Detecting domain-specific information needs in conversational search dialogues

Alexander Frummet, David Elsweiler, and Bernd Ludwig

University of Regensburg, Germany
{alexander.frummet, david.elsweiler, bernd.ludwig}@ur.de

Abstract. As conversational search becomes more pervasive, it becomes increasingly important to understand the user's underlying needs when they converse with such systems in diverse contexts. We report on an insitu experiment to collect conversationally described information needs in a home cooking scenario. A human experimenter acted as the perfect conversational search system. Based on the transcription of the utterances, we present a preliminary coding scheme comprising 27 categories to annotate the information needs of users. Moreover, we use these annotations to perform prediction experiments based on random forest classification to establish the feasibility of predicting the information need from the raw utterances. We find that a reasonable accuracy in predicting information need categories is possible and evidence the importance of stopwords in the classification task.

Keywords: Conversational Search, Information Needs, Prediction, Cooking

1 Introduction and Motivation

Voice-based interaction systems are changing the way people seek information, making search more conversational [14, 35]. Spoken queries are very different to typed queries [14] and by mining spoken interaction data, intelligent assistance can be provided [16]. Voice-based digital assistants such as Amazon Echo and Google Home show that information seeking conversations now take place in diverse situations embedded in users' everyday lives. They utilise both knowledge from research on fields at information retrieval and NLP. One crucial feature for this kind of assistant is the ability to understand and infer user needs. With conversational search tipped to dominate search in the future [9], it is crucial to understand how conversations vary in these diverse domains.

Many challenges remain for both the interactive information retrieval and the NLP community to allow systems to be developed to support the complex tasks suited to this mode of interaction [24]. A recent SWIRL workshop breakout group identified key challenges for conversational search including 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 [9]. Our focus 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 [34]. We study information needs in the domain of home cooking, which, based on the literature, we believed would be a fertile context for the kinds of complex needs suited to conversational search [12, 10] and a situation where users simultaneously perform practical, sometimes cognitively challenging tasks that make searching in the traditional sense problematic.

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 conversational agents literature where intent recognition is crucial for determining and planning the next steps of an agent in a dialogue. Moreover our initial 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 the raw speech utterances. Note, however, that we are reporting preliminary findings and plan to extend our analyses in the future.

2 Related Work

Our work relates to research contributions across diverse fields of the computer and information sciences especially at the intersection of natural language understanding and artificial intelligence. Here we link the fields by highlighting contributions on conversation, conversational agents and understanding and predicting user needs and goals.

2.1 Conversational Agents

Being able to detect and process user intents is a crucial and challenging part in the development of conversational agents. Typically, natural language understanding is performed with using a dialogue manager that processes user input in a way that the agent understands what to do next [15]. One important aspect for understanding user intent is maintaining the context. For this purpose, user models are generated (e.g. [33], [28]) but also linguistic concepts such as dialogue acts (e.g. [17], [29]), meaning relations [21] and sentiment analysis [18] are relevant facets that need to be considered for intent recognition.

With the growing popularity of conversational search systems which are defined as systems "retrieving information that permits a mixed-initiative back and forth between a user and agent, where the agent's actions are chosen in response to a model of current user needs within the current conversation, using both short- and long-term knowledge of the user" [24] the aforementioned concepts become relevant for user needs elicitation. This definition highlights the importance of memory, where the system can recall past interactions and reference these explicitly during conversations as it is done with dialogue managers in conversational agents. Such systems have emerged not only due to hardware developments, but because traditional search systems are unsuited to the complex tasks people perform [24].

2.2 Predicting information needs conversationally

Understanding and algorithmically predicting user needs can be useful for many reasons: different results can be shown [8], results can be presented differently [32] or answers can be presented directly in the results page [3]. Conversations with the user are one means of detecting such information needs. Automated conversational agents can provide personalization and support users in selecting an item [33] or talking about areas of interests [28] and have been applied in scenarios such as in trauma therapy [22]. This often requires systems to exhibit the memory property referred to earlier in order to maintain an understanding of context [17, 2].

In conversational search preliminary work has utilised user speech utterances as a means to identify information needs. Shiga et al. [27] classify needs along two dimensions, the first of which uses Taylor's levels of specification [31] and the second, which delineates type based on a classification derived from the literature. They, moreover, incorporate an aspect of task hierarchy, where a main task (e.g. booking a holiday) can be viewed as consisting of sub-tasks (e.g. findings a destination, comparing flight schedules etc.). Their work shows that information need categories can be distinguished using machine learning approaches. This work represents an excellent contribution and 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. 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 Shiga et al.'s work we test similar approaches in a home cooking context.

3 Methods

3.1 Data Collection

To establish a corpus of naturalistic conversational data large enough to perform machine learning prediction, we devised an in-situ user study. We simulated a natural cooking situation by gifting a box of ingredients to participants in their own kitchen at meal time. Participants were tasked with cooking a meal which they had not cooked before based on as many of the contained ingredients as possible, although these could be supplemented with the contents of their own pantry. To assist the process they could converse with the experimenter who would answer questions and needs using any resource available to him via the Web. The experimenter provided the best answer he could and communicated this orally in a natural human fashion (arguably the optimal behaviour for a conversational system). No time constraints were imposed for the task. Concretely, for each participant, the procedure comprised 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.

Fig. 1: Some example meals cooked during the experiments.

3.2 Ingredients

To ensure divergent recipes and conversations the ingredient boxes varied across participants. The ingredients typically had a value of around €10 and were chosen based on guidelines by the German Nutritional Society [13], which suggest 7 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 and a source of protein (e.g. eggs). Participants prepared diverse meals using the ingredients, a selection of which can be found in Figure 1.

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 [5] claim leads to more informal and open speech. Our impressions confirmed relaxed and natural behaviour in the experiments. Second, it allowed 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 but, to increase response rate, ingredients were gifted.

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. 37 had never used conversational agents before, while four 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 said they cook seldom 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 and Identification of Needs

In total, 38.75 hours of material were collected with the language spoken being German. The recorded conversations were transcribed and annotated by a trained linguist, who was also the experimenter, using the recommendations by Dresing and Pehl [11]. This involved translating any dialectual expressions into standard German – a step necessary to employ word embeddings (see section 4.2). The syntax of the utterances remained unchanged by this process. Thereafter, the utterances were split into queries. Table 1 provides examples of different kinds of utterances treated as a query in our analyses. In general, one or several questions in a row were counted as one query as long as the experimenter was of the opinion that the utterances represented the same information need. Otherwise, the utterances were split and counted as separate information needs. As can be seen in the examples in Table 1, a direct question has the form of an interrogative clause. Indirect questions, however, do not exhibit this grammatical form but can clearly be interpreted as a query or question to the system. Implicit/explicit actions, do not have the grammatical shape of a question at all, but can be interpreted by a human as such. This is strongly connected to the surrounding context and as such, the identification was performed by a human, in this case, a trained linguist. The example in Table 1 illustrates that despite the fact that the utterance does not exhibit the form of an interrogative clause, the user implicitly requires an answer to this utterance. The follow-up type consists of a query, which results in another query after the system has answered the first. These two queries illustrated in Table 1 were counted as two separate queries in the corpus. Based on these rules of counting, trials yielded on average 36.93 queries ($min = 7$, $x_{.25} = 22$, $\tilde{x} = 36$, $x_{.75} = 50$, $max = 73$, $sd = 17.48$, $skewness = 0.26$, $curtosis = 2.19$). The overall number of queries extracted was $N_q = 1662$.

Type	Example
Direct question	"What is the cooking time of asparagus?" (part. 42)
Indirect question	"Er – Alex, tell me how I need to cook red lentils." (part. 29)
Implicit/explicit action	"Ok, then this is similar to couscous" (part. 34)
Follow up	"So, at first the water and then? – System: Put in the asparagus. – Oh, right from the start? – System: No." (part.3)

Table 1: Examples of different means by which participants formulated questions taken from the transcripts.

4 Analyses

We analysed the collected data both qualitatively and quantitatively. First, using methods akin to content analysis, we examine the information needs identified to establish the variation of needs that occurred. This results in a classification scheme and a set of information needs annotated with an appropriate category. We continue to report on quantitative experiments, which establish the feasibility of automatically categorising the queries (information needs) using machine learning with the raw utterance text.

4.1 Coding Scheme for Information Needs

As with the previous processing of the transcribed utterances, the qualitative analysis was performed by a trained linguist familiar with dealing with such data. The starting point for the coding scheme was the set of categories derived for cooking related questions posted on the Google Answers forum in [10]. Out of the examples provided by Cunningham and Bainbridge we derived category definitions. Then, in a process akin to the coding process by Strauss and Corbin [30], each query was taken in turn, and a category from the existing scheme was attempted to be applied. When none of the existing categories were suitable, a new category was derived and a corresponding definition was created. Whenever a new category was established, all existing definitions were carefully reassessed to avoid potential overlap. On occasion an utterance included more than one information need, in which case more than one information need was assigned. This is reasonable given the fact that conversational search systems are generally expected to be able to understand pragmatics [24]. The process was iterative and tested repeatedly until the researcher was satisfied a consistent classification was achieved. The outcome of this classification were 27 different information need categories.

These were used to label all queries. The frequency distribution (see figure 2) of queries per category is heavily right skewed ($\overline{x} = 61.56$, $min = 1$, $x_{.25} = 3.5$, $\tilde{x} = 13$, $x_{.75} = 60.5$, $max = 506$, $sd = 111.77$, $skewness = 2.79$, $curtosis = 10.79$). The 10 most frequently assigned categories account for 93%

of all utterances. For space reasons, we limit our descriptions to five categories[1]. The prediction experiments reported are only concerned with the top 10 categories. Next, the categories are explained in descending order of relative frequency (given next to the label). To assist the reader's understanding we additionally provide a query example for each category. Quotes of transcripts are translated from German to English.

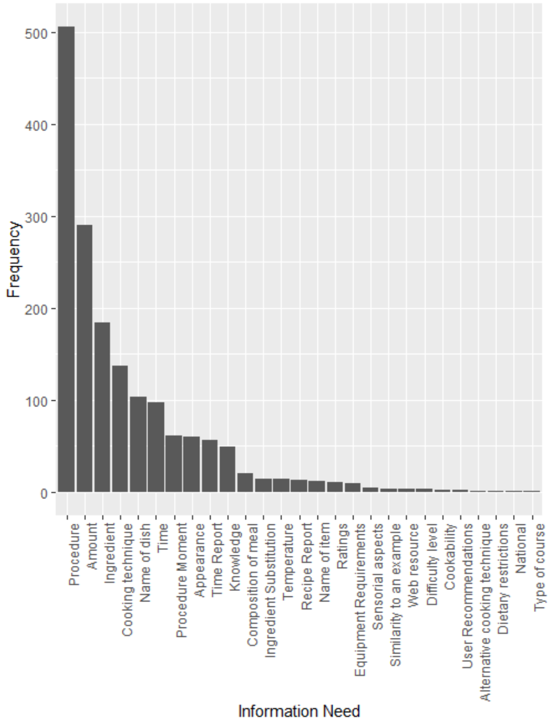

Fig. 2: Information need frequencies.

Procedure – 30.45% Utterances were labelled with this category when queries related to a particular step of the recipe (as opposed to general cooking techniques, see below). An examples is "What's next after bringing to the boil?" (part. 42)

Amount – 17.45% The label *Amount* was used, to code queries from participants who wanted to know about the quantity of an ingredient needed, e.g. "How much egg yolk is needed?" (part. 2)

[1] We plan to publish the full coding scheme and descriptions, as well as the anonymised transcriptions, both in German and English, as an open dataset to the community.

Ingredient – 11.07% Whenever questions regarding which ingredients were necessary for a particular recipe occurred, these utterances were tagged with label *Ingredient*. A typical example was "Which ingredients are needed?" (part. 1).

Cooking Technique – 8.24% Utterances/Queries were labelled this way when participants requested information about preparing ingredients that was not made explicit in the steps of the recipe. For example, a recipe would state "cook the asparagus". Participants not knowing how to cook asparagus would then ask "How does one actually cook asparagus? Can you look this up, please?" (part. 2).

Name of Dish – 6.20% was used in cases where participants searched for recipes they would like to prepare as their main dish. They often used ingredients as search items in such cases, e.g. "Then I'd like to have a dish with lentils, chickpeas and tomatoes" (part. 10)

4.2 Predicting Information Needs

The quantitative analysis was formulated as a prediction task i.e. given a set of features derived from the raw conversational utterances and context information, is it possible to predict the category of information need. We employed a random forest classifier for this purpose because it turned out to be an effective approach in Shiga et al.'s work [27]. All experiments reported below are using the Python package *scikit-learn* [23] and are based on 10-fold cross validation. Table 2 presents the result of experiments reporting average accuracy including $95\% - CIs$ based on 100 replications. The variance in accuracy converged after 30 replications.

As the use of word embeddings was shown to be beneficial for predicting information need categories in [27], we used these word embeddings as a baseline feature for all classification experiments. To this end, we employed 200-dimensional word embeddings trained on 2 million German language Wikipedia articles [7] for the classification task. Using these 200-dimensional word embeddings as features yielded an average accuracy of .4024 [.3355;.4697].

Feature Combinations	Avg. Accuracy	95%-CI
Word Embeddings	.4024	[.3355;.4697]
Word Embeddings + Previous Needs	.4080	[.3377;.4726]
Word Embeddings + Previous Needs + Normalized Sequence IDs	.4149	[.3545;.4901]
Approach		
Stopword removal	.3289	[.2548;.4065]
Top 50 words only	.4480	[.3797;.5186]
Resampling	.6391	[.5313;.7467]

Table 2: Performance of different feature combinations and approaches measured by avg. accuracy over 10 folds.

Several additional features were developed to improve this baseline performance. First, to incorporate the idea of memory into the system, the previous information need was operationalized as a predictive feature.

Second, the normalized sequence ID was added as a context feature. The sequence ID represents the position of an information need in a cooking session with the information need with ID 1 being the first to occur, ID 5 being the fifth and so on. Normalizing the sequence ID was necessary because some sessions were considerably longer than others (see section 3.4).

Next, we examined the vocabulary used more closely. We employed stopword removal using using the German language stopword list available as part of the nltk Python package [4]. As a result, the accuracy significantly decreased to an average of .3289 [.2548;.4065], indicating that stopwords are in fact meaningful and relevant in the cooking task context. In a second run, only the top 50 words were used because these represent $\approx 50\%$ of all words in the corpus and the collection frequency strongly decreases after the 50th word. This results in a small increase in average accuracy to .4480 [.3797;.5186] compared to the baseline.

In a final round of experimentation, we analyzed the impact of resampling on the prediction accuracy. As described above (see section 4.1), the distribution of classes was heavily skewed. We performed oversampling using SMOTE [6] and NearMiss-2 [37] as undersampling approach. We employed imbalanced-learn [19] as resampling library, which led to a significant increase in average accuracy (.6391 [.5313;.7467]).

5 Discussion

In this work we performed an in-situ cooking study where participants were given a box of ingredients and charged with cooking a meal of their choice. The study provided a corpus of conversations whereby diverse information needs were communicated to an experimenter simulating a personal assistant in natural language. This corpus provided the basis for us to study the information needs, which can occur in this context and run prediction experiments to determine if the type of need could be automatically predicted. Discussing the results we gain from the analysis is done along four different lines: peculiarities of conversations in the cooking domain, the importance of memory in predicting information needs, aspects of conversational style and eliciting information needs in the domain of cooking.

What is special about conversations in this domain? We identified 27 fine grained information need categories, ten of which were sufficient to label 93% of all queries. The information need taxonomy presented [10] was used as a starting point for the qualitative analysis. Comparing our results to those in [10] yields differences in terms of occurrence and distribution of information needs. Only *Ingredient* and *Name of dish* are frequent in both studies. While Cunningham and Bainbridge report *Name of dish/item*, *Ingredient*, *Type of dish, course, meal* and *Ethnic/National/Region* being most common, *Procedure* is the most frequently used label in the data we collected. Indeed, *Type of dish, course, meal* and *Ethnic/National/Region* are rarely applied in our corpus – and vice versa for *Cooking Technique*. The differences can largely be explained by the fact that in [10] questions were not posed while actually cooking. Thus, categories like *Amount*, *Time*, *Time Report* and *Knowledge* did not occur in their study. These information needs tend to be more related to actual cooking tasks than being descriptors for text-based search for recipes. In terms of information need categories we find two commonalities with [27] – despite the difference in domains between their study and ours. First, some of the information need categories in the cooking task scenario show the hierarchical relationship of main and subtasks. Categories *Procedure* and *Procedure Moment* are good examples: While the first refers to all steps needed to provide a meal, the latter is concerned with a specific step throughout the cooking process. Second, the different levels of task (with the exception of "search") are mirrored in our data. We find queries relating to topical knowledge about the cooking task (see category *Procedure*) as well as those relating to problem solving (see category *Cooking Technique*) and situation (see category *Time Report*).

Link between natural dialogue and memory The fact that adding the previous information need as a feature did not increase the accuracy values achieved is a surprising result. This is in contrast to the importance of memory which can be derived from theoretical work (see e.g. [24]) and also some task conditions in [27]. One reason for this result may be a lack of data to gain the expected results. Future work will, consequently, be dedicated to gathering more data and a reassessment of the effect memory has. A second possible explanation might be the existence of user subgroups. Users who cook on a regular basis may have different sequences of needs than those who prepare meals less frequently.

The use of conversational style By running experiments with and without stopword removal we provide empirical evidence that the most heavily used words are most important to elicit information needs. This is in line with findings in the domain of very short text retrieval (e.g. [20]) when stopwords are removed. In the context of cooking tasks many stopwords may be discourse cues, which have important functions for text comprehension, including easing the reconstructing of the line of argumentation [1], signaling misunderstanding [26] and facilitate recall in information processing [36]. It makes sense that when the stopwords are removed prediction performance decreases as the line of argumentation in the discourse is no longer observable as it is destroyed by removing the cues.

One compelling area of future research would, thus, be to compile corpus-specific stopword lists (e.g. for different domains), which is e.g. suggested in the domain of sentiment analysis [25].

The need to understand and elicit information needs in a particular domain The results obtained by our prediction experiments show that the queries issued to the conversational search system are useful for distinguishing different information needs. Generally speaking, our results suggest that information need categories during conversation can be predicted with average accuracy values achieved of up to 64% when resampling is used. Even the non-resampled performance of ≈ 40% are significantly larger than chance (which would be 10% with ten classes). A major reason for the misclassification found is the inhomogeneous distribution of queries over the various information need categories. *Procedure* was miss-classified as the dominant category in almost each of the remaining classes. The impact of this class on the accuracy result can be seen in form of a low average precision (≈ 32%). A second aspect explaining miss-classifications may be the (to some extent strong) semantic similarity between individual information need categories, e.g. between *Procedure* and *Procedure Moment* as well as between *Time* and *Time Report*. Grouping such categories might lead to higher accuracies. However, detail information gets lost.

Having said this, we identify several challenges imposed to conversational search systems throughout data collection and preparation. All of these relate to resolving information needs. Spoken language interactions pose, first, the challenge of understanding the pragmatics of dialect use. Dialect – with a variety of types and levels – was used by almost all participants. Translating these expressions to standard German is a major problem for speech recognition systems because it is more than a mere word-by-word translation. On many occasions a wealth of pragmatics was needed to fully comprehend the queries issued by participants. A second, related challenge was the fact that information needs were often not clearly defined. This means, slicing an utterance into information needs requires a large amount of world-knowledge which poses major challenges on conversational search systems.

6 On the problems with prediction and response to information needs

The assistance a conversational search system provides will indeed benefit from the capability to predict user needs. This can be illustrated along three lines, all of which are grounded in encounters from our corpus. The system can, first, focus on the demands of the particular situation – instead of sticking to a static programme flow. The following excerpt from our corpus is one example:

> Er – can you read the ingredients list out loud to me, so I can get them? – System: (reads ingredient list **slowly** and **waits for confirmation** by the participant that s/he got a particular ingredient), part. 14

A system showing predictive capabilities can, second, provide feedback with respect to deviations from anticipated actions, e.g. from a recipe's default procedure. Depending on what the user said, i.e. based on the discourse markers present that the system can observe, it can adapt to the new situation.

> I would have added some tomato paste or something like that, so that it isn't so dry... but if it's not in the recipe. – System: (explain that it would be possible to add this), part. 40

Third, most conversational agents (see [17], [2]) use a rule-based approach to extract the user's intent. However, they only analyze the "surface" of an utterance and make decisions based on keywords. This, however, does not necessarily reflect the true information need by the user which leads to misclassifications. Our prediction task offers the possibility to not only investigate the surface but to go deeper into semantics using word embeddings. Information needs can be detected more accurately by using these and a system can thus provide the information the user really wants.

> Can you search for a recipe for Sauce Hollandaise, please? – System: Sure. (searches and reads the ingredients out loud), part. 2

One possible solution for solving the aforementioned problems might be to include more features than just the information need for the classification task. Our hypothesis is that a multidimensional vector including several linguistic features such as dialogue acts and the current task state might improve the performance when predicting the user's information need and intent. Currently, we are working on relabeling the corpus across various dimensions going beyond cooking-specific information needs. By doing this, we aim to gain a tree-structured coding of the data that might help us to analyse the conversational structure on different levels. We hope that this will improve the classification performance, too.

7 Conclusion and Future Work

Our preliminary results shed light on the information needs which occur in a home cooking context and indicate the feasibility of identifying needs automatically. This pilot study emphasizes the feasibility and value of this kind of approach.

Future Work will collect additional naturalistic data, to gain more generalizable results and thus promote research in the conversational search domain. Also, our results showed that a more detailed classification of user utterances is necessary to classify user intent. Thus, other linguistic dimensions such as dialogue acts will be incorporated in the ongoing turn annotation of this corpus. Based on our feasibility study, similar experiments in the cooking domain or other domains can be conducted to gain higher representativity. Our study focused on the utterances made by users. However, the utterances made by the conversational system are equally important as it needs to be capable of generating utterances suitable for the current context. The utterances made by the experimenter can thus employed in future work to investigate this issue.

References

1. Allbritton, D., Moore, J.: Discourse cues in narrative text: Using production to predict comprehension. In: AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems (1999)
2. Allen, J., Ferguson, G., Stent, A.: An architecture for more realistic conversational systems. In: Proceedings of the 6th International Conference on Intelligent User Interfaces. pp. 1–8. IUI '01, ACM, New York, NY, USA (2001). https://doi.org/10.1145/359784.359822, http://doi.acm.org/10.1145/359784.359822
3. Bernstein, M.S., Teevan, J., Dumais, S., Liebling, D., Horvitz, E.: Direct answers for search queries in the long tail. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 237–246. ACM (2012)
4. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc., Sebastopol, CA (2009)
5. Castellá, V.O., Abad, A.Z., Alonso, F.P., Silla, J.P.: 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), 141 – 159 (2000). https://doi.org/https://doi.org/10.1016/S0747-5632(00)00012-1, http://www.sciencedirect.com/science/article/pii/S0747563200000121
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. J. Artif. Int. Res. **16**(1), 321–357 (Jun 2002), http://dl.acm.org/citation.cfm?id=1622407.1622416
7. Cieliebak, M., Deriu, J.M., Egger, D., Uzdilli, F.: A twitter corpus and benchmark resources for german sentiment analysis. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. pp. 45–51. Association for Computational Linguistics (2017), http://aclweb.org/anthology/W17-1106
8. Craswell, N., Hawking, D., Robertson, S.: Effective site finding using link anchor information. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 250–257. ACM (2001)
9. Culpepper, J.S., Diaz, F., Smucker, M.D.: Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (SWIRL 2018). SIGIR Forum **52**(1), 34–90 (2018). https://doi.org/10.1145/3274784.3274788, https://doi.org/10.1145/3274784.3274788
10. Cunningham, S.J., Bainbridge, D.: An Analysis of Cooking Queries: Implications for Supporting Leisure Cooking Ethnographic Studies of Cooks and Cooking. In: iConference 2013 Proceedings. pp. 112–123 (2013). https://doi.org/10.9776/13160
11. Drehsing, T., Pehl, T.: Praxisbuch Interview, Transkription & Analyse. Anleitungen und Regelsysteme für qualitativ Forschende. Dr. Dresing und Pehl GmbH, Marburg, 6th. edn. (2015)
12. Elsweiler, D., Trattner, C., Harvey, M.: Exploiting food choice biases for healthier recipe recommendation. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 575–584. ACM (2017)
13. für Ernährung, D.G.: Die deutsche gesellschaft für ernährung e.v. (dge) (2018), https://www.dge.de/wir-ueber-uns/die-dge/

14. Guy, I.: 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. pp. 35–44. ACM (2016)
15. Jurafsky, D., Martin, J.H.: Speech and Language Processing (2Nd Edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA (2009)
16. Kiseleva, J., Williams, K., Hassan Awadallah, A., Crook, A.C., Zitouni, I., Anastasakos, T.: Predicting user satisfaction with intelligent assistants. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 45–54. ACM (2016)
17. Kopp, S., Gesellensetter, L., Krämer, N.C., Wachsmuth, I.: A conversational agent as museum guide: Design and evaluation of a real-world application. In: Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P., Rist, T. (eds.) Intelligent virtual agents: Proceedings 5th International Conference. LNAI, 3661, pp. 329–343. Springer, Berlin (2005)
18. Leggeri, S., Esposito, A., Iocchi, L.: Task-oriented conversational agent self-learning based on sentiment analysis. In: Proceedings of the 2nd Workshop on Natural Language for Artificial Intelligence (NL4AI 2018) co-located with 17th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2018), Trento, Italy, November 22nd to 23rd, 2018. pp. 4–15 (2018)
19. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research **18**(17), 1–5 (2017), http://jmlr.org/papers/v18/16-365.html
20. Leveling, J.: On the Effect of Stopword Removal for SMS-Based FAQ Retrieval. In: Bouma, G., Ittoo, A., Métais, E., Wortmann, H. (eds.) Natural Language Processing and Information Systems. pp. 128–139. Springer, Berlin, Heidelberg (2012)
21. Mondal, P.: Lexicon, meaning relations, and semantic networks. In: Proceedings of the 2nd Workshop on Natural Language for Artificial Intelligence (NL4AI 2018) co-located with 17th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2018), Trento, Italy, November 22nd to 23rd, 2018. pp. 40–52 (2018)
22. Morbini, F., Forbell, E., DeVault, D., Sagae, K., Traum, D.R., Rizzo, A.A.: A mixed-initiative conversational dialogue system for healthcare. In: Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 137–139. SIGDIAL '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012), http://dl.acm.org/citation.cfm?id=2392800.2392825
23. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (Nov 2011), http://dl.acm.org/citation.cfm?id=1953048.2078195
24. Radlinski, F., Craswell, N.: A theoretical framework for conversational search. In: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. pp. 117–126. CHIIR '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3020165.3020183, http://doi.acm.org/10.1145/3020165.3020183
25. Saif, H., Fernández, M., He, Y., Alani, H.: On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014. p. 810–817 (2014)
26. Schober, M.F., Bloom, J.E.: Discourse cues that respondents have misunderstood survey questions. Discourse processes **38**(3), 287–308 (2004)

27. Shiga, S., Joho, H., Blanco, R., Trippas, J.R., Sanderson, M.: Modelling information needs in collaborative search conversations. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 715–724. SIGIR '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3077136.3080787, http://doi.acm.org/10.1145/3077136.3080787
28. Spillane, B., Gilmartin, E., Saam, C., Su, K., Cowan, B.R., Lawless, S., Wade, V.: Introducing adele: A personalized intelligent companion. In: Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents. pp. 43–44. ISIAA 2017, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3139491.3139492, http://doi.acm.org/10.1145/3139491.3139492
29. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C.V., Meteer, M.: Dialogue act modelling for automatic tagging and recognition of conversational speech. Computational Linguistics **26**, 339 – 373 (2000)
30. Strauss, A., Corbin, J.: Grounded Theory: Grundlagen qualitativer Sozialforschung. Beltz, Weinheim (1996)
31. Taylor, R.S.: The process of asking questions. American Documentation **13**(4), 391–396 (1962). https://doi.org/10.1002/asi.5090130405, https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090130405
32. Teevan, J., Cutrell, E., Fisher, D., Drucker, S.M., Ramos, G., André, P., Hu, C.: Visual snippets: summarizing web pages for search and revisitation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 2023–2032. ACM (2009)
33. Thompson, C.A., Göker, M.H., Langley, P.: A personalized system for conversational recommendations. J. Artif. Int. Res. **21**(1), 393–428 (Mar 2004), http://dl.acm.org/citation.cfm?id=1622467.1622479
34. Trippas, J.R., Spina, D., Cavedon, L., Joho, H., Sanderson, M.: Informing the design of spoken conversational search: Perspective paper. In: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval. pp. 32–41. CHIIR '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3176349.3176387, http://doi.acm.org/10.1145/3176349.3176387
35. Trippas, J.R., Spina, D., Cavedon, L., Sanderson, M.: How do people interact in conversational speech-only search tasks: A preliminary analysis. In: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. pp. 325–328. CHIIR '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3020165.3022144, http://doi.acm.org/10.1145/3020165.3022144
36. Winterboer, A., Ferreira, F., Moore, J.D.: Do discourse cues facilitate recall in information presentation messages? In: INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008. p. 543. ISCA (2008)
37. Zhang, J., Mani, I.: KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In: Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets (2003)