Building and Using a Corpus of Shallow Dialog Annotated Meetings

Andrei Popescu-Belis, Maria Georgescul, Alexander Clark, Susan Armstrong

ISSCO/TIM/ETI, University of Geneva
40, bd. du Pont d'Arve – CH.1211 Geneva 4 – Switzerland
{andrei.popescu-belis, maria.georgescul, alexander.clark, susan.armstrong}@issco.unige.ch

Abstract

In this paper we provide a framework for shallow dialog annotations (SDA), and for their use in the context of the processing and retrieval of multimodal meeting recordings. The SDA model groups the following elements: dialog segmentation into utterances and episodes, detection of dialog acts and adjacency pairs, and detection of referring expressions and coreference links, including references to documents. An instantiated XML annotation model based on boundaries, labels and links, is provided. The use of SDA data in a meeting retrieval interface is also described.

1. Meeting processing and retrieval (MPR)

One of the current challenges to speech and language understanding is the situation where verbal communication is used interactively. Two main research domains deal with this issue: on one side, spoken dialog systems support interaction between people and computers, and on the other, human dialog understanding systems track dialogs between humans.

The understanding of human dialogs would enable many useful applications, among which we focus here on *automatic meeting processing and retrieval (MPR)*. This application would enable people who did not attend a meeting (e.g. staff or business meeting), or people who want to review a past meeting, to search for a particular piece of information connected to the meeting.

The paper proceeds as follows: we define a model for *shallow dialog analysis (SDA)* in Section 2, which integrates segmentation, dialog acts, and coreference information, and is a result of the trade-off between robustness of extraction and relevance to dialog understanding. The annotation process and its results are described in Section 3, and then the use of the data for MPR is outlined in Section 4. Perspectives are given in Section 5.

2. Description of the Annotations: SDA

We focus on spoken language transcripts, rather than multimodal data. We consider separate transcripts for every speaker present at a meeting, or for each recording channel (individual microphone). The use of manual transcriptions as reference input data means that shallow dialog analysis operates on highly accurate data. Automated speech recognizer systems would have a word error rate of 30% or more in such an environment (Morgan et al., 2003), hence they could only *help* the production of accurate transcriptions, including word-level timestamps.

Our model for shallow dialog analysis (SDA) is aimed at robust extraction of significant information from spoken dialogs. We have identified a set of promising elements from natural language processing: (cf. overview in Table 1): detection of utterances (UT) and thematic episodes (EP / TD), detection and resolution of referring expressions (RE / RT and CO), and recognition of dialog acts (DA) and adjacency pairs (AP) (Popescu-Belis, 2003b).

2.1. Segmentation: Utterances and Topics

Starting from time-stamped transcribed speech, boundaries are inserted at two key levels: individual utterances, and topic-coherent episodes. Word boundaries (spaces), are already present in the transcription.

An utterance (UT) is a coherent, contiguous series of words from a given channel, which serves a precise function in the dialog (or sometimes more than one), labelled with a dialog act (Traum, 2000). An utterance can often be equated with a proposition or a sentence, but in spoken language, utterances do not always correspond to well-formed or completed propositions.

Utterances are the building blocks of dialog structure, and constitute the minimal units that are of interest for dialog retrieval. Their identification also plays a significant role in the formatting of recorded dialogs, namely for capitalization and punctuation. Many cues have been used for this task, such as: lexical markers, (shallow) syntactic structure, and prosody (Stolcke and Shriberg, 1996).

A dialog is also decomposed in thematic episodes (TE), with a short topic description (TD) assigned to each episode. Episodes represent a "flat" structure that cuts across channels, and therefore they require minimal theoretical assumptions about discourse structure, as opposed to more complex hierarchical structures. Automatic thematic segmentation studies are based on different (probabilistic) lexical cohesion methods (Choi et. al, 2001) and/or on the use of different learning mechanisms to combine multiple features such as cue phrases and prosodic features (Passonneau and Litman, 1997). These methods have shown a P_k error rate (Beeferman et. al, 1999) of 10-15% on written texts and spoken monologues. There are few studies on topic segmentation of multi-party conversations: error rates are about 23% (Galley et. al, 2003), which proves that this task is more difficult.

Assigning descriptions to each topic (TD), e.g. in terms of keywords, is a still more arduous task, since the definition of the "correct answer" remain problematic. The more informative a topic description is for dialog understanding and MPR (e.g. short titles as opposed to keywords), the more difficult it is to evaluate the performance of a system on such a task.

2.2. Dialog Acts and Adjacency Pairs

Much was written about the structure of dialog, but little is subject to general agreement, and even less is automatically detectable by a program. For SDA, a minimum is to label utterances with dialog acts (DA), and to detect utterances that are functionally related. The DA is the role of an utterance for the progression of the dialog, such as "question", "statement" or "backchannel" (Traum, 2000).

CODE	NAME	TYPE	SCOPE	OTHER FEATURES
UT	utterances	boundary	intra-channel	non-partitioning
DA	dialog acts	label	on UT	closed vocabulary
AP	adjacency pairs	link	between UT	mostly inter-channel, no labels
EP	episodes	boundary	global	partitioning
TD	topic description	label	on EP	open vocabulary
RE	referring expressions	boundary	intra-channel	non-partitioning
TR	types of REs	label	on RE	closed vocabulary
CO	coreference links	link	between RE	intra-and inter-channel, no labels (or XPath to meeting documents)

Table 1: Annotations for the various SDA elements

Several lists of DAs exist – cf. (Klein et al., 1998) for a comparison – and though there are many correspondences between lists, it is not easy to find a common denominator. For meeting recordings, we have analyzed and simplified the set proposed by ICSI (Dhillon et al., 2004), itself an extension of the DAMSL tagset, with 12 DA basic types and more refinements.

Automatic DA detection has been attempted using either rule-based or statistical systems (Stolcke et. al, 2000). Our experiments with machine learning systems trained on the Switchboard corpus (two-party telephone conversations, DAMSL tagset) reproduced current performance, i.e. ca. 70% accuracy (Clark and Popescu-Belis, 2004).

Adjacency pairs (AP) are functional links between pairs of utterances such as question/answer, invite/accept, offer/ decline, etc. Such links prefigure more complex dialog structures, but even at this "flat" level, their detection by a program has never been evaluated. We focus here on links between questions and answers, as well as orders (or proposals, suggestions) and acceptance or refusal.

2.3. Coreference Information

The detection of referring expressions (RE), that is, chunks of an utterance that point to, or denote an object, person, place, concept, etc., is another component of SDA. Named entities (proper names, dates, times, amounts, etc.) have been a target in the MUC and ACE campaigns[1], and we propose here to follow the MUC-7 guidelines. Therefore, the type of an RE (TR) is annotated along with the RE boundaries. In the future, referring acts in other modalities (e.g., pointing gestures) could also be annotated and related to linguistic REs.

Coreference links (CO) between REs connect those REs that denote the same entity. Here again the MUC-7 guidelines provide a reasonable definition of the targeted task, despite some problematic issues (Van Deemter and Kibble, 2000). Performance levels for this task, using robust methods in the case of (relatively) unrestricted domains, reach 60-70%, with either of the existing evaluation methods (Popescu-Belis, 2003a).

Also, as a pilot experiment, references made in the dialog transcript to meeting documents are annotated. These are references to a closed set of entities, which are available as elements in the logic-based representation of the documents. Their resolution is the focus of an ongoing study (Lalanne and Popescu-Belis, *in preparation*). Only meetings for which the documents are available in electronic format are subject to this annotation.

2.4. Towards an Integrated SDA Parser

The various components of the SDA model were selected not only for their informativeness and their possibility to be detected reliably, but also because they are conceptually interrelated. Indeed, there are strong correlations between boundaries and links in an SDA structure, e.g., coreference relations are sparser *between* thematic episodes than *within* episodes. These correlations are represented in Figure 1, where full lines stand for strict dependency, and dashed lines for preferences.

We are at present in the process of assembling, converting, or writing components for an SDA parser, or rather tagger. The development cycle presupposes a precise definition of the task, i.e. data annotation guidelines, as well as human annotated data to serve as reference for evaluation and subsequent training.

A blackboard-style mechanism for the SDA parser has been envisaged. The shallow analysis proceeds by incremental XML annotation of the input data (in the various channels), with the constraint that a component can only add annotation, but not delete it or change it. This is quite restrictive, but avoids infinite loops. The main loop is:

(1) execute each SDA component once;
(2) loop through the components with the test: if something has changed in the annotation since the previous execution of the component, then execute it again;
(3) stop when no component can add further annotations.

3. SDA Annotation of Meeting Data

This section outlines the data model that lies behind the SDA annotations, with an application to XML annotation. We then specify the annotation tools, and the data that is currently available. We defined a data model for the SDA phenomena, and, based on it, an XML annotation format that is used as a pivot format for all our meeting resources.

3.1. SDA Data Model

3.1.1. Data Structures

For an adequate representation of the phenomena discussed above, we have identified the need for three types of annotations: (1) boundaries; (2) labels on bounded segments; and (3) links between bounded segments. A fourth category could be added: the links could also have a *type* (a label), but for the moment, this is unnecessary since AP links are implicitly typed by the DA labels of their constituents.

[1] Message Understanding Conferences: *http://www.itl.nist.gov/iaui/894.02/related_projects/muc/*. Automatic Content Extraction : *http://www.itl.nist.gov/iad/894.01/tests/ace/*.

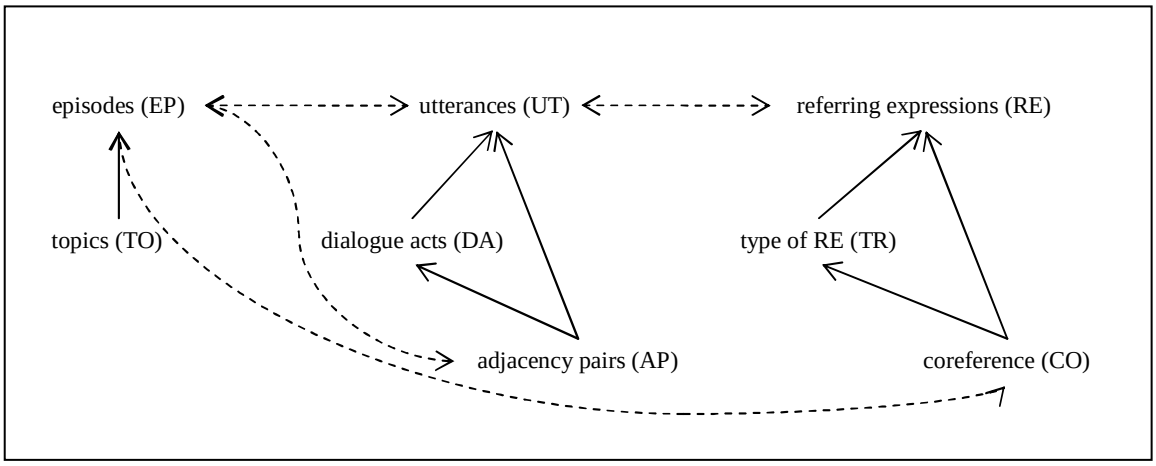

Figure 1: Dependencies between components of SDA

For CO links, only "identity of reference" types are used, as opposed to non-identity coreference, such as "part-of" or "person-function" relations. Links between episodes could, in the future, provide a better structured thematic analysis of the meetings.

The SDA phenomena require the annotation elements summarized in Table 1 with their abbreviations. There are three types of *boundaries*: intra-channel ones for utterances (UT) and referring expressions (RE), and inter-channel ones for episodes (EP). There are three types of *labels*: dialog acts (DA), type of REs (RT), and topics (TO), the first two being defined by a closed vocabulary (fixed list of labels). There are two types of *links*: adjacency pairs (AP) and coreference (CO). The dependencies between the data structures are represented in Figure 1.

3.1.2. XML Annotation

For the XML annotation of boundaries, intra-channel ones are straightforward to annotate as XML elements (opening/closing tags). But the multi-channel structure of the data must also be considered, and cross-channel boundaries must make reference to a global time-scale.

Labels are annotated as XML attributes, while links are annotated as separate XML elements, using the indexes of the bounded segments (types of links could be annotated as attributes on these elements). These considerations parallel the Annotation Graphs (AG) formalism[2], but since we plan to reuse existing non AG-compliant interfaces, we did not adopt this formalism. A full DTD that defines the annotation and enables us to validate resources has been developed (Popescu-Belis, 2003b).

3.2. Annotation tools

The production of meeting data annotated by humans (for the training and/or evaluation of SDA component taggers) can be done by reusing existing tools, i.e. annotation interfaces, as described below.

The consideration of boundaries, labels and links indicates that two interfaces are sufficient. Transcriber[3] is used for the insertion of utterance and topic boundaries (UT, EP) and their labels (DA, TE) – as well as for transcription and revision. For the annotation of referring information, the MMAX[4] program is used, which provides an interface specific to RE and CO. MMAX can also be parameterized to annotate adjacency pairs (AP).

Several data formats are used before the dialogs are fed into our final dialog database, depending on the transcription tool. For instance, the ICSI-MR data (see 3.3) is delivered in multi-channel Transcriber format, or in TableTrans (AGTK) format, or in a CSV-style format (for DAs). We prefer to use standard Transcriber format, and transcribe/annotate channels separately, then transform EP and TO into cross-channel annotation. MMAX uses other XML tags than Transcriber, therefore extensive use of XSLT is necessary.

We also use XSLT (eXtensible Stylesheet Language) to convert XML-exported data between interfaces, and also to generate HTML tables for a user-friendly representation of the dialog information, and to generate the final tabular format that is fed into the database.

3.3. Available Data

Complete annotation of SDA from scratch is a time consuming task (to say nothing about transcription itself). Therefore, reuse of existing resources is a priority. Within the IM2 project[5], three main sites provide transcribed meeting recordings: IDIAP, Martigny, the University of Fribourg, and ICSI, Berkeley. The first two provide transcriptions and some UT and EP annotation for, respectively, ca. 60 and ca. 20 short meetings (5'-15'), although a larger corpus is currently being recorded at IDIAP (McCowan et al., 2003). The ICSI-MR project provides about 75 meetings annotated with UT, DA (especially) and AP information (Shriberg et al., 2004), which we validated and converted to partial SDA (Clark and Popescu-Belis, 2004).

Stylesheets were written and conversion methods were defined for these resources, which await complete annotation of the missing SDA components, in particular EP+TO and RE+RT+AP annotation.

---

[2] See *http://agtk.sourceforge.net*.

[3] Transcriber, a tool for transcribing and labelling speech, is available at: *http://www.etca.fr/CTA/gip/Projets/Transcriber/*.

[4] MMAX is available freely from: *http://www.eml.villa-bosch.de/english/Research/NLP/Downloads*.

[5] See *http://www.im2.ch*.

4. Accessing the Annotated Data

The XML annotations corresponding to the different SDA components are converted in tabular format by using XSLT stylesheets, and then stored in a PostgreSQL database. The database and the interface described below will be demonstrated at the LREC 2004 conference.

4.1. Interface to the Database

The consultation of the database (containing the SDA annotations) is realized as a client-server application, using SOAP (Simple Object Access Protocol) as a communication protocol. The database access using SOAP is based on Web services and has been successfully implemented for the connection to a multimodal interface. Using a communication protocol allows access to databases under different operating systems and hardware platforms without changing applications. The SDA graphical user interface gives access to most of the fields of the database. Based on the user input parameters, the SQL query to the database is dynamically generated.

4.2. Multimedia Rendering

The result of a query is retrieved by the interface as a set of utterances. However, it is likely that users would prefer to retrieve the context of each utterance too. Therefore, we defined the following mechanism: clicking on the utterance transcription from the database gives access to the whole meeting transcription in a new frame, centred on the utterance. Moreover, it is possible to listen to that particular section (and, if available, to view the video) by another mouse click. This mechanism has a standalone implementation, using only HTTP links, inserted in the XML transcription using XSLT stylesheets, and an embedded player. Hence, the mechanism does not require the installation of any additional software such as streaming servers or Java applications.

5. Perspectives and Conclusion

The shallow dialog analysis model presented in this paper extracts a set of useful features from human dialogs recorded during meetings. A prototype that makes use of the SDA annotations is already running and is being refined. The automation of the various SDA component taggers is under way. The evaluation of the SDA parser will show the current performances on the joint SDA task. Our project also includes a user-based study of the relevance of SDA features to the meeting retrieval application.

6. Acknowledgements

The work presented here is part of the Swiss NCCR on 'Interactive Multimodal Information Management' (IM2), funded by the Swiss National Science Foundation (*http://www im2.ch*). The work pertains specifically to the IM2.MDM module, 'Multimodal Dialog Management' (*http://www.issco.unige.ch/projects/im2/mdm*). We would like to thank the authors of the ICSI-MR corpus, in particular Barbara Peskin and Liz Shriberg, for providing valuable resources and advice.

7. References

Armstrong, S., Clark, A., Coray, G., Georgescul, M., Pallotta, V., Popescu-Belis, A., Portabella, D., Rajman, M. and Starlander, M. (2003). "Natural Language Queries on Natural Language Data: a Database of Meeting Dialogs". Proc. of NLDB'2003, Burg/Cottbus, Germany.

Beeferman, D., Berger, A., Lafferty, J. (1999). "Statistical Models for Text Segmentation". Machine Learning 34(1-3), p. 177-210.

Dhillon, R., Bhagat, S., Carvey, H., and Shriberg, E. (2004). Meeting Recorder Project: Dialog Act Labeling Guide. ICSI Technical Report TR-04-002, Berkeley, CA, USA, February 9, 2004.

Choi, F. (2000). "Advances in domain independent linear text segmentation". Proc. of NAACL'2000, Seattle, WA, USA.

Clark, A. and Popescu-Belis, A. (2004) - Multi-level Dialog Act Tags. Proc. of SIGDIAL'04, Cambridge, MA.

Galley, M., McKeown, K., Fosler-Luissier E. Hongyan, J. (2003). "Discourse Segmentation of Multy-Party Conversation". Proc. of ACL'03, Sapporo, Japan.

Klein, M., Bernsen, N. O., Davies, S., Dybkjær, L., Garrido, J., Kasch, H., Mengel, A., Pirrelli, V., Poesio, M., Quazza, S., and Soria, C. (1998). "Supported Coding Schemes". MATE Project LE4-8370, Deliverable D1.1, http://mate.nis.sdu.dk/about/D1.1/.

McCowan, I., Bengio, S., Gatica-Perez, D., Lathoud, G., Monay, F., Moore, D., Wellner, P., and Bourlard, H. (2003). "Modeling Human Interaction in Meetings". Proc. of ICASSP 2003, Hong Kong, China.

Morgan, N., Baron, D. Bhagat, S., Carvey, H., Dhillon, R., Edwards, J. A., Gelbart, D., Janin, A., Krupski, A., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A, and Wooters, C. (2003). "Meetings about meetings: research at ICSI on speech in multiparty conversations". Proc. of ICASSP'03, Hong Kong, China.

Passonneau, R.J. and Litman, D.J. (1997). "Discourse Segmentation by Human and Automated Means". Computational Linguistics, 23(1), p. 103–140.

Popescu-Belis, A. (2003a). "Evaluation-Driven Design of a Robust Reference Resolution System". Natural Language Engineering, 9(2), p. 1–26.

Popescu-Belis, A. (2003b). Shallow Dialog Analysis: Definition, Annotation, Visualisation. Technical Report IM2.MDM-07, July 2003.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., Carvey, H. (2004). "The ICSI Meeting Recorder Dialog Act Corpus". Proc. of SIGDIAL'04, Cambridge, MA.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Meteer, M., and Van Ess-Dykema, C. (2000). "Dialog Act Modeling for Automatic Tagging and Recognition of Conversational Speech", Computational Linguistics, 26(3), p. 339–371.

Stolcke, A., and Shriberg, E. (1996). "Automatic Linguistic Segmentation of Conversational Speech". Proceedings of ICSLP-96, Philadelphia, PA, p. 1005–1008.

Traum, D.R. (2000). "20 Questions for Dialog Act Taxonomies". Journal of Semantics, 17(1), p. 7–30.

Van Deemter, K., and Kibble, R. (2000). "On Coreferring: Coreference in MUC and Related Annotation Schemes". Comp. Linguistics, 26(4), p. 629–637.