@ IWSDS 2018

Third Workshop on Chatbots and Conversational Agent Technologies

WOCHAT Shared Task Update

Luis F. D'Haro, Rafael E. Banchs

Singapore, May 15, 2018

Collocated with International Workshop on Spoken Dialogue Systems (IWSDS) 2018

Shared Task Objectives

- **To collect chat-oriented dialogue data that can be made available for research purposes**
  - Human-chatbot and human-human dialogue sessions
  - Covering a variety of (1) chatbot technologies and approaches, and (2) languages and cultural backgrounds
- **To develop a framework for the automatic evaluation of chat-oriented dialogue systems**
  - Subjective evaluation of chatting-sessions at the turn level
  - Crowdsource multiple annotations for the same utterance
  - Apply ML approaches aiming at reproducing annotations (human subjective evaluations)

Shared Task Activities

- **Task 1: Chat Data Collection**
  - Generation of human-chatbot dialogue sessions
  - Satisfaction from users for each dialogue session
- **Task 2: Subjective Evaluation at Turn Level**
  - Manually evaluate a selection of the generated dialogues according to subjective evaluation metrics and guidelines
  - Multiple evaluations are collected by crowdsourcing
- **Task 3: Subrogated Metrics**
  - Participants attempt to model the manually generated subjective evaluation metrics by using machine learning techniques

Shared Task Participation Roles

- **Chatbot provider**
  - The participant owns a chatbot engine and wants to provide access to it either by distributing a standalone version of it or by giving access to it via a webservice or interface.
- **Data generator**
  - The participant is willing to use one or more of the provided chatbots to generate dialogue sessions.
- **Data provider**
  - The participant owns or has access to a chatbot but cannot provide access to it. However, she/he can generate dialogue sessions with it and can provide the generated datasets.
- **Data annotator**
  - The participant is willing to annotate some of the generated and provided dialogue sessions

Available Chatbots

- **ALANA:** proactive and well-informed social bot who knows about recent news and general world knowledge
- **CHATBOL:** a Spanish conversational agent for providing information about the Spanish Football League "La Liga"
- **JOKER**: an example-based system that uses a database of semantically indexed dialogue examples to manage dialogue.
- **IRIS:** (Informal Response Interactive System): a chat-oriented dialogue system based on the vector space model framework.
- **pyEliza**: a Python-based stand-alone version of the famous Eliza chatbot created by Weizenbaum in 1966.
- **SARAH**: a version of Alice bot, developed by Dr. Wallace in 1995. It is based on the AIML framework and it is accessible through the pandorabots platform.
- **TickTock**: a chatbot with a goal to engage users in a everyday conversation. It is a based retrieval system with engagement conversational strategies.
- **SAMMY**: is a chatbot based on the public "small-talk" domain available at dialogflow.com. She is conversant in either English, French or Italian.

Annotation Guidelines

Appropriateness score:

- **VALID**: the response is semantically and pragmatically valid given the previous utterance and recent dialogue context.
  - Examples of VALID responses to "how old are you?" would be:
  - "I am 25", "older than you", "I am quite young", etc.
- **ACCEPTABLE**: the response is not necessarily semantically valid but can be acceptable from a pragmatic point of view.
  - Examples of ACCEPTABLE responses to "how old are you?" would be:
  - "let us better talk about food", "how old are you?", "what did you say before?", etc.
- **INVALID**: the response is definitively invalid given the previous utterance and the recent dialogue context.
  - Examples of INVALID responses to "how old are you?" would be:
  - "he goes to the supermarket every Saturday", "I do not like pizza", "you seem to be running out of money", etc.

Annotation Guidelines (continuation)

Additional tags (optional):

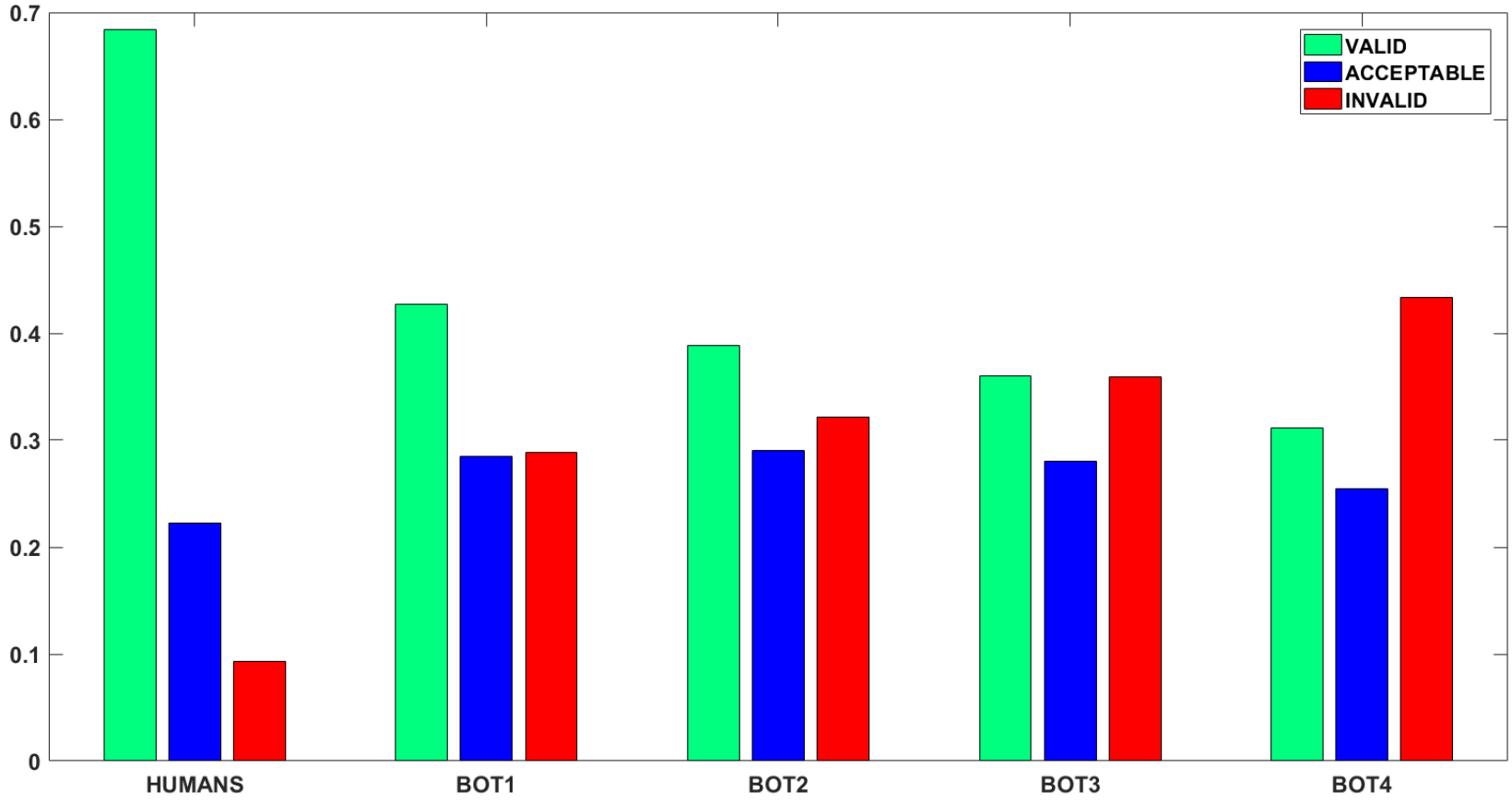
- **POSITIVE:** to indicate positive polarity of the response.
- **NEGATIVE:** to indicate negative polarity of the response.
- **OFFENSIVE:** to indicate inappropriate offensive response, which does not necessarily contain swear words.
- **SWEARLANG:** to indicate the explicit presence of inappropriate language in the given turn, regardless whether it is offensive or not.
- **ISMACHINE:** this tag might be used for assessing the annotator impression on whether the utterance has been produced by a chatbot (if the identities of the interlocutors are hidden to the annotators)

Shared Task Data and Annotations

- Data collected and annotated so far:
  - Over 1,000 dialogue sessions
  - Comprising about 30,000 turns
  - With around 15,000 turn level annotations

- Recent activities:
  - Research work on models for score prediction
  - Dialogue Breakdown Challenge @ DSTC6

Appropriateness Score Distributions



Next Steps…

- Continue promoting the shared task activities….
  - More data generators and providers are needed
  - More data annotators are needed
- Improve the current chatbot ecosystem
  - Developing APIs for better connectivity with the chatbots
  - Integrating into centralized platforms (Webchat, Dialport)
- Future workshop editions and other events
  - Appropriateness Score Prediction Task @ WOCHAT/DBD
  - JSALT Summer Workshop proposal
  - Next WOCHAT Workshop and/or Special Session