

Crowd-sourced Collection of Task-Oriented Human-Human Dialogues in a Multi-Domain Scenario

Norbert Braunschweiler¹, Panagiotis Papadakos², Margarita Kotti¹,
Yannis Marketakis², and Yannis Tzitzikas²

¹ Toshiba Research Europe Limited, Cambridge Research Laboratory, Cambridge, UK
{norbert.braunschweiler, margarita.kotti}@crl.toshiba.co.uk

² Institute of Computer Science, FORTH-ICS, Heraklion, Crete, Greece
{papadako, marketak, tzitzik}@ics.forth.gr

Abstract. There is a lack of high-quality corpora for the purposes of training task-oriented, end-to-end dialogue systems. This paper describes a dialogue collection process which used crowd-sourcing and a Wizard-of-Oz set-up to collect written human-human dialogues for a task-oriented, multi-domain scenario. The context is a tourism agency, where users try to select the more desired hotel, restaurant, museum or shop. To respond to users, wizards were assisted by an exploratory system supporting Preference-enriched Faceted Search. An important aspect was the translation of user intent to a number of actions (hard or soft-constraints) by wizards. The main goal was to collect dialogues as realistic as possible between a user and an operator, suitable for training end-to-end dialogue systems. This work describes the experiences made, the options and the decisions taken to minimize the human effort and budget, along with the tools used and developed, and describes in detail the resulting dialogue collection.

Keywords: dialogue collection, crowd-sourcing, wizard-of-oz, end-to-end, exploratory search, dialogue systems

1 Introduction

One key factor in the development of neural network based dialogue systems is the availability of suitable training material, both in content and volume. More training material is usually associated with better models, but has to be accompanied by sufficient variety and coverage. Despite great progress in this field, especially when it comes to non-task oriented dialogue systems [14, 16, 19], there is still a lack of high-quality corpora for the purposes of training task-oriented, end-to-end trainable dialogue systems. One challenge in this dialogue collection, is the problem of getting sufficiently realistic dialogues to cover the wide range of types and styles which are simultaneously influenced or directed by accessing knowledge sources.

The dialogue collection presented here, is designed to provide sufficient training material to train a task-oriented dialogue system in an end-to-end manner. Criteria for the data collection included: (a) realistic dialogues between a user and an operator, (b) the number of dialogues should be a figure in the thousands, (c) multiple-domains, (d) usage of a knowledge base and an expressive interaction paradigm to retrieve and explore information about domains.

To achieve these requirements a dialogue collection was conducted which used a Wizard-of-Oz set-up, in which the role of the dialogue system is played by a human (the “wizard”), and a crowd-sourcing platform to gather a wide range of subjects acting as dialogue system users. A number of trained wizards acted as the dialogue system response generators and used their access to an exploratory search system over a knowledge base, supporting the expressive Preference-Enriched Faceted Search (PFS), to guide their answers.

The contribution of this paper is that (a) it details a dialogue collection process that exploits an expressive interaction model, (b) it explains the selection of certain tools or platforms, and (c) it provides details about the content of the final corpus and the required effort. The rest of this paper is organized as follows: Section 2 describes related work, Section 3 describes the dialogue collection process, the tools used, and provides some statistics over the collected dialogue corpus. Finally, Section 4 discusses the methodology and the results, and Section 5 concludes the paper.

2 Related Work

Below, some recently collected datasets are discussed, closely related to the presented dialogue collection here, either in their collection style or content. For a summary of available corpora for building data-driven dialogue systems see [15].

The Maluuba Frames³ corpus [1], offers roughly 1.3k human-human dialogues in a task-oriented scenario in which users are aiming to book a trip by conversing with an operator who searches a database to find suitable trips. While this collection also uses a Wizard-of-Oz set-up, it includes just 12 participants and it covers only a single domain.

As part of a challenge regarding end-to-end trainable dialogue models [5], Microsoft released a corpus of human-human dialogues collected by crowd-sourcing⁴. The corpus contains 3 domains and about 10k dialogues (movie=2890, restaurant=4103 taxi=3094). The corpus was fully annotated with dialogue intents and slot values, however first challenge results showed modest performances⁵.

Another dialogue collection of similar type is [2]. This corpus of written human-human dialogues contains 7 domains and both dialogue belief states/actions are annotated by selected crowd-sourced labelers. Regarding human-to-human based datasets, [2] mentions that these are most suitable for building a natural conversational system, but that many of the corpora released in the past (e.g. [6, 12, 13]) lack a grounding of the dialogues onto a knowledge base which limits their use for task-oriented systems.

For the dialogue collection described in this paper, the Wizard-of-Oz method [3, 4, 11] was chosen. In this method a user interacts with a human “wizard” who is acting as the dialogue system response generator. It allows gathering large amounts of text-to-text conversations via crowd-sourcing as shown, for instance, by [2] and previously [20]. While [2] and [20] used an asynchronous set-up in which users and wizards did not have to engage in a coherent dialogue of user-wizard turns, but multiple workers

³ <https://datasets.maluuba.com/Frames/>

⁴ https://github.com/xiul-msr/e2e_dialog_challenge

⁵ https://xiul-msr.github.io/e2e_dialog_challenge/slides/MS_dialog_challenge_result_outlook_sungjin.pptx

Table 1. Comparison of 4 similar corpora with the one presented in this study.

Corpus	#Dial.	#Turns	Turns./Dial.	#Domains	#Workers	Labeled	Synchron.	Ref.
FRAMES	>1.3k	<10k	14.6	1	12	✓	✓	[1]
MicrosoftE2E	>10k	>70k	7.5	3	?	✓	?	[5]
WOZ2.0	1.2k	<10k	7.5	1	?	-	X	[7]
MULTIWOZ	>10k	>100k	13.7	7	1249	✓	X	[2]
ToshWOZ	>3k	>30k	10.4	4	327 + 9 wiz.	-	✓	-

contributed to the same dialogue, the current set-up was synchronous, ensuring more coherent dialogues. Table 1 shows a comparison of the aforementioned four corpora with the one presented in this study (ToshWOZ).

3 Dialogue Collection Method

To collect dialogues between humans in a goal-oriented setting covering multiple domains, a tourism agency scenario was chosen. Crowd-sourced users were given tasks to find a particular place (e.g. a restaurant or a museum) by written interaction with an operator. The operators were trained agents, who accessed an exploratory system that supported both hard and soft-constraints (i.e. preferences) over a knowledge base, to guide their responses to users. The training of the operators, who were in-house experts, included the usage of the exploratory system while interacting with users and the operation of the dialogue platform, i.e. copying text from users, entering it into the exploratory system, storing it, translating user requests to appropriate hard or soft-constraints in the exploratory system, formulating a response to users, storing it in the knowledge base and submitting it to the web-based dialogue interface.

Using a human-to-human set-up involving many different crowd-sourced workers aims to capture the vast variety of language usage as well as the variability in strategies to achieve a certain information seeking task. To support variety in content during dialogues, 4 domains are covered and for each domain a set of task scenarios was created. The 4 domains are: hotels, museums, restaurants and shops in 4 cities of Japan (Kobe, Kyoto, Osaka, Tokyo). The number of task scenarios are: 11 in hotels, 8 in both museums and restaurants, and 5 in shops.

Each scenario describes the profile and preferences of a customer wishing to find something particular (e.g. a hotel in Kyoto as shown in Fig. 1). Most of the scenarios were created using the information in the knowledge base. However, some scenarios require knowledge not existing in the knowledge base. These scenarios were introduced so that the actions of the wizards operating the knowledge base could be recorded. In addition, workers were encouraged to describe their own preferences.

One of the important aspects here, is the translation of user intents into a number of actions by wizards. These actions can be hard or soft-constraints while wizards are accessing the knowledge base and may be helpful in the training of models mapping user intents to expressive actions. The knowledge base contains structured information linked to these 4 domains. Information is structured into facets, which can be set-valued and can support hierarchically organized values, and labeled or non-labeled intervals as values. The types of their values can be boolean (e.g. free WiFi), numerical (e.g. rat-

ings), geographical for describing actual location, or free text. More information about the way facets are constructed can be found in Section 3.3.

The dialogue collection platform had the following 3 main components introduced in detail below: 1) crowd-sourcing platform to access a large user audience, 2) web-based dialogue interface enabling dialogues between workers and wizards and 3) PFS-based exploratory search system over the knowledge base exploited by wizards.

3.1 Crowd-Sourcing Platform

A crowd-sourcing platform can offer a large number of users to which jobs can be disseminated online. Typically, other services are offered as well, such as the selection of user groups with certain features and handling of workers payments. There are many crowd-sourcing platforms available of which some are described in [9, 10, 18].

The crowd-sourcing platform MicroWorkers⁶ was chosen here instead of the more widely used Amazon Mechanical Turk⁷ platform, which was unavailable in Greece, where the data was collected. MicroWorkers appeared to be a viable option since it:

1. Allows a detailed specification of jobs (estimated duration, # jobs per worker, etc.).
2. Allows to reject workers after a quality check.
3. Allows workers to submit a token as evidence that they carried out the task.
4. Comprises of a large community of users (>1.300.000).
5. Allows selection of specific groups of users, e.g. by regions, worker qualification.
6. Allows tasks on external web-pages.
7. Supports VCODE verification for task validation and payment of workers.
8. Allows to pause, resume and dynamically change the speed of a campaign.

VCODE verification was important to link tasks on external web-pages and workers conducting these tasks, plus quality control and payment of workers. The crowd-sourcing platform provides access to a large user base, but communication between users and wizards needs to be enabled and recorded, which is described next.

3.2 The Web Interface for Collecting Dialogues: Workers and Wizards

The web-interface enabling dialogues between users and wizards resembles a chat platform, in which the communication is synchronous and done by text messages. The free *tawk.to*⁸ platform was chosen to connect users and wizards.

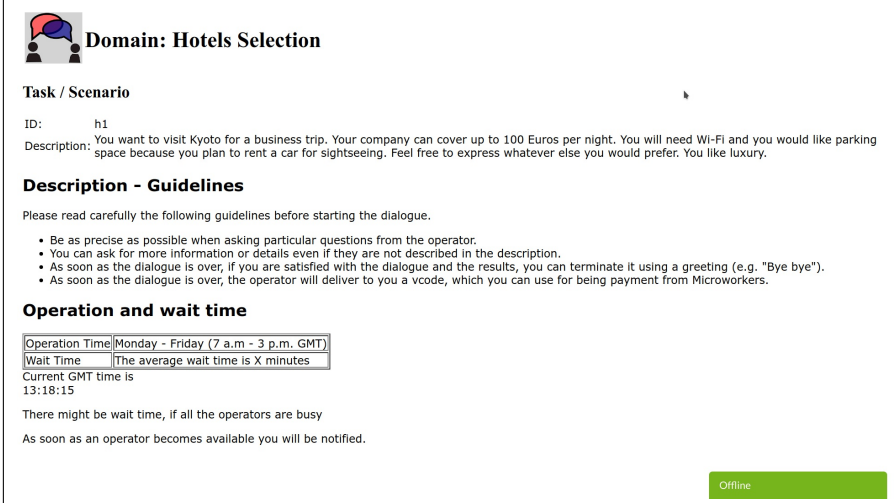
Figure 1 shows the web-interface seen by the worker. A small web application was developed, which loads tasks (either particular or random ones) for each of the different domains. The application shows users a description of the task and some general guidelines for the execution of the dialogue. By clicking on the widget at the bottom right of the window, a chat box opened to initiate the discussion with one of the wizards.

When a worker starts a new dialogue, all the wizards are notified. One (or more) wizards can accept the request and start a dialogue with the user. At that time all the

⁶ <https://www.microworkers.com>

⁷ <https://www.mturk.com>

⁸ <https://www.tawk.to>



Domain: Hotels Selection

Task / Scenario

ID: h1
 Description: You want to visit Kyoto for a business trip. Your company can cover up to 100 Euros per night. You will need Wi-Fi and you would like parking space because you plan to rent a car for sightseeing. Feel free to express whatever else you would prefer. You like luxury.

Description - Guidelines

Please read carefully the following guidelines before starting the dialogue.

- Be as precise as possible when asking particular questions from the operator.
- You can ask for more information or details even if they are not described in the description.
- As soon as the dialogue is over, if you are satisfied with the dialogue and the results, you can terminate it using a greeting (e.g. "Bye bye").
- As soon as the dialogue is over, the operator will deliver to you a vcode, which you can use for being payment from Microworkers.

Operation and wait time

Operation Time	Monday - Friday (7 a.m - 3 p.m. GMT)
Wait Time	The average wait time is X minutes

Current GMT time is
13:18:15

There might be wait time, if all the operators are busy
 As soon as an operator becomes available you will be notified.

Offline

Fig. 1. The chat interface for workers showcasing a task scenario.

required information for generating a valid VCODE (i.e. the ID of the worker, the ID of the task, etc.), were passed as information tags to the wizards. An indicative screenshot of the wizard's view is shown in Figure 2. *tawk.to* can be embedded into a website enabling a live chat functionality.

3.3 Exploring the Knowledge Base: Hippalus

To offer an efficient and easy access to the knowledge base, an exploratory search system called Hippalus⁹ [8] was used. Hippalus enabled fast and efficient access to domain specific information, which is an important aspect for task-oriented dialogue systems. Hippalus allows wizards to explore a knowledge base using the Preference-enriched Faceted Search (PFS) [17] interaction paradigm and overview the information space (e.g. hotels) based on their attributes/values and count information. PFS supports actions with hard and soft-constraints enabling users to order facets, values, and objects. Wizards are able to express hard or soft-constraints (i.e. preferences) over attributes that can be multi-valued, intervals with labeled or non-labeled values, or whose values can be hierarchically organized, and Hippalus automatically resolves any conflicts through preference inheritance. Hard-constraints limit the object space to the desired objects, while soft-constraints provide an ordering of the objects/values/attributes. The actual object space is represented in the *Resource Description Framework Schema (RDFS)*¹⁰ and can be realized either by static collections or is the result of SPARQL queries, a query language of the Semantic Web. Apart from the dialogues, the interest was also to collect the corresponding hard or soft-constraints that the wizards performed in Hippalus.

⁹ <http://www.ics.forth.gr/isl/Hippalus>

¹⁰ <http://www.w3.org/TR/rdf-schema>

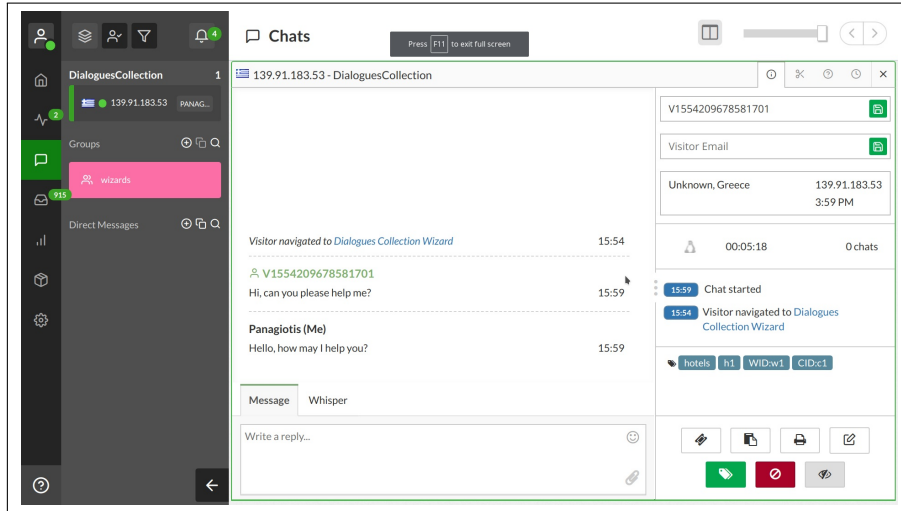


Fig. 2. The chat interface for wizards.

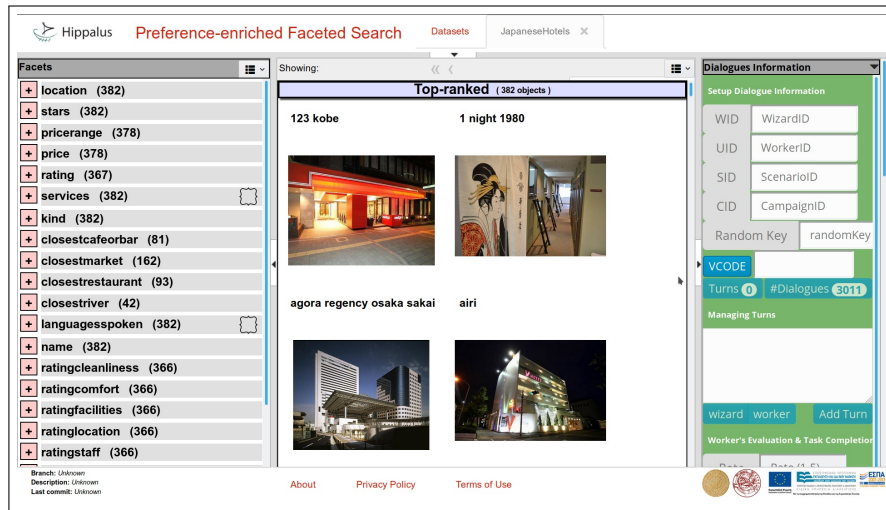


Fig. 3. User interface of Hippalus, enhanced to support the recording of dialogues

The performed preference actions are internally translated to statements in the preference language described in [17] and the respective preference bucket order is computed. Finally, the ranked list of objects according to preference is displayed in the user's browser. For the dialogue collection, Hippalus was extended to record all aspects of a dialogue including the turns (i.e. the narratives), the actions performed in Hippalus (i.e. preferences, restrictions), the restricted ranked objects, information about dialogues (i.e. wizard and worker ID's, scenarios), and evaluation of the dialogue from the perspective of both wizard and worker.

Hippalus was modified to record dialogues, allow wizards to generate a unique token for each dialogue (i.e. UUIDs), generate VCODE tokens which were communicated to workers when dialogues were successfully finished, include tools for inserting turns from workers and wizards, and store all related information and meta-data of each dialogue. Figure 3 shows the enhanced user interface of Hippalus. The right panel shows part of the widgets developed for collecting dialogues, the left panel shows some facets in the hotel domain, while the middle panel shows part of the ranked list of available objects. During dialogue collection, Hippalus enabled wizards fast and efficient access to domain specific data, an important aspect for task-oriented dialogue systems.

3.4 The Dialogue Collection Process

The dialogue collection process includes six major steps shown in Figure 4.

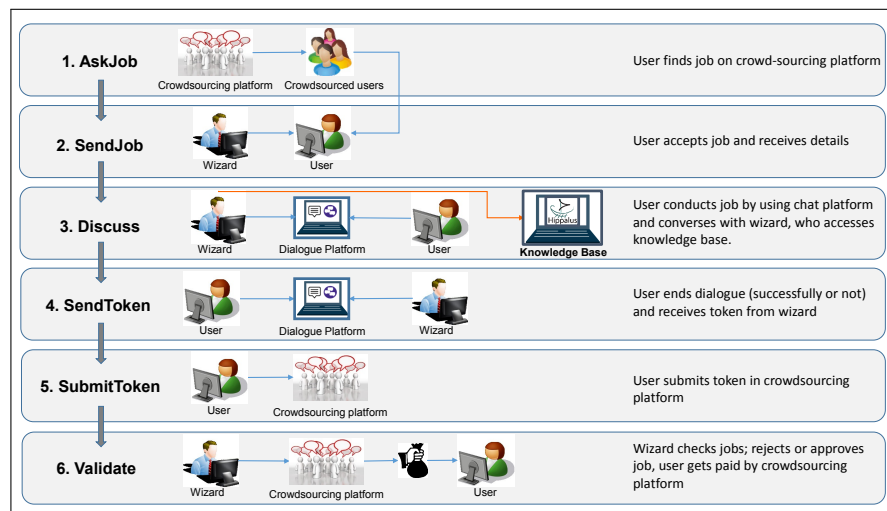


Fig. 4. The processing steps during the crowd-sourced dialogue collection.

To collect dialogues, first the corresponding campaigns had to be created in MicroWorkers. A campaign is a self-contained task submitted to the crowd-sourcing platform. Since MicroWorkers allows creating campaigns for specific groups of workers, a first attempt was made by creating separate campaigns engaging different groups including highly-rated workers from the UK, North-Europe, South-Europe, etc. Unfortunately, these campaigns turned out to be unsuccessful, because they either did not

manage to attract workers, or it took a lot of time to get the attention of workers. The same problem remained, despite increasing the amount of money that workers would receive, or reducing the time period in which they would receive the money.

As a result, international campaigns were started without any restrictions, since there were no available highly-rated groups for international workers. Each campaign was referring to a specific scenario in a particular domain. Workers could participate only once in each campaign, ensuring that many different workers would join, to collect different dialogue strategies and as much linguistic variation as possible.

Another issue was the idle time during dialogues when workers did not respond immediately. To ensure a coherent dialogue, it was meant to be carried out in a synchronous manner. However, it turned out to be important to consider potential delays from workers or wizards. In many cases, wizards had to wait for a response from workers. To minimize the occurrence of such idle periods, wizards tried to respond to the initial request from workers, as soon as they initiated the dialogue. Nine persons were trained as wizards, all in-house experts.

In total 49 campaigns were created; 17 for hotels, 11 for restaurants, 11 for museums, and 10 for shops. The average payment per worker in campaigns was approximately \$0.50 (ranging from \$0.25 – \$1.00). MicroWorkers charged a basic fee of 7.5% for each successful submission + \$0.75 fee per campaign. For “Hire Groups”, which enables one to restrict jobs only to specific workers, a task assignment cost that is 10% of the total cost of the campaign had to be paid.

3.5 Validating and Cleaning Dialogues

A validation step was conducted to ensure that the collected dialogues were of sufficient quality, e.g. not including (a) incomplete dialogues, (b) non-sense dialogues, (c) missing turns, and (d) text that was difficult to comprehend (e.g. typos, grammatical errors, incorrect punctuation). As a result, about 1k dialogues were discarded, most of them incomplete dialogues because workers abandoned the task and additionally some dialogues which were collected during debugging stages.

Also, a number of dialogues had (1) missing turns (usually only one or two), (2) inappropriate user input and (3) grammatical errors, typos, non-sense words and incorrect punctuation. As a result, a final correction and cleaning stage was conducted which included: (1) filling in missing turns, (2) correcting or deleting clearly incorrect user input, and (3) correcting typos in words (e.g. “meseum”, “dishses”), incorrect punctuation (e.g. “are there any. hotels”), nonsense words (e.g. “wellwith”, “facilitiesas”), slang words (e.g. “thnx”, “yw”) and ungrammatical text. All valid dialogues had to be inspected manually and corrected if needed, which took about 4 hours per 100 dialogues.

3.6 Dialogue Data Representation

Dialogues were stored in a relational database using the schema shown in Figure 5. The tables of the schema are: **Scenarios**: Description of the scenarios with their corresponding domain; **Dialogue**: Meta-data information about the dialogue; **Turns**: Text of the dialogue, either from worker or wizard; **Actions**: Actions conducted by the wizards in Hippalus per turn. Each action is described in human-friendly format, the type of the

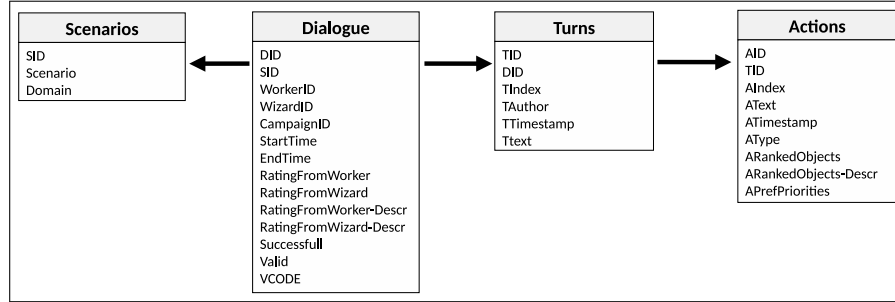


Fig. 5. The schema of the dialogues collection database.

Table 2. Statistics of the collected dialogue corpus.

# dialogues	3010
in domain: hotel	778 (25.8%)
in domain: restaurant	757 (25.1%)
in domain: museum	740 (24.6%)
in domain: shop	735 (24.4%)
Avg. dialogue duration	8 min 52 sec
Avg. # of turns/dialogue	10.4
Avg. # User turns/dialogue	5.7
Avg. # Wizard turns/dialogue	4.7

action and the ranked objects after performing it. For subsequent processing, e.g. in a machine learning framework, dialogues can be exported in JSON format.

3.7 Statistics of the Collected Dialogue Corpus

Table 2 shows the statistics of the collected corpus. More than 3k dialogues were collected which are roughly evenly distributed across the 4 domains. The average duration for the collection of a dialogue was about 9 minutes. The average number of turns in a dialogue is more than 10 and there are more than 31k turns in total. The collection was done in a period of 4.5 months. In total, 590 workers participated for all valid and non-valid dialogues. This number is reduced to 327 for the valid dialogues (workers that provided invalid dialogues of very-low quality were black-listed from our campaigns).

Table 3 shows an example of a collected dialogue from the hotel domain with the corresponding scenario listed on top of the table. Wizards were free to decide which action to use (i.e. hard or soft-constraints) in order to provide better feedback to the users. In the given example, the wizard used a hard-constraint for the *freewifi* service, which is available in all hotels and used a preference for the *freeparking* action. By not making a hard-constraint over the *freeparking* service of the one hotel, the wizard can explore a bigger variety of other services offered by the available hotels.

Table 3. Example of a dialogue in the corpus plus selected Hippalus actions.

Scenario: *You want to visit Kyoto for a business trip. Your company can cover up to 100 Euros per night. You will need WiFi and you would like parking space because you plan to rent a car for sightseeing. Feel free to express whatever else you would prefer. You like luxury.*

Actor	Text of turn	Hippalus actions
WORKER	<i>Hello. I am looking for a hotel room in Kyoto for a business trip with a price up to 100 euros. How many options are there?</i>	Focus=location: kyoto_prefecture; Add preference action=objects order: term price-range...{very_cheap, cheap, moderate} best
WIZARD	Hello, It seems that there are around 47 hotels with the criteria that you mentioned. Would you like to search using some other criteria?	
WORKER	<i>Well, I would need WiFi and a parking space. Any other available features are also welcome.</i>	Focus=services: freewifi; Add preference action=objects order: term services... freeparking best
WIZARD	All of them offer free WiFi, however only one offers free parking. As regards to other facilities I can find many hotels that have a swimming pool and restaurant inside the hotel. Are you interested in these facilities?	
WORKER	<i>Well I think I will take the one with the free parking. Can you please give me the address and the telephone number?</i>	Focus=services: freeparking
WIZARD	It is the rihga royal kyoto. The address is shimogyo-ku higashihorikawa-dori shiokoji-sagaru taimatsu-cho 1, japan. Unfortunately I do not have information about the telephone number.	
WORKER	<i>Ok. Thank you very much. Bye Bye</i>	
WIZARD	Bye	

4 Discussion

One of the lessons learned during the dialogue collection process was certainly the time consuming aspect of the actual dialogue collection process. The original estimation for the time period needed to collect one dialogue was about 5 minutes. In reality, it took on average about 9 minutes, i.e. almost twice the duration originally estimated, mainly due to workers latency to respond and the aim to collect dialogues in a synchronous manner. This also had an impact on the number of dialogues which could be collected in the given time period. However, the chosen set-up ensures a coherent, synchronous dialogue between users and wizards and avoids potential in-coherence which may occur in asynchronous dialogue collection methods such as the one used in [2] and [20].

Conducting a crowd-sourced dialogue collection also depends heavily on the availability and language skills of the workforce. Workers from English-speaking countries do not seem to be readily available, and even when they are, it does not guarantee good quality language skills, since they might not be native speakers of those countries. On the other hand there were some workers from non-native English countries like India, that had excellent language skills, who participated in most of the campaigns, and seem to make a living from crowd-sourcing platforms. Further, the fact that there was a rather large number of “invalid” dialogues and that a cleaning step had to be conducted, shows one of the drawbacks of crowd-sourced data collections: costs to collect relatively large amounts of data can be relatively low, but the quality can also vary significantly.

Analyzing the cost for collecting the dialogues collection showed that 15% of the total budget was spent for development activities, 30% at setting-up, validating and documenting the platforms, datasets and scenarios, 45% for the in-house trained experts that played the role of the wizards, 10% was the cost of the crowd-sourced workers, and finally, 10% was spent for cleaning and validating the collected dialogues. It needs to be mentioned, that a number of workers abandoned the dialogue before completion, and as

a result the expensive effort of the experienced wizards was spent for no results. Consequently, reducing the idle time of the wizards (i.e. by early ending non-active dialogues or black-listing low-quality workers) and increasing the percentage of completed, valid and good quality dialogues (i.e. by increasing the percentage of highly rated workers which is a non-trivial task) is important for reducing the total cost of such efforts. Another option is to train workers from the crowd-source platform as wizards which might be a cheaper option. This option was not chosen, because the presented method allowed to better control the efficiency and quality of the wizards results.

Finally, the fact that the current corpus was collected by written communication might introduce a certain bias towards specific ways of interaction and behaviours, as mentioned in [15]. Further, the time needed to post a response is larger for both workers and wizards in written rather than oral communication and the dialogues should be validated and cleaned as discussed previously. The obvious solution is to collect spoken human-human conversations, which is a path to consider for future work.

5 Conclusion

This paper presented the experiences and the lessons learned from the process of collecting expressive and synchronously written human-human dialogues (and their associated meta-data), for a task-oriented multi-domain scenario. The context is a travel agency environment and the dialogues cover four domains (hotels, museums, restaurants, and shops) in Japan. Users were asked to converse with wizards to achieve their tasks (i.e. select the more appropriate resource) via a chat platform, and wizards formulated their responses while accessing a knowledge base via an exploratory system based on Preference-enriched Faceted Search (*PFS*).

More than 4,000 dialogues were collected in 4.5 months. Due to the crowd-source nature of the set-up and the variation in quality of workers input, a significant percentage was invalid or required a clean-up process, leaving at the end a corpus of 3,010 valid dialogues. Each valid dialogue lasted almost 9 minutes and had on average 10.4 turns. The effort for collecting the above was analyzed, and it turned out, that the cost of the in-house experts that played the role of the wizards was almost 4.5 times more than the cost of the workers for a crowd-sourced campaign.

One distinctive feature of the corpus, is the fact that it recorded the actions taken by wizards, both *hard* and *soft*-constraints, that not only *fill* a slot or restrict the objects, but also *rank* the objects. Consequently the corpus can be exploited for more “refined” training. One future step is to annotate the corpus with dialogue acts, and then use the corpus for training an end-to-end, as well as modular, neural network-based dialogue system, and assessing the value of the corpus.

References

1. Asri, L.E., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., Mehrotra, R., Suleman, K.: Frames: A Corpus for Adding Memory to Goal-Oriented Dialogue Systems. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (2017)

2. Budzianowski, P., Wen, T., Tseng, B., Casanueva, I., Ultes, S., Ramadan, O., Gasic, M.: MultiWOZ - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. [Dataset] (2018)
3. Dahlbäck, N., Jönsson, A., Ahrenberg, L.: Wizard of Oz Studies: Why and How. In: Proceedings of the 1st International Conference on Intelligent User Interfaces. pp. 193–200. IUI '93, ACM, New York, NY, USA (1993)
4. Kelley, J.F.: An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)* 2(1), 26–41 (1984)
5. Li, X., Panda, S., Liu, J.J., Gao, J.: Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems (2018)
6. Lowe, R., Pow, N., Serban, I., Pineau, J.: The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In: Proceedings of the 16th Annual SIGdial Meeting on Discourse and Dialogue (2015)
7. Mrkšić, N., Ó Séaghdha, D., Wen, T.H., Thomson, B., Young, S.: The neural belief tracker: Data-driven dialogue state tracking. In: ACL. Vancouver, Canada (2017)
8. Papadakis, P., Tzitzikas, Y.: Hippalus: Preference-enriched Faceted Exploration. In: EDBT/ICDT Workshops. vol. 172 (2014)
9. Peer, E., Brandimarte, L., Samat, S., Acquisti, A.: Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70, 153–163 (2017)
10. Peer, E., Samat, S., Brandimarte, L., Acquisti, A.: Beyond the Turk: An Empirical Comparison of Alternative Platforms for Online Behavioral Research. *SSRN Electronic Journal* (2015)
11. Petrik, S.: Wizard of Oz Experiments on Speech Dialogue Systems. Diploma thesis, Technical University of Graz (2004)
12. Ritter, A., Cherry, C., Dolan, W.B.: Unsupervised modelling of Twitter conversations. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 172–180 (2010)
13. Schradang, N., Alm, C., Ptucha, R., Homan, C.: An Analysis of Domestic Abuse Discourse on Reddit. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2577–2583 (2015)
14. Serban, I., Sordani, A., Bengio, Y., Courville, A.C., Pineau, J.: Hierarchical neural network generative models for movie dialogues. *ArXiv e-prints* (2015)
15. Serban, I.V., Lowe, R., Henderson, P., Charlin, L., Pineau, J.: A Survey of Available Corpora for Building Data-Driven Dialogue Systems: The Journal Version. *Dialogue & Discourse* 9(1), 1–49 (2018)
16. Shang, L., Lu, Z., Li, H.: Neural Responding Machine for Short-Text Conversation. In: ACL. pp. 1577–1586. Beijing, China (2015)
17. Tzitzikas, Y., Papadakis, P.: Interactive exploration of multi-dimensional and hierarchical information spaces with real-time preference elicitation. *Fundamenta Informaticae* 122(4), 357–399 (2013)
18. Vakharia, D., Lease, M.: Beyond Mechanical Turk: An analysis of paid crowd work platforms. Proceedings of the iConference (2015)
19. Vinyals, O., Le, Q.V.: A Neural Conversational Model. In: ICML Deep Learning Workshop. Lille, France (2015)
20. Wen, T.H., Vandyke, D., Mrkšić, N., Gasic, M., Rojas Barahona, L.M., Su, P.H., Ultes, S., Young, S.: A Network-based End-to-End Trainable Task-oriented Dialogue System. In: EACL. pp. 438–449. Valencia, Spain (2017)