Decalog 2007: Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue, pages 185–186.
Trento, Italy, 30 May – 1 June 2007. Edited by Ron Artstein and Laure Vieu.

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

Christian Raymond◇, Giuseppe Riccardi◇, Kepa Joseba Rodríguez♣, Joanna Wisniewska♠
◇Department of Information and Communication. University of Trento.
{christian.raymond|riccardi}@dit.unitn.it
♣Piedmont Consortium for Information Systems (CSI-Piemonte)
KepaJoseba.Rodriguez@csi.it
♠Institute of Computer Science. Polish Academy of Science
jwisniewska@poczta.uw.edu.pl

Abstract

The LUNA corpus is a multi-domain multi-lingual dialogue corpus currently under development. The corpus will be annotated at multiple levels to include annotations of syntactic, semantic and discourse information and used to develop a robust natural spoken language understanding toolkit for multilingual dialogue services[1].

1 Introduction

LUNA is a project focused on the problem of real-time understanding of spontaneous speech in context of next generation dialogue systems[2].

Three steps will be considered for the Spoken Language Understanding (SLU) interpretation process: generation of semantic concept tags, semantic composition into conceptual structures and 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 in Italian, French and Polish.

The corpus is currently being collected with a target to collect 1000 human-human and 8100 human-machine dialogues in Italian, Polish and French. The dialogues will be collected in the following application domains: travel information and reservation, public transportation information, IT help desk, telecom customer care and financial information and transaction.

2 Segmentation and Transcription

The first step is the segmentation of the speech signal into dialogue turns. The turns will be annotated with time information, speaker identity and gender, and marked where speaker overlap occurs.

The next step is the transcription of the speech signal, using conventions for the orthographic transcription and for the annotation of non-linguistic acoustic events.

3 Multi-level annotation

Semantic interpretation involves several aspects, like the meaning of tokens referred to a domain or the relation between different semantic objects in the utterance and discourse level. In order to capture these different aspects we decided to implement a multi-dimensional annotation scheme. The annotation of some levels is mandatory for all the dialogues of the corpus. The annotation of the other levels is recommended.

The first levels of the annotation are related to the preparation of the corpus for the semantic annotation, and include segmentation of the speech signal in dialogue turns, transcription and syntactic pre-processing with Part of Speech (POS) tagging and shallow parsing.

The next level consist of the annotation of domain information using attribute value pairs. The annotation of this level is mandatory, as the annotation of the other levels depends on it.

The other levels of the annotation are the predicate structure, coreference and anaphoric relations and dialogue acts.

[1] This research was performed under LUNA project funded by the EC, DG Infso, Unit E1.

[2] The members of the consortium are: Piedmont Consortium for Information Systems (IT), University of Trento (IT), Loquendo SpA (IT), RWTH-Aachen (DE), University of Avignon (FR), France Telecom R&D Division S.A. (FR), Polish-Japanese Institute of Information Technology (PL) and the Institute for Computer Science of the Polish Academy of Sciences (PL). http://www.ist-luna.eu

4 POS-tagging and Chunking

The transcribed material will be annotated with POS tags, morphosyntaectic information and segmented based on syntactic constituency. For the POS-tags and morphosyntactic features, we will follow the recommendations made in EAGLES (EAGLES, 1996), which allows us to have a unified representation format for the corpus, independently of the tools used for each language.

5 Domain attribute level

Semantic segments are produced by concatenation of the semantic chunks. A semantic segment is a unit that corresponds unambiguously to a concept of the dictionary described bellow.

Semantic segments are annotated with attribute-value pairs following an approach similar to the used for the annotation of the French MEDIA corpus (Bonneau-Maynard and Rosset, 2003). We specify domain knowledge in domain ontologies that are used to build domain-specific concept dictionaries. Each dictionary contains:

- Concepts corresponding to classes of the ontology and attributes of the annotation.
- Values corresponding to the individuals of the domain.
- Constraints on the admissible values for each concept.

6 Predicate structure

For the annotation of predicate structure we decide to use a FRAMENET-like approach (Baker et al., 1998).

Based on the domain ontology, we define a set of frames for each domain. The frame elements are provided by the named entities, and for all the frames we introduce the negation as default frame element.

For the annotation first of all we annotate the entities with a frame and a frame element. If the target is overt realized we make a pointer from the frame element to the target. The next step is putting all the frame elements and the target (if overt realized) in a set.

7 Coreference

Coreference and anaphoric relations will be annotated in the LUNA corpus using an scheme close to the one used in ARRAU (Artstein and Poesio, 2006).

The first step is the annotation of the information status of the markables with the tags `given` and `new`. If the markables are annotated with `given` the annotator will select the most recent occurrence of the object and add a pointer to it. If the markable is annotated with `new`, we distinguish between markables that are related to a previously mentioned object, the so called associative references, or don't have such a relation.

If there is more that a unique interpretation, the annotator can annotate the markable as `ambiguous` and add a pointer to each of the possible antecedents.

8 Dialogue acts

In order to associate the intentions of the speaker with the propositional content of the utterances, the segmentation of the dialogue turns in utterances is based on the annotation of predicate structure. Each set of frame elements will be correspond with a utterance.

We use a multi-dimensional annotation scheme partially based on the DAMSL scheme (Allen and Core, 1997) and on the proposals of ICSI-MRDA (Dhillon et al., 2004). We have selected nine dialogue acts from the DAMSL scheme as initial tagset, that can be extended for the different application domains. Each utterance will be annotated with as many tags as applicable.

References

J. Allen and M. Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers.

R. Artstein and M. Poesio, 2006. *ARRAU Annotation Manual (TRAINS dialogues)*. Univerity of Essex, U.K.

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL*. Association for Computational Linguistics.

H. Bonneau-Maynard and S. Rosset. 2003. A semantic representation for spoken dialogues. In *Proceedings of Eurospeech*, Geneva.

R. Dhillon, S. Bhagat, H. Carvez, and E. Shriberg. 2004. Meeting Recorder Project: Dialog Act Labeling Guide. Technical report, TR-04-002 ICSI.

EAGLES. 1996. Recomendations for the Morphosyntactic Annotation of Corpora. EAGLES Document EAG-TCWG-MAC/R.