Shared Task on Data Collection and Annotation

WOCHAT SHARED-TASK REPORT

Luis F. D'Haro[1], Bayan Abu Shawar[2], Zhou Yu[3]

[1]Institute for Infocomm Research, [2]Arab Open University, [3]Carnegie Mellon University
luisdhe@i2r.a-star.edu.sg, bshawar@yahoo.com, zhouyu@cs.cmu.edu

Abstract. This report presents and describes the shared task on "Data Collection and Annotation" conducted in WOCHAT, the Second Workshop on Chatbots and Conversational Agent Technologies. We present a quick review of the proposed series of shared tasks and summarize the results of this second edition in terms of chatbot platforms made available for it and the amount of chatting sessions collected and annotated.

Keywords. Chat-oriented Dialogue, Data Collection, Data Annotation.

1 Introduction

As part of the activities of the workshop, WOCHAT[1] (Second Workshop on Chatbots and Conversational Agent Technologies) has continued the shared task on "Data Collection and Annotation" initiated in the first edition of the workshop. The main objective of the shared task continues to be developing and testing a new evaluation framework for non-goal-oriented dialogue systems.

The rest of the paper is structured as follows. First, a brief review of the basic objectives and intended roadmap for the shared task is presented in section 2. Then, the chatbot platforms made available for the shared task are briefly described in section 3 and, finally, a summary of the collected data and annotations is presented in section 4, followed by the conclusions and future work proposal in section 5.

2 Main Objectives and Road Map

As already presented during the first edition of the workshop [7], this shared task is part of a larger scope initiative, which main objectives are:

- to collect large volumes of chat-oriented dialogue data that can be made available to the research community for research purposes, and
- to develop a framework for the automatic evaluation of chat-oriented dialogue systems.

[1] http://workshop.colips.org/wochat/

This effort comprises three interdependent tasks:

- **Chat Data Collection**: involves the collection of human-chatbot and human-human chat-oriented dialogue sessions.
- **Subjective Evaluation:** involves the manual scoring and annotation, at the turn level, of the collected dialogue sessions by following the proposed subjective evaluation criteria[2].
- **Subrogated Metric Generation**: involves the use of machine learning techniques to generate models able to reproduce and automatically generate the manual scoring and annotations.

Similar to the previous edition of the shared task, the current edition has focused only on the first two tasks, as the third task will be addressed in future editions of the workshop after enough annotated data has been generated to make feasible the use of machine learning approaches.

Again, four different ways of participation in the shared tasks were defined:

- **Chatbot provider:** includes participants that own a chatbot engine and want to provide access to it either by distributing a standalone version of it or by giving access to it via a webservice or web interface.
- **Data generator:** includes participants willing to use one or more of the provided chatbots to generate dialogue sessions with it.
- **Data provider:** includes participants that own or have access to a chatbot but you cannot provide access to it. However, they can generate dialogue sessions with it and share the generated datasets.
- **Data annotator:** includes participants that are willing to annonate some of the generated and/or shared dialogue sessions by following the provided annotation guidelines.

In addition to the 14 volunteers that registered for the first edition of the shared tasks, a total of 15 new volunteers registered for participating in this second edition. These new volunteers accounted for a total of 2 chatbot providers, 15 data generators, 4 data providers and 7 data annotators.

3 Chatbot Platforms Available

In addition to the five chatbots available during the first edition of the shared task, four new chatbots have been made available for this second edition. Three of these new chatbots (Zen, Iesha and Rude) [9] are based on a NLTK library implementation, which uses regular expressions. These three chatbots have been integrated into the WebChat online platform [10]. The other new chatbot, Sammy [2], is based on the public "small-talk"[3] domain available api.ai. Three different versions of Sammy are available, each of which is conversant in a different language: English, French and Italian. Next, all nine chatbots available for the shared task are briefly described.

[2] http://workshop.colips.org/wochat/documents/Annotation_Guidelines.pdf
[3] https://docs.api.ai/docs/small-talk

- **Joker.** An example-based system that uses a database of indexed dialogue examples automatically built from a television drama subtitle corpus to manage social open-domain dialogue [6].
- **IRIS** (Informal Response Interactive System). Implements a chat-oriented dialogue system based on the vector space model framework [1].
- **Py-Eliza.** A Python-based stand-alone version of the famous Eliza chatbot created by Weizenbaum in 1966 [8].
- **Sarah.** Upgraded version of Alicebot, created by Dr. Wallace in 1995 [4].
- **TickTock.** A chatbot to engage users in everyday conversations. Keyword based retrieval system with engagement conversational strategies [12].
- **Sammy.** A friendly virtual assistant based on the public "small-talk" domain of api.ai. Sammy is conversant in English, French or Italian [2].
- **Zen.** A regular expression based conversational agent. This agent mostly provides philosophical advices following the Zen philosophy [9].
- **Ieasha.** A regular expression based conversational agent. This is a teenager chatbot that discusses abut anime and talks colloquial terms [9].
- **Rude.** A regular expression based conversational agent. This agent behaves as a cynical man, giving the user snarky answers [9].

Most of these chatbots are available via online interfaces or as standalone systems for collecting chatting interactions with the registered participants. As initiated in the first edition of the shared task, the plan is to keep these systems available on a continuous basis and grow the number of systems on future editions of the shared task.

4 Data Collection and Annotation

The same data generation and annotation guidelines from the first edition of the shared task were provided to participants this time [7]. Table 1 shows a summary of the newly collected dialogue sessions by the time this report was written.

Table 1. Summary of collected dialogues during this second edition of the shared-task.

Joker	IRIS	Py-Eliza	Sarah	TickTock	Sammy	Zen	Iesha	Rude
62	59	10	5	74	50*	2	2	3

** Sammy 50 dialogues include: English (38), French (8) and Italian (4)*

Regarding subjective evaluations, in addition to the user satisfaction surveys (a total of 19 user satisfaction surveys were collected), two teams have prepared comprehensive reports describing their user experiences and subjective evaluations [5,11]. For the case of turn-level annotations, 48 new dialogue annotations have been manually generated by shared task participants. Additionally, an exploratory evaluation on the feasibility of generating turn-level annotations by means of crow-sourcing resources was conducted [3]. As a result of this study, it was concluded that the use of a majority voting strategy over three independent crowd-sourced annotations can confidently replace expert annotations. Based on this results 40 new dialogues annotations have been generated. All data is available through WOCHAT's website.

5 Conclusions

This report described the second edition of the shared task on "Data Collection and Annotation" conducted in WOCHAT. We have reviewed the main road map envisaged for the shared task series and summarized the main results of the shared task, in terms of chatbot platforms, chatting sessions and annotations. A smaller number of sessions have been collected and annotated, compared to the first edition of the shared task, which might be due to the short period of time in between both workshops.

As future work for next editions of the shared task, we propose to continue with both manual annotations by participants and crowd-sourced annotation schemes, as well as to continue working in the consolidation of a centralized data collection and annotation platform for chat-oriented dialogue.

6 Acknowledgements

We want to thank all the volunteers who contributed to the shared task activities: Lin Lue, Sara Falcone, Amanda Cercas, Diego Pedro, Chan Kah Leong, Shirley Gabber, Eugenia Hee, Carla Gordon, Elisabeth Fritzsch, Rafael Schulman, Jessica Tin, Jeremy Brown, Cristian Cepeda, Guillaume Dubuisson Duplessis, Rafael E. Banchs.

7 References

1. Banchs, R.E., Li, H. (2016) IRIS – Informal Response Interactive System, in Proceedings of RE-WOCHAT, LREC 2016, Shared Task Report.
2. Banchs, R.E. (2016) Sammy: a Friendly English/French/Italian Assistant, in Proceedings of WOCHAT, IVA 2016, Shared Task Report.
3. Banchs, R.E. (2016) Expert-generated vs. Crowd-sourced Annotations for Evaluating Chatting Sessions at the Turn Level, in Proc. of WOCHAT, IVA 2016, Shared Task.
4. Bayan, A.S. (2016) Sarah Chatbot, in Proc. of RE-WOCHAT, LREC 2016, Shared Task.
5. Cercas, A., Rieser, V. (2016) WOCHAT Participation Report, in Proceedings of WOCHAT, IVA 2016, Shared Task Report.
6. Dubuisson Duplessis, G., Letard, V., Ligozat, A.L., Rosset, S. (2016b) Joker Chatterbot, in Proceedings of RE- WOCHAT, LREC 2016, Shared Task Report.
7. D'Haro, L.F.; Bayan, A.S.; Yu, Z. (2016). Shared Task on Data Collection and Annotation, in Proceedings of RE-WOCHAT, LREC 2016, Shared Task Report.
8. D'Haro, L.F. (2016) Py-Eliza: A Python-based Implementation of the Famous Computer Therapist, in Proceedings of RE-WOCHAT, LREC 2016, Shared Task Report.
9. D'Haro, L.F., Lue, L. (2016) Regular Expression Based Agents for Online Collection of Human-Chatbot interactions, in Proc. of WOCHAT, IVA 2016, Shared Task Report.
10. D'Haro, L.F., Lue, L. (2016) An Online Platform for Crowd-sourcing Data from Interactions with Chatbots, in Proceedings of WOCHAT, IVA 2016, Shared Task Report.
11. Gordon, C., Tin, J., Brown, J., Fritzsch, E., Gabber, S. (2016) Wochat Chatbot User Experience Summary, in Proceedings of WOCHAT, IVA 2016, Shared Task Report.
12. Yu, Z., Xu, Z., Black A.W., Rudnicky A.I. (2016) TickTock, in Proceedings of RE-WOCHAT, LREC 2016, Shared Task Report.