Standoff Coordination for Multi-Tool Annotation in a Dialogue Corpus

Kepa Joseba Rodriguez♣, Stefanie Dipper♠, Michael Götze♠, Massimo Poesio□, Giuseppe Riccardi♦, Christian Raymond♦, Joanna Wisniewska‡

♣Piedmont Consortium for Information Systems (CSI-Piemonte), ♠Department of Linguistics. University of Potsdam, Center for Mind/Brain Sciences. University of Trento. ♦Department of Information and Communication Technology. University of Trento. ‡Institute of Computer Science. Polish Academy of Science.

1 Motivation

When creating a dialogue corpus with multi-level annotations, the corpus developers must take a number of important decisions, e.g.:

- **Annotation Tools**: Develop new, adapt, or use existing ones?
- **Representation Format**: Develop a new one, adapt, or use an existing one?

Scarce resources favor the use of existing tools and formats, but the required functionality is often not available. Toolkits such as the NITE XML Toolkit allow for adapting both tools and annotation formats, but still require programming skills and effort.

We sketch an approach to corpus development, in which specialized off-the-shelf tools are used for creating annotations, which are subsequently merged into one standoff data representation, the **PAULA Interchange Format for Linguistic Annotation**.

Until recently, this approach was implemented for text corpora only. With the **LUNA dialogue corpus**, the first speech corpus is modeled with PAULA.

2 The LUNA Project

Focus

- Real time understanding of spontaneous and unconstrained speech in conversational systems.

Scientific objectives

- Language modeling for speech understanding.
- Semantic modeling for speech understanding.
- Multilingual portability of Spoken Language Understanding (SLU) components.

Three steps are considered for the **SLU interpretation process**:

- Generation of semantic concept tags.
- Composition into conceptual structures.
- 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.

3 The LUNA Corpus

Target: Collection and annotation of

- 3000 Human-Human and
- 8100 Human-Machine dialogues
- in French, Italian and Polish.

French subcorpus

- Application domains: travel information and reservation, IT help desk, telecom costumer care and financial information transaction
- Human-Machine dialogues: 7100

Italian subcorpus

- Application domain: IT helpdesk
- Human-Human dialogues: 2500
- WOZ dialogues: 500

Polish subcorpus

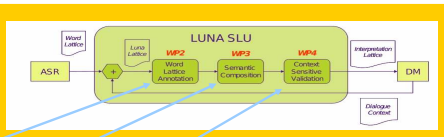
- Application domain: public transportation information
- Human-Human dialogues: 500
- WOZ dialogues: 500

4 Annotation Levels and Workflow

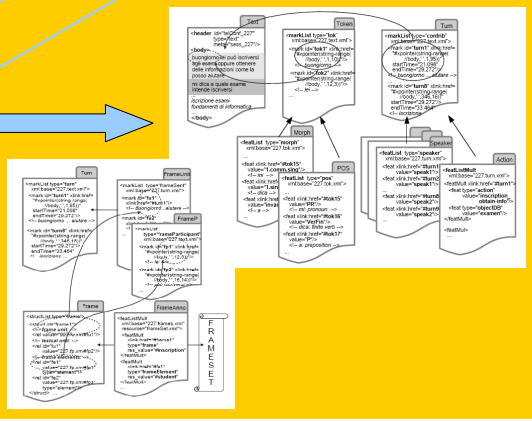
- **Word transcription / orthographic annotation**
- **Morphosyntactic annotation: POS and syntactic chunks.**
- **Domain attribute level**
  - Attribute-value pairs representation
  - Tagset of attribute-value specified using domain ontologies
- **Predicate structure**
  - The corpus is annotated using 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 us to have more than one 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 in several dimensions.

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



PAULA Interchange Format



Example dialogue annotation

[Wizard:] buon giorno lei può inscrivirsi agli esami oppure ottenere delle informazioni come la posso aiutare

Good morning, you can register for the exam or obtain information. How can I help you?

[Caller:] iscrizione essami
Inscription examen.

ANNIS

5 PAULA (Potsdam Interchange Format for Linguistic Annotation)

PAULA annotation scenario:

- Use of specialized **off-the-shelf annotation tools**
- **Merging** of different annotations into one representation
- **Further processing with ANNIS** (for visualization and querying) and other tools (e.g. for statistical analysis)

PAULA realizes a **standoff-architecture**, i.e. separate XML-files for:

- Text; Tokens; Meta-Information
- Markables (segments); structures (trees); features (annotations)
- Linking by XLink and XPointer expressions

Other features:

- import (EXMARaLDA, TIGER-XML, MMAX, RSTTool) and export functionality (WEKA etc.)
- support of various data structures (graphs, trees, pointer), discontinuous constituents and conflicting hierarchies; reference to externally defined tagsets
- *PAULA-inline*: an integrated version of the standoff-annotations for efficient querying and further processing

6 Representing Speech and Dialogue

Extending PAULA for dialogue and speech, annotations can additionally refer to points and spans of the timeline.

In dialogue annotations, this often requires representation of partial information:

(a) si allora **[noise=noise]+**tredici zero ottantasei
yes, 13 0 86
*(*noise *overlapping with the start of the token 'tredici')*

(b) allora avrei bisogno dell' **[lex=filler]** RWS del PC
so I need the RWS of the computer
*(*lex=filler *as an isolated annotation)*

In the examples, the exact alignment of the annotations with the time-line is not specified. Moreover, these annotations are no „standard" annotations for the token level neither. (a) only *overlaps* with the beginning of the token, and (b) is located between two tokens.

With PAULA, we decided not to represent the *overlap* explicitly, leaving it to the semantics of the tag. Case (b) is encoded qua convention referring to the last character of the previous token, with an extension of zero:

PAULA (a): <feat xlink:href="#tok_34" value="noise"/>
PAULA (b): <mark xlink:href="#xpointer(
string-range(//body,'',59,0))"/>

7 Related Work

MEDIA (France)

- Goal: Evaluation of understanding capabilities of dialogue systems.
- Annotation of words, acoustic events, semantic segments with attribute-value pairs, coreference.
- *Main Features:* Annotation with only one tool: Semantizer; Representation: all levels together in one file.

NITE XML Toolkit (Edinburgh)

- AMI Meeting Corpus, SAMMIE Corpus, etc.
- *Main Features:* open-source libraries for richly annotated corpora; support for multimodal and dialogue corpora; query support; Java API and media support; Rich Corpus Meta Specification

ELAN/ACM (MPI Nijmegen)

- Corpora in the Dobes („Documentation of Endangered Languages");
- *Main Features:* audio and video support, import and export facilities, other off-the-shelf tools for data management, metadata management, visualization and querying.

Links

LUNA: http://www.ist-luna.eu
PAULA: http://www.sfb632.uni-potsdam.de/projects/d1/paula/doc/