Wizard-of-Oz Support Using a Portable Dialogue Corpus

Masashi Inoue and Hiroshi Ueno

Yamagata University,
3-16, 4 Jyonan, Yonezawa, Yamagata, Japan

Abstract. This paper presents a Wizard of Oz (WOZ) data collection method that uses dialogue examples (or utterances) from one domain for use in a target domain. Providing a text-based dialogue system with empathy requires providing the system with a wide range of expressions, with expressions corresponding best to users. However, there are few dialogue examples available and the variation of utterances is limited. We have to collect wider range of example utterances. A typical method to collect dialogue data is the WOZ method. The use of WOZ for dialogue data collection often requires substantial cognitive load for participating wizards. To alleviate this problem, an utterance suggestion mechanism using a portable corpus is introduced. We investigated differences in the response times of a wizard when utterance suggestions from a portable corpus are offered. We also evaluated the ratio of utterance suggestions selected versus free utterances. The experimental results indicate that using a portable dialogue corpus to suggest utterances for wizards has a potential to be helpful in data collection.

1 Introduction

Providing a text-based dialogue system with empathy requires a wide range of expressions corresponding best to users. However, in most dialogue scenarios or domains, there are currently few dialogue corpora for obtaining utterance variations in constructing example-based dialogue models. Dialogue samples (or utterances) must be collected from either human-to-human or human-to-machine interactions. The Wizard-of-Oz (WOZ) method is often used for such data collection in human-to-machine situation [1]. In this method, a participant, or a wizard, simulates the behaviour of the dialogue system being developed and interacts with the users through utterances. A problem with the use of this method is that the cognitive load of a wizard is often considerably high, when composing and typing utterances in text-based dialogue systems. In addition, it is required to collect diverse interactions to create the dialogue corpus suitable for training data of statistical dialogue systems. If a wizard is not the system designer, it might be difficult to predict what dialogue would benefit a future system. As a result, wizards may find it difficult to generate appropriate utterances. On the wizards' cognitive load problem, user interfaces for wizards have been improved to mitigate the cognitive load [8]. In addition to the interface, we further

consider dialogue content to overcome the cognitive load problem by preparing utterance suggestions, or candidates. By selecting and clicking one of the candidates, wizards can bypass the process of generating utterances and preparing textual output. The problem is then how to prepare the collected utterance candidates. The size of utterance candidates may still significantly large to be created manually and dialogue samples must been collected automatically.

To create an utterance candidate set automatically, in this paper, we propose a method for importing a corpus that has been developed in a domain other than the current target domain. The imported corpus after modification is called a portable corpus. We experimentally compared the effectiveness of using the portable corpus as the source of utterance candidates during the WOZ data collection. We found that role-play became easier for wizards with utterances suggested from the portable corpus.

2 Corpus Insufficiency

2.1 Corpus Porting

The scarcity of dialogue corpora has been an issue in the field of dialogue understanding in which systems have been developed by using large-scale training data. There have been several attempts to mitigate corpus insufficiency. For example, domain and language portability were sought by using machine translation and concept mapping [5]. In contrast to the dialogue understanding task, in which it is important to handle what is said, in the dialogue generation task, in which it is important how to say something, the portability needs to be carried out by preserving expressions rather than by skimming the statistical essence of the original corpora [4]. By focusing on the exact phrases used in one domain, researchers have investigated the utility of cue phrases in one domain for classifying the dialogue acts in another domain [9].

2.2 Alternative Dialogue Corpus

Considering the scarcity of recorded dialogue data, other types of language resources have been investigated for alternative dialogue corpora. For example, twitter conversation has been considered as the corpus. In reality, there are few direct interactions between tweets. Therefore, a pseudo conversation has been investigated. Two tweets that are similar in content are considered a pseudo conversation [2]. Another approach is the use of gamification. Crowdsourcing has been combined with gamification to obtain dialogue data for non task-oriented dialogue [3]. The game used was the dialogue skill test in which participants got higher score if they select typical utterances than unusual ones. These methods were used for generic chitchat. In this study, we investigate another approach that incorporates actual dialogue samples in different domains for example-based utterance generation for specific dialogue tasks.

3 Dialogue Corpus

3.1 Target Corpus

We consider three kinds of corpora: source corpus, target corpus, and portable corpus. First, we describe the target corpus that consists of dialogue samples relevant to the dialogue systems being developed. Our target corpus is the collection of dialogues for the LAST MINUTE task [7]. In this task, users were asked to pack travel items into a suitcase during summer break. The travel was planned without prior appointment, and a taxi had been called to pick the users up in approximately a half hour. Through the interaction with a dialogue system, the users had to complete packing their items in a limited time. The items were grouped in several categories. There are multimodal dialogues for the task collected by using the WOZ method. The original corpus consists of three dialogues in German, and we translated one dialogue into Japanese. We use the translated dialogue as an example from the target corpus. Most of the utterances in the example are to progress the task such as "Next category". Part of the dialogue that is translated into English is shown in Table 5.

3.2 Source Corpus

We use the NICT Kyoto tour dialogue corpus[6] as the source corpus. This corpus contains the dialogues exchanged between simulated travellers and a professional tour guide for determining a one-day sightseeing plan in the city of Kyoto. They talked about where to visit and which means of transportation should be used. The corpus is a large-scale dataset and consists of 100 dialogues each of which contains 300-700 utterances. A total of $42,673$ utterances were available and used in our experiment. All participants were native speakers of Japanese. An example dialogue from this source corpus is shown in Table 6.

3.3 Portable Corpus

The portable dialogue corpus is derived from a source dialogue corpus by removing domain-dependent or task-dependent expressions from the utterances. The process of converting a source corpus into a target corpus is called porting. Given an utterance $\mathbf{u} = \{w_1, w_2, ..., w_m..., w_M\}$ where w_m represents a word or morpheme in the utterance, and assuming that each word has its domain dependency score s_m, we remove all the utterances that contain w_m with $s_m \geq t$ where t denotes a threshold value. In the following experiment, we set $s_m = 1$ when w_m is either a proper noun or a numeral; otherwise, $s_m = 0$. The threshold value t was set to 0. For example, assume that the following three utterances are in the source corpus: 1) 'Do you want to go to Tenryuji temple?', 2) 'Yes, how long does it take?', 3) 'It takes about 30 minutes.' The first utterance contains the proper noun w_7 'Tenryuji' with $s_7 = 1$ and deleted. Similarly, the third utterance includes a number w_4 '30' with $s_4 = 1$ and removed. Then, the remaining second utterance is stored in the portable corpus. As the result of processing the entire

target corpus, we obtained 5,868 utterances for the portable corpus. Although the LAST MINUTE task focuses on the items for travelling while the NICT Kyoto tour dialogue task focuses on the schedule, there are some expressions that are commonly used in both settings. Example utterances in the portable corpus are shown in Table 7. Note that the utterances in portable corpora are sampled and used independently; They do not appear in the original order when used in the experiment.

3.4 Difference Between the Corpora

The difference between source and target corpora can be characterized in the following two aspects. The first aspect is the expected use of the corpora. The source corpus was created for the development of a statistical dialogue system, and a large number of dialogues of human-to-human communication in a face-to-face setting were collected. Different tourists have different preferences, and their destinations may vary. The target corpus was created to understand the user behaviour when interacting with dialogue systems. The task is designed for collecting emotionally elicited multimodal reactions by users under stress. The dialogues proceed according to a fixed procedure toward the completion of the task. Therefore, it is relatively easier to collect dialogue samples of similar content from diverse participants than a free-structured dialogue. This allows the designers to explore the nature of multimodal emotional reactions of different users.

The second aspect is the type of interaction. Both source and target corpora are created from task-oriented dialogue records, and are related to travel situations. The dialogue task in the source corpus involved planning a schedule through human-to-human communication in a face-to-face setting. On the other hand, the dialogue task with the system in the target corpus is concise and operational. Therefore, although the tasks are related to the travel preparations in both source and target corpora, the characteristic of utterances are not similar except the ones used as portable corpora.

4 WOZ System

4.1 System Overview

The WOZ dialogue system consists of user input window, wizard input window, utterance candidate window, and task progress assist window. The task progress assist window is similar to the utterance candidate window but contains predefined utterance candidates for making topic shift toward the task goal. Users and wizards interact through user input text and wizard input text. The current dialogue logs are shown to both wizards and users. Utterance candidates and progress assist list are shown to wizards only. Progress assist list consists of utterances aligned in line with the progress of LAST MINUTE packing task. Wizards can select utterances from the list at any point to move the dialogue to the next stages.

4.2 Utterance Candidates

To reduce the load for wizards in the WOZ data collection process, wizards can select utterances from the list rather than composing utterances by themselves. The utterance candidates are generated by using a ranking model that is statistically trained by using dialogue samples. For the rank-learning algorithm to find relevant responses from the samples, the ListNet algorithm was used[10]. Five highest ranked utterances were presented to wizards as the response candidates. The candidates were taken from the portable corpus, not from the target corpus. If the data collection process continues, we can add utterances from the collected target dialogues to enrich sample utterances.

5 Experiment

5.1 Experimental Conditions

We conducted a WOZ task in the LAST MINUTE scenario described in 3.1 for the purpose of evaluation, measuring the degree of cognitive load. The progress of the dialogue task was supported by the system. The subjects or wizards were ten students majoring in computer science, most with limited travel experience. Note that the wizards, not the users of the system, were the subjects of the experiments. The person playing the role of a user of the dialogue system was fixed and proceeded with the dialogue in the same manner as in the example dialogue in the corpus, independently of the participating wizards. Therefore, even though the user knew that the counterpart was wizards and not automatic dialogue system, it did not influence the interaction. We provided information on the dialogue, and asked the wizards to utter freely with the person playing the role of the user. When the user uttered, the system selected five utterance candidates for the wizards from the portable corpus based on the trained language model. The subject can select a relevant utterance as the system utterance from the five candidates shown. If none were relevant, the wizard composed a system utterance. The dialogues ended when time ran out, or when the packing task was completed. The number of candidate utterances was determined based on usability, which depended on the window size.

We estimate the degree of a wizard's load reduction in terms of response time and the ratio of utterances taken from the candidates. The response time was the sum of the user time (a subject was selecting or editing responses) and the system time (the WOZ system was retrieving utterance candidates). We had an assumption that the content of the utterances are not influenced by the use of candidates.

5.2 First Experiment

In the first experiment, the user was the system designer, and the wizard was a volunteer student. The LAST MINUTE task was performed using a text-based

Table 1. Median response time measured from the end of previous utterance.

Candidate unused	80.24 sec.
Candidate used	41.72 sec.

Table 2. Ratio of candidate usage at each dialogue phase (%).

Introduction	14.29
Packing	6.45
Closing	25.00

dialogue interface. In the system, the wizard had three options to input utterances: editing text freely, selecting utterances from the candidates and modifying them if needed, and selecting task progression utterances to move to the next topic. The user followed as similar path as that of the example dialogue from the target corpus. All participants were native speakers of Japanese and used Japanese as their language of communication. The response time result is summarized in Table 1 and the candidate usage result is summarized in Table 2. From Table 1, it can be concluded that the utterances using candidates were generated quicker than the manually edited utterances. Moreover, utterance candidates were used more often during the beginning and ending phrases of the dialogue as shown in Table 2. In those phases, there were greetings and informational utterances about users themselves. The wizard used candidate utterances more in those introduction and closing stages to diversify the interaction because the users could not develop the topic if the system responded with simple back-channelling. During the packing task in the middle of the dialogue, if the system responded with simple back-channelling, the user did not have to respond but could proceed with the task. This characteristic led to a lower rate of candidate usage.

5.3 Second Experiment

The entire LAST MINUTE session lasts about an hour and is expensive. Therefore, we conducted a shorter experiment by using the first introduction phase of the session. The user uttered dialogues freely to interact with the wizards. As in the first experiment, the user was fixed and eight wizards participated. The closing utterance was pre-defined as in the previous experiment, but other interactions were conducted freely. We found that the use of candidates varied among participants as shown in Table 3. Some participants used only the candidate utterances as the wizards while other participants did not use the candidates at all. Two wizards, 4 and 5, used both utterance candidates and their own utterances. We compared their median response times in both conditions as shown in Table 4. Although the absolute response time differed among wizards, we could observe that the use of candidates reduced their response time.

Table 3. Ratio of candidate usage (%).

Participant index	1	2	3	4	5	6	7	8
Usage ratio	100	100	100	71.43	75	0	100	0

Table 4. Median response time measured from the end of previous utterance.

Participant index	4	5
Candidate unused	46.6 sec.	38.0 sec.
Candidate used	31.9 sec.	24.1 sec.

6 Conclusion

The WOZ method is often used for collecting realistic dialogue data between users and the dialogue system in a particular domain. In WOZ, however, a wizard may find it troublesome to generate and input appropriate utterances for data collection purpose. In this study, we investigated a method to suggest utterances taken from a dialogue corpus in order to reduce the burden of the wizards. Since large dialogue corpora usually do not exist in the target domain, we modified an existing dialogue corpus from another domain (source corpus) into a reusable form (portable corpus). The utterance candidates were then suggested from the portable corpus. In the experiment using a text-based dialogue interface with the utterance suggestion functionality, the utterance candidates were frequently used during the WOZ data collection process. In addition, we found that the utterance input time was reduced by suggesting utterances. These results suggest the utility of utterance candidates taken from the portable dialogue corpus.

To increase confidence in the above mentioned benefits of the proposed method, we need to conduct a further experiment. In this study, we fixed a user and varied wizards for the purpose of comparison. If we compare multiple users, we can examine the influence of different interaction styles. Also, the dialogue experience of users were not quantitatively evaluated. With multiple users, we can measure the quality of the utterances generated by the wizards. That is, we can test that the decrease of cognitive loads of the wizards did not degrade their response quality.

A possible improvement to the proposed method would be to create a more sophisticated porting procedure for creating the portable corpus. Another direction of improvement is the combination of several source corpora to build a larger portable corpus. In this study, we used quantitative evaluation measure: usage ratio of utterance candidates and response time. Qualitative evaluation such as protocol analysis may help understanding the utterances selection criteria used by the wizards.

Acknowledgements

This research was partially supported by Grant-in-Aid for Scientific Research 24500321. The initial WOZ system was developed by Takahiro Sekino. Part of the experiment was conducted by Kodai Takahashi.

References

1. Dahlbäck, N., Jönsson, A., Ahrenberg, L.: Wizard of Oz studies - why and how. Knowledge-Based Systems 6(4), 258–266 (1993)
2. Higashinaka, R., Kawamae, N., Sadamitsu, K., Minami, Y., Meguro, T., Dohsaka, K., Inagaki, H.: Building a conversational model from two-tweets. In: IEEE Workshop on Automatic Speech Recognition and Understanding. pp. 330–335. Waikoloa, HI (2011)
3. Inaba, M., Iwata, N., Toriumi, F., Hirayama, T., Enokibori, Y., Takahashi, K., Mase, K.: Constructing a non-task-oriented dialogue agent using statistical response method and gamification. In: 7th International Conference on Agents and Artificial Intelligence (ICAART2014). pp. 14–21 (2014)
4. Inoue, M., Matsuda, T., Yokoyama, S.: Web resource selection for dialogue system generating natural responses. In: HCI International 2011 – Posters' Extended Abstracts. Communications in Computer and Information Science, vol. 173, pp. 571–575. Springer Berlin Heidelberg (2011)
5. Mostefa, L.F., Besacier, D., Esteve, L., Quignard, Y., M. Camelin, N., e.a.: Leveraging study of robustness and portability of spoken language understanding systems across languages and domains: The PORTMEDIA corpora. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey (2012)
6. Ohtake, K., Misu, T., Hori, C., Kashioka, H., Nakamura, S.: Dialogue acts annotation for NICT kyoto tour dialogue corpus to construct statistical dialogue systems. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valetta, Malta (2010)
7. Rösner, D., Frommer, J., Friesen, R., Haase, M., Lange, J., Otto, M.: LAST MINUTE: a multimodal corpus of speech-based user-companion interactions. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey (2012)
8. Schlögl, S., Schneider, A., Luz, S., Doherty, G.: Supporting the wizard: Interface improvements in Wizard of Oz studies. In: BCS-HCI '11 Proceedings of the 25th BCS Conference on Human-Computer Interaction. pp. 509–514. Swinton, UK. (2011)
9. Webb, N., Liu, T.: Investigating the portability of corpus-derived cue phrases for dialogue act classification. In: The 22nd International Conference on Computational Linguistics (Coling 2008). pp. 977–984. Manchester, UK. (2008)
10. Z.Cao, T.Qin, T.Y.Liu, M.F.Tsai, H.Li: Learning to rank: From pairwise approach to listwise approach. In: Proceedings of the 24th International Conference on Machine Learning. pp. 129–136. ICML '07 (2007)

Appendix

Table 5. Dialogue example derived from the LAST MINUTE corpus. W and U indicate the wizard and user utterances, respectively. The transcripts here were translated from Japanese. The Japanese transcripts were translated from German.

	Utterance
W	A taxi will arrive in a few minutes.
W	Next, you can select reading materials.
U	I'll take a newspaper with me.
W	Newspaper has been added.
U	Next category.
W	In this category, you can choose devices.
U	Audio player.
W	Audio player has been added.
U	Adapter.
W	Adapter has been added.
U	And, a camera.
W	Maximum weight has been exceeded and the camera has not been added.
U	I'll take out the adapter.
W	Adapter has been removed.

Table 6. Dialogue example derived from the Kyoto sightseeing speech corpus. G and U indicate the guide and user utterances, respectively. The transcripts here were translated from Japanese.

	Utterance
U	Excuse me. The entrance fee for Tenryuji temple is ...
G	The entrance fee is 500 yen, yes.
U	500 yen, I see.
G	Yes. Or it will be around noon then.
U	Yes, it will almost be time.
G	Uh, you can have lunch then.
U	Yes.

Table 7. Example utterances from the portable corpus. These are independent and not a series of utterances.

Yes, how long does it take?
Thank you very much.
Where do you recommend?
Then, it will be evening, yes.
Well, I will finish here.
Uh, that's correct.
So, I cannot make it, yes.
Yes, yes, yes.