The SAMMIE Multimodal Dialogue Corpus Meets the Nite XML Toolkit

Ivana Kruijff-Korbayová, Verena Rieser, Ciprian Gerstenberger
Saarland University, Saarbrücken, Germany
vrieser@coli.uni-sb.de

Jan Schehl, Tilman Becker
DFKI, Saarbrücken, Germany
jan.schehl@dfki.de

Abstract

We demonstrate work in progress[1] using the Nite XML Toolkit on a corpus of multimodal dialogues with an MP3 player collected in a Wizard-of-Oz (WOZ) experiments and annotated with a rich feature set at several layers. We designed an NXT data model, converted experiment log file data and manual transcriptions into NXT, and are building annotation tools using NXT libraries.

1 Introduction

In the TALK project[2] we are developing a multimodal dialogue system for an MP3 application for in-car and in-home use. The system should support natural, flexible interaction and collaborative behavior. To achieve this, it needs to provide advanced adaptive multimodal output.

To determine the interaction strategies and range of linguistic behavior naturally occurring in this scenario, we conducted two WOZ experiments: SAMMIE-1 involved only spoken interaction, SAMMIE-2 was multimodal, with speech and screen input and output.[3]

We have been annotating the corpus on several layers, representing linguistic, multimodal and context information. The annotated corpus will be used (i) to investigate various aspects of multimodal presentation and interaction strategies both within and across the annotation layers; (ii) to design an initial policy for reinforcement learning of multimodal clarifications.[4] We use the Nite XML Toolkit (NXT) (Carletta et al., 2003) to represent and browse the data and to develop annotation tools.

Below we briefly describe our experiment setup, the collected data and the annotation layers; we comment on methods and tools for data representation and annotation, and then present our NXT data model.

2 Experiment Setup

24 subjects in SAMMIE-1 and 35 in SAMMIE-2 performed several tasks with an MP3 player application simulated by a wizard. For SAMMIE-1 we had two, for SAMMIE-2 six wizards. The tasks involved searching for titles and building playlists satisfying various constraints. Each session was 30 minutes long. Both users and wizards could speak freely. The interactions were in German (although most of the titles and artist names in the database were English).

SAMMIE-2 had a more complex setup. The tasks the subjects had to fulfill were divided in two classes: with vs. without operating a driving simulator. When presenting the search results, the wizards were free to produce mono- or multimodal output as they saw fit; they could speak freely and/or select one of four automatically generated screen outputs, which contained tables and lists of found songs/albums. The users also had free choice between unconstrained

[1] Our demonstration results from the efforts of a larger team including also N. Blaylock, B. Fromkorth, M. Grác, M. Kaißer, A. Moos, P. Poller and M. Wirth.

[2] TALK (Talk and Look: Tools for Ambient Linguistic Knowledge; http://www.talk-project.org), funded by the EU 6th Framework Program, project No. IST-507802.

[3] SAMMIE stands for Saarbrücken Multimodal MP3 Player Interaction Experiment.

[4] See (Kruijff-Korbayová et al., 2006) for more details about the annotation goals and further usage of the corpus.

natural language and/or selecting items on the screen. Both wizard and user utterances were immediately transcribed. The wizard's utterances were presented to the user via a speech synthesizer. To simulate acoustic understanding problems, the wizard sometimes received only part of the transcribed user's utterance, to elicit CRs. (See (Kruijff-Korbayová et al., 2005) for details.)

3 Collected Data

The SAMMIE-2 data for each session consists of a video and audio recording and a log file.[5] The gathered logging information per session consists of Open Agent Architecture (Martin et al., 1999) (OAA) messages in chronological order, each marked by a timestamp. The log files contain various information, e.g., the transcriptions of the spoken utterances, the wizard's database query and the number of results, the screen option chosen by the wizard, classification of clarification requests (CRs), etc.

4 Annotation Methods and Tools

The rich set of features we are interested in naturally gives rise to a multi-layered view of the corpus, where each layer is to be annotated independently, but subsequent investigations involve exploration and automatic processing of the integrated data across layers.

There are two crucial technical requirements that must be satisfied to make this possible: (i) stand-off annotation at each layer and (ii) alignment of base data across layers. Without the former, we could not keep the layers separate, without the latter we would not be able to align the separate layers. An additional equally important requirement is that elements at different layers of annotation should be allowed to have overlapping spans; this is crucial because, e.g., prosodic units and syntactic phrases need not coincide.

Among the existing toolkits that support multi-layer annotation, it was decided to use NXT (Carletta et al., 2003)[6] in the TALK project. The NXT-based SAMMIE-2 corpus we are demonstrating has been created in several steps: **(1)** The speech data was manually transcribed using the Transcriber tool.[7] **(2)** We automatically extracted features at various annotation layers by parsing the OAA messages in the log files. **(3)** We automatically converted the transcriptions and the information from the log files into our NXT-based data representation format; features annotated in the transcriptions and features automatically extracted from the log files were assigned to elements at the appropriate layers of representation in this step.

Manual annotation: We use tools specifically designed to support the particular annotation tasks. We describe them below.

As already mentioned, we used Transcriber for the manual transcriptions. We also performed certain relatively simple annotations directly on the transcriptions and coded them in-line by using special notation. This includes the identification of self-speech, the identification of expressions referring to domain objects (e.g., songs, artists and albums) and the identification of utterances that convey the results of database queries.

For other manual annotation tasks (the annotation of CRs, task segmentation and completion, referring expressions and the relations between them) we have been building specialized tools based on the NXT library of routines for building displays and interfaces based on Java Swing (Carletta et al., 2003). Although NXT comes with a number of example applications, these are tightly coupled with the architecture of the corpora they were built for. We therefore developed a core basic tool for our own corpus; we modify this tool to suite each annotation task. To facilitate tool development, NXT provides GUI elements linked directly to corpora elements and support for handling complex multi-layer corpora. This proved very helpful.

Figure 4 shows a screenshot of our CR annotation tool. It allows one to select an utterance in the left-hand side of the display by clicking on it, and then choose the attribute values from the pop-down lists on the right-hand side. Cre-

[5] For 19 sessions the full set of data files exists.

[6] `http://www.ltg.ed.ac.uk/NITE/`

[7] `http://trans.sourceforge.net/`

ating relations between elements and creating elements on top of other elements (e.g., words or utterances) are extensions we are currently implementing (and will complete by the time of the workshop). First experiences using the tool to identify CRs are promising.[8] When demonstrating the system we will report the reliability of other manual annotation tasks.

Automatic annotation using indexing: NXT also provides a facility for automatic annotation based on NiteQL query matches (Carletta et al., 2003). Some of our features, e.g., the dialogue history ones, can be easily derived via queries.

5 The SAMMIE NXT Data Model

NXT uses a stand-off XML data format that consist of several XML files that point to each other. The NXT data model is a multi-rooted tree with arbitrary graph structure. Each node has one set of children, and can have multiple parents.

Our corpus consists of the following layers. Two base layers: words and graphical output events; both are time-aligned. On top of these, structural layers correspond to one session per subject, divided into task sections, which consist of turns, and these consist of individual utterances, containing words. Graphical output events will be linked to turns at a featural layer.

Further structural layers are defined for CRs and dialogue acts (units are utterances), domain objects and discourse entities (units are expressions consisting of words). We keep independent layers of annotation separate, even when they can in principle be merged into a single hierarchy.

Figure 2 shows a screenshot made with Amigram (Lauer et al., 2005), a generic tool for browsing and searching NXT data. On the left-hand side one can see the dependencies between the layers. The elements at the respective layers are displayed on the right-hand side.

Below we indicate the features per layer:

- **Words:** Time-stamped words and other sounds; we mark self-speech, pronunciation, deletion status, lemma and POS.
- **Graphical output:** The type and amount of information displayed, the option selected by the wizard, and the user's choices.
- **Utterances:** Error rates due to word deletion, and various features describing the syntactic structure, e.g., mood, polarity, diathesis, complexity and taxis, the presence of marked syntactic constructions such as ellipsis, fronting, extraposition, cleft, etc.
- **Turns:** Time delay, dialogue duration so far, and other dialogue history features, i.e. values which accumulate over time.
- **Domain objects and discourse entities:** Properties of referring expressions reflecting the type and information status of discourse entities, and coreference/bridging links between them.
- **Dialogue acts:** DAs based on an agent-based approach to dialogue as collaborative problem-solving (Blaylock et al., 2003), e.g., determining joint objectives, finding and instantiating recipes to accomplish them, executing recipes and monitoring for success. We also annotate propositional content and the database queries.
- **CRs:** Additional features including the source and degree of uncertainty, and characteristics of the CRs strategy.
- **Tasks:** A set of features for estimating user satisfaction online for reinforcement learning (Rieser et al., 2005).
- **Session:** Subject and wizard information, user questionnaire aswers, and accumulating attribute values from other layers.

6 Summary

We described a multi-layered corpus of multimodal dialogues represented and annotated using NXT-based tools. Our data model relates linguistic and graphical realization to a rich set of context features and represents structural, hierarchical interactions between different annotation layers. We combined different annotation methods to construct the corpus. Manual annotation and annotation evaluation is on-going. The corpus will be used (i) investigate multimodal presentation and interaction strategies with respect

[8] Inter-annotator agreement of 0.788 (κ corrected for prevalence).

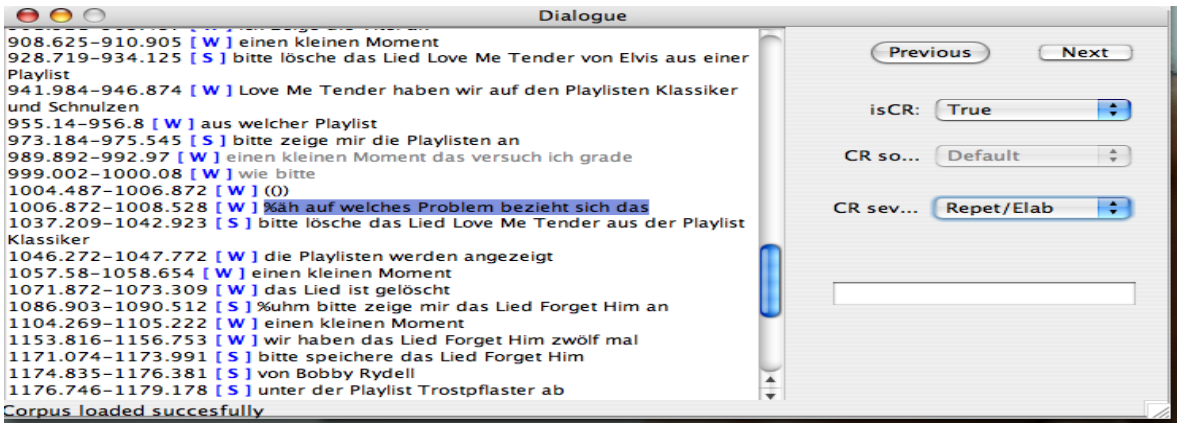

Figure 1: NXT-based tool for annotating CRs

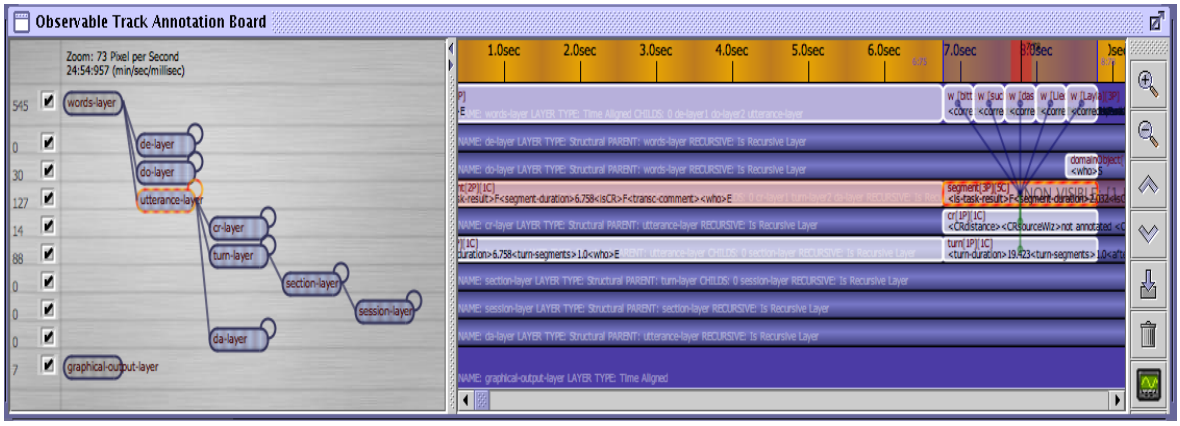

Figure 2: SAMMIE-2 corpus displayed in Amigram

to dialogue context and (ii) to design an initial policy for reinforcement learning of multimodal clarification strategies.

References

[Blaylock et al.2003] N. Blaylock, J. Allen, and G. Ferguson. 2003. Managing communicative intentions with collaborative problem solving. In *Current and New Directions in Discourse and Dialogue*, pages 63–84. Kluwer, Dordrecht.

[Carletta et al.2003] J. Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann. 2003. The NITE XML Toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers, special issue on Measuring Behavior*. Submitted.

[Kruijff-Korbayová et al.2005] I. Kruijff-Korbayová, T. Becker, N. Blaylock, C. Gerstenberger, M. Kaißer, P. Poller, J. Schehl, and V. Rieser. 2005. An experiment setup for collecting data for adaptive output planning in a multimodal dialogue system. In *Proc. of ENLG*.

[Kruijff-Korbayová et al.2006] I. Kruijff-Korbayová, T. Becker, N. Blaylock, C. Gerstenberger, M. Kaißer, P. Poller, V. Rieser, and J. Schehl. 2006. The SAMMIE corpus of multimodal dialogues with an mp3 player. In *Proc. of LREC (to appear)*.

[Lauer et al.2005] C. Lauer, J. Frey, B. Lang, T. Becker, T. Kleinbauer, and J. Alexandersson. 2005. Amigram - a general-purpose tool for multimodal corpus annotation. In *Proc. of MLMI*.

[Martin et al.1999] D. L. Martin, A. J. Cheyer, and D. B. Moran. 1999. The open agent architecture: A framework for building distributed software systems. *Applied Artificial Intelligence: An International Journal*, 13(1–2):91–128, Jan–Mar.

[Rieser et al.2005] V. Rieser, I. Kruijff-Korbayová, and O. Lemon. 2005. A corpus collection and annotation framework for learning multimodal clarification strategies. In *Proc. of SIGdial*.