iLISTEN at EVALITA 2018
Task Guidelines

Nicole Novielli, Pierpaolo Basile
University of Bari Aldo Moro, Italy
{nicole.novielli,pierpaolo.basile}@uniba.it

August 22, 2018

Contents

1 Task Description

iLISTEN (itaLIan Speech acT labEliNg) is the first task on speech act labeling of Italian dialogues and it is open to everyone from industry and academia.

The task consists in **automatically annotating dialogue turns with *speech act labels***, i.e. with the communicative intention of the speaker, such as statement, request for information, agreement, opinion expression, general answer, etc.

Table 1 reports the full set of speech act labels used for the evaluation, with definition and examples.

Table 1: The set of speech act labels employed for the User moves.

Speech Act	Description	Example
`OPENING`	Dialogue opening or self-introduction	*'Ciao, io sono Antonella'*
`CLOSING`	Dialogue closing, e.g. farewell, wishes, intention to close the conversation	*'Va bene, ci vediamo prossimamente'*
`INFO-REQUEST`	Utterances that are pragmatically, semantically, and syntactically questions	*'E cosa mi dici delle vitamine?'*
`SOLICIT-REQ-CLARIF`	Request for clarification (please explain) or solicitation of system reaction	*'Mmm, si ma in che senso?'*
`STATEMENT`	Descriptive, narrative, personal statements	*'Penso che dovrei controllare maggiormente il consumo di dolciumi.'*
`GENERIC-ANSWER`	Generic answer	*'Si'*, *'No'*, *'Non so.'*
`AGREE-ACCEPT`	Expression of agreement, e.g. acceptance of a proposal, plan or opinion	*'Si, so che è importante.'*
`REJECT`	Expression of disagreement, e.g. rejection of a proposal, plan, or opinion	*'Ho sentito tesi contrastanti al proposito.'*
`KIND-ATT-SMALLTALK`	Expression of kind attitude through politeness, e.g. thanking, apologizing or smalltalk	*'Thank you.'*, *'Sei per caso offesa per qualcosa che ho detto?'*

2 Development and Test Data

2.1 A Dataset of Dialogues

The development and test datasets are extracted from a corpus of natural language dialogues collected in the scope of previous research about Human-ECA interaction [1]. The corpus contains overall transcripts of 60 dialogues, 1,576 user dialogue turns, 1,611 system turns and about 22,000 words.

The dialogues were collected using a Wizard of Oz tool as dialogue manager. Sixty subjects (aged between 21–28) involved in the study, in two interaction mode conditions: thirty of them interacted with the system in a written-input setting, using keyboard and mouse; the remaining thirty dialogues were collected with users interacting with the ECA in a spoken-input condition.

During the interaction, the ECA played the role of an artificial therapist and the users were free to interact with it in natural language, without any particular constraint: they could

simply answer the question of the agent or taking the initiative and ask questions in their turn, make comments about the agent behavior or competence, argument in favor or against the agent's suggestion or persuasion attempts. The Wizard, on his behalf, had to choose among a set of about 80 predefined possible move. As such, the *the pre-defined system moves are provided only as a context information but will not be subject to evaluation and will not contribute to the final ranking of the participant systems.* Conversely, systems will be evaluated on the basis of the performance observed for the user dialogue turns.

Dialogues collected using the spoken interaction mode were manually transcribed based on audio-recording of the dialogue sessions. Information about interaction mode (spoken vs. written) are encoded in the **id** of each dialogue turn where 'T' and 'S' indicate text- and speech-based interaction, respectively.

In Table 2 we provide an excerpt from a dialogue from our gold standard. The system moves (dialogue moves and corresponding speech act labels) will be only as a context information. Conversely, the final ranking of the participating systems will be performed based on the performance observed only on the users' move, with respect to the set of users' label provided in Table 1. Please, note that the set of agent's moves will include also speech acts (such as persuasion attempts) that are observed only for the agent, given its caregiver role in the dialogue systems.

Table 2: An excerpt from a text-based dialogue from our dataset. The 'T_' in the id indicate the text-based interaction mode.

Actor	**ID**	**Speech Act**	**Dialogue Move**
SYSTEM	T_5-S1	OPENING	Ciao, il mio nome è Valentina. Sono qui per darti dei suggerimenti su come migliorare la tua dieta
USER	T_5-U1	OPENING	Ciao, il mio nome è Oriana.
SYSTEM	T_5-S2	INFO-REQUEST	Quali sono le tue abitudini alimentari?
USER	T_5-U2	STATEMENT	Ho delle abitudini disordinate, mangio anche fuori orario.
SYSTEM	T_5-S3	INFO-REQUEST	Ti piace mangiare?
USER	T_5-U3	STATEMENT	Si, ma dipende dal periodo: se sono sotto stress mangio meno, quando sono più rilassata mangio molto
SYSTEM	T_5-S4	INFO-REQUEST	Segui una dieta variata?
USER	T_5-U3	GENERIC-ANSWER	No.

2.2 Data Format

We provide both the training and testing dialogues in the XML format following the structure proposed in 1.

Each participating team will initially have access to the training data only. Later, the unlabeled test data will be released.

3 Submission Format

The participants must provide results in a plain text file with comma-separated fields. **Only the dialogue turns of the *User***, marked as U in the id (as in T_5_U4) will be subject

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<dialogues>
  <dialogue>
    <speechAct act="OPENING" actId="13" id="T_5_S1">Ciao, il mio nome e' Valentina. Sono
qui per darti dei suggerimenti su come migliorare la tua dieta.</speechAct>
    <speechAct act="OPENING" actId="10" id="T_5_U1">ciao il mio nome è oriana</speechAct>
    <speechAct act="QUESTION" actId="12" id="T_5_S2">Quali sono le tue abitudini
alimentari?</speechAct>
    <speechAct act="STAT-ABOUT-SELF" actId="5" id="T_5_U2">ho delle abitudini un pò
disordinate,nel senso che mangio anche fupri orario</speechAct>
    <speechAct act="QUESTION" actId="12" id="T_5_S3">Ti piace mangiare?</speechAct>
    <speechAct act="STAT-ABOUT-SELF" actId="5" id="T_5_U2">si,ma dipende dal periodo,se
sono sotto stress mangio meno,quando sono più rilassata mangio molto</speechAct>
    <speechAct act="QUESTION" actId="12" id="T_5_S4">Segui una dieta variata?</speechAct>
    <speechAct act="GENERIC-ANSWER" actId="7" id="T_5_U4">no</speechAct>
    ...
```

Figure 1: Data format

to evaluation and should be returned. In table 2 following, we report an example of a what a submitted run should look like. Please, note that the id in the first column (in **bold**) should be the same provided for each User dialogue turn in the test set, while the speech act label in the second column (in *italic*) is the prediction of your system.

```
id,act
T_5_U1,OPENING
T_5_U2,STAT-ABOUT-SELF
T_5_U4,GENERIC-ANSWER
...
```

Figure 2: Format of submission run.

To encourage participants to experiment novel approaches as well as more traditional ones, we allow two submissions per team. However, we strongly recommend participants to submit two runs only if they implement **substantially different approaches**. If you rather want to fine-tune the performance of your system by varying the features included in your classifier, we suggest submitting only one run and to describe the feature engineering activity in the final report.

4 Evaluation

Regarding the evaluation procedure, we will assess the ability of each system to issue the correct speech act label for the user moves. The speech act label are those included in the taxonomy used for annotation of the user move and reported in 1.

Specifically, we will compute precision, recall and F1-score (macroaveraging) with respect to our gold standard. This approach, while more verbose than a simple accuracy test, arise from the need to correctly address the unbalance distribution of labels in the dataset. Furthermore, by providing detailed performance metrics, we aim at enhancing interesting conclusions on

the nature of the problem and the data, as they might emerge from the participants' final reports.

As a baseline, we will use the most frequent label for the user speech acts.

5 How to submit your runs

The test data will be distributed on **September 17th, 2018**. Once you have run your system over the test data, you will have to send your predictions to us following these recommendations:

- choose a team name and name the file containing your run in the following way:
  $ilisten2018.teamName.systemID.csv$
- send the file to the email address: ilisten.evalita2018@gmail.com using the subject:
  $ilisten - teamName$
- in the body of the email please include a short description of the system by specifying the methodology and all the resources used in building your system

References

[1] G. Clarizio, I. Mazzotta, N. Novielli, and F. De Rosis. Social attitude towards a conversational character. pages 2–7, 2006.