Generating Usable Formats for Metadata and Annotations in a Large Meeting Corpus

Andrei Popescu-Belis and Paula Estrella
ISSCO/TIM/ETI, University of Geneva
40, bd. du Pont-d'Arve
1211 Geneva 4 - Switzerland
{andrei.popescu-belis, paula.estrella}@issco.unige.ch

Abstract

The AMI Meeting Corpus is now publicly available, including manual annotation files generated in the NXT XML format, but lacking explicit metadata for the 171 meetings of the corpus. To increase the usability of this important resource, a representation format based on relational databases is proposed, which maximizes informativeness, simplicity and reusability of the metadata and annotations. The annotation files are converted to a tabular format using an easily adaptable XSLT-based mechanism, and their consistency is verified in the process. Metadata files are generated directly in the IMDI XML format from implicit information, and converted to tabular format using a similar procedure. The results and tools will be freely available with the AMI Corpus. Sharing the metadata using the Open Archives network will contribute to increase the visibility of the AMI Corpus.

1 Introduction

The AMI Meeting Corpus (Carletta and al., 2006) is one of the largest and most extensively annotated data sets of multimodal recordings of human interaction. The corpus contains 171 meetings, in English, for a total duration of ca. 100 hours. The meetings either follow the remote control design scenario, or are naturally occurring meetings. In both cases, they have between 3 and 5 participants.

Perhaps the most valuable resources in this corpus are the high quality annotations, which can be used to train and test NLP tools. The existing annotation dimensions include, beside transcripts, forced temporal alignment, named entities, topic segmentation, dialogue acts, abstractive and extractive summaries, as well as hand and head movement and posture. However, these dimensions as well as the implicit metadata for the corpus are difficult to exploit by NLP tools due to their particular coding schemes.

This paper describes work on the generation of annotation and metadata databases in order to increase the usability of these components of the AMI Corpus. In the following sections we describe the problem, present the current solutions and give future directions.

2 Description of the Problem

The AMI Meeting Corpus is publicly available at `http://corpus.amiproject.org` and contains the following media files: audio (headset mikes plus lapel, array and mix), video (close up, wide angle), slides capture, whiteboard and paper notes. In addition, all annotations described in Section 1 are available in one large bundle. Annotators followed dimension-specific guidelines and used the NITE XML Toolkit (NXT) to support their task, generating annotations in NXT format (Carletta and al., 2003; Carletta and Kilgour, 2005). Using the NXT/XML schema makes the annotations consistent along the corpus but more difficult to use without the NITE toolkit. A less developed aspect of the corpus is the metadata encoding all auxiliary information about meetings in a more structured and informative manner. At the moment, metadata is spread implicitly along the corpus data, for example

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 93–96,
Prague, June 2007.

it is encoded in the file or folder names or appears to be split in several resource files.

We define here annotations as the time-dependent information which is abstracted from the input media, i.e. "higher-level" phenomena derived from low-level mono- or multi-modal features. Conversely, metadata is defined as the static information about a meeting that is not directly related to its content (see examples in Section 4). Therefore, though not necessarily time-dependent, structural information derived from meeting-related documents would constitute an annotation and not metadata. These definitions are not universally accepted, but they allow us to separate the two types of information.

The main goal of the present work is to facilitate the use of the AMI Corpus metadata and annotations as part of the larger objective of automating the generation of annotation and metadata databases to enhance search and browsing of meeting recordings. This goal can be achieved by providing plug-and-play databases, which are much easier to access than NXT files and provide declarative rather than implicit metadata. One of the challenges in the NXT-to-database conversion is the extraction of relevant information, which is done here by solving NXT pointers and discarding NXT-specific markup to group all information for a phenomenon in only one structure or table.

The following criteria were important when defining the conversion procedure and database tables:

- Simplicity: the structure of the tables should be easy to understand, and should be close to the annotation dimensions—ideally one table per annotation. Some information can be duplicated in several tables to make them more intelligible. This makes the update of this information more difficult, but as this concerns a recorded corpus, changes are less likely to occur; if such changes do occur, they would first be input in the annotation files, from which a new set of tables can easily be generated.
- Reusability: the tools allow anyone to recreate the tables from the official distribution of the annotation files. Therefore, if the format of the annotation files or folders changes, or if a different format is desired for the tables, it is quite easy to change the tools to generate a new version of the database tables.
- Applicability: the tables are ready to be loaded into any SQL database, so that they can be immediately used by a meeting browser plugged into the database.

Although we report one solution here, there are other approaches to the same problem relying, for example, on different database structures using more or fewer tables to represent this information.

3 Annotations: Generation of Tables

The first goal is to convert the NXT files from the AMI Corpus into a compact tabular representation (tab-separated text files), using a simple, declarative and easily updatable conversion procedure.

The conversion principle is the following: for each type of annotation, which is generally stored in a specific folder of the data distribution, an XSLT stylesheet converts the NXT XML file into a tab-separated text file, possibly using information from one or more annotations. The stylesheets resolve most of the NXT pointers, by including redundant information into the tables, in order to speed up queries by avoiding frequent joins. A Perl script applies the respective XSLT stylesheet to each annotation file according to its type, and generates the global tab-separated files for each annotation. The script also generates an SQL script that creates a relational annotation database and populates it with data from the tab-separated files. The Perl script also summarizes the results into a log file named `<timestamp>.log`.

The conversion process can be summarized as follows and can be repeated at will, in particular if the NXT source files are updated:

1. Start with the official NXT release (or other XML-based format) of the AMI annotations as a reference version.
2. Apply the table generation mechanism to XML annotation files, using XSLT stylesheets called by the script, in order to generate tabular files (TSV) and a table-creation script (`db_loader.sql`).
3. Create and populate the annotation database.
4. Adapt the XSLT stylesheets as needed for various annotations and/or table formats.

4 Metadata: Generation of Explicit Files and Conversion to Tabular Format

As mentioned in Section 2, metadata denotes here any *static information* about a meeting, not directly related to its content. The main metadata items are: date, time, location, scenario, participants, participant-related information (codename, age, gender, knowledge of English and other languages), relations to media-files (participants vs. audio channels vs. files), and relations to other documents produced during the meeting (slides, individual and whiteboard notes).

This important information is spread in many places, and can be found as attributes of a meeting in the annotation files (e.g. start time) or obtained by parsing file names (e.g. audio channel, camera). The relations to media files are gathered from different resource files: mainly the `meetings.xml` and `participants.xml` files. An additional problem in reconstructing such relations (e.g. files generated by a specific participant) is that information about the media resources must be obtained directly from the AMI Corpus distribution web site, since the media resources are not listed explicitly in the annotation files. This implies using different strategies to extract the metadata: for example, stylesheets are the best option to deal with the above-mentioned XML files, while a crawler script is used for HTTP access to the distribution site. However, the solution adopted for annotations in Section 3 can be reused with one major extension and applied to the construction of the metadata database.

The standard chosen for the explicit metadata files is the IMDI format, proposed by the ISLE Meta Data Initiative (Wittenburg et al., 2002; Broeder et al., 2004a) (see `http://www.mpi.nl/IMDI/tools`), which is precisely intended to describe multimedia recordings of dialogues. This standard provides a flexible and extensive schema to store the defined metadata either in specific IMDI elements or as additional key/value pairs. The metadata generated for the AMI Corpus can be explored with the IMDI BC-Browser (Broeder et al., 2004b), a tool that is freely available and has useful features such as search or metadata editing.

The process of extracting, structuring and storing the metadata is as follows:

1. Crawl the AMI Corpus website and store resulting metadata (related to media files) into an XML auxiliary file.
2. Apply an XSLT stylesheet to the auxiliary XML file, using also the distribution files `meetings.xml` and `participants.xml`, to obtain one IMDI file per meeting.
3. Apply the table generation mechanism to each IMDI file in order to generate tabular files (TSV) and a table-creation script.
4. Create and populate metadata tables within database.
5. Adapt the XSLT stylesheet as needed for various table formats.

5 Results: Current State and Distribution

The 16 annotation dimensions from the public AMI Corpus were processed following the procedure described in Section 3. The main Perl script, `anno-xml2db.pl`, applied the 16 stylesheets corresponding to each annotation dimension, which generated one large tab-separated file each. The script also generated the table-creation SQL script `db_loader.sql`. The number of lines of each table, hence the number of “elementary annotations”, is shown in Table 1.

The application of the metadata extraction tools described in Section 4 generated a first version of the explicit metadata for the AMI Corpus, consisting of 171 automatically generated IMDI files (one per meeting). In addition, 85 manual files were created in order to organize the metadata files into IMDI corpus nodes, which form the skeleton of the corpus metadata and allow its browsing with the BC-Browser. The resources and tools for annotation/metadata processing will be made soon available on the AMI Corpus website, along with a demo access to the BC-Browser.

6 Discussion and Perspectives

The proposed solution for annotation conversion is easy to understand, as it can be summarized as “one table per annotation dimension”. The tables preserve only the relevant information from the NXT

Annotation dimension	Nb. of entries
words (transcript)	1,207,769
named entities	14,230
speech segments	69,258
topics	1,879
dialogue acts	117,043
adjacency pairs	26,825
abstractive summaries	2,578
extractive summaries	19,216
abs/ext links	22,101
participant summaries	3,409
focus	31,271
hand gesture	1,453
head gesture	36,257
argument structures	6,920
argumentation relations	4,759
discussions	8,637

Table 1: Results of annotation conversion; dimensions are grouped by conceptual similarity.

annotation files, and search is accelerated by avoiding repeated joins between tables.

The process of metadata extraction and generation is very flexible and the obtained data can be easily stored in different file formats (e.g. tab-separated, IMDI, XML, etc.) with no need to repeatedly parse file names or analyse folders. Moreover, the advantage of creating IMDI files is that the metadata is compliant with a widely used standard accompanied by freely available tools such as the metadata browser. These results will also help disseminating the AMI Corpus.

As a by-product of the development of annotation and metadata conversion tools, we performed a consistency checking and reported a number of to the corpus administrators. The automatic processing of the entire annotation and metadata set enabled us to test initial hypotheses about annotation structure.

In the future we plan to include the AMI Corpus metadata in public catalogues, through the Open (Language) Archives Initiatives network (Bird and Simons, 2001), as well as through the IMDI network (Wittenburg et al., 2004). The metadata repository will be harvested by answering the OAI-PMH protocol, and the AMI Corpus website could become itself a metadata provider.

Acknowledgments

The work presented here has been supported by the Swiss National Science Foundation through the NCCR IM2 on Interactive Multimodal Information Management (`http://www.im2.ch`). The authors would like to thank Jean Carletta, Jonathan Kilgour and Maël Guillemot for their help in accessing the AMI Corpus.

References

Steven Bird and Gary Simons. 2001. Extending Dublin Core metadata to support the description and discovery of language resources. *Computers and the Humanities*, 37(4):375–388.

Daan Broeder, Thierry Declerck, Laurent Romary, Markus Uneson, Sven Strömqvist, and Peter Wittenburg. 2004a. A large metadata domain of language resources. In *LREC 2004 (4th Int. Conf. on Language Resources and Evaluation)*, pages 369–372, Lisbon.

Daan Broeder, Peter Wittenburg, and Onno Crasborn. 2004b. Using profiles for IMDI metadata creation. In *LREC 2004 (4th Int. Conf. on Language Resources and Evaluation)*, pages 1317–1320, Lisbon.

Jean Carletta and al. 2006. The AMI Meeting Corpus: A pre-announcement. In Steve Renals and Samy Bengio, editors, *Machine Learning for Multimodal Interaction II*, LNCS 3869, pages 28–39. Springer-Verlag, Berlin/Heidelberg.

Jean Carletta and Jonathan Kilgour. 2005. The NITE XML Toolkit meets the ICSI Meeting Corpus: Import, annotation, and browsing. In Samy Bengio and Hervé Bourlard, editors, *Machine Learning for Multimodal Interaction*, LNCS 3361, pages 111–121. Springer-Verlag, Berlin/Heidelberg.

Jean Carletta, Stefan Evert, Ulrich Heid, Jonathan Kilgour, Judy Robertson, and Holger Voormann. 2003. The NITE XML Toolkit: flexible annotation for multimodal language data. In *Behavior Research Methods, Instruments, and Computers*, special issue on Measuring Behavior, 35(3), pages 353–363.

Peter Wittenburg, Wim Peters, and Daan Broeder. 2002. Metadata proposals for corpora and lexica. In *LREC 2002 (3rd Int. Conf. on Language Resources and Evaluation)*, pages 1321–1326, Las Palmas.

Peter Wittenburg, Daan Broeder, and Paul Buitelaar. 2004. Towards metadata interoperability. In *NLPXML 2004 (4th Workshop on NLP and XML at ACL 2004)*, pages 9–16, Barcelona.