Datenbank Spektrum (2020) 20:37–41
https://doi.org/10.1007/s13222-020-00333-z

SCHWERPUNKTBEITRAG

Comparing Wizard of Oz & Observational Studies for Conversational IR Evaluation

Lessons Learned from These two Diverse Approaches

David Elsweiler[1] · **Alexander Frummet**[1] · **Morgan Harvey**[2]

Received: 25 November 2019 / Accepted: 17 January 2020 / Published online: 10 February 2020

Abstract
Systematic and repeatable measurement of information systems via test collections, the Cranfield model, has been the mainstay of Information Retrieval since the 1960s. However, this may not be appropriate for newer, more interactive systems, such as Conversational Search agents. Such systems rely on Machine Learning technologies, which are not yet sufficiently advanced to permit true human-like dialogues, and so research can be enabled by simulation via human agents. In this work we compare dialogues obtained from two studies with the same context, assistance in the kitchen, but with different experimental setups, allowing us to learn about and evaluate conversational IR systems. We discover that users adapt their behaviour when they think they are interacting with a system and that human-like conversations in one of the studies were unpredictable to an extent we did not expect. Our results have implications for the development of new studies in this area and, ultimately, the design of future conversational agents.

Keywords Conversational search · Evaluation

1 Introduction

The field of Information Retrieval (IR) has a long and proud tradition of empirical scientific evaluation. The Cranfield paradigm, developed by Cleverdon and colleagues in the 1960s, permits systematic and repeatable measurement of retrieval system performance [6] and has served the community well for over half a century. Over time this approach has been adapted to fit different types of search problem [8], however, as IR systems have become increasingly interactive in nature, in some cases it is now reaching its limits. One such modern approach to search that may test the Cranfield paradigm to its breaking point is that of Conversational Search. Recent progress in Machine Learning technologies has permitted advances in automated natural language comprehension to the extent that many tasks can now be achieved by entering into a two-way dialogue, in either text or spoken form, with a virtual agent.

Cranfield-based IR evaluations typically have only few queries per topic, while in conversational search queries are free-form and vary from typical system-like queries to long and rich, human like descriptions. Another challenge is that, although considerable progress has been made in recent years, many of the technologies that would be required to fulfil the vision of true conversational search, such as accurate speech recognition or dialogue modelling, do not yet work sufficiently [4]. As such, if IR researchers wish to study an aspect such as search result utility in a conversational context, much of the experience must be simulated.

One means to achieve such simulations is to apply a so-called Wizard of Oz (WoZ) study where test participants interact with a system that they believe to be automated but that is, unbeknown to them, actually remotely operated by a human. The approach itself can vary from so-called "slot-filling", which is highly procedural, to fully conversational systems that react spontaneously without restriction, like a human would [2]. This can be taken further still by placing the human wizard in full view of the participant (in-situ).

✉ David Elsweiler
david@elsweiler.co.uk

[1] University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

[2] Northumbria University, Newcastle upon Tyne, NE1 8ST, UK

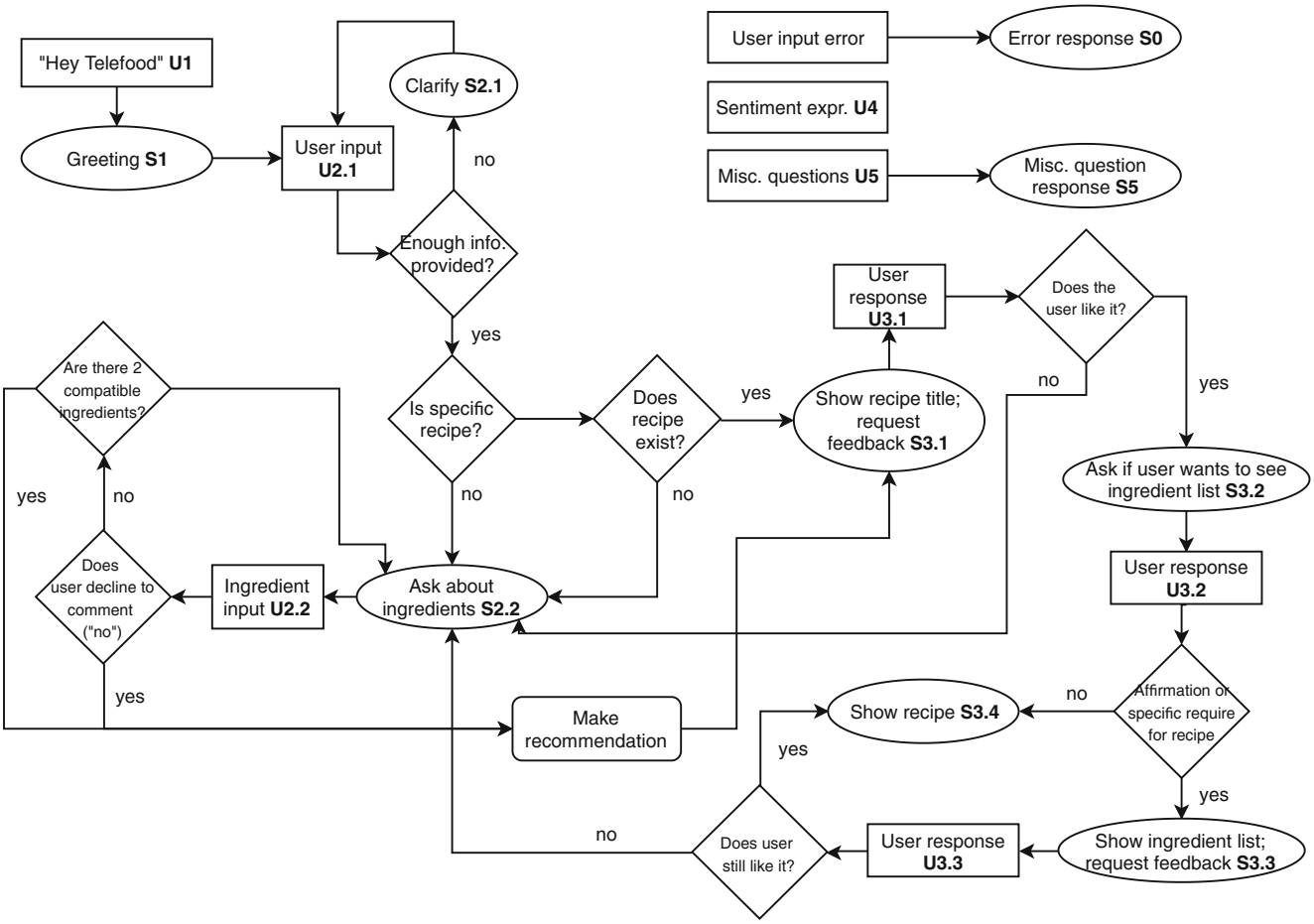

Fig. 1 Conversational framework

Such observational studies have been used in conversational search research in IR and typically involve using pairs of human participants who converse with each other [7].

Although substantial literature exists describing individual studies of these different kinds individually, typically it is difficult to compare and contrast the methods as the contexts studied are so different. Here, we share our experiences of performing two comparable studies: one with a tightly-controlled Wizard and a second with an assistant present in-situ. Both relate to a conversational assistant in the kitchen. By annotating and analysing the utterances collected in the different setups, we can establish not only the similarities and differences that occur, but also learn lessons of when different types of study should be performed and derive insights to inform future design of such systems.

2 Data Collection Approaches

2.1 Study 1: Wizard of Oz

The first study employed a Wizard of Oz (WoZ) methodology, meaning that participants were unaware that the system used was being controlled by another human. Participants were told they were interacting with a bot and the style of interaction – what the bot communicated to the participant, how and when – supported this assumption. Although the Wizard did not employ a fully-scripted, slot-filling approach, the conversation flow and Wizard responses were tightly controlled.

We developed a *conversational* framework in the form of a flowchart (see Fig. 1) describing how an idealised dialogue between a human user and the virtual agent (named *Telefood*) should proceed to provide recipe recommendations. Nodes in the flowchart represent user utterances (preceded by *U*) and system responses and queries (preceded by *S*). If the user asked a question or made a statement that did not adhere to the framework at the point in the dialogue that had been reached, the wizard either generated an error mes-

sage (*S0*), such as "I didn't understand that" and requested the user try again or, if the request was relevant to another part of the framework, the Wizard jumped to the appropriate node and proceeded from there. Expressions of sentiment by the user, either to a recipe suggestion (e.g. "super, thank you") or task completion (e.g."thank you, Telefood") were assigned to *U4]*.

Conversations were designed to flow in what we believed to be a plausible fashion. The participant has the opportunity to first describe the need for the recipe and the context *(U2.1)* with the system prompting for context if little is provided. If the Wizard has sufficient information a recommendation is made, otherwise the user is prompted for ingredient preferences *(U2.2)*. After a suggestion has been made, the participant can respond to control the flow of the process *(U3.1)*. Note that while the framework does not allow all user input to be handled as an actual human would, there is no restriction with respect to what participants can say.

To motivate conversations, participants were provided with recipe finding tasks. Each completed 3 such tasks, interacting with the Wizard either with spoken or typed utterances depending on the condition they were assigned to using a between-groups design. 28 participants (15 females; $\widetilde{x}_{age}$ = 22 years, min_{age} = 18 years, max_{age} = 33 years) provided 999 utterances, 514 of which were from the audio condition.[1]

2.2 Study 2: In-situ Study

The second study, which was performed independently by a researcher who was not involved in the first, established a somewhat comparable corpus of conversational data by means of an in-situ user study. Here we simulated a naturalistic cooking situation by gifting a box of ingredients to participants in their own kitchen at meal time. Participants were tasked with cooking a meal they had not cooked before using the contained ingredients, which 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). Although many of the utterances related to user needs beyond those captured in the WoZ study (e.g. queries relating to cooking steps or techniques), every participant conversed with the researcher in order to establish a suitable recipe to cook, making these utterances a fair point of comparison.

45 participants (22 females, $\overline{x}_{age}$ = 24 years, min_{age} = 19 years, max_{age} = 71 years) provided 38.75 hours of spoken dialogue, which was subsequently transcribed and analysed qualitatively.[2] The process resulted in 1,662 participant utterances. To allow fair comparison we remove any utterance after a recipe has been selected (i.e. which relate to actually preparing the dish). In cases where an additional recipe is subsequently sought (e.g. a side-salad), we again remove utterances labelled as *U5*, which by definition could not occur in the WoZ study. 5 participants in the in-situ study did not actually require a recipe. These were removed from the dataset. In sum, 464 in-situ utterances were analysed.

2.3 Annotating both Corpora

To establish whether the corpora are comparable, the WoZ Framework was used as a means to annotate the utterances from both studies. We focused on the stages *U2.1,U2.2 and U3.1*. Out of context utterances were marked with *U5*. In the in-situ data this meant that most utterances after the selection of recipes were labelled as *U5*. In some cases, participants were required to return to the recipe recommendation stage when they realised that the selected recipe could not be achieved due to, for example, missing ingredients or equipment. For the WoZ corpus, coding was performed by two researchers, who first worked together to code 25% of the data set, resolving any disagreements through discussion. They then worked separately, each coding half of the remaining data. A random sample of 100 rows of these was selected and re-coded by the other researcher in order to assess inter-rater reliability. Cohen's kappa values were obtained indicate almost perfect agreement (κ = 0.836, z = 20.7, p-value $\ll$ 0.01). A similar process was undertaken for the in-situ corpus. First, three researchers coded the utterances for four test participants, chosen at random. On average all three coders achieved 90.02% agreement, in 100% cases at least two coders agreed (κ = 0.76, z = 27.1, p-value $\ll$ 0.01).

3 Results & Discussion

The first result of note is that the inter-rater agreement statistics evidence that the corpora are indeed comparable. Table 1, which shows the distribution of codes, however, reveals differences in the conversations in these different studies that had comparable aims (i.e. find a recipe to cook). Clearly, the in-situ conversations are heavily skewed towards code *U5* in the framework, with over half of all

[1] Due to space limitations, we cannot provide full details of the study. These can be found in [1].

[2] Space limitations mean we cannot provide full details of the study. These can be found in [3].

Table 1 Distribution of codes by experimental source

	U2.1		U2.2		U3.1		U5	
	n	*%*	***n***	*%*	***n***	*%*	***n***	*%*
In-Situ	36	7.7	64	13.7	123	26.5	241	51.9
WoZ	283	32.5	141	16.2	187	21.5	259	29.8

utterances being assigned to this code, while the early information-providing code *U2.1* is hardly ever visited. This contrasts with the WoZ results, where the distribution over framework codes is much more uniform and where a large number of utterances provide information about the wished-for recipe (*U2.1*. There does not seem to be a large difference in the percentage of *U3.1* visits meaning participants responded to roughly the same number of recommendations in both conditions.

The in-situ utterances were longer, with a median of 10 words per utterance, compared with a median of only 2 words for the WoZ study utterances. They also contained more turns per task (in-situ: $\mu = 11.6$; WoZ: $\mu = 10.36$) and had much less of a clearly defined order – 25% started with 2.1 in the in-situ experiments, 85,7% were on second position in the WoZ corpus (after "Hey Telefood", which was code *U1*).

These statistics endorse the researchers' impression of more human-like conversation in the in-situ study. Users seem to adapt their behaviour when they think they are interacting with a system. There were examples of human-like conversation in the WoZ corpus, such as use of politeness markers (e.g. "please" and "thank you", function words (i.e. non-content carrying) and indirect request strategies using modal verbs. Often, however, WoZ utterances were stripped down to query-like utterances – text adapted to what a system needs[3]. A further observation was that, even though the WoZ framework was designed to be representative of a typical conversational pattern, in the naturalistic in-situ setting this pattern almost never occurred.

One reason for the bias toward state *U2.2* as a starting state in the in-situ study was likely to be an artefact of providing ingredient boxes. An interpretation for the lack of *U2.1* utterances in the in-situ study could be that participants may not have felt the need to communicate the context because they had a human user who was experiencing it along with them and implicitly knew the details of why the recipe was being cooked, who for, etc.. This does not account for the fact that when conversations are free, participants provide very little context for their preferences when they are not explicitly pressed for these. There were very few examples such as "I really like Chinese food" or "I'm not a fan of spicy stuff" in the in-situ data. Due to the use of the framework – where users were explicitly asked for their preferences – these sorts of utterances were very common in the WoZ study data and participants often provided data as simple ordered lists of ingredients or requirements, much like they would when writing queries. Lots of the in-situ utterances rely on a human being in the room being able to use their senses (e.g. "What can I do with these ingredients?" TP13); or require a lot of interpretation, i.e. implicit intentions: "a little bit too hot for a bake" TP13.

The lack of structure in conversations highlights just how challenging supporting a completely free-form conversational bot would be. As people do not provide helpful information by default, a passive agent such as in the in-situ study is less helpful. As Radlinski and Craswell suggest, conversational assistants should be mixed initiative systems, where information is exchanged [5]. This was true in both studies, although the balance offered by the WoZ study, where the initiative lay primarily with the Wizard, provided the system with more information and structured dialogue, which would be easier to support technologically.

4 Conclusions

By performing two variant experiments where human users interact with an agent (either a human in the room or a human disguised as a system) we were able to compare such methods as a means to learn about and evaluate conversational IR systems. We discovered that users seemed to adapt their behaviour when they thought they were interacting with a system and that human-like conversations in the in-situ study were unpredictable to an extent we did not expect.

All this builds towards the conclusion that, while in-situ studies provide interesting insights to human interaction, they do not – at least in our case – lead to utterances that you would get with an actual system. Our WoZ study offered much more standardised interaction patterns that would be much easier to support from a technical perspective. There was no evidence that this was detrimental to the user experience.

Funding Open Access funding provided by Projekt DEAL.

[3] More detailed explanations, examples and count information can be found in [1].

References

1. Barko-Sherif S, Elsweiler D, Harvey M (2020) Conversational agents for recipe recommendation. In: 2020 Conference on Human Information Interaction and Retrieval. ACM, New York
2. Dubiel M, Halvey M, Azzopardi L, Daronnat S (2018) Investigating how conversational search agents affect user's behaviour, performance and search experience. In: The second international workshop on conversational approaches to information retrieval
3. Frummet A, Elsweiler D, Ludwig B (2019) Detecting domain-specific information needs inconversational search dialogues. In: Natural language for artificial intelligence
4. Luger E, Sellen A (2016) Like having a really bad pa: the gulf between user expectation and experience of conversational agents. In: Proceedings of the 2016 CHI conference on human factors in computing systems. ACM, New York, pp 5286–5297
5. Radlinski F, Craswell N (2017) A theoretical framework for conversational search. In: Proceedings of the 2017 conference on conference human information interaction and retrieval. ACM, New York, pp 117–126
6. Robertson S (2008) On the history of evaluation in ir. J Inf Sci 34(4):439–456
7. Shiga S, Joho H, Blanco R, Trippas JR, Sanderson M (2017) Modelling information needs in collaborative search conversations. In: Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 715–724
8. Voorhees EM (2019) The evolution of cranfield. In: Information retrieval evaluation in a changing world. Springer, Berlin Heidelberg, pp 45–69