"What Can I Cook with these Ingredients?" - Understanding Cooking-Related Information Needs in Conversational Search

ALEXANDER FRUMMET, DAVID ELSWEILER, and BERND LUDWIG,
University of Regensburg

As conversational search becomes more pervasive, it becomes increasingly important to understand the users' underlying information needs when they converse with such systems in diverse domains. We conduct an in situ study to understand information needs arising in a home cooking context as well as how they are verbally communicated to an assistant. A human experimenter plays this role in our study. Based on the transcriptions of utterances, we derive a detailed hierarchical taxonomy of diverse information needs occurring in this context, which require different levels of assistance to be solved. The taxonomy shows that needs can be communicated through different linguistic means and require different amounts of context to be understood. In a second contribution, we perform classification experiments to determine the feasibility of predicting the type of information need a user has during a dialogue using the turn provided. For this multi-label classification problem, we achieve average F1 measures of 40% using BERT-based models. We demonstrate with examples which types of needs are difficult to predict and show why, concluding that models need to include more context information in order to improve both information need classification and assistance to make such systems usable.

CCS Concepts: • **Information systems** → **Collaborative search;**

Additional Key Words and Phrases: Conversational search, information needs, conversational agents

1 INTRODUCTION

Voice-based interaction systems are changing the way people seek information by making search more conversational [40, 90]. Spoken queries are known to be very different from typed queries [40]. They are longer and use richer semantics and more natural language phrases compared with text queries [40]. These qualities make it possible for users to express more complex and abstract concepts than are typically used in search. Human-to-human conversation, nevertheless, goes further, typically featuring follow-up and clarification questions that lead to better shared

understanding of needs and perspectives. If systems could replicate this process, conversing as humans do, systems could move beyond simply providing results to offering intelligent assistance [50].

Understanding how this process works in human-to-human conversation is a prerequisite to implementing this kind of functionality into conversational assistants. Despite fields, such as Information Science, having a rich tradition of studying conversations with librarians to learn about user information needs (e.g., see [37, 80, 81]) and their behaviour to address these (e.g., see [11, 15, 75]), voice-based applications on mobile devices and dedicated hardware, such as Amazon Echo and Google Home, mean that information-seeking conversations now take place in diverse situations embedded in users' everyday lives that are very different from a library visit. Our understanding of how information-seeking conversations unfold is very limited in domains beyond the library. Recent research meetings focusing on conversational search have identified the need to accurately elicit information needs, correct user misconceptions and provide the right amount of information at the right time across all possible domains (e.g., [8, 28]). Our focus in this work is on the first of these challenges — need elicitation — specifically on understanding and predicting user information needs, which are important for systems to conversationally identify what a user requires, facilitate appropriate retrieval, and attain relevance feedback [89]. Although domain-independent taxonomies already exist describing, for example, conversational moves for search [91] or conversations for domain-agnostic planning tasks (e.g., [45, 77]), no domain-specific schemata have been published to date.

Prominent researchers in this area emphasise that developing new information need taxonomies derived from studies in diverse realistic settings is crucial to developing usable real-world systems (e.g., [74, 91]). With this in mind, we study information needs in one particular domain, home cooking, which, based on the literature, we believe provides a fertile context for the kinds of complex needs suited to conversational search [29, 34, 63]. Cooking, moreover, offers a situation in which users simultaneously perform practical, sometimes cognitively challenging tasks, which make searching in the traditional sense problematic. By studying conversational information-seeking behaviour in this setting, we can learn not only about the diverse information needs that occur but also how information needs are communicated and resolved in this specific scenario. Being able to predict these needs is a first, yet crucial, step towards a truly conversational search system.

Concretely, our contributions are the following:

- We perform an in situ study that facilitates a naturalistic cooking situation, resulting in the organic development of information needs.
- We analyse the collected data qualitatively to learn about the diverse types of information needs which can occur in this context.
- We utilise machine learning approaches to classify needs using the raw transcription of participant utterances.

In doing so, our findings add to the information-seeking literature in which information needs have been studied in many contexts for many reasons but rarely in the context of home cooking. The results are insightful for the future development of conversational search systems, as they show that within this context it is possible to detect the kind of need a user has based on raw speech utterances. For systems of practical use to be developed, however, performance needs to be improved. Our results suggest that this may be achieved by providing algorithms with more context information.

2 RELATED WORK

Our work relates to research contributions across diverse fields of the computer and information sciences, ranging from information seeking to natural language understanding to artificial

intelligence. Here, we link the fields by first reviewing appropriate and relevant literature on conversations in interactive information retrieval (IIR), and understanding and predicting user needs and goals, before relating this research to appropriate contributions on natural language understanding more generally.

2.1 Conversations in Interactive Information Retrieval

Studying conversations is not new in information seeking. Many of the key models which have guided research for decades were conceived by analysing user conversations (e.g., [11, 52, 80]). Belkin's research used dialogues to model information-seeking behaviour; later dialogue structure inspired the design of IIR systems [12, 101]. Prekop [65] identified patterns describing prototypical behaviours, actions, and interactions of participants in collaborative tasks. The patterns were "information seeking by recommendation" [p. 543], "direct questioning" [p. 544], and "advertising information paths" [p. 545]. Other publications looked at the language used in conversations. Thomas et al. [84] distinguish several conversational styles and discovered that divergent styles between agents can be strenuous for users. One take-away from the work is that conversational systems should be capable of adapting to the user's style to reduce workload. Trippas et al. [90] identified patterns resembling Prekop's when pairs of users conversed during collaborative tasks. They also found that integrating discourse in conversational search models is necessary to cope with the complexity of the information-seeking process [91]. For example, grounding, that is, making mutual knowledge and intentions clear, is a means of relevance feedback in the conversational search domain [91] which can help reduce complexity. This strategy was also found to be important by Aliannejadi et al. [5] in the sense that clarifying questions led to better search results since they help to resolve complex information needs. Zamani et al. [102], who investigated which kind of clarifying questions are preferred by users, also found that such questions are typically employed in ambiguous and long natural language user interactions. This indicates that the information needs are hard to express in such situations and, thus, might be complex. This aligns with the findings of Ruthven [70], which indicate that early-stage, that is, hard to express, information needs are expressed with more words and greater narrative. The work by Trippas, Aliannejadi, and their colleagues shows that conversational search supports users by allowing natural dialogue to address complex goals. The underlying information needs of these goals, however, are linguistically hard to express for users. For this reason, an in-depth understanding of information needs would be beneficial, as it would offer the potential for conversational search systems to adapt to user needs and behaviour.

2.2 Understanding User Information-seeking Goals

The aforementioned properties of information-seeking dialogues — adaptation of the language style [84, 85], the usage of clarifying questions [5, 102] and grounding [91] — are important aspects that help in understanding information-seeking goals. However, understanding user information needs is not a new field of investigation.

Conversational search fits with the progression in IR, from the matching of queries and textual documents to understanding and supporting user needs. Classifications of search tasks are one way of thinking about user information needs. In IIR, such classifications are numerous and often delineated based on type or characteristic [48]. Type-based examples focus on what people are trying to achieve (e.g., see [87]). Characteristic-based classifications, on the other hand, delineate tasks by property, such as specificity [69] or complexity [22]. Such classifications are useful because they inform the design of IIR systems [48, 97], help to study and model search behaviour [41], and provide the basis for the evaluation of search systems and user interfaces [33]. Classifications for specific domains — including web [19], mobile [25], and personal search

[33] — have been published. The closest in the literature to our work is [29], which establishes categories based on 678 Google Answer Forum queries relating to cooking, identifying 17 types of information needs. However, these are not based on conversational speech data; thus, they are only to some extent transferable to conversational search systems. Trippas et al. [91] suggest that domain-specific taxonomies need to be created to better understand the complexity of information needs and user information-seeking goals. Extracting these domain-specific information needs from the user can be done by mixed-initiative conversational search systems that retrieve information by user-agent conversations and help users reveal their underlying information needs [67]. To achieve this goal technically, systems must be able to automatically predict and detect the user's information-seeking goal, that is, information needs.

2.3 Predicting Information Needs Conversationally

Understanding and algorithmically predicting user needs can be useful for many reasons: different results can be shown [27], results can be presented differently [83], or answers can be presented directly in the results page [13]. Conversations with the user is one method of detecting such information needs. Automated conversational agents can provide personalisation and support users in selecting an item [86] or talking about areas of interests [76] and have been applied in scenarios such as in trauma therapy [61]. This often requires systems to utilise a memory function in which systems can reference past statements [67] in order to maintain an understanding of context [6, 51].

In conversational search, preliminary work has utilised user speech utterances as a means to identify user intents. Intent classification models are also used in other conversational domains, including generic agents (e.g., see [49, 54, 57]). Focussing on the conversational search domain, Ahmadvand [1] attempts to capture context with the use of topic sequence models in order to suggest topics in open-domain dialogues. Qu et al. [66] predict generic user intents with the help of dialogue acts, while Boteanu et al. [16] focus on the prediction of latent intents in text-based product queries. In addition to predicting user intents, there have been efforts to improve information need prediction. For example, Aliannejadi et al. [4] found that contextually relevant utterances lie at a position in the conversation that is close to the current utterance which informs the design of the memory and context property of such systems. Ren et al. [68] attempt to understand complex information needs and provide appropriate answers in a conversational search/agent setting. They emphasise the importance of context, for example, via conversational history, which they argue must be modelled to fully understand a current turn. Shiga et al. [74] classify needs along two dimensions: one uses Taylor's levels of specification [80] and the other delineates type based on a classification derived from the literature. Moreover, they incorporate an aspect of task hierarchy in which a main task (e.g., booking a holiday) can be viewed as consisting of sub-tasks (e.g., finding a destination, comparing flight schedules, etc.). Their work shows that information need categories can be distinguished using machine learning approaches. This work is the closest to our own research in terms of motivation and approach. However, the categories of needs predicted are very high level and domain unspecific, which is also the case in the study by the authors of [66]. One could imagine that conversations and the types of support required across domains could be quite different. If systems could identify specific need types within specific domains, conversational systems could provide much more appropriate assistance. Thus, building on the work of Shiga et al. [74], we test similar approaches in a home cooking context.

2.4 Understanding Dialogue

Theoretical foundations and technical solutions for conversational search are closely related to research on dialogue understanding in linguistics and artificial intelligence. In the vast literature

on this topic, two complementary trends are discernible. The first (and historically earlier) trend relates to discourse analysis. It aims to generate computational features from all linguistic, cognitive, and (discourse) pragmatic processes that are active — often in parallel — while dialogue participants interact. Macagno and Bigi [59] present a recent overview that positions conversational search in the class of *information-seeking dialogues*. While this strand of research develops extensive linguistic models for the pragmatics of dialogues, it does not focus on developing effective algorithms for them. The second major trend does exactly this, based on recent successes of machine learning algorithms (e.g., see [73, 96]): Deep neural networks are leveraged to implicitly model the current state of any active process using the hidden and recurrent layers of the employed network architecture. Dialogue systems based on these machine learning approaches are regularly evaluated in the dialogue state tracking challenge [99] on a class of so-called *open-domain* dialogues [2, 43] or *multi-domain* dialogues [39]. Whereas algorithms for open-domain dialogues aim to imitate human-to-human conversations as naturally as possible without necessarily trying to complete a task, other single-domain dialogues can be task based (e.g., [10]), in which the user provides rational interactions trying to complete a certain goal (e.g., booking a flight). In this class of dialogue systems, the task knowledge is formalized in terms of slots [10] whose values must be extracted from natural language user input using appropriate natural language understanding (NLU) algorithms. If such problems are to be extended to multi-domain dialogues, a small number of tasks may need to be completed in parallel (i.e., booking a flight, reserving a room in a hotel, and retrieving a public transport connection from the airport to the hotel). In either case, machine learning–based dialogue systems are trained on huge corpora of mostly publicly available datasets [99].

Such systems are trained to predict values for slots to be filled from current user input and a current latent state of the task inferred from the sequence of dialogue moves (i.e., classified user utterances) observed so far (see [2]). Since dialogue move sequences result from changes in the state of the task to be completed, the training process tends to result in models fitted strongly to the specific task. This, in turn, means that models perform poorly in contexts outside the training scenario [46].

Building a task-oriented conversational search system for a new domain, that is, a domain for which no corpus is available, remains a huge challenge despite the recent progress in machine learning. Therefore, only a few cross-context systems exist due to the lack of data available and the high costs involved in annotating such data (e.g., see [92, 95]).

2.5 Key Differentiators

The summarized related work indicates several open issues in the literature. Next, we highlight these issues and formulate concrete research questions that we will answer with our empirical work.

To implement working conversational search systems, it is crucial to develop taxonomies or classification schemes to understand the information needs commonly occurring in one specific domain such that systems can employ appropriate strategies to elicit the user's underlying need. As, to date, no such taxonomies exist for conversational search in the cooking domain, **RQ1** aims to answer the question of **what kind of information needs arise when cooking with a conversational assistant.**

Detecting such needs in a dialogue still remains a huge challenge for current systems due to difficulties in understanding discourse pragmatics. In **RQ2**, we want to investigate **how information needs are communicated** to identify linguistic features in dialogue that raise difficulties in interpreting information needs.

Finally, systems must be capable of predicting user information needs to satisfy them accordingly. Since the previous literature has focused only on predicting broad and domain-agnostic

needs, we evaluate in **RQ3** whether **cooking-related information needs can be predicted in a conversational assistance scenario**.

3 METHODOLOGY

3.1 Data Collection

We designed an experiment to collect naturalistic dialogues that involved the communication of and attempts to solve information needs. To simulate a naturalistic cooking scenario, an in situ experiment was performed whereby participants were gifted a box of ingredients and tasked to cook a meal that they had not cooked previously using as many of the contained ingredients as possible. Participants were allowed to supplement the supplied ingredients with the contents of their own pantry if necessary. To assist this process, participants could converse with the experimenter, who would answer any questions and needs arising using any resources available via the Web. The experimenter provided the best answer possible and communicated this orally in a natural human fashion, which is arguably the optimal behaviour for a conversational system. The setup was equivalent for all participants who worked in their own kitchens with the experimenter mimicking a conversational search system, sitting close by with access to the Internet via laptop. No time constraints were imposed for the task. Each experiment comprised the following six steps:

(1) The instructions were read to the participant.
(2) Participants signed a consent form explaining how the collected data would be stored and used in the future.
(3) The ingredient box was provided.
(4) The recording device was tested.
(5) Participants started the cooking task and the full dialogue between experimenter and participant was recorded.
(6) After the task, the experimenter thanked the participant and gifted the remaining ingredients.

To encourage a naturalistic setting, all experiments took place in each participant's home. As such, the participants were more familiar with their surroundings and materials available. Feeling comfortable in the environment is advantageous, as it is known to leverage informal speech [23].

3.2 Ingredients

To ensure diversity in the recipes and, thus, variety in subsequent conversations, the ingredient boxes varied across participants. The ingredients mostly had a value of around 10€ and were chosen based on the guidelines by the German Nutritional Society, which suggests that seven categories of ingredient are required for a balanced meal. Typically, the box contained some kind of grain or starch (e.g., potatoes or rice), a selection of vegetables, a source of protein (e.g., eggs), and at least one ingredient that was seasonal (e.g., asparagus) or that we felt might not have been commonly cooked with by all participants. We did this in order to stimulate conversation and, thus, to gain as many potential utterances as possible. Participants prepared diverse meals using the ingredients. To ensure freshness, all ingredients were bought on the day they were used in an experiment.

3.3 Participants

Participants were recruited using a snowball sampling technique with a convenience sample providing the first group of candidates. These participants, in turn, were willing to recruit friends and relatives, and so on. This method offers two advantages. First, it generates a basis for trust among the participants and the experimenter, which leads to more informal and open dialogue

[23]. Our impressions confirmed relaxed and natural behaviour in the experiments. Second, it allows a relatively large sample to be achieved. The only requirements for potential participants were a kitchen and Internet connection. Participants were not paid for participation. However, to increase response rate, the ingredients were gifted. A total of 45 participants (22 females, $\overline{x}_{age}$ = 24 years, min_{age} = 19 years, max_{age} = 71 years, 20% non-students) were tested between May 7, 2018 and June 28, 2018. A total of 37 had never used conversational agents before, while 4 used either Alexa or Google Home. Asked about their cooking experience, six participants reported cooking multiple times per week or on a daily basis, 18 reported seldom cooking or not at all, and one person regarded cooking as her hobby. The remaining 20 participants stated that they cooked but not on a regular basis.

3.4 Transcription

In total, 38.75 hours of material was collected, with the language spoken being German. The recorded conversations were transcribed by a trained linguist. Compared with our pilot study [38], this new transcription represents a much more rigorous and work-intensive approach, providing additional detail to systems and future research, that is, pauses and intonation.

The transcription approach was inspired by previous work dealing with comparable data, that is, naturalistic conversational search dialogues in German dialect [38]. Since many participants spoke in dialect (Bavarian German), we followed the recommendation in Dresing and Pehl [32] to translate the dialectal expressions into standard German so that other German readers can fully understand what was said. This is also important for the natural language processing steps that take place afterwards, which necessitate Standard German text. However, syntactic errors and the word order were maintained [32]. Building on past work from the conversational search literature, we applied the guidelines for transcribing spoken conversational search dialogues that Trippas et al. [88] adopted from [60]. These recommendations include the suggestions that transcription rules should be independent in the sense that independent transcribers or third parties can easily understand them, that the transcripts should be suitable for both human and computer use, and that the transcription rules should be elegant, thus, minimising the number of rules and keeping them easy to learn.

More concretely, inspired by Dresing and Pehl [32], Trippas et al. [88], Braun and Clarke [18], and McLellan et al. [60], we followed these guidelines:

- Recorded audio files were transcribed verbatim (following the guidelines outlined earlier). Every dialogue related to the cooking session was transcribed. Only private chatter and small talk during the experiment were not transcribed due to ethical issues. These passages were marked with a (smalltalk) placeholder.
- Noisy, incomprehensible passages were marked with (unv.), which means "unverständlich", the German word for incomprehensible. Other notes were also marked in rounded brackets.
- There is no punctuation in the transcript. Instead, intonation was marked in the following way: (_) and (;) mark a low intonation, whereby (;) is not as low as (_). (-) indicates that intonation is on the same level, that is, neither going down nor up. (,) and (?) indicate a high intonation, where (?) is higher than (,). For further investigation with the usage of intonation, it is highly recommended to group the high and low intonation levels.
- Pauses were added with the following pattern: (.) is a short break of about one second, (..) is a break of about two seconds, (...) is a three-second pause, and ([number]) is a pause lasting [number] seconds.
- Numbers up to 12 were transcribed as words, higher numbers were transcribed numerically.

- URLs were transcribed as they were spoken (e.g., chefkoch.de was transcribed as chefkoch d e).
- Abbreviations were written as said.
- Overlaps in talking were marked with double slashes: //
- Strong accentuation was marked with capital letters.

3.5 Annotation

One primary goal is to achieve a detailed understanding of the information needs that occur during the cooking process as well as how these originate and are expressed conversationally. We previously published a preliminary analysis of the collected data using less rigorous transcription and analysis approaches [38]. The results of the preliminary work were sufficient to convince us of the value of insights that could be provided using a more detailed analysis. More detail in the transcription and analysis processes are necessary, as the information needs categories in our pilot study were very broad and, therefore, difficult to distinguish using machine learning approaches. Moreover, hierarchical relationships and semantic similarity between categories led to misclassifications. By creating a more precise, tree-structured representation of the information needs in the cooking context, our aim was to develop more accurate classifiers. By doing this, we aim to improve the performance of classifiers when distinguishing between different types of information needs.

The process comprised the following steps outlined by Braun and Clarke [17]: (Step 1) familiarising self with data, (Step 2) generating initial codes, (Step 3) searching for themes, (Step 4) reviewing themes, (Step 5) defining and naming themes, and (Step 6) producing the report. Furthermore, we followed the recommendations by Charmaz [24] on line-by-line coding, which helps to remain open during the coding process and to detect more nuances in the data. One crucial element for comparison and consistency validation was the strong usage of memos where decisions taken and new ideas were noted as suggested by Corbin [26]. Using the *f4analyse*[1] software made iterative comparison and refinement convenient due to support for qualitative data analysis (QDA), which allows utterances from multiple classes to be compared. This makes the identification of commonalities and differences easier. The whole annotation process was conducted by the first author, who is also a trained linguist. As outlined by Braun and Clarke [17], the process was conducted iteratively and tested repeatedly until the researcher was satisfied that a consistent information need taxonomy was achieved. The following steps were taken to ensure the development of a consistent coding scheme:

(1) Identifying which participant spoke, that is, identifying turns. For turns in spoken conversational search, the definition by Trippas et al. [88] was employed: "Taking the initiative equals taking the turn". Thus, a turn could consist of multiple moves or conversational goals.
(2) Assigning codes to each turn and taking notes and memos for later comparison and to understand why certain decisions have been made.
(3) Combining codes to themes for further analysis and merging indistinguishable labels.
(4) Checking for consistency across participants and themes to ensure that the labels actually are distinct.

Selective coding was employed to systematically relate codes, validate the relationships between them, and add and refine further categories as suggested by Strauss and Corbin [78]. By doing this, we gained the tree-like structure of our information needs taxonomy. This helps us to better understand the relationships between the information needs.

---

[1]See https://www.audiotranskription.de/en/f4analyse/.

4 RESULTS

The results are presented in three subsections, each of which addresses one of the research questions formulated in Section 2.5. In Section 4.1, we address RQ1 by presenting a hierarchical information need coding scheme along with its top-level categories. These categories are described by means of examples in Table 2. Section 4.2 addresses RQ2 and illustrates how information needs were communicated linguistically by participants. This allows us to identify dialogue-specific challenges that need to be considered when implementing systems for information need detection. To address RQ3, Section 4.3 reports on experiments designed to establish the extent to which the class of information need (based on the level 1 categories) can be successfully predicted given participant utterances.

4.1 Information Need Taxonomy

We first offer a broad overview of the different levels of information needs detected in our corpus and provide a sample description of the utterances issued to the experimenter. We subsequently present a more detailed description of the tree structure and the taxonomy codes to offer a more detailed understanding of the information needs that occur when cooking. We refer to our annotated dataset as the *CookversationalSearch* dataset.

4.1.1 The CookversationalSearch Dataset. contains, overall, N_t = 2376 turns describing one or multiple user information needs that were extracted from the dialogues.[2] A total of 2102 (88.5%) of these express one information need only, 249 (10.5%) of all turns represent two information needs, and 25 (1.05%) turns express three different needs. Each row in the dataset represents one turn and its corresponding information need. Thus, some turns appear multiple times, that is, in multiple rows, as they can represent multiple needs. This results in a corpus with N_t = 2675 turns in total, with each representing an information need and consisting of one or more sentences. We also refer to such turns as queries. Thus, the terms *query* and *turn* will be used interchangeably for the rest of this article. Each trial yielded an average of 52.80 queries (min = 8, $x_{.25}$ = 33, $\tilde{x}$ = 45, $x_{.75}$ = 70, max = 120, sd = 28.73, $skewness$ = 0.57, $curtosis$ = −0.60). The coding scheme used to describe these queries is structured as a tree. The tree consists of two top level (= level 0) branches, each of which descends, at deepest, to a 5th level. Each level provides additional detail describing the user's information need. To illustrate this process, we provide samples from the dataset in Table 3. Readers can additionally access the full coding scheme.[3] Turns were coded at the lowest level possible. Figure 1 shows the number of codes distributed over the different information need levels. It can be seen that in addition to the two level 0 categories, 11 information need categories occurred on level 1, 93 information needs occurred on level 2, 75 on level 3, 18 on level 4, and 2 on level 5. All 2376 turns were assigned a level 0, level 1, and level 2 code. Thus, all turns were coded to at least level 2. A total of 1848 queries were assigned with a level 3 code, 395 with level 4, and 97 with level 5. More details about the different levels and information needs will be provided in the following sections.

4.1.2 Inter-rater Reliability. To check the consistency of the annotation scheme, inter-rater reliability with a second researcher who is also an experienced coder was measured. Before calculating agreement, small samples from the corpus were coded by both annotators in multiple iterations. Problems were discussed and the codebook was refined after each round. Utterances representing an information need were labelled, with annotators having access to full dialogues and context to better interpret an utterance. In the final pass, a stratified sample of 10% of all corpus utterances

---

[2]https://github.com/AlexFrummet/CookversationalSearch.
[3]https://github.com/AlexFrummet/CookversationalSearch/blob/master/annotation/InfoNeedTaxonomy.svg.

Table 1. Cohen's κ Values for Each
Level 1 Information Need Category

Information Need	Cohen's κ
Amount	0.96
Cooking technique	0.72
Equipment	0.89
Ingredient	0.59
Knowledge	0.66
Meal	0.67
Miscellaneous	0.86
Preparation	0.62
Recipe	0.90
Temperature	1.00
Time	0.97

was selected among level 1 categories. Stratifying the sample allowed us to test whether agreement could be achieved for all information need categories.

To measure inter-rater agreement, unweighted Cohen's κ was chosen. Artstein and Poesio [9] suggest this measure for dialogue data in which category labels are equally distinct from one another. In the final pass, the annotators achieved a Cohen's κ value of $\kappa = 0.69$ using codes at level 2 (93 codes) and a value of $\kappa = 0.75$ using codes at level 1 (11 codes), which, according to Landis and Koch [53], both represent substantial agreement. Although there are many more codes on level 2 than on level 1, this did not relate to a substantially lower κ value. The results across need types are shown in Table 1.

The agreement achieved is higher than what has been achieved in other conversational search corpora that were created in much more controlled contexts. For example, when evaluating the SCoSAS annotation schema, a κ score of $\kappa = 0.59$ was achieved [91]. Moreover, on the widely used Maptask dialogue corpus — which was collected in a very controlled situation — the same level of inter-rater reliability ($\kappa = 0.59$) was attained according to Isard and Carletta [44].

4.1.3 Description of the Communicated Information Needs. This section provides a broad outline of the information needs as derived from the qualitative analysis. For space reasons, we limit our descriptions to the two topmost levels, that is, the two information need types on level 0 and the 11 information need categories on level 1.[4] Level 0 distinguishes between *Fact-based* needs, which relate to information about recipes, and *competence* needs, which refer to knowledge about how to prepare recipes. Level 1 delineates needs more precisely and includes codes such as *Recipe,* which relates to the user seeking a recipe to cook, and *Ingredient*, which relates, in some way, to ingredients that are contained within the recipe being cooked. To assist the reader's understanding, we provide several examples of queries for each class of information need at level 0 and 1 in Table 2. Figure 2 shows the frequency of queries over the different types of information needs. Here, we first illustrate the hierarchical annotation scheme with a short example. We then describe the defining characteristics of different information need types in greater detail.

Hierarchical annotation: As described in the previous section, the annotation scheme has a tree-like structure. A visual representation of this taxonomy can be found in the accompanying repository.[5] To illustrate the hierarchical annotation process, we provide some examples from the dataset

---

[4]The codebook with a detailed description of every information need class is available online: https://github.com/AlexFrummet/CookversationalSearch/.

[5]https://github.com/AlexFrummet/CookversationalSearch/blob/master/annotation/InfoNeedTaxonomy.svg.

Table 2. Level 1 Information Need Categories with their Corresponding Level 0 Code

Level 1 Code	Level 0 Code	Frequency	Definition	Examples
Preparation	Competence	768 (28.71%)	Utterances were labelled with this category when queries related to a particular step of the recipe (as opposed to general cooking techniques; see below).	"What's next after bringing to the boil?", "What should I do after the potatoes are prepared?", "OK, that means that I now have to add the flour, right?"
Ingredient	Fact	658 (24.60%)	Whenever questions regarding which ingredients were necessary for a particular recipe occurred, these utterances were tagged with the label *Ingredient*.	"Which ingredients are needed?", "Please read the ingredients aloud.", "What [ingredients] goes in a vegetarian casserole?"
Amount	Fact	476 (17.79%)	The label *Amount* was used to code queries from participants who wanted to know about the quantity of an ingredient needed.	"How much egg yolk is needed?", "How much of the couscous do I need for one person?", "150 grams of tomatoes?"
Cooking Technique	Competence	210 (7.85%)	Utterances/queries were labelled this way when participants requested information about preparing ingredients that was not made explicit in the steps of the recipe.	"But it has become a bit too watery. It would be good if…How can I reduce the liquid?", and "OK, how do you prepare potatoes properly?"
Recipe	Fact	205 (7.66%)	*Recipe* was used in cases in which participants searched for recipes they would like to prepare as their main dish. They often used ingredients as search items in such cases.	"Then I'd like to have a dish with lentils, chickpeas, and tomatoes", "OK, search for a recipe with couscous and vegetables.", "Is there a recipe for Jerusalem-Artischocke and celery soup?"
Time	Fact	182 (6.80%)	Any utterances regarding cooking time needed were labelled with the category *Time*.	"OK, how long does it take? 10 minutes or 20 minutes?", "Er – how long do I need to bubble the, the, the lentils?"

(Continued)

Table 2. Continued

Level 1 Code	Level 0 Code	Frequency	Definition	Examples
Knowledge	Fact	49 (1.83%)	*Knowledge* was applied to all queries/questions relating to definitions. It was also assigned to general knowledge questions, i.e., questions that did not relate to the implementation of a cooking step but were of a more general nature.	"Can lemons go bad?", "Teaspoons are the small ones, right?", "What happens to sour cream when it is heated?"
Miscellaneous	Fact	44 (1.64%)	Needs and commands that did not fit into any of the other information need classes and occurred rarely were gathered here.	"Alex, name some good pizza delivery services.", "Am I allowed to modify the recipe?", "OK - er - do you have an integrated timer?"
Equipment	Fact	40 (1.50%)	Utterances were labelled with *Equipment* when queries/questions regarding the cooking utensils needed occurred.	"Can I use a pot for this?", "How big should the pot be? Small, medium, large?"
Temperature	Fact	31 (1.16%)	Any utterances regarding the heating type and the temperature at which ingredients and meals should be cooked received this label.	"Do I need to preheat the oven?", "At which temperature?", "I need to heat that, right?"
Meal	Fact	12 (0.45%)	*Meal* includes information needs that are related to the current meal but from a high-level perspective. This category does not include information needs that are particularly related to the preparation steps or the ingredients used in the recipe. While the recipe class is focused on the process of recipe selection, this class includes turns after a recipe has been chosen.	"Which drinks go with chili?", "Can this be compared to stew?"

Table 3. Coding Samples from the Dataset

| Turn	Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
"Um can you find me dishes with asparagus with many dairy products." (part. 1)	Fact	Recipe	Recipe Retrieval	Recipe Request	Recipe Request with Ingredients	Explicit
"How much sugar?" (part. 2)	Fact	Amount	Amount of Ingredients	–	–	–
"Um – How do you prepare bulgur?" (part. 34)	Competence	Cooking technique	Cooking technique — Ingredient	–	–	–
"Then I need to add the onion, right?" (part. 40)	Competence	Preparation	Preparation Steps	Preparation Step —Explicit Reassurance	–	–
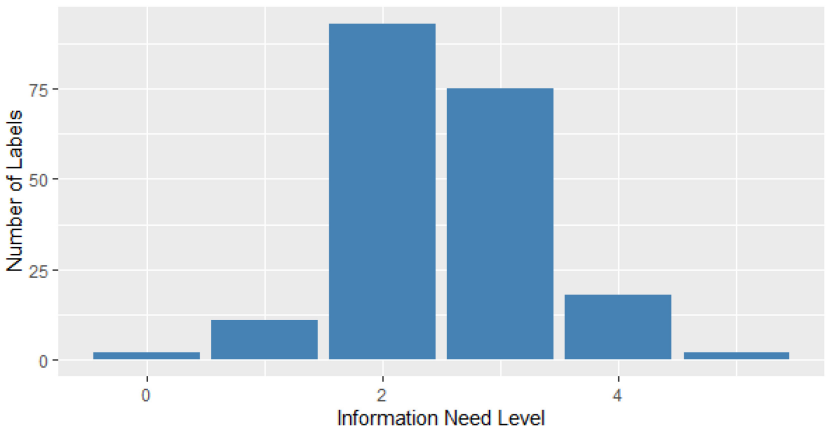
Fig. 1. Information need distribution over levels.

in Table 3. Lower-level codes subsume the code from the higher-level branches. The turn "How much sugar?" (part. 2) from Table 3, for instance, is coded as *Amount of Ingredients* on the lowest level and, thus, inherits the *Amount* category from level 1 and *Fact* from level 0.

Fact-based Information Needs. refers to information about the recipe being cooked. One prototypical fact-based information need is the level 1 category *Amount*, which includes turns that target the quantity of ingredients and recipe portions, for example. Information needs of this type are easily identifiable due to their specific vocabulary and syntax. The majority of queries that were labelled with *Amount* start with keywords such as "How many..." or "How much..." as in "How many onions do I need?" (part. 25). This makes it straightforward for the annotator to label such utterances correctly. The inter-rater agreement of $\kappa = 0.96$ for this category is much higher than the average $\kappa = 0.75$ over all categories. It is similarly straightforward to annotate the information
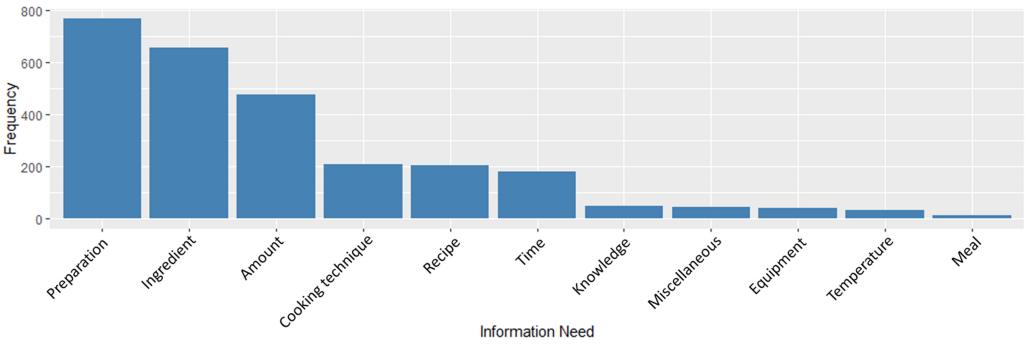
Fig. 2. Level 1 Information need frequencies.

need categories *Equipment*, *Ingredient*, *Temperature*, and *Time*. *Time*, for example, contains many phrases, such as "How much longer does the bulgur need?" (part. 25), and "How many minutes are remaining?" (part. 29). Time-related words, such as "minutes", and time phrases, such as "How long...", occur frequently, which makes it easy to code utterances as time-related information needs. This impression is evidenced by the relatively high inter-rater agreement in the Time category ($\kappa = 0.97$).

Keywords and phrases which are specific for certain information needs made it easier for the coder to identify needs. One commonality of fact-based needs is that little context is needed to identify this kind of information need in the dialogue.

Competence-oriented Information Needs, in contrast, were more challenging to annotate because context was often required to decide the appropriate label. Competence-oriented information needs involve the user requesting information from the system on how to perform certain cooking actions. Categories associated with this type of need are *Preparation* and *Cooking technique*. Turns in the *Preparation* category relate to particular steps of the recipe being cooked. In such cases, a user needs assistance regarding what to do next and how to do it. When turns occurred that relate to more basic cooking processes —often not described in the recipe — the system (i.e., the human mimicking a system) made use of a search engine to provide assistance. If this was the case, then utterances were labelled as *Cooking technique.* For utterances in the *Preparation* category, the user is looking for information regarding the steps to perform, for example, what they should do next. Being able to differentiate between these two categories requires an understanding of the context in which a query was issued. As an annotator, one needs to understand the current task state of the cooking process and know which steps have already been completed. In contrast to fact-based information needs, the task state needs to be considered, which makes it more challenging to correctly label these kinds of needs. Unlike fact-based needs, it was hardly possible to distinguish competence needs based on keywords alone. This can be illustrated by the examples "Can I add them before the water boils?" (part. 12, labelled as *Cooking technique*) and "So I need to let the water boil first?" (part. 10, labelled as *Preparation*). The first query was related to a more general cooking process; thus, it was labelled as *Cooking technique.* The second query was concerned with the order of steps; therefore, it was coded as *Preparation*. It was only after the context of situation was considered that this distinction could be made. The difficulties that arise when coding such needs is reflected in the achieved κ values ($\kappa = 0.72$ for *Cooking technique* and $\kappa = 0.62$ for *Preparation*) which are below the average that was achieved for fact-based needs.

To further investigate why some information need categories were easier to detect than others, we performed a keyword analysis to assess the importance of certain words for different

Table 4. Keywords of Level 1 Information Need Categories *Amount*, *Time*, *Temperature*, *Cooking Technique* and *Preparation*

Amount	Time	Temperature	Cooking Technique	Preparation
wie *how**	lange *long**	hitze *heat**	man *one**	dann *then**
viel *much**	minuten *minutes**	auf *on**	spargel *asparagus**	schritt *step**
gramm *gram**	wie *how**	mittlerer *middle**	schälen *peel**	jetzt *now**
zwei *two**	lang *long**	stufe *level**	am *at**	und *and**
liter *liter**	kochen *cook**	heiß *hot**	ob *if*	weiter *further*
esslöffel *tbspn**	zehn *ten**	welcher *which**	kann *can*	erst *first*
milliliter *milliliter**	dauert *takes**	ofen *oven**	besten *best*	muss *need*
hälfte *half**	bis *to**	niedrigster *lowest**	kocht *cook*	die *the*
personen *persons**	braucht *need**	ober *upper**	wasser *water*	nächstes *next*
eine *one**	uhrzeit *time**	umluft *circul. air**	grünen *green*	machen *do*
wasser *water**	zwanzig *twenty**	unterhitze *b. heat**	schneidet *cut*	fertig *ready*
viele *many**	fünf *five**	vorheizen *preheat**	mach *do*	nächste *next*
für *for**	ungefähr *about**	grad *degree*	weißen *white*	zuerst *first*
halber *half**	zwölf *twelve*	mittlere *middle*	zu *to*	als *as*
menge *amount**	zeit *time*	runter *down*	abschälen *peel*	der *the*

Note: Arranged in descending order along log-likelihood values, with * : $p < 0.05$ (Bonferroni adjusted).

information need categories. To measure this, we employed log-likelihood analysis using odds ratio as effect-size measure and Bonferroni adjusted p values as the statistical threshold. The results for the words with highest log-likelihood values among several information need categories are provided in Table 4. The results highlight that fact-based information needs *Amount*, *Time,* and *Temperature* have many more significant keywords than competence-oriented information needs *Cooking technique* and *Preparation*. While the most distinctive words for *Cooking technique* and *Preparation* are stopwords such as *you* and *then*, these kind of words are not found to be discriminatory in *Amount*, *Time* and *Temperature*. The most relevant words for *Amount* are *how*, *much*, and *gram*, which is plausible for this information need. Similarly, it is plausible that *long* and *minutes* are especially relevant for *Time* and *heat* is an important keyword for *Temperature*. These observations corroborate our impression that fact-based needs are easier to distinguish based on keywords than competence needs.

Section 4.1 has presented a robustly coded taxonomy with different levels of information needs. At the highest level, information needs can be broadly divided into two types – fact- and competence oriented – which require different types of assistance from the system and different amounts of context information to be interpreted correctly.

4.2 Communication of Information Needs

A second research aim was to investigate how information needs are communicated. An understanding of how users verbally pose queries to conversational systems would allow us to better extract information needs from these turns by focusing on certain linguistic features. We can derive implications for the system design by knowing which linguistic aspects in the dialogue make detection and interpretation of turns as particular types of information needs difficult. Since the speech used in this corpus is natural and communicated as humans typically do, this might help us to learn about what features are relevant for information need detection for humans to make these also available for system-side information need detection. The following stylistic peculiarities of issuing an information need could be observed from the data:
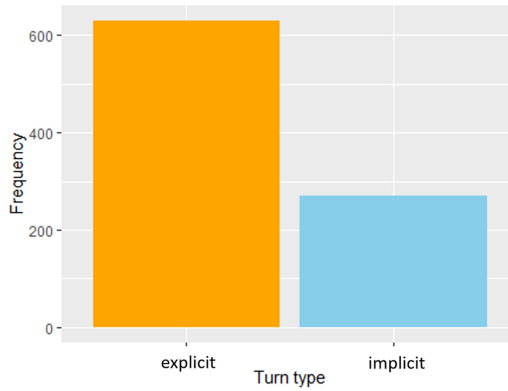
Fig. 3. Distribution of implicit and explicit coded requests over all classes.

4.2.1 Explicit Versus Implicit Queries. We observed both explicitly and implicitly expressed queries. Explicit queries took the form of a direct question, including an increase in intonation when the query was uttered. Explicit queries also exhibited the grammatical form of questions, for example, "Which ingredients are needed?" (part. 1). Implicit queries, on the other hand, largely do not possess an interrogative particle and the speaker tends to go down in intonation, as would be expected in a typical declarative sentence. An example from the corpus is "OK. And I forgot the other ingredients." (part. 9). The difference between these two types of queries was so pronounced that a distinction in the annotation schema was made to reflect this *Ingredients in Recipe — implicit* and *Ingredients in Recipe — explicit.* Figure 3 illustrates the distribution of explicit and implicit labels. It should be noted that this distinction was made only when it was obvious. Otherwise, explicit queries did not receive a separate code. This is also the reason why Figure 3 does not split all queries into one of these two categories. What it does show, however, is that explicit queries are more frequent than implicit ones. Implicit queries, nevertheless, represent a significant portion of utterances in the corpus. Thus, they need to be further analysed to identify patterns which help identify them in a dialogue. As outlined in Section 2.4, conversational agents often analyse only the surface structure of utterances, which makes it hard to fully understand dialogue pragmatics. The human annotators found it harder to identify implicit queries due to their declarative and indirect shape compared with a direct question. For example, "I will just cut it in pieces." (part. 8) illustrates this challenge. In this case, the *Preparation Step —Implicit Reassurance* label was assigned to this turn. Some people might not interpret this as an implicit query but rather as an example of simply thinking aloud. However, this utterance is embedded in a reading situation of preparation steps. Thus, it was interpreted as an implicit query to reassure the next action. This illustrates that previous turns, that is, contextual information, needs to be taken into account, especially when implicit queries should be detected.

As already mentioned, using both explicit and implicit ways of expressing queries was a common pattern observed in the corpus. Thus, dealing with both explicit and implicit queries should be taken into consideration for systems design and will be further discussed in Section 5.2.

4.2.2 Reassurances. A further means that participants used to communicate information needs was to make use of reassurances. These were used extensively by participants to determine whether they had correctly interpreted the instruction provided by the assistant. Linguistically, these are marked either with explicit queries, such as "Meanwhile, I have to let it cook, right?" (part. 1) or with implicit queries like "Er — steeping it." (part. 2). Implicit queries offer more scope

for interpretation, which made it difficult to determine whether something should be treated as a query or not, as already discussed in the section before. When using reassurances in this way, participants expected an appropriate response from the assistant. Thus, this is an interaction pattern which should be considered for system design. To present a satisfying response to the participant, the assistant had to track the task state to inform the participant as to whether the currently performed task is correct or not. Here, again, context information is an important feature for response generation.

   4.2.3   Discourse Markers and Linguistic Feedback. Discourse markers, such as "Mhm" (part. 1) and "Ähm" (part. 26) were frequently used in the cooking experiments. This was particularly noticeable in the classes *Next ingredient* and *Next preparation step.* Discourse markers relate segments pragmatically [36]. In our corpus, they were mostly used to signal that the system instruction had been understood and that it is appropriate for the assistant to move to the next step in the process. In general, in this corpus, a lot of discourse markers occur since they pragmatically link conversational units in spoken language [21]. The following example from the corpus illustrates their usage:
   "Participant: OK. Good. (Alarm bell rings). Now we need to add zucchini and corn, right? Zucchini, corn and —
   Assistant: Exactly. You need to mix corn, zucchini, and chickpeas.
   Participant: **Mhm?**
   Assistant: Fill it up with half a liter of water and add the tomato paste." (part. 5)
   This example shows that the discourse marker, highlighted in bold, signals the assistant to read out the next step in the recipe. The example also demonstrates that discourse markers require context and cannot be interpreted in isolation. "Mhm?" in the example can only be interpreted as *Next Preparation Step* if the content of previous turns, that is, the reassurance question of whether mixing zucchini and corn is correct, is known. Deriving this information from the context was quite simple during annotation since this is a common situation in everyday dialogues, where Discourse Markers such as *Mhm?* are used as linguistic feedback [7]. Its basic functions include an understanding of the currently heard and the reaction of the conversational partner to go on to the next point in the conversational flow [7]. This stylistic property should be taken into account in future systems design, as it frequently occurred in this corpus and brings structure into conversation [71]. For this reason, it is a relevant linguistic aspect to deal with as a system to better understand natural language.

This section has illustrated how three important linguistic features shaped conversational descriptions of needs. The use of explicit and implicit queries, reassurances as a means to request feedback from the assistant, and discourse markers all featured heavily in user behaviour.

So far, we have analysed which information needs can occur in cooking conversational search (see Section 4.1) and how these are communicated throughout the dialogue (see Section 4.2). One theme that is relevant for both information need understanding and detection is **context**. It is important for understanding competence needs but is also highly relevant for identifying implicit queries, reassurances, and discourse markers in a dialogue. Since context helps humans with interpretation of dialogue and needs, in the following section we want to test whether systems are able to utilise context information to predict information needs automatically.

4.3   Predicting Information Needs

Now that we understand more about what information needs occur in cooking situations and how they are expressed, in this section we explore whether the turns contain enough information for

systems to determine the kind of information need. To test this, we formulate the problem as a multi-label prediction task since a single turn may express one or more information need: Given a set of features derived from the raw conversational turns, is it possible to predict the level 1 category of information need?

4.3.1   Dataset. For the classification task, the *CookversationalSearch* dataset described earlier was employed. In our quantitative analysis, all N_t = 2675 turns with their corresponding information need label were used. As outlined in Section 4.1.1, each turn consists of one or multiple sentences. On average, one level 1 category contains 243 turns (min = 12, $x_{.25}$ = 42, $\tilde{x}$ = 182, $x_{.75}$ = 343, max = 768, sd = 269.29, $skewness$ = 0.85, $curtosis$ = −0.96).

4.3.2   Experimental Conditions. The multi-label classification problem is treated as a binary classification task whereby we attempt to classify needs at level 1 in the taxonomy. Each target turn is classified as either an information need type or not, that is, 11 classifiers (*[information need]* vs. *Other*) were trained — one for each need type. Since we found context to be an important factor for identifying information needs throughout the coding process and in the literature (see [5, 68, 91]), we tested the following conditions which model context information in the following ways:

No context — We consider only the target turn, ignoring what has been uttered previously.
One previous participant turn — We add the previous information need turn that was issued by the participant to the target turn. The previous turn and the target turn are simply concatenated.
All previous participant turns — We include all participant turns preceding the target turn. The collected turns are added as text before the target.

We account for the turns uttered by the participant only, since this has been shown to be very effective in similar prediction tasks (e.g., [35]). To avoid overfitting, we employed 10-fold cross-validation. Due to the highly skewed and imbalanced dataset, we further used stratified sampling to ensure a fair distribution in training and test sets. We also tested hierarchical classification methods but found them to significantly lower prediction performance and, thus, do not report them here. The observation that hierarchical classifiers lower prediction performance has also been made by Fornaciari et al. [35].

4.3.3   Models.
Baseline Models: To generate a baseline for our experiments, we employed some of the models, which have been shown to be effective in past experiments predicting information needs in conversational settings [38, 72, 74]. For our baseline models, we used 300-dimensional FastText embeddings [47] combined with BytePairEmbeddings provided by Flair [3], which are able to capture out of vocabulary words using sub-word information [42]. This feature was the main reason for choosing FastText embeddings, as it was considered helpful when dealing with spoken dialogue data where out of vocabulary words might occur more likely due to its natural style. For the baseline experiments, we use Gaussian Naive Bayes (=GNB or GaussianNB), Random Forest and SVM classificators provided by the Python package *scikit-learn* [64].

BERT-based models: Since Schwabl [72] showed that information need detection in the cooking domain benefits from using BERT-based models, we employed them for comparison. BERT was chosen due to its success in natural language understanding problems [98], such as Named Entity Recognition and Text Classification. For our experiments, we leveraged pre-trained BERT models:

GermanBERT[6] and the multilingual BERT models mBERT [7] and XLM-RoBERTa.[8] Due to the small amount of data, 85% was kept as training and 15% as test data to ensure that there are still samples available for the test dataset even for categories with few data. For each BERT model, we chose the $BERT_{BASE}$ architecture which was available for all three models in contrast to $BERT_{LARGE}$. Since fine-tuning pre-trained BERT models on downstream tasks has shown to improve classification performance [79], this technique was also applied in our experiments. All BERT-related experiments were conducted using the transfer-learning library FARM.[9] To fine-tune each of the 11 binary classifiers, a TextClassficationHead[10] was added on top of the BERT language model. The TextClassificationHead consists of a one-layer feed-forward neural network (FFNN) of dimension (768,2) and a softmax activation function. This follows the recommendations made by Sun et al. [79] and Devlin et al. [31]. The classification head receives the final hidden state of the [CLS] token which represents the turn. Stopwords were not removed since removing stopwords decreases the effectiveness of the BERT model [30]. Also, resampling was not conducted, as BERT handles imbalanced datasets well [82]. Class weights were adjusted by FARM's DataSilo, which also ensured stratified sampling for cross-validation.[11] We followed the suggestions by Sun et al. [79] and Devlin et al. [31] regarding the parameters to choose for fine-tuning. To avoid the "catastrophic forgetting problem", we selected a lower learning rate (2e-5) [79]. The classifiers were trained for 4 epochs with a dropout probability of 10% and a batch size of 32. To avoid overfitting, early stopping was included. The models were trained on three GeForce GTX 1080 Ti GPUs. Due to computing limitations, we had to choose a maximum sequence length of 256, which would have resulted in clipping down 63% of all samples to 256 tokens for the *all prev turns* condition. As preceding turns were added as raw text before each target turn in this condition, leaving the text unchanged would have led to cutting out the actual target turn. Thus, as a preprocessing step, only the last 256 tokens which also include the current target turn were considered. The impact on the classification results will be discussed in the limitations section.

4.3.4 Results. The results are illustrated in Table 5. We report precision, recall, and the F-measure. Accuracy was not taken into consideration since it would not be a reliable measure for such imbalanced datasets. In this subsection, we assess the performance of the different models that were employed. In the subsequent sections, we present the results that were achieved regarding the different context conditions and information needs.

Model performance: In Table 5, we report the values achieved when removing stopwords for classification. While no significant difference could be observed when stopwords were removed for the Random Forest and SVM classifiers, removing stopwords when using the GaussianNB classifier significantly improved classification performance ($p < 0.05$) compared with the results in which stopwords are kept. GaussianNB also outperforms Random Forest and SVM regarding the F-Measure in detecting information needs ($p < 0.001$ for both). Thus, it will be the baseline to test against BERT-based models.

To assess overall model performance, a one-way ANOVA as well as post-hoc tests with Bonferroni-adjusted values were conducted. These indicate that the GermanBERT model performs significantly better than mBERT ($p < 0.001$) and XLM-RoBERTa ($p < 0.001$). As shown in Table 6,

---

[6]Taken from https://huggingface.co/bert-base-german-cased.

[7]Taken from https://huggingface.co/bert-base-multilingual-cased.

[8]Taken from https://huggingface.co/xlm-roberta-base.

[9]See https://github.com/deepset-ai/FARM.

[10]See https://farm.deepset.ai/api/modeling.html#farm.modeling.prediction_head.TextClassificationHead.

[11]See https://farm.deepset.ai/api/data_handling.html#farm.data_handler.data_silo.DataSiloForCrossVal.

Table 5. Experiment Results After 10-fold Cross-Validation Grouped by Models and Conditions

Model	Condition	Precision	Recall	F-Measure	95%- CI
GaussianNB	no context	19.20%	60.72%	23.68%	[20.44%; 26.91%]
	1 prev turn	19.14%	60.67%	25.21%	[21.96%; 28.47%]
	all prev turns	14.09%	45.99%	19.51%	[15.61%; 23.41%]
Random Forest	no context	34.97%	13.02%	18.30%	[14.44%; 22.17%]
	1 prev turn	27.51%	04.35%	07.21%	[05.12%; 09.30%]
	all prev turns	02.79%	03.86%	02.69%	[01.51%; 03.87%]
SVM	no context	32.41%	14.51%	19.13%	[13.92%; 24.34%]
	1 prev turn	25.50%	07.11%	10.49%	[07.34%; 13.63%]
	all prev turns	00.16%	01.18%	00.11%	[00.00%; 00.38%]
mBERT	no context	34.42%	36.84%	34.18%	[27.51%; 40.84%]
	1 prev turn	33.76%	38.56%	34.46%	[27.85%; 41.06%]
	all prev turns	21.26%	32.09%	22.82%	[17.59%; 28.05%]
XLM-RoBERTa	no context	26.58%	34.92%	27.61%	[21.01%; 34.20%]
	1 prev turn	19.32%	29.94%	20.36%	[14.70%; 26.01%]
	all prev turns	15.37%	33.59%	18.64%	[14.18%; 23.09%]
GermanBERT	no context	41.95%	43.14%	41.43%	[34.57%; 48.28%]
	1 prev turn	42.28%	43.71%	42.11%	[35.19%; 49.03%]
	all prev turns	38.05%	40.88%	37.87%	[30.71%; 45.03%]

Table 6. Overall Average F1 Scores Model Performance

Model	F-Measure	95%-CI
GaussianNB	22.80%	[20.80%; 24.80%]
mBERT	**30.48%** ***	[26.89%; 34.08%]
XLM-RoBERTa	22.20%	[18.95%; 25.45%]
GermanBERT	**40.47%** ***	[36.48%; 44.46%]

In bold: the significant results against GNB, with $***: p < 0.001$.

mBERT and GermanBERT perform significantly better ($p < 0.001$) compared to baseline models. However, no difference was found between XLM-RoBERTa and GaussianNB.

Fine-tuning BERT with a language specific model, GermanBERT, on the text classification task yielded strong results compared with previous approaches with word embeddings that were, for example, employed in Frummet et al. [38]. The main reason for this is that BERT can capture contextual information much better than, for instance, word2vec or FastText embeddings [31]. Among the BERT models we tested, GermanBERT was found to yield the best results. One reasonable explanation for this is the more accurate tokenisation by GermanBERT, which leads to better understanding of words by the BERT model.[12]

Context Conditions: Since the GermanBERT model performs best, this model was employed to further assess classification performance regarding context conditions and information need prediction. Figure 4 illustrates that while the *1 prev turn* condition ($M = 42.11\%$) performs slightly better than *no context* ($M = 41.43\%$) and *all prev turns* ($M = 37.87\%$), no significant differences were detected after conducting one-way ANOVA analysis ($F = 0.42, p = 0.69$). The average F1 score achieved for GermanBERT among all conditions was 40.47%.

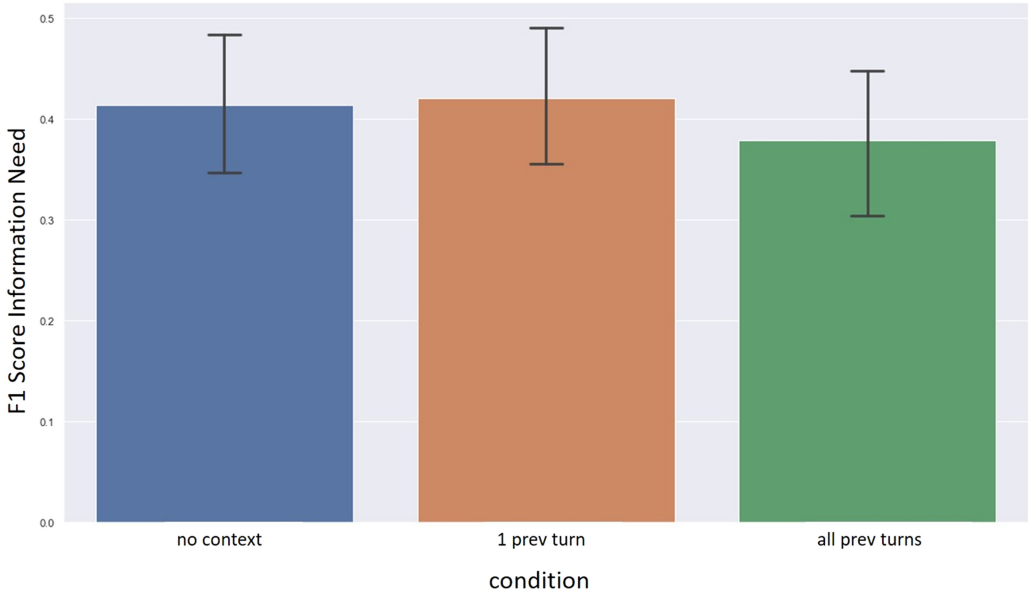[12]see https://www.deepset.ai/german-bert.

Fig. 4. Average F1 score for each condition over all information needs using the GermanBERT model.
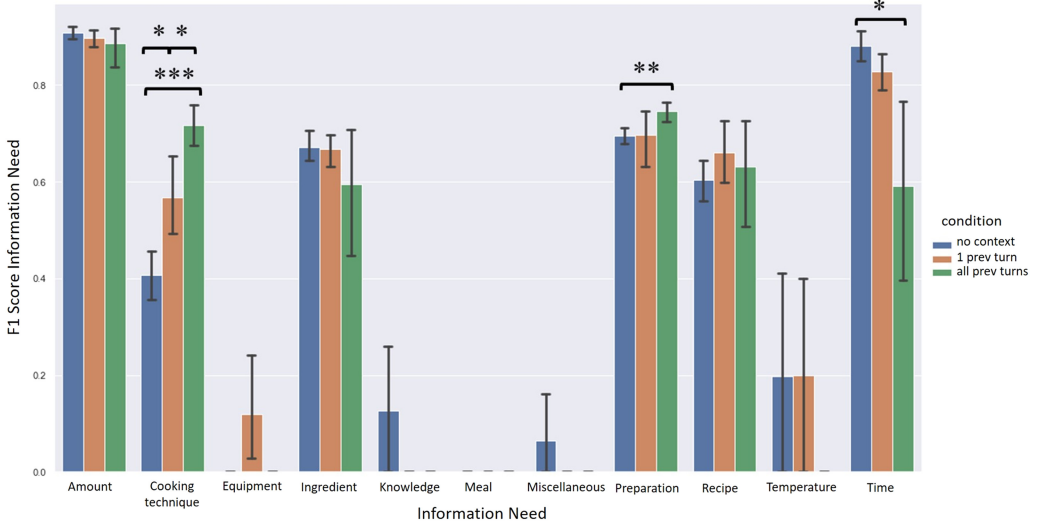

Fig. 5. F1 score for each information need grouped by condition using the GermanBERT model, with $*$: $p < 0.05$; $**$: $p < 0.01$; $***$: $p < 0.001$.

Information Need prediction: The achieved F1 scores, grouped by information need and condition for the GermanBERT model, as well as an overview of which conditions lead to significantly better results within information need categories, are illustrated in Figure 5.

While some information need categories, for example, *Amount* and *Time*, performed quite well, *Equipment, Knowledge, Meal, Miscellaneous,* and *Temperature* achieved F1 scores below 20%. This is likely due to the lack of data for these categories since the aforementioned needs relate to, on average, only 35 turns. While BERT can usually deal quite well with limited data [82], limits seem to

be reached with this dataset. The lack of data for these categories means that the results achieved cannot be interpreted reliably. Therefore, we will focus on the interpretation of the remaining information need categories in which it is possible to identify several patterns.

For the fact-based information needs *Amount*, *Time*, and *Ingredient*, the no-context condition performs best. The results of the *Time* information need indicate a significantly better performance for the condition where no-context information was taken into account compared with the *all prev turns* condition. This aligns with the impression gained during the qualitative analysis that annotating these needs was straightforward as little context needs to be accounted for. Similar patterns were observed for the competence-oriented information needs *Cooking technique* and *Preparation*. For both needs, the *all prev turns* condition performed significantly better than the no-context condition. Turns labelled as *Cooking technique* performed significantly better when more context was modelled. The *1 prev turn* condition showed significantly better results compared with the *no-context* one and *all prev turns* showed significant better performance than *1 prev turn*. This, again, aligns with the findings from Section 4.2.

The classification at level 1 gives the impression that fact-based needs are easier to predict, which aligns with the experiences of the manual annotation as explained in Section 4.1.3. To test this, we formulated a similar prediction task with the same setup but for level 0 needs. We employed the GermanBERT model and fine-tuned this as it was done in the best performing model in the first round of experiments. Fact-based information needs are indeed significantly easier to detect than competence needs ($M_{fact} = 86.68\%$, $M_{competence} = 74.78\%$, $t = 4.20, p < 0.001$).

Our experiments demonstrate that fact-based needs such as *Amount* and *Time* can be predicted with high confidence and require little context information. On the other hand, classifiers for competence-oriented needs, that is, *Cooking technique* and *Preparation*, perform better when more context information is added but are generally less accurate than classifiers for fact-based needs.

5 DISCUSSION

In this section, the results reported earlier will be discussed. In Section 5.1, we focus on our information need taxonomy. We compare it with other coding schemes from the literature and illustrate the additional value that is offered by our *CookversationalSearch* taxonomy. In Section 5.2, we discuss the implications that our results have on the design of conversational assistants and extrapolate what needs to be considered when implementing a conversational search system. Finally, we elaborate on the relevance of context for information need classification and discuss why some needs are easier to predict than others.

5.1 Comparing CookversationalSearch with other Conversational Taxonomies

To assess the value of our dataset and our taxonomy for the research community, we discuss differences in cooking- and non-cooking-related taxonomies and annotation schemes.

5.1.1 Comparison with Domain-agnostic Taxonomies. A prominent domain-agnostic taxonomy is the Dialogue Act ISO annotation scheme that was introduced by Bunt et al. [20]. This hierarchical annotation scheme contains, at the highest level, 26 communicative functions that can be used to label turns, for example, *Accept Suggest* and *Decline Suggest*. Some of these labels relate to codes in our taxonomy. For example, the *Accept Suggest* move can be observed in the *Recipe Selection* need since participants can accept or reject suggested recipes. Other similar dialogue-related functions can also be observed in datasets, such as VERBMOBIL [45], in which codes such as *Clarify* are assigned to turns that aim to make taken actions explicit and seek reassurance. *Reassurance*s form part of our information need taxonomy; this relates to the *Clarify* dialogue act, which is also part of the taxonomy of Bunt et al. A difference between these domain-agnostic coding schemes and

the one we present is that they lack a description of the associated content. Knowing what is being reassured would help the system to provide better assistance in the sense that systems then know how they should react. For example, *Amount of Ingredient — Reassurance* (level 3 code) specifies that participants want to be reassured with respect to the amount of a certain ingredient. A system can then react accordingly and clarify whether the amount of an ingredient has been understood correctly.

Our coding scheme also goes beyond that proposed by Shiga et al. [74] which, in contrast to the aforementioned coding schemes, offers a stronger relation to information needs. They employed Taylor's information needs [81] and combined them with high-level task-related needs, such as *Topic*, *Problem Solving*, and *Search* in a travel planning task. Although — like our work — the coding scheme is applied in a conversational search setting, the way the needs are classified was coarse-grained and abstract. We argue that these labels are not specific enough to adequately describe travel planning content such that systems can provide assistance. Our taxonomy, on the other hand, illustrates that two different kinds of assistance should be provided: fact based, in which information is provided, and competence oriented, which revolves around teaching skills. Moreover, the tree-like structure allows needs to be distinguished based on content. This, again, helps the system tailor the information provided. To summarise, the coding scheme we developed offers utility beyond domain-agnostic taxonomies (e.g., see Bunt et al. [20]) and the information need classifications (e.g., Taylor [81] and Shiga et al. [74]) since it offers finer-grained descriptive qualities that are content specific. In the next section, we compare our contribution with previously published taxonomies related to cooking.

5.1.2 Comparison with Cooking-related Taxonomies. Our coding scheme goes beyond those previously collated in the cooking domain. For example, Nouri et al. [62] created a very controlled environment and had a very small group of users (eight participants) that used smart speakers privately and were cooking pre-selected recipes. In a similar environment, Vtyurina and Fourney [94] analysed intents that occur during cooking, again with only 10 participants. In our study, we had 45 participants and the recipe search was part of the natural dialogue (i.e., uncontrolled). Moreover, participants in our study conversed in a more natural manner with the system, that is, the experimenter. Due to the small sample size, Nouri et al. identified only a small number of cooking-related intents, which we list in Table 7. Table 7 shows how the level 1 categories in our taxonomy relate to those reported in Nouri et al. [62] and Cunningham and Bainbridge [29]. For intents such as ingredient_quantity, ingredient_list_all, and step_next, *Amount*, *Ingredient*, and *Preparation* can be found as equivalents. Our taxonomy, however, goes beyond their limited list of cooking-related intents to offer a broader and finer-grained description of needs that can occur during cooking.

A second obvious point of comparison is the taxonomy created by Cunningham and Bainbridge [29], which is perhaps the closest to our work. As outlined in Section 2, their taxonomy is based on Google Answer posts that were created by hobby chefs and distinguishes 13 different needs that were uttered in these posts. Most frequently, posters requested information about products and recipes but were also curious about definitions and cooking techniques. These are needs similar to the ones detected in our corpus, that is, *Recipe*, *Knowledge* (akin to *definitions*) and *Cooking technique*. However, Cunningham and Bainbridge's taxonomy covers needs of only one specific group, hobby chefs, which is the reason why needs such as food trivia and food history, which are less likely to occur in an everyday cooking situation, are part of their taxonomy. Our coding scheme is more applicable to practical situations in which assistance may be required, which is reflected in additional codes such as *Time* being prominent.

Table 7. Comparison of Selected Cooking User Intents and Needs from Nouri et al. [62],
Cunningham & Bainbridge [29] and our Work (Right Column)

Intents (Nouri et al.)	Response Types (Cunningham & Bainbridge)	Level 1 info. need
ingredient_quantity	–	Amount
ingredient_list_all	–	Ingredient
step_next	–	Preparation
recipe_name	Recipes	Recipe
–	Definitions	Knowledge
–	Cooking Techniques	Cooking technique
–	–	Time
–	–	Equipment

For demonstration purposes, we highlight just some of the categories that are in our taxonomy and not in past works.

This section has illustrated the value of our taxonomy for the research community. It is the first conversational search-related taxonomy for the cooking domain that represents information needs to such a level of detail and, at the same time, contains codes that allow a broader range of needs — ones that are likely to require assistance in everyday cooking — to be described.

5.2 Implications for the Interaction Design of Conversational Assistants

In this section, we discuss which aspects are considered relevant when designing the interaction with conversational assistants. With this aim in mind, we elaborate on the role of conversational cues in conversational search dialogues, as well as to what extent our findings may be generalised beyond the scenario of assisting cooking.

5.2.1 The Role of Conversational Cues. In Section 4.2, we identified conversational cues, which seem to be relevant in the context of information need communication and, thus, should be considered when designing conversational systems. We observed the use of both explicit and implicit queries to convey information needs. Conversational agents often process only those turns that exhibit the shape of explicit queries (see [6, 51]), as is the case in this example: "Which ingredients are needed?". Queries of this type make it easier for systems to interpret utterances as system input that needs to be processed since less context modeling is required [68]. Existing dialogue management systems such as *RavenClaw* [14], which are employed in task-oriented dialogue contexts, are good at handling explicit requests. However, they have trouble coping with the implicit queries. If such systems are to interact as humans do, they would need to deal with the kind of implicit requests that were outlined in Section 4.2. Our results align with findings from Vtyurina and Fourney, with both studies finding a similar proportion ($\approx 25\%$) of implicit queries. In their analyses, Vtyurina and Fourney [94] also found that grounding behaviour was exhibited by means of verbatim repetitions of what has been understood — a pattern described as *Reassurance* in our study. For example, with a turn such as "Covered?" (part. 15), a participant is seeking reassurance as to whether the taken action is correct and, therefore, grounds one's behaviour by requesting feedback from the system. Grounding and relevance feedback have also been considered appropriate for conversational systems by Trippas et al. [91].

To summarise, our findings reinforce patterns observed in previous studies in that they underline — even in this specific cooking context context — that conversational cues are important for assistants to identify and provide appropriate assistance. Accounting for concepts such as implicit queries, grounding, and reassurance is imperative if human-like interaction patterns are to be supported.

5.2.2 Generalisability of our Findings. The presented study was performed in the fairly specific context of home cooking. Nevertheless, there are aspects of the task performed that might mean that our findings transfer to other situations. Cooking tasks are highly procedural, goal-oriented tasks. They involve completing a sequence of individual steps to achieve a larger goal. There are other situations which would be comparable. For example, we could imagine that tasks such as bicycle maintenance, for example, fixing a punctured tyre, or building flat-pack furniture would have considerable overlap with the cooking examples studied here. We believe, for instance, that since such tasks are similarly sequential, one could expect that similar usage of discourse markers such as "Mhm" or "OK" indicate task completion.

Furthermore, we would anticipate a similar high-level delineation of information need types in these situations to fact-based/competence-based distinctions we present at level 0 in our taxonomy. For example, when building a shelf, a user might ask "How many screws do I need to fix the bottom plate?" (fact-based). On the other hand, competence-oriented needs could also be expected, for example, "How do I fix the bottom plate to the wall?" (competence-oriented).

We would argue that it is not only the high-level codes that would transfer. It seems plausible that some level 1 needs may also apply in other domains in which procedural tasks take place. The turn "How many screws do I need to fix the bottom plate?", for example, expresses an *Amount*-related need that can be used in both the cooking domain and the DIY domain, for instance. Also, utterances asking for system capabilities, such as "Do you have an integrated timer?" (labelled as *Miscellaneous*), seem to be generic and applicable to other domains. Similarly, turns such as "What do I have to do next?", which are labelled as *Preparation* in our corpus, seem to generalise beyond the cooking context.

One could easily perform a comparable study in a do-it-yourself (DIY) context to test to what extent our codes and the classification performance transfer to this domain. The Alexa Prize TaskBot challenge 2021,[13] in which teams develop conversational assistants for both cooking and DIY tasks, underlines how closely related to each other these domains are.

In this section, we have discussed the relevance of conversational cues for system design and arguing, for example, that implicit requests are important, particularly in procedural tasks, to replicate human to human–style interaction. We have also argued that there is potential for our results to transfer into other domains that feature procedural tasks. We named DIY assistance as one example.

5.3 On the Relevance of Context

Both the qualitative annotation process and the prediction experiments underlined the importance of context information needs. The results from our prediction experiment revealed that fact-based needs perform significantly better than competence-oriented needs. One reason for this is that fact-based information needs contain more information need-discriminating words (illustrated in Table 4). The annotation process revealed that the competence-oriented categories require more context to be taken into account given the lack of distinctive features, that is, words, which makes it challenging for both human coders and classifiers to distinguish classes. This assertion is supported by the classification experiment results. While fact-based needs, such as *Amount* and *Time*, achieved F1 scores of ≈90%, *Cooking technique* and *Preparation* achieved results of ≈70% in their best performing condition. Including the context in the form of previous turns, however, significantly improved results in competence-oriented needs *Cooking technique* and *Preparation*.

---

[13]See https://www.amazon.science/academic-engagements/ten-university-teams-selected-to-participate-in-alexa-prize-taskbot-challenge.

With respect to the development of future systems, we argue that while we can reliably detect *Amount* and *Time* needs and, to some extent, other fact-based needs — such as *Temperature*, which achieved low results due to lack of data — other strategies need to be established to identify competence needs. Similar to our results, Ren et al. [68] and Aliannejadi et al. [5] showed that including more context, that is, conversational history, improves the classification performance of the current turn. Recently, this problem was investigated by Nouri et al. [63], who tried to enable step-wise recommendations for current steps in a controlled cooking task, including different contextual information.

Our results demonstrate the feasibility of predicting some types of need. To further improve the detection of the classes that are trickier to identify, we will need to find other ways for augmenting and representing the cooking process and the context. One way to achieve this may be to reformulate queries by adding missing context from the conversation history to the current turn query [93]. In our study, we employed a simple approach, modelling information need turns with different amounts of context by adding previous queries as plain text to the current query. However, more elaborate methods exist to better model context. See, for example, Voskarides et al. [93], Yu et al. [100], and Lin et al. [55, 56]. In future work, more effort can be put into this by building on prior work.

In summary, our results — both of the annotation and classification — highlight the importance of context, particularly in identifying competence needs. Applying more elaborate strategies to achieve a better representation of context information will be part of our future work.

6 LIMITATIONS

In this section, we reflect on some of the limitations of this study. When performing data collection with 45 participants, we cannot exclude learning effects by the experimenter mimicking a conversational system. From session to session, the experimenter increased in confidence and developed strategies for interacting with participants. This affected, for example, the amount of information that was provided when participants issued queries. The means of interacting with participants was certainly not completely consistent throughout the study. However, we do not believe that this detracts from the quality of the data since the interaction patterns were always plausible, human dialogue. Moreover, we did not observe different information needs or particular phraseology at different time points when transcribing or analysing the data. It is also possible that having the experimenter act as the assistant led to specific outcomes (experimenter bias). However, this is unlikely, as the experimenter did not have any preconceived ideas of what assistance should be provided and the conversations were largely driven by the participants with utterances arising naturally from the cooking process.

A further point worthy of discussion is that since we conducted an in situ study, both experimenter and participant were able to see and interact with each other using non-oral communication (e.g., body language). This raises the question of whether the physical environment, that is, gestures and visual aspects, might have influenced the dialogue interaction. While we fully relied on speech-only interaction and the experimenter did not accept queries that were not possible in such a scenario, minor environmental effects cannot be fully excluded. When participants employed anaphoras to refer to ingredients, as in "Is **this** enough?" seeing what the participant was doing might have unconsciously simplified the interpretation of utterances which would be less clear in a spoken-only Wizard-of-Oz scenario.

One limitation of our quantitative experiment relates to the representation of the multi-label classification as a binary classification task. While we learned about how well information needs can be identified based on an utterance, we cannot measure whether co-occurrences of certain

needs influence one another, which is a common issue for binary relevance classification (see Luaces et al. [58] and Zhang et al. [103]). The influence of co-occurring information needs on the classification results can be investigated in future studies.

Finally, when modelling context, we employed the plain text from previous turns. However, this text can be represented in more elaborate ways, as outlined in Section 5.2. Moreover, the impact of linguistic features, such as intonation and the explicit-implicit dimension on the classification results, could have been investigated in further detail and will be part of our future work.

7 CONCLUSION

Our results provide an in-depth understanding of the information needs occurring in the domain of home cooking. RQ1 raised the question of what kind of information needs arise when cooking with a conversational assistant. To answer this, we provided a detailed hierarchical information need taxonomy that shows us a variety of different needs that can occur in a cooking scenario. These needs can broadly be categorised in different types, fact based and competence oriented, which require different types and degrees of assistance. Addressing RQ2, these needs can be communicated through different linguistic means, both as explicit and implicit queries, and discourse markers are often used. Furthermore, our findings indicate the feasibility of identifying needs automatically (RQ3). In this respect, our experiments demonstrate some success, with fact-based information needs being particularly feasible. We have also shed light on the problems that need to be solved to successfully detect and process other, more challenging information need types in an automated scenario. These mainly include finding methods to better incorporate context and conversational history. Future work will focus on creating methods to better incorporate context information in the classification and assistance process.

REFERENCES

[1] Ali Ahmadvand. 2020. User intent inference for web search and conversational agents. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, Houston, TX *(WSDM'20)*. ACM, New York, NY, 911–912. https://doi.org/10.1145/3336191.3372187

[2] Ali Ahmadvand, Jason Ingyu Choi, and Eugene Agichtein. 2019. Contextual dialogue act classification for open-domain conversational agents. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Paris, France *(SIGIR'19)*. ACM, New York, NY, 1273–1276. https://doi.org/10.1145/3331184.3331375

[3] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Minneapolis, Minnesota*. 54–59.

[4] Mohammad Aliannejadi, Manajit Chakraborty, Esteban Andrés Ríssola, and Fabio Crestani. 2020. Harnessing evolution of multi-turn conversations for effective answer retrieval. In *Proceedings of the Conference on Human Information Interaction and Retrieval*, Vancouver BC, Canada *(CHIIR'20)*. ACM, New York, NY, 33–42. https://doi.org/10.1145/3343413.3377968

[5] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval,* Paris, France) *(SIGIR'19)*. ACM, New York, NY, 475–484. https://doi.org/10.1145/3331184.3331265

[6] James Allen, George Ferguson, and Amanda Stent. 2001. An architecture for more realistic conversational systems. In *Proceedings of the 6th International Conference on Intelligent User Interfaces,* Santa Fe, New Mexico, *(IUI'01)*. ACM, New York, NY, 1–8. https://doi.org/10.1145/359784.359822

[7] Jens Allwood, Joakim Nivre, and Elisabeth Ahlsen. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics* 9, 1 (01 1992), 1–26. https://doi.org/10.1093/jos/9.1.1 arXiv:https://academic.oup.com/jos/article-pdf/9/1/1/9836977/1.pdf.

[8] Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2019. Conversational search (Dagstuhl seminar 19461). *Dagstuhl Reports* 9, 11 (2019), 34–83. https://doi.org/10.4230/DagRep.9.11.i

[9] Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 4 (2008), 555–596. https://doi.org/10.1162/coli.07-034-R2

[10] Vevake Balaraman and Bernardo Magnini. 2020. Domain-aware dialogue state tracker for multi-domain dialogue systems. *CoRR* abs/2001.07526 (2020). arXiv:2001.07526. https://arxiv.org/abs/2001.07526.

[11] Nicholas J. Belkin, Helen M. Brooks, and Penny J. Daniels. 1987. Knowledge elicitation using discourse analysis. *International Journal of Man-Machine Studies* 27, 2 (1987), 127–144.

[12] Nicholas J. Belkin, Colleen Cool, Diane Kelly, S.-J. Lin, S. Y. Park, J. Perez-Carballo, and C. Sikora. 2001. Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing & Management* 37, 3 (2001), 403–434.

[13] Michael S. Bernstein, Jaime Teevan, Susan Dumais, Daniel Liebling, and Eric Horvitz. 2012. Direct answers for search queries in the long tail. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, Texas*. ACM, 237–246.

[14] Dan Bohus and Alexander I. Rudnicky. 2009. The RavenClaw dialog management framework: Architecture and systems. *Computer Speech & Language* 23, 3 (2009), 332–361. https://doi.org/10.1016/j.csl.2008.10.001

[15] J. Boose. 1993. A survey of knowledge acquisition techniques and tools. *Knowledge Acquisition* 1 (1993), 39–56.

[16] Adrian Boteanu, Emily Dutile, Adam Kiezun, and Shay Artzi. 2020. Subjective search intent predictions using customer reviews. In *Proceedings of the Conference on Human Information Interaction and Retrieval, Vancouver BC, Canada (CHIIR'20)*. ACM, New York, NY, 303–307. https://doi.org/10.1145/3343413.3377987

[17] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa.

[18] V. Braun and V. Clarke. 2013. *Successful Qualitative Research: A Practical Guide for Beginners*. SAGE Publications. https://books.google.de/books?id=EV_Q06CUsXsC.

[19] Andrei Broder. 2002. A taxonomy of web search. *SIGIR Forum* 36, 2 (Sept. 2002), 3–10.

[20] Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017. *Dialogue Act Annotation with the ISO 24617-2 Standard*. Springer International Publishing, Cham, 109–135. https://doi.org/10.1007/978-3-319-42816-1_6

[21] Hadumod Bußmann. 2008. *Lexikon der Sprachwissenschaft*. Kröner.

[22] Katriina Byström and Kalervo Järvelin. 1995. Task complexity affects information seeking and use. *Information Processing & Management* 31, 2 (March 1995), 191–213.

[23] V. Orengo Castellá, A. M. Zornoza Abad, F. Prieto Alonso, and J. M. Peiró Silla. 2000. The influence of familiarity among group members, group atmosphere and assertiveness on uninhibited behavior through three different communication media. *Computers in Human Behavior* 16, 2 (2000), 141–159. https://doi.org/10.1016/S0747-5632(00)00012-1

[24] K. Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. SAGE Publications. https://books.google.de/books?id=v1qP1KbXz1AC.

[25] Karen Church, Mauro Cherubini, and Nuria Oliver. 2014. A large-scale study of daily information needs captured in situ. *ACM Trans. Comput.-Hum. Interact.* 21, 2, Article 10 (February 2014), 46 pages. DOI : https://doi.org/10.1145/2552193

[26] Juliet Corbin. 2016. Taking an analytic journey. In *Developing Grounded Theory: The Second Generation*, J. M. Morse, P. N. Stern, J. Corbin, B. Bowers, K. Charmaz, and A. E. Clarke (Eds.). Taylor & Francis.

[27] Nick Craswell, David Hawking, and Stephen Robertson. 2001. Effective site finding using link anchor information. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information retrieval, New Orleans, Louisiana, USA (SIGIR'01)*. Association for Computing Machinery, New York, NY, USA, 250–257. DOI : https://doi.org/10.1145/383952.383999

[28] J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. 2018. Research frontiers in information retrieval: Report from the 3rd Strategic Workshop on Information Retrieval in Lorne (*SWIRL'18*). *SIGIR Forum* 52, 1 (2018), 34–90. https://doi.org/10.1145/3274784.3274788

[29] Sally Jo Cunningham and David Bainbridge. 2013. An analysis of cooking queries: Implications for supporting leisure cooking ethnographic studies of cooks and cooking. In *Conference 2013 Proceedings, Fort Worth, Texas, USA*. 112–123. https://doi.org/10.9776/13160

[30] Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 985–988. https://doi.org/10.1145/3331184.3331303

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, MN, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[32] Thorsten Drehsing and Thorsten Pehl. 2015. *Praxisbuch Interview, Transkription & Analyse. Anleitungen und Regelsysteme für qualitativ Forschende* (6th. ed.). Dr. Dresing und Pehl GmbH, Marburg.

[33] David Elsweiler and Ian Ruthven. 2007. Towards task-based personal information management evaluations. In *Proceedings of SIGIR'07* (Amsterdam, The Netherlands). ACM, New York, NY, 23–30.

[34] David Elsweiler, Christoph Trattner, and Morgan Harvey. 2017. Exploiting food choice biases for healthier recipe recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan*. ACM, 575–584.

[35] Tommaso Fornaciari, Federico Bianchi, Massimo Poesio, and Dirk Hovy. 2021. BERTective: Language models and contextual information for deception detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 2699–2708. https://aclanthology.org/2021.eacl-main.232.

[36] Bruce Fraser. 1999. What are discourse markers? *Journal of Pragmatics* 31, 7 (1999), 931–952. https://doi.org/10.1016/S0378-2166(98)00101-5

[37] B. French. 1990. User needs and library services in agricultural sciences. *Library Trends* 38 (1990), 415–441.

[38] Alexander Frummet, David Elsweiler, and Bernd Ludwig. 2019. Detecting domain-specific information needs in conversational search dialogues. In *Proceedings of the 3rd Workshop on Natural Language for Artificial Intelligence, Rende, Italy*.

[39] Gabriel Gordon-Hall, Philip John Gorinski, Gerasimos Lampouras, and Ignacio Iacobacci. 2020. Show us the way: Learning to manage dialog from demonstrations. *CoRR* abs/2004.08114 (2020). arXiv:2004.08114 https://arxiv.org/abs/2004.08114

[40] Ido Guy. 2016. Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, Pisa, Italy (SIGIR'16)*. Association for Computing Machinery, New York, NY, USA, 35–44. DOI:https://doi.org/10.1145/2911451.2911525

[41] Jiyin He and Emine Yilmaz. 2017. User behaviour and task characteristics: A field study of daily information behaviour. In *Proceedings of CHIIR'17* (Oslo, Norway). ACM, New York, NY, 67–76.

[42] Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1473.

[43] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.* 38, 3, Article 21 (April 2020), 32 pages. https://doi.org/10.1145/3383123

[44] Amy Isard and Jean Carletta. 1995. Replicability of transaction and action coding in the map task corpus. In *AAAI 1995 Spring Symposium Series: Empirical Methods in Discourse Interpretation and Generation*. 60–66.

[45] Susanne Jekat, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and J. Joachim Quantz. 1995. Dialogue acts in VERBMOBIL. *Verbmobil Report* 1998, 65 (1995).

[46] Megha Jhunjhunwala, Caleb Bryant, and Pararth Shah. 2020. Multi-action dialog policy learning with interactive human teaching. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 1st virtual meeting, 290–296. https://www.aclweb.org/anthology/2020.sigdial-1.36.

[47] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).

[48] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. 2015. Development and evaluation of search tasks for IIR experiments using a cognitive complexity ramework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval, Northampton, MA (ICTIR'15)*. Association for Computing Machinery, New York, NY, USA, 101–110. DOI:https://doi.org/10.1145/2808194.2809465

[49] Talaat Khalil, Kornel Kiełczewski, Georgios Christos Chouliaras, Amina Keldibek, and Maarten Versteegh. 2019. Cross-lingual intent classification in a low resource industrial setting. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, Hong Kong, China, 6419–6424. https://doi.org/10.18653/v1/D19-1676

[50] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, Pisa, Italy (SIGIR'16)*. Association for Computing Machinery, New York, NY, USA, 45–54. DOI:https://doi.org/10.1145/2911451.2911521

[51] Stefan Kopp, Lars Gesellensetter, Nicole C. Krämer, and Ipke Wachsmuth. 2005. A conversational agent as museum guide: Design and evaluation of a real-world application. In *Intelligent Virtual Agents: Proceedings of the 5th International Conference. Lecture Notes in Artificial Intelligence, Vol. 3661*, Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier, and Thomas Rist (Eds.). Springer, Berlin, 329–343.

[52] Carol C. Kuhlthau. 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science* 42, 5 (1991), 361–371.

[53] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174. http://www.jstor.org/stable/2529310.

[54] Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 1311–1316. https://www.aclweb.org/anthology/D19-1131.

[55] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy J. Lin. 2020. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. *ArXiv* abs/2004.01909 (2020).

[56] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy J. Lin. 2020. Query reformulation using query history for passage retrieval in conversational search. *ArXiv* abs/2005.02230 (2020).

[57] Han Liu, Xiaotong Zhang, Lu Fan, Xuandi Fu, Qimai Li, Xiao-Ming Wu, and Albert Y. S. Lam. 2019. Reconstructing capsule networks for zero-shot intent classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, Hong Kong, China, 4799–4809. https://doi.org/10.18653/v1/D19-1486

[58] O. Luaces, J. Díez, José Barranquero, J. J. Coz, and A. Bahamonde. 2012. Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence* 1 (2012), 303–313.

[59] Fabrizio Macagno and Sarah Bigi. 2017. Analyzing the pragmatic structure of dialogues. *Discourse Studies* 19, 2 (2017), 148–168. https://doi.org/10.1177/1461445617691702 arXiv:https://doi.org/10.1177/1461445617691702

[60] Eleanor McLellan, Kathleen M. MacQueen, and Judith L. Neidig. 2003. Beyond the qualitative interview: Data preparation and transcription. *Field Methods* 15, 1 (2003), 63–84. https://doi.org/10.1177/1525822X02239573 arXiv:https://doi.org/10.1177/1525822X02239573

[61] Fabrizio Morbini, Eric Forbell, David DeVault, Kenji Sagae, David R. Traum, and Albert A. Rizzo. 2012. A mixed-initiative conversational dialogue system for healthcare. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue,* Seoul, South Korea *(SIGDIAL'12)*. Association for Computational Linguistics, Stroudsburg, PA, 137–139. http://dl.acm.org/citation.cfm?id=2392800.2392825.

[62] Elnaz Nouri, Adam Fourney, Robert Sim, and Ryen W. White. 2019. Supporting complex tasks using multiple devices. In *Proceedings of Task Intelligence Workshop @ WSDMConference (WSDM'19)*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.475/1234

[63] Elnaz Nouri, Robert Sim, Adam Fourney, and Ryen W. White. 2020. Step-wise recommendation for complex task support. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval,* Vancouver, BC, Canada *(CHIIR'20)*. ACM, New York, NY, 203–212. https://doi.org/10.1145/3343413.3377964

[64] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12 (Nov. 2011), 2825–2830. http://dl.acm.org/citation.cfm?id=1953048.2078195.

[65] Paul Prekop. 2002. A qualitative study of collaborative information seeking. *Journal of Documentation* 58, 5 (2002), 533–547. https://doi.org/10.1108/00220410210441000 arXiv:https://doi.org/10.1108/00220410210441000

[66] Chen Qu, Liu Yang, W. Bruce Croft, Yongfeng Zhang, Johanne R. Trippas, and Minghui Qiu. 2019. User intent prediction in information-seeking conversations. In *Proceedings of the Conference on Human Information Interaction and Retrieval,* Glasgow, Scotland, UK *(CHIIR'19)*. ACM, New York, NY, 25–33. https://doi.org/10.1145/3295750.3298924

[67] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the Conference on Conference Human Information Interaction and Retrieval,* Oslo, Norway *(CHIIR'17)*. ACM, New York, NY, 117–126. https://doi.org/10.1145/3020165.3020183

[68] Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten de Rijke. 2020. Conversations with search engines. *CoRR* abs/2004.14162 (2020). arXiv:2004.14162 https://arxiv.org/abs/2004.14162.

[69] Jean-François Rouet. 2003. What was I looking for? The influence of task specificity and prior knowledge on students' search strategies in hypertext. *Interacting with Computers* 15, 3 (2003), 409–428.

[70] Ian Ruthven. 2019. The language of information need: Differentiating conscious and formalized information needs. *Information Processing & Management* 56, 1 (2019), 77-90. https://doi.org/10.1016/j.ipm.2018.09.005

[71] Deborah Schiffrin. 1987. *Discourse Markers: Contextual Coordinates of Talk*. Cambridge University Press, 312–330. https://doi.org/10.1017/CBO9780511611841.011

[72] Patrick Schwabl. 2021. Classifying user information needs in cooking dialogues: An empirical performance evaluation of transformer networks. Retrieved December 16, 2021 from https://epub.uni-regensburg.de/46308/.

[73] Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative deep neural networks for dialogue: A short review. *CoRR* abs/1611.06216 (2016). arXiv:1611.06216 http://arxiv.org/abs/1611.06216.

[74] Sosuke Shiga, Hideo Joho, Roi Blanco, Johanne R. Trippas, and Mark Sanderson. 2017. Modelling information needs in collaborative search conversations. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval,* Shinjuku, Tokyo, Japan *(SIGIR'17).* ACM, New York, NY, 715–724. https://doi.org/10.1145/3077136.3080787

[75] S. Shute and P. J. Smith. 1993. Knowledge-based search tactics. *Information Processing & Management* 29 (1993), 29–46.

[76] Brendan Spillane, Emer Gilmartin, Christian Saam, Ketong Su, Benjamin R. Cowan, Séamus Lawless, and Vincent Wade. 2017. Introducing ADELE: A personalized intelligent companion. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents,* Glasgow, UK *(ISIAA'17).* ACM, New York, NY, 43–44. https://doi.org/10.1145/3139491.3139492

[77] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modelling for automatic tagging and recognition of conversational speech. *Computational Linguistics* 26 (2000), 339- 373.

[78] Anselm Strauss and Juliet Corbin. 1996. *Grounded Theory: Grundlagen Qualitativer Sozialforschung.* Beltz, Weinheim.

[79] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification?. In *Chinese Computational Linguistics*, Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu (Eds.). Springer International Publishing, Cham, 194–206.

[80] Robert S. Taylor. 1962. The process of asking questions. *American Documentation* 13, 4 (1962), 391–396. https://doi.org/10.1002/asi.5090130405 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.5090130405.

[81] Robert S. Taylor. 1968. Question-negotiation and information seeking in libraries. *College & Research Libraries* 29, 3 (1968), 178–194. https://doi.org/10.5860/crl_29_03_178

[82] Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2019. Cost-sensitive BERT for generalisable sentence classification on imbalanced data. In *Proceedings of the 2nd Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda.* Association for Computational Linguistics, Hong Kong, China, 125–134. https://doi.org/10.18653/v1/D19-5018

[83] Jaime Teevan, Edward Cutrell, Danyel Fisher, Steven M. Drucker, Gonzalo Ramos, Paul André, and Chang Hu. 2009. Visual snippets: Summarizing web pages for search and revisitation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA (CHI'09).* ACM, New York, NY, 2023–2032. DOI: https://doi.org/10.1145/1518701.1519008

[84] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and alignment in information-seeking conversation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, New Brunswick, NJ, USA.* ACM, New York, NY, 42–51.

[85] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2020. Expressions of style in information seeking conversation with an agent. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval,* virtual event, China *(SIGIR'20).* ACM, New York, NY, 1171–1180. https://doi.org/10.1145/3397271.3401127

[86] Cynthia A. Thompson, Mehmet H. Göker, and Pat Langley. 2004. A personalized system for conversational recommendations. *J. Artif. Int. Res.* 21, 1 (March 2004), 393–428. http://dl.acm.org/citation.cfm?id=1622467.1622479.

[87] Elaine G. Toms, Heather O'Brien, Tayze Mackenzie, Chris Jordan, Luanne Freund, Sandra Toze, Emilie Dawe, and Alexandra MacNutt. 2008. *Task Effects on Interactive Search: The Query Factor.* Springer, Berlin, 359–372.

[88] Johanne Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. A conversational search transcription protocol and analysis. In *Proceedings of ACM SIGIR Workshop on Conversational Approaches for Information Retrieval, August 2017 (CAIR'17).* Tokyo, Japan.

[89] 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 Conference on Human Information Interaction & Retrieval,* New Brunswick, NJ *(CHIIR'18).* ACM, New York, NY, 32–41. https://doi.org/10.1145/3176349.3176387

[90] 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 Conference Human Information Interaction and Retrieval,* Oslo, Norway *(CHIIR'17).* ACM, New York, NY, 325–328. https://doi.org/10.1145/3020165.3022144

[91] Johanne R. Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a model for spoken conversational search. *Information Processing & Management* 57, 2 (2020), 102162. https://doi.org/10.1016/j.ipm.2019.102162

[92] Andrea Vanzo, E. Bastianelli, and Oliver Lemon. 2019. Hierarchical multi-task natural language understanding for cross-domain conversational AI: HERMIT NLU. In *Proceedings of the SIGDial 2019 Conference, Stockholm, Sweden.* Association for Computational Linguistics, 254–263.

[93] Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, virtual event, China.* Association for Computing Machinery, New York, NY, USA, 921–930. DOI:https://doi.org/10.1145/3397271.3401130

[94] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the role of conversational cues in guided task support with virtual assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems,* Montreal, QC, Canada *(CHI'18).* ACM, New York, NY, 1–7. https://doi.org/10.1145/3173574.3173782

[95] Cheng Wang, Hui Li, and Mathias Niepert. 2020. RecSys-DAN: Discriminative adversarial networks for cross-domain recommender systems. *IEEE Transactions on Neural Networks and Learning Systems* 31 (2020), 2731–2740.

[96] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers.* Association for Computational Linguistics, Valencia, Spain, 438–449. https://www.aclweb.org/anthology/E17-1042.

[97] Barbara Wildemuth, Luanne Freund, and Elaine G. Toms. 2014. Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation* 70, 6 (07 10 2014), 1118–1140.

[98] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:cs.CL/1910.03771

[99] Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, R. Chulaka Gunasekara, Walter S. Lasecki, Jonathan K. Kummerfeld, Michel Galley, Chris Brockett, Jianfeng Gao, Bill Dolan, Xiang Gao, Huda AlAmri, Tim K. Marks, Devi Parikh, and Dhruv Batra. 2019. Dialog system technology challenge 7. *CoRR* abs/1901.03461 (2019). arXiv:1901.03461 http://arxiv.org/abs/1901.03461.

[100] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, China.* Association for Computing Machinery, New York, NY, USA, 1933–1936. DOI:https://doi.org/10.1145/3397271.3401323

[101] Xiaojun Yuan and Nicholas J. Belkin. 2014. Applying an information-seeking dialogue model in an interactive information retrieval system. *Journal of Documentation* 70, 5 (2014), 829–855.

[102] Hamed Zamani, Bhaskar Mitra, Everest Chen, Gord Lueck, Fernando Diaz, Paul N. Bennett, Nick Craswell, and Susan T. Dumais. 2020. Analyzing and learning from user interactions for search clarification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval,* virtual event, China*(SIGIR'20).* ACM, New York, NY, 1181–1190. https://doi.org/10.1145/3397271.3401160

[103] Min-Ling Zhang, Yukun Li, Xu-Ying Liu, and Xin Geng. 2017. Binary relevance for multi-label learning: An overview. *Frontiers of Computer Science* 12 (2017), 191–202.