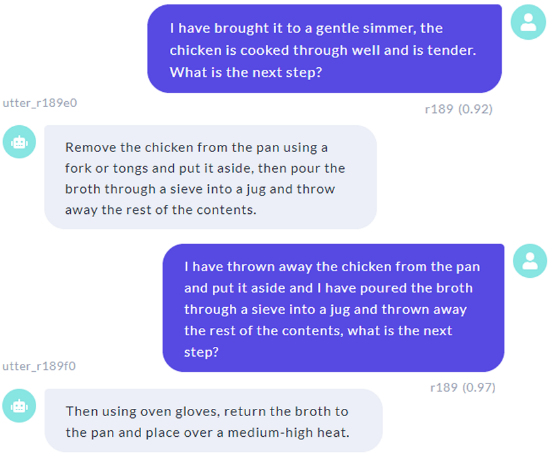
The Task2Dial Dataset: A Novel Dataset for Commonsense-enhanced Task-based Dialogue Grounded in Documents

Carl Strathearn and Dimitra Gkatzia
Edinburgh Napier University
{c.strathearn,d.gkatzia}@napier.ac.uk

Abstract

This paper describes the Task2Dial dataset, a novel dataset of document-grounded task-based dialogues in the food preparation domain, where an Information Giver (IG) provides instructions to an Information Follower (IF) so that the latter can successfully complete the task. In this novel setting, the IF can ask clarification questions which might not be able to be grounded in the underlying document and might require commonsense knowledge to be answered. The Task2Dial dataset poses new challenges: (1) its human reference texts show more lexical richness and variation than other document-grounded dialogue datasets; (2) generating from this set requires paraphrasing as instructional responses have been modified from the underlying recipe; (3) and commonsense knowledge, since questions might not necessarily be grounded in the document; (4) generating requires planning based on context, as recipe steps need to be provided in order. As such, learning from this dataset promises more natural, varied and less template-like system utterances. The dataset contains dialogues with an average 18.15 number of turns and 19.79 tokens per turn, as compared to 12.94 and 12 respectively in existing datasets. Finally, we also provide a data statement, and we discuss the challenges associated with this novel task/dataset.



I have brought it to a gentle simmer, the chicken is cooked through well and is tender. What is the next step?

Remove the chicken from the pan using a fork or tongs and put it aside, then pour the broth through a sieve into a jug and throw away the rest of the contents.

I have thrown away the chicken from the pan and put it aside and I have poured the broth through a sieve into a jug and thrown away the rest of the contents, what is the next step?

Then using oven gloves, return the broth to the pan and place over a medium-high heat.

Figure 1: Excerpt from dialogue showing the commonsense handling of hot objects in the Task2Dial dataset

1 Introduction

Goal and task oriented dialogue systems enable users to complete tasks, such as restaurant reservations and travel booking, through natural language (Chen et al., 2017). Traditionally, goal-oriented dialogue is based on domain-specific database schemas (Shah et al., 2018). However, encoding all domain information can be prohibitive. Instead, most domain knowledge exists in some unstructured format, such as documents (Feng et al., 2020). Grounding dialogue in documents is a promising direction for several tasks. Here, we propose a new task for document-grounded dialogue, `Task2Dial`. The proposed task aims at generating task instructions grounded in a document, for a user to complete a recipe. This task requires following steps in a pre-specified order, and invokes every day communication characteristics, such as asking for clarification, questions or for advice on safe practice, which might require the use of commonsense knowledge. The proposed task is different to existing document-grounded tasks such as CoQA (Reddy et al., 2019) in the sense that it is intended for *real-world practical scenarios* that are executed in highly variable conditions. This requires enhanced and context aware dialogue so that the conversation can be as natural and concise as possible. Similar tasks involve the building of furniture, repairing things at home, troubleshooting and many more practical scenarios that involve following instructions provided by a domain expert.

Inspired by previous work on document-grounded dialogue (Feng et al., 2020) (Hu et al., 2016), (Stoyanchev and Piwek, 2010), commonsense-enhanced natural language generation (NLG) (Lin et al., 2020) and task-based/instructional dialogue (Gargett et al., 2010), we aim to capture two different types of knowledge: (1) document-level procedural context, i.e. what is the next step in the recipe; (2) commonsense, i.e. answering questions that are not available in the document, such as recommendations about replacing an ingredient with another or requesting information on the storage and handling of common tools, devices and utensils, as demonstrated in Figure 1. We consider the recipe-following scenario with an information giver (IG) and an information follower (IF), inspired partly by the GIVE challenge (Gargett et al., 2010). The IG has access to the recipe and gives instructions to the IF. The IG might choose to omit irrelevant information, simplify the wording in the recipe or provide it as is. The IF will either follow the task or ask for further information. For instance, the IF might not have access to an ingredient and may request a recommendation for a substitute or information on the common storage locations, or query the handling/use of mentioned entities. The IG will have to rely on information outside the given document, in other words the IG will rely on their common sense (e.g. replacing butter with olive oil) to enhance understanding and success of the task.

This paper follows a theoretical framework which combines a background literature review with the development and challenges of Task2Dial dataset (§2). The proceeding sections cover the data curation methodology (§3), analysis of the dataset (§4), related work (§5), discussion (§6) and future work on real-world applications of the Task2Dial dataset (§7).

2 Theoretical Framework

2.1 Task and Goal-oriented dialogue

In dialogue management, a task-oriented approach focuses on the successful completion of the individual stages of a task, towards achieving an end goal (Hosseini-Asl et al., 2020). Comparatively, goal-oriented approaches are focused on comparing the outcome or overall performance against a gold standard (Ham et al., 2020). To put this into a cooking context, some people focus more on following the stages of a recipe, and others more on what the outcome will be against an image or previous example. Task and goal oriented dialogue systems are common in domains such as booking and reservation systems for businesses (Zhang et al., 2020). However, virtual business models are typically goal-oriented as the instructions are minimal and the focus is on the outcome (selling a product or service) (Ilievski et al., 2018). The Task2Dial dataset is task-oriented as practical scenarios are more complex and require adaptability, additional information, clarification and natural conversation in order to enhance understanding and success.

2.2 Dialogue State Tracking and Planning

Traditional task-based and goal oriented dialogue systems require the user and artificial agent to work synergistically by following and reciting instructions to achieve a goal. Zamanirad et al. (2020) define these methods in human-bot conversational models as

- **Single intent and single turn policy:** relies solely on question and answer pairs without dialogue state tracking.
- **Single intent and multi-turn policy:** missing and historic information is extracted and used to structure data.
- **Multi-intent and multi-turn policy:** the information continuously changes depending on the context.

Real-world scenarios must accommodate knowledge and variability outside of a linear deterministic model as practical tasks and environments are more complex than tasks in virtual environments. For example, in human-human scenarios there is no restriction on the amount of variability introduced into a task, such as alternate methods, commonsense knowledge and objects that change the structure and information within the dialogue. Variability is significantly reduced in human-machine scenarios as the IG is limited in knowledge and the IF in asking questions or for clarification (Shum et al., 2018). This has an effect on the natural interaction between the IF and IG, as the IF will give shortened responses and not ask questions on aspects of the task (Byrne et al., 2019). This approach neglects to ensure that the IF has understood the IGs directions, which may produce irregular outcomes or result in an unfinished task. Therefore, capturing and emulating natural variability within the dialogue is crucial for creating a robust and reliable human-machine IF/IG conversational system for real-world scenarios.

Existing task-based datasets such as Multi-Domain Wizard-of-Oz (MultiWOZ)

(Budzianowski et al., 2018), Taskmaster-1 (Byrne et al., 2019), Doc2dial (Feng et al., 2020) and the Action-Based Conversations Dataset (ABCD) (Chen et al., 2021) are designed for virtual tasks, such as making bookings and appointments. These datasets assume that the user has prior knowledge of the task, components and the outcome which is different than real-world scenarios, such as cooking a meal for the first time as the user may not have prior knowledge of the task, methods or the outcome. Therefore, the dialogue needs to accommodate this uncertainty and allow for questions and clarifications on different aspects of the task to complete the task successfully. Furthermore, previous task-based datasets focus on short sample utterance's that do contextualise or capture the natural flow of a human-human conversation (Majumdar et al., 2019). This restricts the IFs ability to provide detailed answers that demonstrate to the IG that the IF has understood and completed stages of a task. For example, in the Doc2dial dataset there are a significant number of one word sample utterances such as 'yes, no, ok, next' to classify intents. These one word responses are not indicative of the natural flow of a conversation and can lead to diffusion (Zamanirad et al., 2020). For example, in a cooking scenario, the IG instructs 'Cook 200g of kale in boiling water for approximately 5 minutes, until they go soft'. A more natural response to this command than 'ok' or 'next', would be: 'I have boiled the kale and they are soft, what do I do now?'. This methodology is crucial for state / multi-state tracking as document-level procedural context provides key words for defining turns and paths.

Using document-grounded subroutines to capture intents that change the direction of a task broadens the interaction between the IG and IF (Chen et al., 2021). However, an issue with this approach is that the user is limited to the path of the subroutine making the interaction seem template-like and unnatural. In human-human scenarios, the IF can ask the IG questions at any stage of the task, regardless of the position within a given sequence and then return to that position after the question is fulfilled. For example, in a cooking scenario the IF may ask the IG how to use a certain kitchen utensil. The IG would need to answer this question, then return to the correct stage in the recipe in order to continue the sequence. Finally, directions given in instructional documents such as recipes frequently neglect important information such as identifying appropriate alternatives, common cooking utensils, tools and devices.

Additionally, there is a lack of consistency in instructional documents as the tools needed to complete a task are frequently described at the start of a document and presume that the user knows how to use that tool for different jobs throughout the task. This is particularly important in cooking scenarios as utensils like forks and knifes are multipurpose and can be used for different tasks such as mixing, dicing and preparing ingredients as well as consuming them. Thus, clarity and consistency in task-based dialogue is vital for enhancing understanding, efficiency and natural interaction between the IF and IG. This in turn increases the probability of successfully completing a task.

2.3 Document-grounded dialogue

Document-grounded dialogue systems (DGDS) classify unstructured, semi-structured and structured information in documents to aid AI in understanding human knowledge and interactions, creating greater naturalistic human-computer interaction (HCI) (Zhou et al., 2018). The aim of DGDS is to formulate a mode of conversation from the information (utterances, turns, context, clarification) provided in a document(s) (Ma et al., 2020). DGDS are particularly useful in task-oriented and goal-oriented scenarios as they emulate natural dialogue flow between the IG and IF. Doc2Dial is a multi-domain DGDS dataset for goal-oriented dialogue modelled on hypothetical dialogue flows and dialogue scenes to simulate realistic interactions between a user and machine agent in information seeking settings (Feng et al., 2020). Although the Doc2Dial dataset is highly effective for form centric applications, the system is not grounded in practical task scenarios that would affect the dialogue flow or change the outcomes of a task. This consideration is vital in the development of a real-world IF-IG task-based or goal-oriented conversational agent as the pipeline needs to accommodate external variables that reflect real-world conditions. Therefore, high quality dialogue is needed that authentically emulates practical interactions between a IF and IG based on real-life events.

2.4 Commonsense Enhanced Dialogue

Commonsense reasoning is the innate understanding of our surroundings, situations and objects, which is essential for many AI applications (Ilievski et al., 2021). Simulating these perceptual processes in task and goal oriented DGDS generates greater context and grounding for more human-like comprehension. An example of commonsense dialogue in a practical task-based scenario is understanding the common storage locations of objects, or the safe handling and use of objects from their common attributes i.e. a handle, knob or grip. Commonsense dialogue is highly contextual. In Question Answering in Context (QuAC) (Choi et al., 2018), dialogues are constructed from Wikipedia articles interpreted by a teacher. A student is given the title of the article and asks the teacher questions on the subject from prior knowledge, the teacher responds to the students' questions using the information in the document. This mode of question answering (Q&A) development is more naturalistic and grounded than previous methods as the challenges of understanding the information is ingrained in the dialogue from the underlying context. However, it does not capture all the information and questions that could be extracted from the original article as understanding is limited to the students' knowledge of the subject. Therefore, although this method shows promise in capturing commonsense knowledge within the dialogue, it is not appropriate for task-based IF/IG scenarios where all the information from the original document is needed to complete a task. Similarly, the Conversational Question Answering Challenge (CoQA) dataset (Reddy et al., 2019) is formulated on a rationale, scenario and conversation topic, and the Q&As pairs are extracted from this data. This methodology is used in the Task2Dial dataset as it provides greater co-reference and pragmatic reasoning within the dialogue for enhanced comprehension as shown in Figure 1.

2.5 Commonsense Enhanced Actions

In human-human IG/IF tasks, the IG may have prior knowledge of appropriate alternative methods, components and tools that can be used in a task that are not mentioned in the instructions. This information is vital if the IF has missing components or requires clarification on aspects of the task that are not clearly represented in the document. Variability is problematic to capture in DGDS alone as hypothetical scenarios in documents cannot account for all the potential issues in practice (Li et al., 2019). Thus, the ability to ask questions that are not available in the document is crucial when conducting real-world tasks due to the changeable conditions, complexity of the task and availability of components. This is particularly important in cooking tasks as the user may not have all the ingredients stated in a recipe, but may have access to alternative food items that can be used instead. This approach can also be used in other domains such as maintenance or construction tasks if the user does not have a specific tool, but has access to a suitable alternative tool without knowing it.

Therefore, moving away from the limited knowledge base/s in DGDs, into incorporating multiple sources of information has the potential to broaden knowledge bases, adaptability and application of DGDs (Ni et al., 2021). To explore this concept further an additional document was created that listed alternative ingredients to those listed in the metadata from the original recipes. Appropriate alternative ingredients were collected and verified using certified online cooking resources that provide food alternatives. A series of custom actions where created using the list of alternative ingredients. Further to this, clarification of the correct handling and explanations of kitchen utensils and tools was managed as custom actions. This is important as many tools look similar, have different storage locations, and multiple uses and names that may be unknown to the IF. Thus, the ability to ask for more information and clarification is crucial for completion of a task.

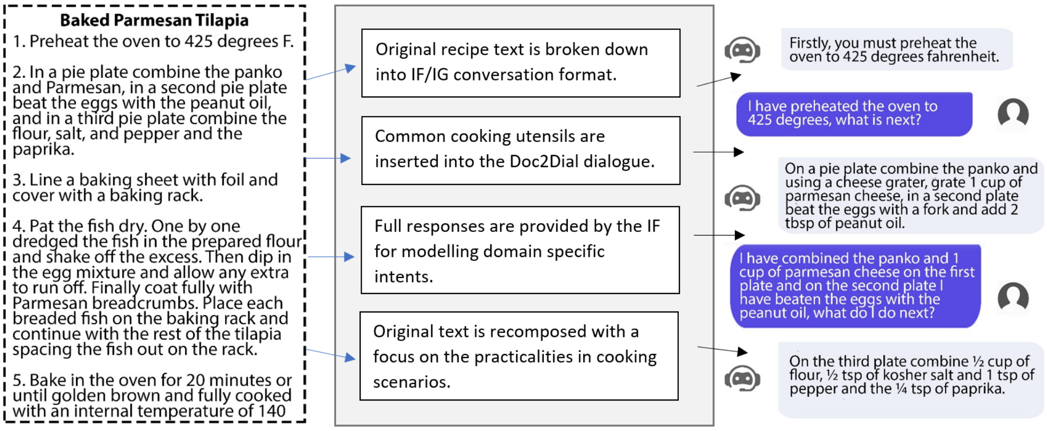
3 Task2Dial

We introduce `Task2Dial`, a new dataset that includes (1) a set of recipe documents; and (2) conversations between an IG and an IF, which are grounded in the associated recipe documents. Figure 2 presents sample utterances from a dialogue along with the associated recipe. It demonstrates some important features of our dataset, such as mentioning entities not present in the recipe document; re-composition of the original text to focus on the important steps; and the break down of the recipe into manageable and appropriate steps. Following recent efforts in the field to standardise NLG research (Gehrmann et al., 2021), we will make the dataset available via HuggingFace.

3.1 Data Collection Methodology

The overall data collection methodology is show in Figure 3 and is described in detail below.

Pilot Data Collection Prior to data collection, we performed two small pilot studies. In the first, two participants assumed the roles of IG and IF



Baked Parmesan Tilapia

1. Preheat the oven to 425 degrees F.

2. In a pie plate combine the panko and Parmesan, in a second pie plate beat the eggs with the peanut oil, and in a third pie plate combine the flour, salt, and pepper and the paprika.

3. Line a baking sheet with foil and cover with a baking rack.

4. Pat the fish dry. One by one dredged the fish in the prepared flour and shake off the excess. Then dip in the egg mixture and allow any extra to run off. Finally coat fully with Parmesan breadcrumbs. Place each breaded fish on the baking rack and continue with the rest of the tilapia spacing the fish out on the rack.

5. Bake in the oven for 20 minutes or until golden brown and fully cooked with an internal temperature of 140

Original recipe text is broken down into IF/IG conversation format.

Common cooking utensils are inserted into the Doc2Dial dialogue.

Full responses are provided by the IF for modelling domain specific intents.

Original text is recomposed with a focus on the practicalities in cooking scenarios.

Firstly, you must preheat the oven to 425 degrees fahrenheit.

I have preheated the oven to 425 degrees, what is next?

On a pie plate combine the panko and using a cheese grater, grate 1 cup of parmesan cheese, in a second plate beat the eggs with a fork and add 2 tbsp of peanut oil.

I have combined the panko and 1 cup of parmesan cheese on the first plate and on the second plate I have beaten the eggs with the peanut oil, what do I do next?

On the third plate combine ½ cup of flour, ½ tsp of kosher salt and 1 tsp of pepper and the ¼ tsp of paprika.

Figure 2: Original recipe text converted to Task2Dial dialogue

respectively, where the IG provided recipe instructions to the IF. We experimented with two participants talking over the phone, recording the session and then transcribing it. Next, we repeated the process with text-based dialogue through an online platform. The latter study used *self-dialogue* (Byrne et al., 2019), where one member of the team wrote an entire dialogue. Self-dialogue results were proximal to the results of two person studies. However, time and cost was higher for producing two person dialogues, with additional time needed for transcribing and correction, thus, we opted to use self-dialog (see section 4).

Creation of a recipe dataset From the pilot study and preliminary research it was determined that the most effective method for data collection within the four week schedule was to use online cooking resources. Three open-source and creative commons licensed cookery websites [1] were identified for data extraction, which permit any use or non-commercial use of data for research purposes as suggested in previous research (Bień et al., 2020; Marin et al., 2019). As content submission to the cooking websites was unrestricted, data appropriateness was ratified by the ratings and reviews given to each recipe by the public, highly rated recipes with positive feedback were given preference over recipes with low scores and poor reviews (Wang and Kim, 2021). From this, a list of 353 recipes was compiled and divided amongst the annotators for the data collection. As mentioned earlier, annotators were asked to take on the roles of both IF and IG, rather than a multi-turn WoZ approach, to allow flexibility in the utterances. This approach allowed the annotators additional time to formulate detailed and concise responses, including the appropriate use of common kitchen utensils and protective gear for each cooking task.

Participants Research assistants (RAs) from the school of computing where employed on temporary contracts to construct and format the dataset. After an initial meeting to discuss the job role and determine suitability, the RAs were asked to complete a paid trial, this was evaluated and further advice given on how to write dialogues and format the data to ensure high quality. After the successful completion of the trial, the RAs were permitted to continue with the remainder of the data collection. To ensure high quality of the dataset, samples of the dialogues were often reviewed and further feedback was provided.

Instructions to annotators Each annotator was provided with a detailed list of instructions, an example dialogue and an IF/IG template (see Appendix A). The annotators were asked to read both the example dialogue and the original recipe to understand the text, context, composition, translation and annotation. The instructions included information handling and storage of data, text formatting, meta data and examples of quality and poor dialogues. An administrator was on hand throughout the data collection to support and guide the annotators. This approach reduced the amount of low quality dialogues associated with large crowdsourcing platforms that are often discarded post evaluation, as demonstrated in the data collection of the Doc2Dial dataset (Feng et al., 2020).

Time Scale The data collection was scheduled over four weeks. This was to permit additional

[1] 1: www.makebetterfood.com; 2: www.cookeatshare.com; 3: www.bbcgoodfood.com

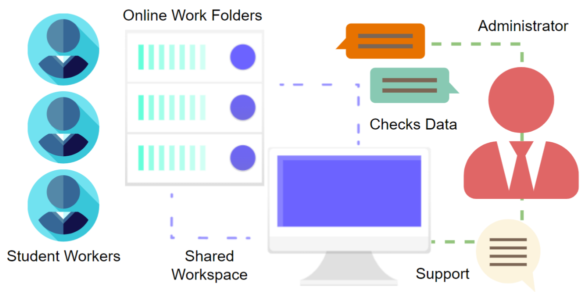


Figure 3: Overview of Task2Dial Dataset Construction

time for the annotators to conduct work and study outside of the project. Unlike crowdsourcing methods, the annotators were given the option to work on the project flexibly in their spare time and not commit to a specific work pattern or time schedule.

Ethics An ethics request was submitted for review by the board of ethics at our university. However, no personal or other data that may pertain to personal or sensitive data that may by used to identify an individual or individuals was not collected in this study. Taking into account ethical concerns surrounding crowd sourcing platforms(Schlagwein et al., 2019), this project employed research assistants with existing temporary contracts to the university to collect and annotate the data (Gleibs).

3.2 Task2Dial Long-form description

Unlike previous task and goal oriented DGDS the Task2Dial corpus is unique as it is configured for practical IF/IG scenarios as demonstrated in Figure 2. Thus, following (Bender and Friedman, 2018), we provide a long-form description of the Doc2Dial cooking dataset here.

Curation Rationale Text selection was dependent on the quality of information provided in the existing recipes. Too little information and the transcription and interpretation of the text became diffused with missing or incorrect knowledge. Conversely, providing too much information in the text resulted in a lack of creativity and commonsense reasoning by the data curators. Thus, the goal of the curation was to identify text that contained all the relevant information to complete the cooking task (tools, ingredients, weights, timings, servings) but not in such detail that it subtracted from the creativity, commonsense and imagination of the annotators.

Language Variety The recipes selected for this dataset were either written in English or translated into English prior to data collection for ease of the annotators, language understanding and training for language models. This made the dataset accessible to all contributors involved in the curation, support and administration framework.

Speaker Demographic The recipes are composed by people of different race / ethnicity, nationalities, socioeconomic status, abilities, age, gender and language with significant variation in pronunciations, structure, language and grammar. This provided the annotators with unique linguistic content for each recipe to interpret the data and configure the text into an IF/IG format. To help preserve sociolinguistic patterns in speech, the data curators retained the underlying language when paraphrasing, to intercede social and regional dialects with their own interpretation of the data to enhance lexical richness (Zampieri et al., 2020).

Annotator(s) Demographic Undergraduate research assistants were recruited through email. The participants were paid an hourly rate based on a university pay scale which is above the living wage and corresponds to the real living wage, following ethical guidelines for responsible innovation (Silberman et al., 2018). The annotation team was composed of two males and one female data curators, under the age of 25 of mixed ethnicity's with experience in AI and computing. This minimised the gender bias that is frequently observed in crowd sourcing platforms (Goodman et al., 2012).

Speech Situation The annotators were given equal workloads in an online folder, allowing them to access their files remotely. Workloads were adjusted accordingly over time per annotator availability to maximise data collection and coordinate with their schedules. The linguistic modality of the dialogue is semi-structured, synchronous interactions as existing recipes were used to paraphrase the instructions for the IG. Following this, the IF responses where created spontaneously following the logical path of the recipe in the context of the task. The intended audience for the Task2Dial dataset is broad, catering for people of different ages and abilities. Thus, the dataset is written in plain English with no jargon or unnecessary commentary to maximise accessibility.

Text Characteristics The structural characteristics of the Task2Dial dataset is influenced by real-world cooking scenarios that provide genre, texture

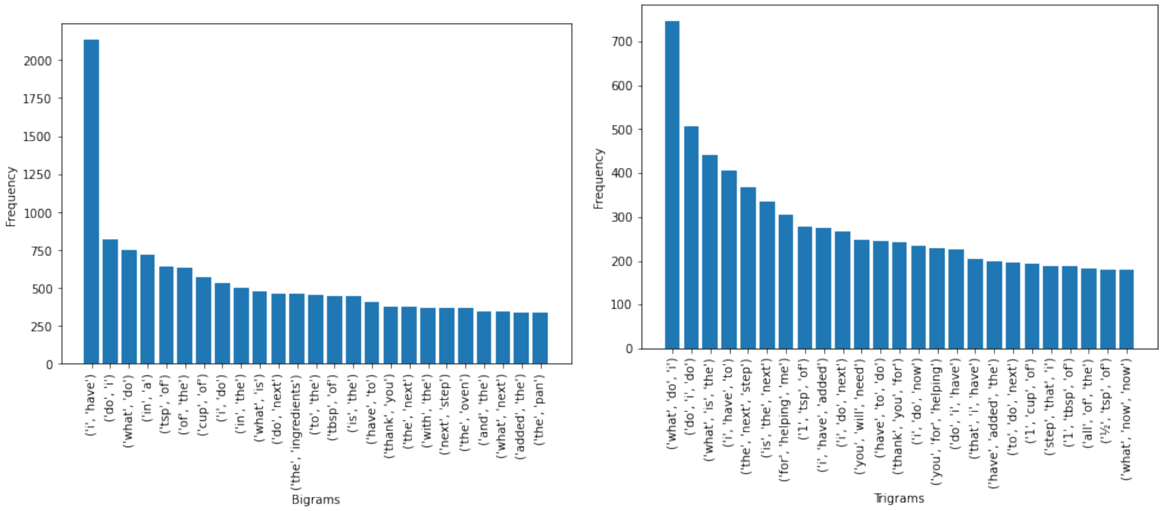


Figure 4: Distribution of the top 25 most frequent bigrams and trigrams in our dataset (left: most frequent bigrams, right: most frequent trigrams).

Dataset	#docs	#Turns	#Tkns/Turn	TTR	MSTTR
TASK2DIAL	353	**18.15**	**19.79**	**0.025**	0.84
DOC2DIAL	487	12.94	12	0.011	0.86

Table 1: Size and Lexical Richness of the dataset.

and structure to the dialogues. This provides two important classifications, utterances and intents that are universal for all task-based datasets and domain specific text that is only relevant for certain tasks. This data is used when training language models as non-domain specific sample utterances such as 'I have completed this step' can be used to speed up the development of future task-based DGDS.

Recording Quality In the pilot study, the conversations between IG/IF took place over the phone, then recorded and transcribed into text. However, as suggested in similar studies, although this method better captures natural conversation, the process is time consuming, error prone and restrictive as two people are required to construct the dialogue simultaneously (Ma et al., 2020). Therefore, the data collection was changed to single user and written text to save time and minimise translation issues.

4 Dataset Analysis

Document Analysis & Dataset Quality To examine natural language, the two-person, spoken dialogues from the pilot were comparatively analysed against the single person WOZ dialogues. Like (Ma et al., 2020) study, there was little notable differences between the two methods, other than repeat single word responses. To negotiate this, a universal intent was created to capture common phrases such as 'yes, next, ok next' that can be used across all dialogues in the model. The rule of this intent is to push the recipe to the next step, compared to the more detailed responses used for complex state tracking such as replacing items and explaining utensils.

Employing undergraduate research assistants to collect and annotate data rather than using crowdsourcing platforms meant that no dialogues were discounted from the dataset. However, a pre-evaluation check was performed on the dataset before statistical analysis to reduce spelling and grammatical errors that may affect the results of the lexical analysis.

Size Table 1 summarises the main descriptive statistics of Task2Dial and Doc2Dial. The dialogues in Task2Dial contain a significantly higher number of turns than Doc2Dial dialogues (18.15 as opposed to 12.94). In addition, Task2Dial utterances are significantly longer than in Doc2Dial, containing on average more than 7 tokens.

Lexical Richness & Variation We further report on the lexical richness and variation (Van Gijsel et al., 2005), following Novikova et al. (2017) and

Perez-Beltrachini and Gardent (2017). We compute both Type-token ratio (TTR), i.e. the ratio of the number of word types to the number of words in a text, and the Mean segmental TTR (MSTTR), which is computed by dividing the corpus into successive segments of a given length and then calculating the average TTR of all segments to account for the fact the compared datasets are not of equal size[2]. All results are shown in Table 1

We further investigate the distribution of the top-25 most frequent bigrams and trigrams in our dataset as seen in Figure 4. The majority of both trigrams (75%) and bigrams (59%) is only used once in the dataset, which creates a challenge to efficiently train on this data. For comparison, in Doc2Dial's 54% of bigrams and 70% of trigrams are used only once. Infrequent words and phrases pose a challenge for the development of data-driven dialogue systems as handling out-of-vocabulary words is a bottleneck.

5 Related Work

This research considers the development of a DGDS for practical tasks. The work is inspired by previous research in DGDS such as Doc2Dial (Ma et al., 2020), and domain specific Q&A modelled DGDS like DoQA (Campos et al., 2020) that demonstrate the effectiveness of mutli-modal and goal-orientated modelling for dialog grounded in documents. Recipe datasets such as RecipeNLG (Bień et al., 2020) and Recipie1M (Marin et al., 2019) provided key information on the organisation of metadata, itemisation and recipe categorisation that was used to develop a format for the Task2Dial dataset. Furthermore, these datasets highlighted the need for more detailed instructions and utterances as vital information such as object definitions and handling of objects are missing from these datasets.

Task-driven datasets such as MultiWoz (Budzianowski et al., 2018), Taskmaster-1 (Byrne et al., 2019) and ABCD (Chen et al., 2021), demonstrate how DGDS can be configured in end-to-end pipelines for task-driven dialog in virtual applications such as online booking systems. These pipelines are adapted in the Task2Dial dataset to accommodate variability and uncertainty that closer emulates real-world conditions and experiences, such as alternative objects and methods. Commonsense enhanced dialog datasets such as QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2019) provided key information on infusing commonsense knowledge in dialog and commonsense actions to instil greater human-like comprehension for artificial agents to operate more effectively in the real-world.

6 Conclusion

In this paper we introduce the Doc2Task dataset of task-based document-grounded conversations with everyday speech characteristics, between an IG and IF during a cooking task. Unlike previous research in DGDS, we consider the challenges and complexity of modelling dialogue for practical tasks that incorporate variability, confirmation, Q&A, state-tracking and commonsense reasoning to manage the unpredictability of real-world tasks. A key contribution of the Task2Dial dataset is that it is significantly larger in domain specific areas, with longer utterances and a higher number of turns than comparable datasets such as Doc2Dial and CoQA, making it more robust for real-world settings. The Task2Dial dataset is more sophisticated than previous task-driven datasets like MultiWOZ and Taskmaster-1 by incorporating commonsense knowledge such as handling of objects, object definitions and common storage locations of objects to enhance user accessibility and understanding of a task. Although, commonsense enhanced DGDS such as QuAC and CoQA infuse contextual commonsense with dialogue, they do not consider how this knowledge may be used in the real-world, which requires variable state tracking and multimodal interactions in order to complete a task.

7 Future Work

We aim to develop a spoken dialog system based on the Task2Dial dataset in a real-world human-robot interaction (HRI) cooking scenario to evaluate the accuracy, naturalness and effectiveness of the system and dataset. This is to be extended into vision and language (V&L) to include visually enhanced dialogue such as recognising gestures for the correct handling of objects and reference to objects that are visible and non-visible in a scene.

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

[2] TTR and MSTTR have been computed using https://github.com/LSYS/LexicalRichness.

Michał Bień, Michał Gilski, Martyna Maciejewska, Wojciech Taisner, Dawid Wisniewski, and Agnieszka Lawrynowicz. 2020. RecipeNLG: A cooking recipes dataset for semi-structured text generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 22–28, Dublin, Ireland. 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. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. *CoRR*, abs/1909.05358.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. Doqa – accessing domain-specific faqs via conversational qa.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3002–3017, Online. Association for Computational Linguistics.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, 19(2):25–35.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Kristina Striegnitz. 2010. The GIVE-2 corpus of giving instructions in virtual environments. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Sebastian Gehrmann, Tosin P. Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondrej Dusek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, and Rubungo Andre Niyongabo. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. *CoRR*, abs/2102.01672.

Ilka H. Gleibs. "are all "research fields" equal? rethinking practice for the use of data from crowdsourcing market places.". *Behavior research methods*, 48:1333.

Joseph K. Goodman, Cynthia Cryder, and Amar Cheema. 2012. Data collection in a flat world: Strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making, Forthcoming*.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, Online. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc.

Zhichao Hu, Michelle Dick, Chung-Ning Chang, Kevin Bowden, Michael Neff, Jean Fox Tree, and Marilyn Walker. 2016. A corpus of gesture-annotated dialogues for monologue-to-dialogue generation from personal narratives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3447–3454, Portorož, Slovenia. European Language Resources Association (ELRA).

Filip Ilievski, Alessandro Oltramari, Kaixin Ma, Bin Zhang, Deborah L. McGuinness, and Pedro Szekely. 2021. Dimensions of commonsense knowledge.

Vladimir Ilievski, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. 2018. Goal-oriented chatbot dialog management bootstrapping with transfer learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4115–4121. AAAI Press.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 12–21, Florence, Italy. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Longxuan Ma, Wei-Nan Zhang, Mingda Li, and Ting Liu. 2020. A survey of document grounded dialogue systems (DGDS). *CoRR*, abs/2004.13818.

Sourabh Majumdar, Serra Sinem Tekiroglu, and Marco Guerini. 2019. Generating challenge datasets for task-oriented conversational agents through self-play.

Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2019. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Trans. Pattern Anal. Mach. Intell.*

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, Vinay Adiga, and Erik Cambria. 2021. Recent advances in deep learning based dialogue systems: A systematic survey.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.

Laura Perez-Beltrachini and Claire Gardent. 2017. Analysing data-to-text generation benchmarks. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 238–242, Santiago de Compostela, Spain. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Daniel Schlagwein, Dubravka Cecez-Kecmanovic, and Benjamin Hanckel. 2019. Ethical norms and issues in crowdsourcing practices: A habermasian analysis. *Information Systems Journal*, 29(4):811–837.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana. Association for Computational Linguistics.

Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From eliza to xiaoice: Challenges and opportunities with social chatbots.

M. S. Silberman, B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar. 2018. Responsible research with crowds: Pay crowdworkers at least minimum wage. *Commun. ACM*, 61(3):39–41.

Svetlana Stoyanchev and Paul Piwek. 2010. Constructing the CODA corpus: A parallel corpus of monologues and expository dialogues. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Sofie Van Gijsel, Dirk Speelman, and Dirk Geeraerts. 2005. A variationist, corpus linguistic analysis of lexical richness. volume 1, pages 1–16.

Yiqi Wang and Jewoo Kim. 2021. Interconnectedness between online review valence, brand, and restaurant performance. *Journal of Hospitality and Tourism Management*, 48:138–145.

Shayan Zamanirad, Boualem Benatallah, Carlos Rodriguez, Mohammadali Yaghoubzadehfard, Sara Bouguelia, and Hayet Brabra. 2020. State machine based human-bot conversation model and services. In *Advanced Information Systems Engineering*, pages 199–214, Cham. Springer International Publishing.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Zheng Zhang, Ryuichi Takanobu, Minlie Huang, and Xiaoyan Zhu. 2020. Recent advances and challenges in task-oriented dialog system. *CoRR*, abs/2003.07490.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.