Annotation Guidelines for WOCHAT Shared Task

Shared Task Description

As part of WOCHAT program activities, the Shared Task on Data Collection and Annotation will continue. In this task, participants are required to generate human-machine and human-human dialogues, as well as to produce turn-level annotations on them. Human-machine dialogues are generated by using online and offline chat engines made available for such purposes. Annotations are generated following these guidelines. The collected dataset will be made publicly available to the research community for further research and experimentation in future editions of the workshop.

Metrics and Resources for Chat-oriented Dialogue Evaluation

The Shared Task in WOCHAT is part of a larger scope initiative aiming at both collecting chat-oriented dialogue data that can be made available for research purposes and 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.

Similar to RE-WOCHAT workshop, the Shared Task in WOCHAT still focuses on Tasks 1 and 2 only. 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.

Chat Data Collection: XML File Formatting

Generated chat-session logs should follow the proposed XML format, which is described below:

```
<dialogue id="[UNIQUE_DIALOGUE_IDENTIFIER]">
    <system_name>[NAME_OF_THE_CHATBOT]</system_name>
    <user_name>[NAME_OR_NICKNAME_OF_THE_USER]</user_name>
    <timestamp>[TIMESTAMP_OF_SESSION_STARTING_TIME]</timestamp>
    <turn id="1">
        <speaker>[SYSTEM|USER]</speaker>
        <utterance>[SYSTEM|USER_UTTERANCE_CONTENT]</utterance>
    </turn>
    <turn id="2">
        <speaker>[USER|SYSTEM]</speaker>
        <utterance>[USER|SYSTEM_UTTERANCE_CONTENT]</utterance>
    </turn>
    ...
</dialogue>
```

Important notes related to chat-session log file formatting:

- The UNIQUE_DIALOGUE_IDENTIFIER is a string (chatbot name or 'human') followed by an integer sequentially assigned from 00000 to 99999 based on the time stamp of the session.
- For human-human dialogues, <system_name> and <user_name> tags should be replaced by <user1_name> and <user2_name>. Similarly, for multiparty chat sessions any arbitrary number of <user[#]_name> tags might be added.
- The valid entries for the <speaker> field in a dialogue session are restricted to the same ones defined in the <[value]_name> tags. For instance, for a dialogue session including <system_name> and <user_name> tags, the only valid entries for <speaker> are SYSTEM and USER.
- Within the utterance contents, any occurrence of speaker names (as defined within the <[value]_name> tags) must be replaced by its corresponding tag. For instance, given the definition <system_name>IRIS</system_name>, the utterance "about you Iris, you are very funny indeed" must be reported as <utterance>about you SYSTEM_NAME, you are very funny indeed</utterance>.
- Additional optional tags can be included inside any <turn> definition such as, for instance, individual turn's timestamps, dialogue context information or metadata, and evaluation-related annotations. Evaluation-related annotations are described in detail in the following section.

Subjective Evaluation: XML Annotation Formatting

The proposed subjective evaluation comprises the assignment of one (out of three) basic subjective score to each turn in a chatting session. The three possible valid tags for the subjective scores are: </VALID>, </ACCEPTABLE> and </INVALID>. These three subjective scores should be used according to the following conventions:

- </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 <utterance>how old are you?</utterance> include, for instance, <utterance>I am 25</utterance>, <utterance>older than you, for sure</utterance> and <utterance>I am quite young</utterance>.
- </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 <utterance>how old are you?</utterance> include, for instance, <utterance>let us better talk about food</utterance>, <utterance>how old are you?</utterance> and <utterance>what did you say before?</utterance>.
- </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 <utterance>how old are you?</utterance> include, for instance, <utterance>he goes to the supermarket every Saturday</utterance>, <utterance>I do not like pizza</utterance> and <utterance>you seem to be running out of money</utterance>.

Additionally, some other optional evaluation annotations can also be included, such as:

- `</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 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 (only if the identities of the speakers are hidden to the annotators)

Subjective evaluations tags are to be included in each `turn` of a chat-session log by using the following format:

```
<turn id="[TURN_NUMBER]">
    <speaker>[SYSTEM|USER]</speaker>
    <utterance>[SYSTEM|USER_UTTERANCE_CONTENT]</utterance>
    <annotator id="[UNIQUE_ANNOTATOR_IDENTIFIER]">
        </[VALID|ACCEPTABLE|INVALID]> [OTHER_OPTIONAL_TAGS]
    </annotator>
</turn>
```

where `UNIQUE_ANNOTATOR_IDENTIFIER` is a unique identifier to be assigned to each of the data annotators participating in the shared task.

For more information visit the workshop and shared task website at: http://workshop.colips.org/wochat/index.html

Shared Task Co-organizers

- Bayan Abu Shawar, Arab Open University, Jordan
- Luis Fernando D'Haro, Agency for Science, Technology and Research, Singapore
- Zhou Yu, Carnegie Mellon University, USA

Workshop Organizers

- Rafael E. Banchs, Institute for Infocomm Research, Singapore
- Ryuichiro Higashinaka, Nippon Telegraph and Telephone Corporation, Japan
- Wolfgang Minker, Ulm University, Germany
- Joseph Mariani, IMMI & LIMSI-CNRS, France
- David Traum, University of Southern California, USA