The LUNA Corpus: an Annotation Scheme for a Multi-domain Multi-lingual Dialogue Corpus

Christian Raymond♣, Giuseppe Riccardi♣, Kepa Joseba Rodríguez♦, Joanna Wiśniewska♠

Department of Information and Communication University of Trento ♣

Piedmont Consortium for Information Systems (CSI-Piemonte) ♦

Institute of Computer Science Polish Academy of Science ♠

LUNA

The LUNA project

The focus of the LUNA project is the real time understanding of spontaneous speech in dialogue systems.
Thee steps will be considered for the Spoken Language Understanding:

1. **Generation of semantic concepts tags.**
2. **Semantic composition into conceptual structures.**
3. **Context sensitive validation using information provided by the dialogue manager.**

The SLU models will be trained and evaluated on the LUNA corpus and applied to different multilingual conversational systems.

The LUNA corpus

The corpus is being collected with the target to collect 3000 human-human and 8100 human-machine dialogues in Italian, French and Polish.
Dialogues will be collected in following application domains: travel information and reservation, public transportation information, IT help desk, telecom costumer care and financial information transaction.

Dialogue example

[Operator:] allora m'ha detto che non riusciva ad accedere al computer e le manca la procedura
so, you have told me that you cannot access to the computer, and that you need the procedure
[Caller:] esatto
exactly
[Operator:] allora avrei bisogno dell' RWS del PC
so I need the RWS of the computer
[Caller:] si allora tredici zero ottantasei
yes, 13 0 86

Levels of annotation

Domain attribute level

- **Attribute-value pairs representation**
- **Tagset of attribute-value specified by domain ontologies**

Predicate structure

The corpus is annotated using a FrameNet-like approach. Based on domain knowledge we define a set of frames for each domain.

Coreference

Different kinds of anaphoric relations like:

- **Identity**
- **bridging : exploiting the relations and properties of the domain ontologies.**
- **set-element**

The annotation scheme allows to have more that a unique interpretation of the coreference.

Dialogue acts

- **Initial tageset: 9 selected dialogue acts from the DAMSL scheme. Extensible for the different application domains.**
- **The utterances are defined based on the predicate structure and annotated with as many tags as possible.**

The annotation on this level will be used to build prototypes in the different application domains.

Transcription

Operator: allora m'ha detto che non riusciva ad accedere al computer *[silence]* **e le manca la procedura** *[pron=unintelligible]*
Caller: esatto
Operator: allora avrei bisogno dell' *[lex=filler]* *[pron=spelled-]* **RWS** *[-pron=spelled]* **del** *[pron=spelled-]* **PC** *[-pron=spelled]*
Caller: si allora tredici zero ottantasei

Domain attribute

[Operator:] allora m'ha detto che [non riusciva] ad [accedere] [al computer] e [le manca] [la procedura]

```
trouble : unable_to
action : access
computer-hardware : pc
trouble : lack_of
computer-software : procedure
```

Caller: esatto
Operator: allora avrei bisogno [dell' RWS] [del PC]

```
concept : code-identificationCode
computer-hardware : pc
```

Caller: si allora [tredici zero ottantasei]

```
code-identificationCode-rws : 13086
```

Predicate structure

Operator: allora m'ha detto che [non riusciva]fe1 ad [accedere]fe2 [al computer]fe3 e le [manca]fe4 [la procedura]fe5

```
frame : access
frame-elements : {user, hardware}
      fe id:fe1 f-element: negation
      fe id:fe2 f-element: target
      fe id:fe3 f-element: hardware
frame : need
frame-elements : {user, requirement}
      fe id:fe4 f-element: target
      fe id:fe5 f-element: requirement
```

Coreference

[Operator:] allora m'ha detto che non riusciva ad accedere [al computer]c1 e le manca [la procedura]c2

```
Coref id:c1 info-status:given …
Coref id:c2 info-status:given …
```

[Caller:] esatto
[Operator:] allora avrei bisogno [dell' RWS]c3 [del PC]c4

```
coref id:c3, info-status: new, related:yes,
      related-phrase:c1,  relation: rwsOf
coref id:c4, inf_status: given,
      single-phrase-atecedent:c1
```

[Caller:] si allora [tredici zero ottantasei]c5

```
coref id:5, info-status: new, related:yes
      related-phrase: c3 relation: instanceOf
```

Dialogs act

[Operator:] allora m'ha detto che [non riusciva ad accedere al computer]u1 e [le manca la procedura]u2

```
u1 + u2: statement, info-request
```

[Caller:] [esatto]u3

```
u3: Statement, answer
```

[Operator:] [allora avrei bisogno dell' RWS del PC]u4

```
u4: statement, info-request
```

[Caller:] [si]u5 [allora tredici zero ottantasei]u6

```
u5: accept
u6: statement, answer
```