Shared Task on Data Collection and Annotation
RE-WOCHAT 2016 – SHARED TASK DESCRIPTION REPORT

Luis F. D'Haro
Human Language Technology,
Institute for Infocomm Research
luisdhe@i2r.a-star.edu.sg

Bayan Abu Shawar
IT department; School of Computing,
Arab Open University
bshawar@yahoo.com

Zhou Yu
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, 15217
zhouyu@cs.cmu.edu

Abstract

This report presents and describes the shared task on "Data Collection and Annotation" conducted with RE-WOCHAT, the first Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents - Development and Evaluation. We describe the main road map envisaged for this and future shared tasks, as well as the proposed collection and annotation schemes. We also summarize the result of the shared task in terms of chatbot platforms made available for it and the amount of collected chatting sessions and annotations.

Keywords: shared task, chat-oriented dialogue, data collection, manual annotation

1. Introduction

As part of the activities of the workshop, RE-WOCHAT[1] (Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents - Development and Evaluation) has accommodated a shared task on "Data Collection and Annotation". The main objective of this shared task is to develop and test a new evaluation framework for non-goal-oriented dialogue engines.

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

2. Shared Task Background

Different from task-oriented dialogue, automatic evaluation of chat-oriented dialogue poses some interesting challenges to due to the specific nature and lack of specific goals in it. Different approaches have been proposed to this end, including time of engagement and user satisfaction (Abu Shawar and Atwell, 2007), Dialogue Coherence Models (Gandhe and Traum, 2008), and comparative evaluations (Banchs and Kim, 2014).

Although data driven approaches have provided a useful means for training evaluation metrics in many other areas of research, one of the main problems related to the development of similar strategies for chat-oriented dialogue is for certain the lack of enough annotated data. In this sense, the main motivation for a shared task on "Data Collection and Annotation" is to provide an experimental platform for the research community to generate data and resources for chat-oriented dialogue research. This must be achieved by a collaborative effort continued over time and expanded to multiple languages and modalities.

3. Main Objectives and Road Map

This shared task is part of a larger scope initiative, which main objectives are (1) collecting chat-oriented dialogue data that can be made available for research purposes and (2) developing a framework for the automatic evaluation of chat-oriented dialogue.

This effort comprises three interdependent tasks:

- **Task 1. Chat data collection**: participating teams will produce dialogues between human users and chat engines, as well as between humans only.
- **Task 2. Subjective evaluation**: participating teams will manually evaluate a selection of the generated dialogues according to different subjective evaluation metrics.
- **Task 3. Subrogated metrics**: participant teams will attempt to model the manually generated subjective evaluation metrics by using machine learning techniques

The current edition of the Shared Task in RE-WOCHAT has focused only on Tasks 1 and 2 described above. Task 3 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.

Four different ways of participation in the shared tasks were defined:

- **Chatbot provider**. Participants owning a chatbot engine and willing to provide access to it either by distributing a standalone version of it or via a web service or web interface.
- **Data generator.** Participants willing to use one or more of the provided chatbots to generate dialogue sessions with it.

[1] http://workshop.colips.org/re-wochat/shared.html

- **Data provider.** Participants owning or having access to a chatbot that are not accessible to the general public willing to generate chatting sessions and share the generated data with other participants.
- **Data annotator.** Participants willing to annotate some of the generated and/or shared dialogue sessions by following the provided annotation guidelines.

A total of 14 volunteers registered for participating in the first edition of the shared task. These 14 volunteers accounted for a total of six chatbot providers, seven data generators, three data providers and eight data annotators.

4. Chatbot Platforms made Available

The six chatbot engines made available for the shared task include:

- 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 (Dubuisson et. al, 2016b).
- IRIS. Informal Response Interactive System, which implements a chat-oriented dialogue system based on the vector space model framework (Banchs and Li, 2016).
- Py-Eliza. A Python-based stand-alone version of the famous Eliza chatbot created by Weizen- baum in 1966 (D'Haro, 2016).
- Sarah. An upgraded version of Alice bot, developed by Dr.Wallace in 1995 (AbuShawar, 2016)
- TickTock. A chatbot with a goal to engage users in an everyday conversation. It is a keyword based retrieval system with engagement conversational strategies (Yu et al, 2016).
- Politician. A question-answering system, which is designed as a chatbot imitating a politician. It answers questions on political issues. (Kuboň and Hladká, 2016).

For more detailed information about each one of these chatbots, the reader can refer to the Shared Task Chatbot Description Reports in the Workshop Proceedings.

Most of these chatbots are available via online interfaces or as standalone systems for collecting chatting interactions with registered participants. 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.

The following tips have been provided to the shared task participants to be taken into account during the data generation phase:

- Use the same nickname when interacting with the different chatbots. As chatting sessions are anonymous, this will be the only way to track all different chatting sessions for the same user.
- Remember these are just chatbots, do not expect too much from them. Please try to converse as much as you can and in the most natural way.
- Generate as much chatting sessions as you can. Ideally, a chatting session should include more than 20 turns but no more than 50 turns.

5. Data Collection and Annotations

To the date this report was written, a total of 554 chatting sessions had been collected since the beginning of the Shared Task. In addition, a total of 61 contributed dialogs were received, out of which 41 dialogs were contributed by the Joker system[2] (Dubuisson et al, 2016a). All these provided chatbot systems are still available for the participants to continue interacting with them, so the data collection is still ongoing. An updated report will be presented the day of the RE-WOCHAT workshop and will be made available at the workshop official website.[3]

Table 1 and Figure 1 show some of the statistics of the dialogs with all the chatbots. Surprisingly chatbots use more vocabulary and longer sentences than humans. This may be because they want to keep engaged humans or because humans tend to be more specific to keep the dialog focused and easier for the chatbot to understand.

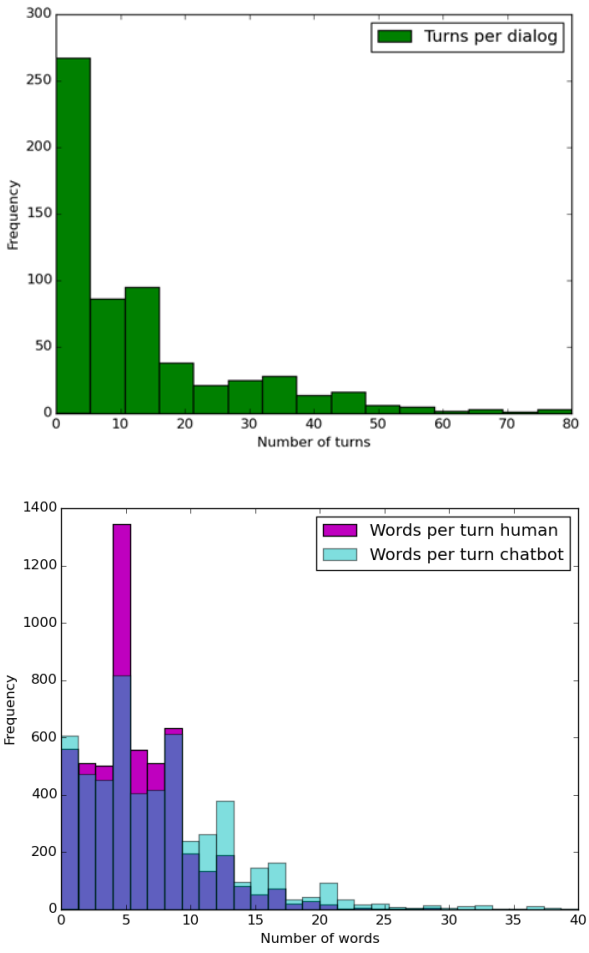

Figure 1. a) Histograms for number of turns per dialog and b) words per turn comparing human vs chatbot turns

[2] https://ucar.limsi.fr
[3] http://workshop.colips.org/re-wochat/index.html

	Chatbot	Human	Total
No. dialogs	-	-	615
No. Turns	-	-	8589
Vocab. Size	4445	4088	
Av. no. Words per sentences	7.66 ±2.64	5.84 ± 0.40	
Polarity	0.08±0.25	0.07±0.28	
Subjectivity	0.22±0.31	0.23±0.33	

Table 1. Basic statistics of the collected dialogue sessions.

On the other hand, we also provide polarity and subjectivity calculated using TextBlob[4], where polarity is within the range [-1.0, 1.0] and subjectivity is a within the range [0.0, 1.0], being 0.0 a very objective sentence and 1.0 a very subjective sentence. Here, we cannot see a dominant trend on either the chatbots or humans, but with a very small difference toward chatbots being more subjective and positive.

A total of 126 of the collected catting sessions have been manually annotated by human evaluators according to the proposed subjective evaluation guidelines (see Table 2). These guide- lines comprise the assignment of one (out of three) basic subjective scores to each turn in a chatting session. The possible valid tags are: VALID, ACCEPTABLE and INVALID, meaning:

- VALID: this score is used to access a response that is semantically and pragmatically valid given the previous utterance as well as the previous recent dialogue context. Some examples of VALID responses to the utterance "how old are you?" include: "I am 25", "older than you" and "I am quite young".
- ACCEPTABLE: this score is used to access a response that is not necessarily semantically valid but can be acceptable, given the previous recent dialogue context, from the pragmatic point of view. Some examples of ACCEPTABLE responses to the utterance "how old are you?" include: "let us better talk about food", "how old are you?" and "what did you say before?"
- INVALID: this score is used to access a response that is definitively invalid given the previous utterance and the recent dialogue context. Some examples of INVALID responses to the utterance "how old are you?" include: "he goes to the supermarket every Saturday" or "I like pizza".

In additions to the three subjective scores described above, annotators were also requested to evaluate the polarity and offensiveness of the utterances, in those cases in which this was possible. These optional annotations were used according to the following conventions:

- POSITIVE: this tag might be used to indicate positive polarity of the response.
- NEGATIVE: this tag might be used to indicate negative polarity of the response.
- OFFENSIVE: this tag might be used to indicate inappropriate offensive response, which does not necessarily contain swear words.
- SWEARLANG: this tag might be used to indicate the explicit presence of inappropriate language, regardless whether it is offensive or not.

In a similar way to the data collection task, annotations are scheduled to continue over time. An updated report will be presented the day of the RE-WOCHAT workshop and will be made available at the workshop website.

Metric		Chatbot	Human	Total
No. of evaluated dialogs		126		
No. of evaluated turns		2723		
Subjective scores	Valid	777	1432	2209
	Acceptable	534	315	849
	Invalid	600	65	665
	Kappa	*0.567*		
Optional annotations	Positive	34	31	65
	Negative	59	57	116
	Offensive	50	32	82
	Swearlang	31	2	33

Table 2. Basic statistics of the annotated dialogue sessions.

6. Conclusions and Future Work

This report has presented and described the shared task on "Data Collection and Annotation" conducted with RE-WOCHAT, the first Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents - Development and Evaluation. We described the main road map envisaged for this and future shared tasks, as well as the proposed collection and annotation schemes used in the shared task activities. We also summarized the result of the shared task in terms of chatbot platforms made available for it and the amount of collected chatting sessions and annotations.

As future work for preparing next editions of the shared task, we plan to consolidate a data collection and annotation platform for chat-oriented dialogue, by centralizing the different available chatbots into the same interface. Additionally we plan to evaluate different gamification strategies to encourage more people to participate and contribute to the shared task activities.

7. Acknowledgements

We want to thank the workshop organizers: Rafael E. Banchs, Ryuichiro Higashinaka, Wolfgang Minker, Joseph Mariani and David Traum for their assistance and support during the organization of the shared task.

Similarly, we also want to thank all the volunteers who contributed to the shared task activities: Andreea Niculescu, Emer Gilmartin, Guillaume Dubuisson Duplessis, Kheng Hui Yeo, Natalia Kocyba, Rafael Banchs, Ryuichiro Higashinaka, Soe Gon Yee Thant, Sophie Rosset, Vincent Letard, and Vlad Maraev.

[4] https://textblob.readthedocs.org/en/dev/

8. References

AbuShawar, B., Atwell, E. (2007). Different measurement metrics to evaluate a chatbot system, in Proceedings of Workshop on "Bridging the Gap: Academic and Industrial Research in Dialogue Technologies". pp. 89—96, NAACL-HLT

AbuShawar, B. (2016) Sarah Chatbot, in Proceedings of RE-WOCHAT, LREC 2016, Shared Task Report.

Banchs, R., Kim, S. (2014). An empirical evaluation of an IR-based strategy for chat-oriented dialogue systems, in Proccedings of APSIPA, Special Session on Chatbots and Conversational Agents

Banchs, R., Li, H. (2016) IRIS – Informal Response Interactive System, in Proceedings of RE-WOCHAT, LREC 2016, Shared Task Report.

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.

Dubuisson Duplessis, G. and Letard, V. and Ligozat, A.-L. and Rosset, S. (2016a) Purely Corpus-based Automatic Conversation Authoring, in Proceedings 10th International Conference on Language Resources and Evaluation (LREC), May 2016. 8 p.

Dubuisson Duplessis, G., Letard, V., Ligozat, A.L., Rosset, S. (2016b) Joker Chatterbot, in Proceedings of RE- WOCHAT, LREC 2016, Shared Task Report.

Gandhe, S., Traum, D. (2008). An evaluation understudy for dialogue coherence models, in Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, pp. 172—181, ACL.

Kuboň, D., Hladká, B. (2016). Politician, in Proceedings of RE-WOCHAT, LREC 2016, Shared Task Report.

Yu, Z., Xu, Z., Black A.W., Rudnicky A.I. (2016) Tick Tock, in Proceedings of RE-WOCHAT, LREC 2016, Shared Task Report.