THE CALO MEETING SPEECH RECOGNITION AND UNDERSTANDING SYSTEM

G. Tur[1] *A. Stolcke*[1,3] *L. Voss*[1] *J. Dowding*[2] *B. Favre*[3] *R. Fernandez*[2]
M. Frampton[2] *M. Frandsen*[1] *C. Frederickson*[1] *M. Graciarena*[1] *D. Hakkani-Tür*[3]
D. Kintzing[1] *K. Leveque*[1] *S. Mason*[1] *J. Niekrasz*[2] *S. Peters*[2] *M. Purver*[2]
K. Riedhammer[3] *E. Shriberg*[1,3] *J. Tien*[1] *D. Vergyri*[1] *F. Yang*[1]

[1] SRI International, Menlo Park, CA
[2] Stanford University, Center for the Study of Language and Information (CSLI), Stanford, CA
[3] International Computer Science Institute (ICSI), Berkeley, CA

ABSTRACT

The CALO Meeting Assistant provides for distributed meeting capture, annotation, automatic transcription and semantic analysis of multi-party meetings, and is part of the larger CALO personal assistant system. This paper summarizes the CALO-MA architecture and its speech recognition and understanding components, which include realtime and offline speech transcription, dialog act segmentation and tagging, question-answer pair identification, action item recognition, and summarization.

1. INTRODUCTION

In most organizations, staff spend many hours each week in meetings, and technological advances have made it possible to routinely record and store meeting data. Consequently, automatic means of transcribing and understanding meetings would greatly increase productivity of both meeting participants and nonparticipants. The meeting domain has a large number of subdomains including judicial and legislative proceedings, lectures, seminars, board meetings, and a variety of less formal group meeting types. All of these meeting types could benefit immensely from the development of automatic speech recognition (ASR), understanding, and information extraction technologies that could be linked with a variety of online information systems.

In this paper we present the meeting recognition and understanding system for the CALO Meeting Assistant (CALO-MA) project. CALO-MA is an automatic agent that assists meeting participants, and is part of the larger CALO [1] effort to build a “Cognitive Assistant that Learns and Organizes” funded by the Defense Advanced Research Projects Agency (DARPA). CALO-MA supports multi-party meetings with a variety of information capture and annotation tools. Meetings are recorded via client software running on participants’ laptop computers. The system is aware of each participant’s identity. Meetings may be geographically distributed as long as a broadband internet connection to the server is avaialable (a phone-based interface is being developed as well). The client software captures the participants audio signal, as well as optional handwriting recorded by digital pens. During the meeting, a real-time transcript is available to the particpants to which annotations may be attached. Real-time chat via keyboard input is also supported. All interations are logged in a database, and at the conclusion of the meeting various further automatic annotation and interpretation technologies are initiated, for later browsing via a web-based interface.

Apart from being highly usable in its present form, the system presents a data collection and research experimentation platform to support ongoing research in natural language and user interface technologies. The nature of multi-party interaction and the extreme variability found in meeting genres makes this domain one of the most difficult challenges for speech and natural language processing today. In the following sections we discuss the speech-based component technologies contributing to CALO-MA, including meeting recognition, dialog annotation, action item detection, and summarization. We conclude by pointing out research challenges and directions for future work.

2. CALO-MA ARCHITECTURE

2.1. Meeting capture

An early goal of the CALO-MA project was to allow lightweight data capture. Because of this, highly instrumented rooms were avoided in favor of running on each individual’s Java Runtime enabled computer. Meeting participants can attend meetings using a desktop or laptop running Windows XP/Vista, Linux, or Mac OS X Leopard. Servers for data transport, data processing, and meeting data browsing run on Windows and Linux environments. If scaling is an issue, additional servers can be integrated into the framework to load balance the various tasks. New efforts will allow participants to conference into a meeting via a bridge between the data transport server and the public switched telephone network (PSTN).

During a meeting, client software sends Voice over Internet Protocol (VoIP) compressed audio data to the server when either energy thresholds are met or when a hold-to-talk mechanism is enabled. The data transport server splits the audio: sending one stream to data processing agents and the other stream to remote meeting participants. Other shared data (text chat, file sharing, digital ink, and collaborative text editing) is handled in a similar manner with data going from client to server and then distributed to both processing agents and other meeting participants. Finally, any processing agents that operate in real-time send their data back to the data transport server which relays the data back to the meeting participants.

2.2. Integration with other CALO components

Both during the live meeting and anytime after the meeting, the meeting data transport server makes available all meeting data to interested parties using XML-RPC interfaces. This allows both local and distributed users and processing agents to access the data in a language neutral way. Meeting processing agents that are order dependent register with a meeting post processor framework to ensure that processing order is enforced (e.g. speech transcription, prosodic feature detection, dialog act recognition, action item detection, decision detection, topic boundary detection, meeting summarization, and email notification to meeting participants) and processing load is balanced.

Any CALO components outside the meeting processing framework (including the meeting browser) can send XML-RPC queries to the meeting data transport server. Those components can then perform further integration with user desktop data to facilitate additional machine learning (a focus of many other CALO processes) or present other visualizations of the data to the user.

2.3. Meeting Browser

After the meeting has been fully processed, an email is sent out to all meeting participants. This email includes a static version of the meeting data and a link to a website where the data can be browsed dynamically from any Internet enabled device. Once connected to the browser, the user can select a meeting to review and browse any of the data: both user-generated (e.g. shared files and notes) and auto-generated (e.g. detected action items and summaries). As all data is time stamped, a user can click on any data element and bring up the corresponding section of the transcript to read what was being discussed at that time. To overcome any speech transcription errors, all transcript segments can be selected for streaming audio playback.

3. MEETING TRANSCRIPTION

The audio stream from each meeting participant is transcribed into text using two separate recognition systems. A real-time recognizer generates "live" transcripts with 5-15 seconds of latency for immediate display (and possible interactive annotation) in the CALO user interface. Once the meeting is concluded, a second, off-line recognition system generates a more accurate transcript for later browsing and serves as the input to the higher-level processing step described in the following sections.

The off-line recognition system is a modified version of the SRI-ICSI NIST meeting recognizer [2]. It performs a total of 7 recognition passes, including acoustic adaptation and language model rescoring, in about 4.2 times real-time (on a 4-core 2.6GHz Opteron server). The real-time recognition systems consists of an online speech detector, causal feature normalization and acoustic adaptation steps, and a sub-real-time trigram decoder. On a test set where the off-line recognizer achieves a word error rate of 26.0%, the real-time recognizer obtains 39.7%.

4. DIALOG ACT SEGMENTATION

Output from a standard speech recognition system typically consists of an unstructured stream of words lacking punctuation, capitalization, or formatting. Sentence segmentation for speech enriches the output of standard speech recognizers with this information. Previous work on sentence segmentation used lexical and prosodic features from news broadcasts and spontaneous telephone conversations [3]. Work on multi-party meetings has been more recent.

Following the similar approaches taken for sentence segmentation (such as [3] we treated the segmentation task as a boundary classification problem. To this end we built hybrid models combining hidden event language models (HELMs) with a discriminative classifier, namely Boosting exploiting additional prosodic features such as pitch, energy, pause, and duration characteristics. The main idea in this hybrid approach is converting the posterior probabilities obtained from Boosting into state observation likelihoods.

In order to exploit the sentence boundary tagged meeting corpora as obtained from other projects such as ICSI and AMI, we also proposed model adaptation [4] and semi-supervised learning [5] techniques for this task.

5. DIALOG ACT TAGGING

A dialog act is a primitive abstraction or an approximate representation of the illocutionary force of an utterance, such as *question* or *backchannel*. Dialog acts are designed to be task independent. The main goal of dialog acts is to provide a basis for further discourse analysis and understanding. For example, dialog acts can be used to extract the action items or question/answer pairs in a meeting as discussed later. Note that dialog acts can be organized in a hierarchical fashion. For instance, statements can be further categorized as *fact* or *opinion*.

In this project, we followed the MRDA standard [6], which especially focuses on multi-party meetings. For example, it includes a set of labels for floor management mechanisms, such as *floor grabbing* and *holding*, which are common in meetings. In total it has 11 general (such as question) and 39 specific (such as yes/no question) dialog act tags.

For the CALO project, dialog act tagging is framed as an utterance classification problem using Boosting. More specifically, we built 3 different taggers: i) for capturing high level dialog act tags (namely, statement, question, disruption, floor mechanism, and backchannel). To build this model, we only used lexical features. ii) for detecting action motivators since they are shown to help action item extraction [7]. For this, we only considered suggestion, command, and commitment dialog act tags using only lexical features. iii) for detecting agreement and disagreement dialog act tags for single word utterances, such as *yeah* or *okay*. For this task we used prosodic and contextual information.

6. ATTENTION, ADDRESSING, AND REFERENCE

An important intermediate step in the analysis of meeting conversations is to determine the entities and individuals to which the particpants are speaking, listening, and referring. This processing step is essential to downstream summarization in that it provides necessary situational and discourse context for doing interpretation. To summarize decisions, for example, detecting requests, promises, and disagreements alone is inadequate – the system must understand who performs these actions, with whom, and in reference to whom.

In contrast to two-party dialogues, multi-party meetings present new challenges to this problem. Consider the following sentence: "*and um if **you** can get that binding point also maybe with an example that would be helpful for **John** and **me***." The challenge here is to use gaze and dialogue context to identify the person marked by "*you*" and to resolve any linguistic or gestural references to individuals or present objects such as "*John*".

The addressee identification task is typically approached as an utterance-level classification where some subset of the participants in the meeting are identified as addressees. [8] used a combination of lexical features of the utterance and gaze features from each participant to detect addressee in 4-person meetings using Bayesian networks. For their system, using cues from multiple modalities proved most effective. But to overcome situations where video is unavailable or gaze tracking is difficult, the CALO system leverages deeper structural, durational, and lexical features taken from the speech transcript only [9]. Additionally, conditional random field (CRF) is used to model discourse context, explicitly modelling forward and backward dependencies in the dialogue.

The need for deeper linguistic understanding of the dialogue is particularly evident with the problem of pronoun referentiality. Especially in human-human conversation, words like "you" and "it" are frequently used in non-referential and indefinite senses (e.g., "it's raining" or "you really need a umbrella in Seattle"). As a preprocessor to a downstream reference resolution system, [10] used a rule induction classifier to detemine whether “it” was referential in meetings from the ICSI corpus. For the CALO system, a CRF is used to perform the related classification of second-person pronouns into referential and generic senses [9].

7. TOPIC IDENTIFICATION AND SEGMENTATION

Identifying topic structure provides a user with the basic information of *what* people talked about *when*, and can also feed into further processing (enabling topic-based summarization, browsing and retrieval). Topic modeling can be seen as two sub-tasks: *segmentation*, dividing the speech data into topically coherent units (answering the “when” question), and *identification*, extracting some representation of the topics discussed therein (the “what”). While both tasks have been widely studied for broadcast news, meetings pose a different problem, being typically more coherent overall with less sharp topic boundaries.

Segmentation is typically approached by tracking changes in lexical distribution (following text-based methods e.g. [11]). Many variants of this approach, either using lexical distribution function directly, or incorporating it into a discriminative classifier, have been applied successfully to meeting transcripts [12, 13, among others].

As there is more to meeting dialogue than the words it contains, performance can often be improved by adding features of the interaction itself, from simple prosodic features to higher-level changes in discourse structure and the behavior of the participants [13, 14, among others] or even exploiting the participants’ note-taking behaviour [15]. The identification problem can then be approached as a separate step after segmentation using supervised discriminative techniques to classify topic segments according to a known list of existing topics.

However, there may be reason to treat the two as joint problems: segmentation can depend on the topics of interest, and these topics are not necessarily known beforehand. [16] investigated the use of Latent Semantic Analysis, learning vector-space models of topics and using them as the basis for segmentation, but accuracy was low.

Instead, we therefore use a generative topic model with a variant of Latent Dirichlet Allocation to learn models of the topics automatically, without supervision, while simultaneously producing a segmentation of the meeting [17]. Segmentation performance is competitive with that of a lexical cohesion approach (P_k between 0.27 and 0.33 on the ICSI Meeting Corpus) and robust to ASR errors, while the topic models learnt can be used to extract lists of descriptive keywords which human judges rate well for coherence.

8. ACTION ITEM AND DECISION EXTRACTION

Amongst the most commonly requested outputs from meetings are lists of the *decisions* made, and the *action items* assigned (public commitments to perform particular tasks). This requires two steps: *detection* of the task or decision discussion, and *summarization* or *extraction* of some concise descriptive representation (for action items, typically the task itself together with the due date and responsible party; for decisions, perhaps the issue involved and the resolved course of action).

One approach to detection might be as binary classification: classifying utterances as decision- or action-item-related or not. In email text, this has shown good performance: F-scores around 0.8 when detecting relevant messages, and 0.6-0.7 for individual sentences [18, 19]. However, applying this to meeting dialogue (see [20] for decisions, [21] for action items) gives poor results, with F-scores around 0.3-0.35. Given the nature of meeting dialogue, this may not be surprising: tasks and decisions tend not to be contained within individual sentences, but are defined incrementally, with commitment being established through the interaction and the related discourse structure.

The CALO system therefore takes a structural approach to detection: utterances are classified according to their role in the commitment process (e.g. task definition, agreement, acceptance of responsibility), and then action item or decision discussions detected from patterns of these roles. This significantly improves detection performance, achieving F-scores around 0.45 for action items [22] and 0.6 for decisions [23].

The extraction problem can be approached by parsing: [18]’s email system builds logical forms from the relevant sentences and then generates descriptions via a realizer. With spoken language and ASR output, the parsing problem is more difficult; a parsing-based approach works well in some cases, but in general gives no improvement over a baseline of returning the 1-best utterance transcripts for the relevant utterances [22]. Better performance is now being shown using a lexical approach, extracting important words using classifiers or sequence models [24].

9. MEETING SUMMARIZATION

The goal of summarization is to create a shortened version of a text or speech while keeping the important points. While textual document summarization is a well studied topic, speech summarization (and in particular meeting summarization) is an emerging research area, and apparently very different from text or broadcast news summarization. While hot spot detection, action item extraction, dialog act tagging, and topic segmentation and detection methods can be used to improve summarization, there are also preliminary studies using lexical, acoustic, prosodic, and contextual information.

In text or broadcast news summarization, the dominant approach is extractive summarization where “important” sentences are concatenated to produce a summary. For meeting summarization it is not clear what constitutes an important utterance. In an earlier study the sentences having the most number of frequent content words are considered to be important [25]. Using the advances in written and spoken document extractive summarization [26], some recent studies focused on feature-based classification approaches [27], while others mainly used maximum marginal relevance (MMR) [28] for meeting summarization [27, 29]. MMR iteratively selects utterances most relevant to a given query q, which is expected to encode the user’s information need, while trying to avoid utterances redundant to the already selected ones. Due to the lack of a query, the common approach for meetings has been using the centroid vector of the meeting as the query [27].

Our summarization work mainly focused on investigating the boundaries of extractive meeting summarization in terms of different evaluation measures[30]. We proposed a method to compute ”oracle” summaries that extract a set of sentences maximizing the ROUGE performance measure [30]. We observed much lower oracle performances for meetings than for text indicating that the extractive approach is unlikely to work as well as for textual news. Moreover, in meetings the information is usually distributed across multiple sentences, making simple extraction ineffective. We also presented a very simple baseline (that extracts the longer sentences) that beats most of the proposed approaches on ROUGE scoring, suggesting that this metric might not be very suitable for meetings and evaluation methods that are specifically designed for meeting summarization can be helpful. Finally, current extractive methods force the user to recontextualize the information by looking at the meeting itself, a rather time consuming task. Keywords or other representations including action-item and decision lists might be more informative in less time. We proposed using a set of keywords and key phrases extracted from meetings, as a query for the MMR algorithm [31]. While this resulted in a better summarization performance, it also allows the users to interactively modify the set of keywords, as well as the length of the summary, according to their information needs.

10. CONCLUSIONS

We have presented a system for automatic processing of tasks involving multi-party meetings. Progress in these tasks, from low-level transcription to higher-level shallow understanding functions, such as action item extraction and summarization, has a potentially enormous impact on human productivity in many professional settings. Further integration of these tasks and multiple potential modalities, such as speech and video, is part of the future work.

Acknowledgments: This material is based upon work supported by the DARPA CALO (NBCHD-030010) funding. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

11. REFERENCES

[1] "SRI Cognitive Agent that Learns and Organizes (CALO) Project", http://www.ai.sri.com/project/CALO.

[2] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, F. Grézl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, "Further progress in meeting recognition: The ICSI-SRI Spring 2005 speech-to-text evaluation system", in S. Renals and S. Bengio, editors, *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005*, vol. 3869 of *Lecture Notes in Computer Science*, pp. 463–475. Springer, 2006.

[3] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics", *Speech Communication*, vol. 32, pp. 127–154, 2000.

[4] S. Cuendet, D. Hakkani-Tür, and G. Tur, "Model adaptation for sentence segmentation from speech", *in Proceedings of the IEEE/ACL SLT Workshop*, Aruba, 2006.

[5] U. Guz, S. Cuendet, D. Hakkani-Tür, and G. Tur, "Co-training using prosodic and lexical information for sentence segmentation", *in Proceedings of the Interspeech*, Antwerp, Belgium, 2007.

[6] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI Meeting Recorder Dialog Act (MRDA) Corpus", *in Proceedings of the SigDial Workshop*, Boston, MA, May 2004.

[7] F. Yang, G. Tur, and E. Shriberg, "Exploiting dialog act tagging and prosodic information for action item identification", *in Proceedings of the ICASSP*, Las Vegas, NV, 2008.

[8] N. Jovanovic, R. op den Akker, and A. Nijholt, "Addressee identification in face-to-face meetings", *in Proceedings of the EACL*, pp. 169–176, Trento, Italy, 2006.

[9] S. Gupta, J. Niekrasz, M. Purver, and D. Jurafsky, "Resolving "you" in multi-party dialog", *in Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, 2007.

[10] C. Müller, "Automatic detection of nonreferential *It* in spoken multi-party dialog", *in Proceedings of the EACL*, pp. 49–56, Trento, Italy, 2006.

[11] M. Hearst, "TextTiling: Segmenting text into multi-paragraph subtopic passages", *Computational Linguistics*, vol. 23, pp. 33–64, 1997.

[12] S. Banerjee and A. Rudnicky, "A TextTiling based approach to topic boundary detection in meetings", *in Proceedings of the ICSLP*, Pittsburgh, PA, September 2006.

[13] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation", *in Proceedings of the ACL*, 2003.

[14] M. Georgescul, A. Clark, and S. Armstrong, "Exploiting structural meeting-specific features for topic segmentation", *in Actes de la 14ème Conférence sur le Traitement Automatique des Langues Naturelles*, Toulouse, France, June 2007. Association pour le Traitement Automatique des Langues.

[15] S. Banerjee and A. Rudnicky, "Segmenting meetings into agenda items by extracting implicit supervision from human note-taking", *in Prooceedings of the International Conference on Intelligent User Interfaces (IUI'07)*, Honolulu, Hawaii, January 2007. ACM.

[16] A. Popescu-Belis, A. Clark, M. Georgescul, D. Lalanne, and S. Zufferey, "Shallow dialogue processing using machine learning algorithms (or not)", in S. Bengio and H. Bourlard, editors, *MLMI, Revised Selected Papers*, vol. 3361 of *Lecture Notes in Computer Science*, pp. 277–290. Springer, 2005.

[17] M. Purver, K. Körding, T. Griffiths, and J. Tenenbaum, "Unsupervised topic modelling for multi-party spoken discourse", *in Proceedings of the COLING-ACL*, pp. 17–24, Sydney, Australia, July 2006. Association for Computational Linguistics.

[18] S. Corston-Oliver, E. Ringger, M. Gamon, and R. Campbell, "Task-focused summarization of email", *in Proceedings of the ACL Workshop Text Summarization Branches Out*, 2004.

[19] P. N. Bennett and J. G. Carbonell, "Combining probability-based rankers for action-item detection", *in Proceedings of the HLT/NAACL*, pp. 324–331, Rochester, NY, April 2007. Association for Computational Linguistics.

[20] P.-Y. Hsueh and J. Moore, "What decisions have you made?: Automatic decision detection in meeting conversations", *in Proceedings of NAACL/HLT*, Rochester, New York, 2007.

[21] W. Morgan, P.-C. Chang, S. Gupta, and J. M. Brenier, "Automatically detecting action items in audio meeting recordings", *in Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pp. 96–103, Sydney, Australia, July 2006. Association for Computational Linguistics.

[22] M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, S. Noorbaloochi, and S. Peters, "Detecting and summarizing action items in multi-party dialogue", *in Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, September 2007.

[23] R. Fernández, M. Frampton, P. Ehlen, M. Purver, and S. Peters, "Modelling and detecting decisions in multi-party dialogue", *in Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pp. 156–163, Columbus, OH, 2008. Association for Computational Linguistics.

[24] R. Fernández, M. Frampton, J. Dowding, A. Adukuzhiyil, P. Ehlen, and S. Peters, "Identifying relevant phrases to summarize decisions in spoken meetings", *in Proceedings of Interspeech'08*, Brisbane, 2008.

[25] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen, "Meeting browser: Tracking and summarizing meetings", *in Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, June 1998.

[26] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization", *in Proceedings of the Interspeech*, Lisbon, Portugal, September 2005.

[27] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings", *in Proceedings of the Interspeech*, Lisbon, Portugal, September 2005.

[28] J. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries", *Research and Development in Information Retrieval*, pp. 335–336, 1998.

[29] S. Xie and Y. Liu, "Using Corpus and Knowledge-Based Similarity Measure in Maximum Marginal Relevance for Meeting Summarization", *in Proc. ICASSP, Las Vegas, USA*, 2008.

[30] K. Riedhammer, D. Gillick, B. Favre, and D. Hakkani-Tür, "Packing the Meeting Summarization Knapsack", *in Proc. Interspeech, Brisbane, Australia*, 2008.

[31] K. Riedhammer, B. Favre, and D. Hakkani-Tür, "A Keyphrase based approach to interactive meeting summarization", *in In submission*, 2008.