Annual Report for Period:09/2001 - 09/2002
Principal Investigator: Morgan, Nelson .
Organization: Internationl CompSci Inst
Title:
ITR/PE+SY:Mapping Meetings: Language Technology to make Sense of Human Interaction

Submitted on: 05/31/2002
Award ID: 0121396

Project Participants

Senior Personnel

Name: Morgan, Nelson
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Ostendorf, Mari
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Stolcke, Andreas
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Ellis, Daniel
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Kirchhoff, Katrin
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Renals, Steve
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Shriberg, Elizabeth
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: McKeown, Katherine
Worked for more than 160 Hours: Yes
Contribution to Project:

Post-doc

Graduate Student

Name: Baron, Don
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Coulis, Costas

E. Shriberg, A. Stolcke, D. Baron, "Can Prosody Aid the Automatic Processing of Multi-Party Meetings? Evidence from Predicting Punctuation, Disfluencies, and Overlapping Speech
", M. Bacchiani, J. Hirschberg, D. Litman, & M. Ostendorf eds. Proc. ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding, Red Bank, NJ
, p. 139, vol. , (2001). Published

D. Baron, E. Shriberg, A. Stolcke, "Automatic Punctuation and Disfluency Detection in Multi-Party Meetings Using Prosodic and Lexical Cues.
", ICSLP-02, p. , vol. , (2002). Submitted

T. Pfau, D.P.W. Ellis, A. Stolcke, "Multispeaker Speech Activity Detection for the ICSI Meeting Recorder
", Proc. IEEE Automatic Speech Recognition and Understanding Workshop, Madonna di Campiglio, Italy.
, p. 0, vol. , (2001). Published

Sonali Bhagat, Ashley Krupski, Raj Dhillon, Elizabeth Shriberg, "Guide for Labeling Dialog Acts in Multi-Party Meetings", ICSI Technical Report, p. , vol. , (2002). In Preparation

Books or Other One-time Publications

Don J. Baron, "Prosody-Based Automatic Detection of Punctuation and Interruption Events in the ICSI Meeting Recorder Corpus
", (2002). Thesis, Accepted
Editor(s): N/A
Bibliography: Master's Thesis, Dept of EE
University of California, Berkeley.

Web/Internet Site

Other Specific Products

Contributions

Contributions within Discipline:

Contributions

Wider use and distribution

Both our prosodic database (including recognition output) and dialog act annotations are being distributed to others both inside and beyond the ITR project. For example, the prosodic data and recognition output is being used by many of the internal sites, as well as by a group on speaker identification at the 2002 Johns Hopkins summer workshop. Dialog act annotations will be distributed when ready, to researchers at UW and Lucent for topic segmentation, and a group in Switzerland interested in dialog modeling.

Contributions to Other Disciplines:

Contributions to Human Resource Development:

Contributions to Resources for Research and Education:

Contributions Beyond Science and Engineering:

Special Requirements

Special reporting requirements: None
Change in Objectives or Scope: None
Unobligated funds: less than 20 percent of current funds
Animal, Human Subjects, Biohazards: None

Categories for which nothing is reported:

Organizational Partners
Activities and Findings: Any Training and Development
Activities and Findings: Any Outreach Activities
Any Web/Internet Site
Any Product
Contributions: To Any Other Disciplines
Contributions: To Any Human Resource Development
Contributions: To Any Resources for Research and Education
Contributions: To Any Beyond Science and Engineering

ACTIVITIES

In the first 9 months of our project, we have begun several studies that will ultimately lead to the kind of "Meeting Maps" that we described in our original proposal. The major categories of this effort have been:

1) Acoustic processing for speaker location, particularly from distant mics in a meeting room

2) Topic segmentation, both manual and automatic

3) Summarization, both building examples and developing automatic approaches

4) Modeling prosodic "beyond the words" events; building a database for 13 meetings, consisting of relevant features for determination of punctuation, disfluency, dialog acts, and overlaps

5) Automatic determination of agreement and disagreement in meetings

6) Development of information retrieval software relevant to a query-dependent content map.

The sections below will elaborate on these themes.

1) Speaker location (Columbia):
We have been investigating the information available from the table top mics used in the ICSI recording setup. We would like to be able to identify each speaker turn based on these signals alone, and we have used the cross-correlation between different mic pairs to get constraints on speaker location. Another problem arising from this data is the cross-talk among the head-mounted microphones: we have been investigating automatic methods for the dynamic estimation of this coupling with a view to cross-cancelling each speaker from the microphones of others. Finally, we have begun some investigations into browsing and visualization tools, to investigate if there are informative patterns in the speaker activity when viewed on the scale of entire meetings.

2) Topic Segmentation (Columbia and UW):

A key issue for generating summaries of segments is to be able to identify topic shifts during meetings (or, equivalently, finding regions with coherent topics). It is not only desirable for indexing purposes, but also needed for automatic summarization. We expect that most summarization techniques would not perform well without identifying topic boundaries, since many meetings span many unrelated subjects. This is the case for most of the meetings in the Meeting Recorder project.

We have begun segmenting meetings by hand (nine so far) and initial results have shown an encouraging agreement between humans. This work helped us to identify interesting features,

including lexical dissimilarity between segments and lexical cues present at boundaries. We have started to prototype a system using segmentation techniques that are known to perform well on expository text (Hearst 1994) In order to develop a system that can perform robustly, we are currently investigating how to integrate information about speaker activity during the meeting. Topic shifts are often paired with a change of speakers who are involved in the discussions. Our preliminary analysis of the meeting data confirmed this behavior.

Previous work on topic segmentation has primarily been on Broadcast News data, and although some of the techniques may transfer, our initial investigation has shown that news is very different in nature from meetings. Topic boundaries are much more clear in news than in meetings, and since the meetings tend to be technical (in our corpus) sub-topic boundaries are not always easy to identify by non-experts. Consequently a first step in the project (in progress) is the definition of topic segmentation and labeling guidelines.

While meetings pose significant challenges for topic segmentation, including much higher ASR error rates that reduce the usefulness of lexical cues (and conversely increases the importance of prosodic cues) they do bring a new dimension to the problem in terms of multi-speaker interactions. New features can be extracted, such as rate of overlaps and average turn lengths. Based on inspection of a small amount of topic-labeled data, our hypothesis is that such cues will be useful, including long monologs serving as a negative cue to topic boundaries. For positive cues, we expect that it will be important to combine such information with speech act labels.

Because of the constantly changing topics in the meetings, we do not want to assume a set of predefined topic labels, but rather will use unsupervised techniques to extract this information. We plan to extend approaches that were applied in BN (Sista et al. 2002). Although past approaches are promising, they have dimensionality problems, so we have begun looking into dimensionality reduction techniques which result into low-dimensional continuous projections. Possibilities include latent semantic analysis (LSA) (Bellegarda 1998) and neural networks.

An important goal of the project is the extraction of information directly from speech, avoiding the current path of speech → text → information. Information extraction (IE) from speech is significantly different than IE from text (which has a rich body of work) for two main reasons: 1) ASR errors can be directly modeled in the IE process, and 2) non-lexical features like prosody, speaker interactions etc. can augment the recognized word sequence. For ASR error modeling, we have begun developing a Bayesian network framework that models the temporal dependency of ASR errors. Our current focus on non-lexical features, as mentioned above, is on speaker interactions.

3. Summarization of meeting transcriptions (Columbia, UW):

The first stage in investigating automatic summarization of meetings is to analyze meeting recordings and determine what kind of summaries we can realistically aim to produce. We have built examples of the two kinds of summaries we plan to generate: high level summaries indicative of the topics discussed during the whole meeting, and informative summaries of a particular segment of interest. We have identified the issues involved in creating such summaries, and investigated possible solutions.

One approach for generated informative summaries of particular segments of the meeting would be to use statistical techniques to extract sentences from the transcription of the meeting, then remove speech disfluencies using existing techniques. We would then need to chunk the text to identify noun phrases and verb phrases; for this purpose, we have performed experiments on shallow parsing, although to date the parsers we have experimented with do not yield good results. Finally, we plan to use language generation to reformulate the wording of the summary.

In addition to the problem of generating summaries from manually transcribed data, we are concerned with the need to generate summaries from errorful transcriptions as produced by ASR. Our approach is to develop a unified model for speech summarization which extracts relevant information from ASR output and produces a readable, coherent summary in one single step. Most previous approaches to speech summarization have concentrated on extracting simple key words or phrases from ASR output; potential users of a meeting browser, however, may require 'executive summaries' of a meeting that are not only informative but also coherent, i.e. grammatical and logically connected. This might be done by first applying an automatic speech recognizer to the signal, filtering the output to extract relevant information and passing the extracted words or phrases to a speech generation component which produces a 'readable' summary.

We have begun to develop an alternative approach which is based on a noisy channel model. Under this approach, the summary and the corresponding ASR output are assumed to be related through a noisy channel. The goal is to recover the most probably message given the ASR output. In order to transform the ASR output to a coherent summary a number of basic operations can be performed, such as additions, deletions, substitutions or permutations of words can be performed. The sequence of these operations can be represented as hidden states in a probabilistic model. The model can be trained on observed sequences of operations, which requires a training corpus where summaries and ASR output have been aligned and transformation operations have been made explicit. The parameters of the hidden states can then be learned via Expectation-Maximization. The learned model is then used on unseen data to produce new summaries.

This application is similar to the use of noisy channel models in machine translation Brown 1993) text summarization (Knight et al. 2000) and question answering (Radev 2002). Its advantage is that the two steps of information extraction and speech generation can be collapsed and performed by a single set of transformations. Moreover, it reduces the need for hand-crafted rules or other forms of expert knowledge by learning the summarization model from data.

As a first step towards producing training data for this model we are currently comparing and evaluating different annotation schemes used for text summarization in order to adapt them for speech summarization annotation.

4. Prosodic modeling of beyond-the-words events (SRI,ICSI):

Automatic processing and understanding of meetings relies on knowledge about how to segment the speech into fluent utterances, how each utterance functions in the dialog, and how speakers

interact with each other. In studies leveraged with the DARPA communicator project, we have studied the use of automatic prosody modeling for a range of tasks on meeting data, including: automatic punctuation, detection of disfluencies, discrimination of very simple dialog acts (such as statements versus questions), and speaker interaction phenomena such as overlap and turn-taking behavior. We have constructed a prosodic database for 33 meetings, that includes pause, duration, stylized pitch, voicing, energy, and a variety of other contextual features--for each word in the database. Features are completely automatically extracted, and normalized for speaker and context. This information is aligned with punctuation, disfluency, minimal dialog act, and overlap information. We computed parallel databases for true and automatically recognized words, to assess effect of word errors in both training and testing. Prosodic features, as modeled by decision trees, were combined with language model features to assess relative contribution of each as well as combined performance. In a second area of research, we have begun to develop a dialog act annotation system for the meeting data, in which all sentence-like units are labeled with respect to both their structure and function in the conversation. We began with a labeling manual developed for the Switchboard data, and are making significant modifications to appropriately represent the much richer types of structures and functions that occur in the meeting data (Bhagat et al. 2002).

5) Automatic determination of agreement and disagreement in meetings (UW):

To support browsing and summarization of automatically transcribed meetings, we developed a classifier to automatically recognize utterances as agreements, disagreements or neither (other). The overall approach to classifying utterances (or, "spurts" of speech) is to extract word-based and prosodic features of the spurt and combine these in a decision tree classifier (Breiman et al. 1984) The word-based features use n-gram language models trained using "lightly supervised" clustering techniques, i.e. keyword assignment rules derived from a small amount of labeled data are used to assign initial clusters to a large unlabeled data set, which are iteratively refined. The prosodic features are taken from the database developed by Shriberg and colleagues at ICSI. To reduce our initial candidate feature set to a smaller, optimal set, we used an iterative feature selection algorithm that involved running multiple decision trees (Shriberg et al. 2000). We showed that the cost of hand-labeling data can be minimized by using unsupervised training on a large unlabeled data set combined with supervised training with a small amount of hand-labeled data. Given accurate transcriptions, the classifier gives only a 5% confusion rate between agreements and disagreements with recovery rates of over 80%.

Current work involves assessing performance of the classifier with automatic speech recognition (ASR) output, with the expectation that prosody will be much more important because ASR errors make lexical cues less reliable. The next step will be to look at co-training techniques to better leverage unlabeled data, and to investigate use of word confusion networks to better model ASR uncertainty.

6) Information retrieval software for query-dependent content maps (ICSI):

We have been exploring the use of a spoken document retrieval system as a basis for the development of a (query-dependent) content map. An existing spoken document retrieval system Renals et al. 2000) has been modified to handle multi-channel audio, characteristic of the

meetings domain. Automatic segmentation, based on an overlapping window, is used (with additional constraints based on speaker overlap), with recombination into longer segments occurring dynamically at query time. Summarization of these segments has been implemented using baseline techniques derived from text retrieval. Additionally by keeping a record of meeting segments marked as relevant by a user (perhaps with adjusted boundaries) we are collecting data that may be usable for training high compression summarizers. A software tool has been constructed for this purpose.

Publications

1. D. Baron, E. Shriberg, A. Stolcke. 2002.Automatic Punctuation and Disfluency Detection in Multi-Party Meetings Using Prosodic and Lexical Cues. Submitted to *ICSLP-02*.

2. J. Bellegarda.. 1998. Exploiting latent semantic information in statistical language modeling. In *Proceedings of the IEEE*. 88:8. Aug. 1998. pp1279-1296.

3. Sonali Bhagat, Ashley Krupski, Raj Dhillon, Elizabeth Shriberg. 2002. *Guide for Labeling Dialog Acts in Multi-Party Meetings*. ICSI Technical Report. in preparation.

4. L. Breiman, J. Freidman, R. Olshen and C. Stone. 1984. *Classification And Regression Trees*. Wadsworth International Group. Belmont, CA

5. P.F. Brown et al. 1993. The mathematics of statistical machine translation: parameter estimation. In *Proceedings of Computational Linguistics*. 19:263-311.

6. M. A. Hearst. 1994. Multi-paragraph segmentation of expository text. *In Proceedings of ACL '94*.pp. 9-16.

7. K. Knight and D. Marcu. 2000. Statistics-based summarization step one: sentence compression. In *Proceedings of AAAI*. pp.703 - 710.

8. T. Pfau, D.P.W. Ellis, A. Stolcke. 2001. Multispeaker Speech Activity Detection for the ICSI Meeting Recorder. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Italy.

9. D. Radev et al. 2002. Probabilistic question answering on the web. In *Proceedings of 2002 WWW Conference.*

10. S. Renals, D. Abberley, D. Kirby, and T. Robinson. 2000. Indexing and Retrieval of Broadcast News. *Speech Communication*. 32. 5--20.

11. E. Shriberg, A. Stolcke, D. Baron. 2001. Can Prosody Aid the Automatic Processing of Multi-Party Meetings? Evidence from Predicting Punctuation, Disfluencies, and Overlapping Speech. In M. Bacchiani, J. Hirschberg, D. Litman, & M. Ostendorf eds., *Proc. ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*. pp. 139-146, Red Bank, NJ.

12. E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur. 2000. Prosody-Based Automatic Segmentation of Speech into Sentences and Topics. In T. Robinson and S. Renals, eds. *Speech Communication*. 32:1-2. Sep. 2000. pp.127-154.

13. S. Sista, R. Schwartz, T. R. Leek, and J. Makhoul. 2002. An Algorithm for Unsupervised Topic Discovery from Broadcast News Stories. In *Proceedings of the Human Language Technology Conference.* to appear.

In the first 9 months of our projects we have made a number of key observations, though it is far too early to be able to make strong pronouncements about their ultimate significance. However, we can briefly list here some of our findings thus far:

1) We found that we were able to determine speaker location in 3-space with reasonable accuracy using 4 tabletop microphones.

2) Finding useful features for meeting segmentation, and correlating topic shift with speaker change.

3) Prosody is important for classification of dialog acts; dialog act sequences in meetings are more complex than in 2-party telephone conversations.

4) Methods developed for Automatic Speech recognition of telephone conversations also work for meeting recordings.

5) Modeling agreement/disagreement can leverage a small amount of labeled data w/ larger amounts of unsupervised data.

The sections below will elaborate on these themes.

1) Speaker location (Columbia):

Using the four microphones placed in the middle of the table top we have been able to solve for speaker locations in 3-space, and use the resulting positions to refine the assumed positions of the microphones in an iterative procedure (see http://www.ee.columbia.edu/~dpwe /mtgrcd /spkrlocs.pdf). We hope this will develop into a high-accuracy speaker activity inference system. Recovering the cross-correlation between head-mounted microphones is proving more difficult: we have been able to process simulated data successfully, but the noise present in the real data is causing problems. Our preliminary experiments with meeting-level turn visualization have been very encouraging: there are clear patterns visible in this data, which we hope to encapsulate in features for automatic coarse-level segmentation and classification of meeting activity (see http://www.ee.columbia.edu/~dpwe/mtgrcdr/mtgactivity.gif).

2) Features for meeting segmentation (Columbia):

We have identified interesting features, including lexical dissimilarity between segments and lexical cues present at boundaries. We have started to prototype a system using segmentation techniques that are known to perform well on expository text (Hearst, 1994). In order to develop a system that can perform robustly, we are currently investigating how to integrate information about speaker activity during the meeting. Topic shifts are often paired with a change of speakers who are involved in the discussions. Our preliminary analysis of the meeting data confirmed this behavior.

3) Prosodic modeling and dialog act annotation (SRI, ICSI):

Our results (to be described in Don Baron's thesis, thesis, as well as in a workshop paper (Shriberg et al., 2001) and an ICSLP submission (Baron et al. 2002)) show that prosody, even with the crude automatic features used on our studies, significantly improves performance over lexical features for both true words and ASR output. Furthermore, for some tasks, prosody alone is a much better cue than words alone. Results also show interestingly that best performance when testing on recognized words comes from training on prosodic data based on true-word alignments. That is, it is better to have reliable word boundaries in training than to match the training and test data for word error rates. Another very interesting finding was that some speakers are simply more prosodically effective than others. By training and testing on individual speaker data for frequent speakers, we found that for these speakers prosody is a better cue to punctuation than are true words, while for other speakers the opposite is true. Results on the effort to annotate dialog acts in meetings have revealed that unlike two-party telephone conversations, meetings involve complex dialog act sequences, such as nesting, multiple simultaneous dialogs, and so on. And because the participants know each other and share significant sophisticated world knowledge, there is a much greater use of irony, emotion, dispreferred responses, and other interesting phenomena in this data.

4) Recognition of conversational speech from meetings (SRI,ICSI):

We verified that the main algorithms and techniques used in conversational telephone speech recognition are equally effective in recognition of meeting speech, especially when recognizing from close-talking microphones. For far-field microphones, we see significantly higher error rates, but were able to develop several techniques to improve our baseline recognizer. The most effective approaches were supervised adaptation to the meeting channel, and noise-filtering of the far-field signal using algorithms borrowed from ICSI's Aurora-2 front end (citation). For the case of unknown speaker segmentation we found that automatic segmentation and clustering resulted in only a small penalty compared to using ideal information. SRI/ICSI was the only team participating in the 2002 NIST meeting recognition evaluations; we obtained a 30.6% word error rate (WER) on close-talking channels, and 61.6% WER on a table-top microphone channel (for a mix of 8 meetings from four different sites).

5) Modeling agreement/disagreement (UW):

We found that it is possible to leverage unsupervised clustering techniques in combination with a small amount of data to greatly reduce hand labeling costs for modeling agreement and disagreement. This is extremely important for generalizing the methods to new sources and styles of meetings, since it is impractical to label large amounts of data.