1 Activities Introduction

In months 10 through 21 of our NSF-ITR on Mapping Meetings, we expanded our efforts significantly over the initial period, which had been used to ramp up the project. Our goals remained roughly the same: to learn to describe spoken language in meetings at varying levels, so that systems for information retrieval, extraction, and summarization could subsequently function most effectively. This year's report will describe efforts at the University of Washington, SRI, ICSI, and Columbia in the following areas:

1. Speaker Separation - separating multiple voices from meetings
2. Multi-speaker Language Models
3. Detecting Important Regions (”hot spots”, emphasis, and agreement/disagreement)
4. Dialog Acts
5. Topic Detection, Segmentation and Classification
6. Discourse Markers
7. Summarization
8. Related Meeting Work

The sections below will elaborate on these themes.

2 Speaker Separation [Columbia]

We have been investigating the problem of separating multiple, overlapping voices in meeting recordings. Our basic approach is to use the constraints implicit in a statistical model of the speech (such as is used in a speech recognizer) to help recover from data lost due to signal overlap. We have looked both at multiple-microphone situations, where the task is to find array-type signal recombination coefficients that optimize the signal features for correct subword state classification [14], and also the more difficult case where multiple voices overlap in a single channel with no opportunity to cancel interference, but only to infer the most likely speech state.

3 Multi-Speaker Language Models (MSLMs) [UW]

3.1 Motivation

In almost all work on language modeling, one uses and produces a probability distribution over words produced by a single speaker. For example, in n-gram modeling, one produces a probabilistic model of the form $p(w_t|w_{t-1},\ldots,w_{t-n+1})$, where w_t is the t^{th} word in a sentence.

In a conversational environment, however, words from one speaker may affect the words from another speaker. In a meeting setting, our belief is that the effect of the words from another speaker can be quite significant. This relationship, however, is not represented by standard n-gram language models, nor by the typical single-speaker exponential models that have also been used. For our meeting project on language modeling, one of our goals is to model the speakers from an entire meeting *jointly*. In particular, we have been investigating the modeling of distributions over words from a speaker given words not only previously spoken by that speaker, but also given words previously spoken by other speakers. We have achieved preliminary results on Switchboard (because of the plethora of data in that corpus) using the Factored Language Model/Generalized Parallel Backoff (FLM-GPB) [3] extensions to the SRI-LM toolkit. We have found that our multi-speaker language modeling (MSLM) yields significant improvements in perplexity. The basic framework and various experiments we have run are described in the next few sections.

3.2 A simple two-stream language model

In this first set of experiments, we use a model where for each word w_t in the current speaker's word stream, the history is not only the two previous words from that same speaker w_{t-2}, w_{t-1}, but also the closest previous word from the other speaker a_t as shown in Figure 1.

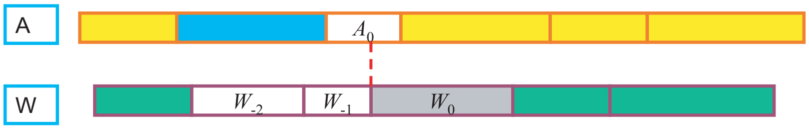

Figure 1: Multi-speaker language models. The probability distribution over the current word W_0 depends directly on the two previous words (W_{-1} and W_{-2}), but also on the word spoken by the other speaker that overlaps with the start time of W_0, which we indicate by A_0.

In other words, the probability model we represent is the following:

$$P(w_i|h_i) \approx P(w_i|w_{i-2}, w_{i-1}, a_i),$$

where a_i is the closest previous word said by another speaker.

Note that in this approach, word tokens are taken as they are irrespective of whether or not the word is a silence token. In other words, we look at the starting time of W_0, and look at the corresponding temporal position in the other word stream. If the other word stream is silence at that point (i.e., a short pause, or the beginning/ending of an utterance), we take a_0 to be silence. If the other stream contains a word at that point, we take a_0 to be that word. In the experiments we ran, non-silence occurred as the value of A_0 approximately 12% percent of the time, meaning that even though the novel information occurs only occasionally, when it does occur it is quite informative, and it interacts well with the language model smoothing procedures we are using (see below).

3.3 Experimental results

We tested this approach on Switchboard-I which has 2.4k conversations. The perplexity results we report are the average in a 5-fold cross-validation setting. As mentioned above, we used the FLM-GPB extensions done by co-PIs Bilmes and Kirchhoff [3] to the SRI language modeling toolkit to produce and represent all of the models presented below.

Since there is no direct temporal order among the variables w_{i-1}, w_{i-2}, a_i (in particular, W_{-1} might or might not come before A_0), there is a choice as to which Backoff path should be used (and which one is best). We tested a number of different possible paths, and the one we found to be best is shown in the highlighted line in Figure 2.

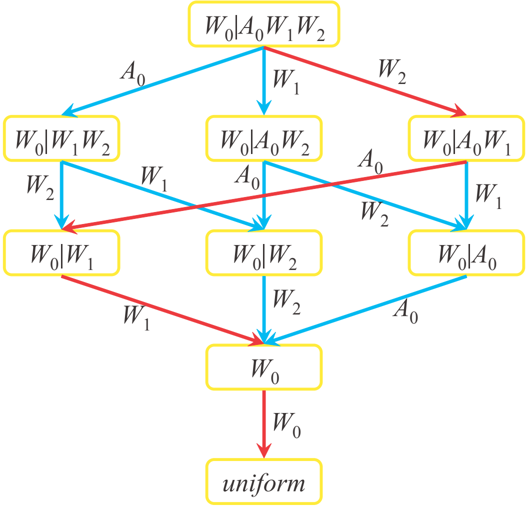

Figure 2: Backoff paths in multi-speaker language models

We also tried different smoothing methods. To be accurate, we include here the model (called MSLM-1) using FLM-GPB notation which yielded the best results:

```
W : 3 W(-1) W(-2) A(0) word.count.gz word.lm.gz 4
  W1,W2,A0  W2 kndiscount interpolate
     W1,A0  A0 kndiscount interpolate
        W1  W1 kndiscount interpolate
         0   0 kndiscount
```

In this model, a word has three conditioning variables, the two previous words from the same stream, W_{-1} and W_{-2}, and one word from another speaker, A_0. During backing-off, we first drop the node W_{-2}, next A_0, and then W_{-1}, and ultimately the unigram and then the uniform distribution. We drop first W_{-2} since A_0 is typically closer to the current word W_0 and therefore may provide more information. Also, W_0 is less dependent on A_0 compared to W_{-1} since it was better to drop W_{-1} last.

We also tested a model that used two words in the other stream instead of just one. We call this language model (MSLM-2):

```
W : 4 W(-1) W(-2) A(0) B(0) word.count.gz word.lm.gz 5
  W1,W2,A0,B0  B0 kndiscount interpolate
     W1,W2,A0  W2 kndiscount interpolate
        W1,A0  A0 kndiscount interpolate
           W1  W1 kndiscount interpolate
            0   0 kndiscount
```

The average perplexity is given in Table 1. As we can see, the perplexity is reduced 12.6% from trigram baseline when adding information from another speaker, which is quite significant. The number is slightly worse when adding two words from another speaker, but is still 10.8% better than the baseline. This indicates that in a conversational environment, speakers are greatly affected by each other and we can gain accuracy when modeling the dependencies, or more generally when jointly modeling their sentence structure.

Table 1: Perplexity of MSLM on swbd.

LM	PPL	abs. impr.	rel. impr.
trigram	84.6		
MSLM-1	73.9	10.7	12.6 %
MSLM-2	75.4	9.2	10.8 %

3.4 Thresholding silences

There are many silence tokens in stream A and they have various lengths. The question is whether the length information in these silence tokens are at all useful. In particular, if there is only a short bit of silence in the other stream before word W_0, it might be beneficial to set A_0 to be the preceding non-silence word rather than silence itself. We test this by putting a threshold on the silence tokens. We set a_0 to be the silence token only when the end time of the preceding non-silence word in the other stream plus some threshold is earlier than the starting time of the current word. Otherwise, we set a_0 to be the token of the previous non-silence word in the other stream. This is shown in Figure 3. In the top case, the corresponding token in the other stream is a non-silence word, so we just use it. In the middle case, the corresponding word is a long silence, so we set a_0 to silence. In the bottom case, the silence is short so we set a_0 to the preceding non-silence word.

The average perplexities are shown for Switchboard in Table 2. From the table, we can see that as the perplexity decreases when the threshold gets smaller. In other words, it appears that it is better in this case to only condition on another word when it overlaps with the current word, and that we want to keep the silence tokens as is. Note, we have so far only tried the same and a single backoff path for each such threshold — it might be the case that it is better to drop a_0 first, since as thresholds get larger, a_0 is

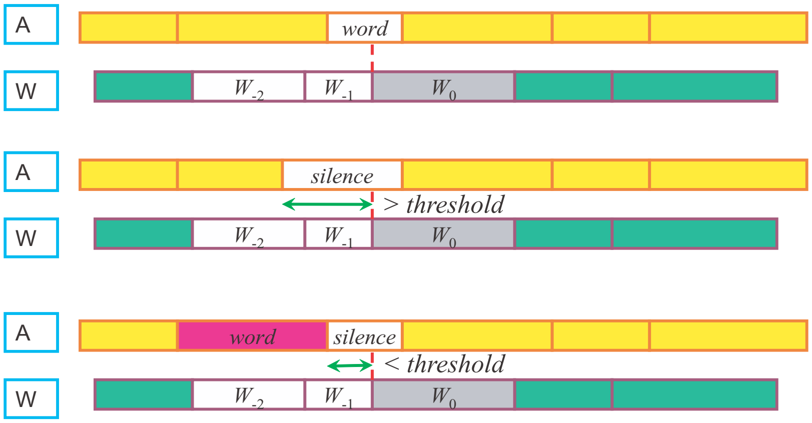

Figure 3: Silence thresholds in multi-speaker language models.

non-silence more often thereby decreasing the counts in the language model. We are currently studying this issue further.

Table 2: Perplexity of MSLM with silence threshold on swbd.

Threshold	0.0 s	0.1 s	0.3 s	0.5 s	1.0 s
MSLM-1	73.9	74.2	74.8	75.5	77.4
MSLM-2	75.4	75.8	76.4	77.1	78.8

3.5 Discussion

We introduced and produced the cross-speaker dependent language modeling (MSLM) for conversational tasks. The perplexity results on Switchboard show that the words from other speakers can significantly improve the accuracy of predicting the words for the current speaker. This alone shows broad implications for language modeling in general. Specifically, what you say is influenced by what other people say. We expect this to be particularly relevant in a meeting environment, where the specific words you choose are such that they (ideally) should align with the topic and mode of the current meeting.

In the next round of research, will be extending the above to the meeting task where A_0 is essentially a vector of words from multiple speakers (as of this writing we have quite promising preliminary results). Moreover, we will be considering models where A_0 are values other than words, specifically constructs from a lexical hierarchy ranging from word classes to broad meeting topics.

4 Detecting Important Regions [ICSI, SRI, UW, and Columbia]

4.1 Hot Spots [ICSI and SRI]

Recent interest in the automatic processing of meetings is motivated by a desire to summarize, browse, and retrieve important information from lengthy archives of spoken data. One of the most useful capabilities such a technology could provide is a way for users to locate "hot spots", or regions in which participants are highly involved in the discussion (e.g., heated arguments, points of excitement, and so on). Such regions are likely to contain important information for users who are browsing a meeting of for applications of information retrieval. We researched the following two questions:

- Can human listeners agree on utterance-level judgments of speaker involvement?
- Do judgments of involvement correlate with automatically extractable prosodic cues?

To address the first question we conducted a study in which human subjects were asked to rate utterances with respect to involvement. We found that despite the subjective nature of the task, raters showed significant agreement in distinguishing involved from non-involved utterances. We also found a difference in ratings depending on whether raters were native or nonnative speakers of the language – which may reflect language differences, cultural differences, or both.

To address the second question, we correlated acoustic features based on F0 and energy with the human ratings of involvement. These acoustic features were extracted and (where applicable) normalized completely automatically based on previous work at ICSI (funded by the DARPA Communicator project). We found remarkably reliable acoustic cues to involvement, based on F0 and energy values. Furthermore, it is likely that this is a general effect over all speakers (rather than a correlation between speaker prosodic values and speaker tendency for involvement), because we found that the most affected features of an individual speaker were similar to the most affected features that were computed over all speakers.

Taken together, these results suggest that hotspots, as defined via involvement level of utterances, can be fairly reliably identified by humans, and thus could be an important construct to label in our meeting maps. The automatic detection of these regions can take advantage of prosodic cues, as we have found in this study and are continuing to pursue. In future work, we are planning to examine lexical cues (using both true and automatically recognized words), looking at hotspot identification direction (rather than via involvement at the utterance level), investigating subclasses of involvement (e.g., arguments versus jokes), and further exploring acoustic cues. By increasing our database of hotspots, we will also be able to conduct machine learning experiments to predict involvement on unseen data.

4.2 Emphasized segment detection [Columbia]

We looked at using voice pitch to detect emphasized segments in discussions. A standard pitch tracker was run over close-mic signals to recover voicing and pitch information for each participant in a portion of a single meeting. Emphasis was considered detected when a certain number of frames in an utterance (turn) had pitch higher than a particular percentile within each individual speaker's pitch distribution, where the thresholds were set on training data. 800 utterances were labelled for emphasis by several transcribers [10]

4.3 Automatic Detection 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) [8]. 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 [4]. The word-based features use n-gram language models (LMs) trained using weakly supervised clustering techniques, i.e. keyword assignment rules automatically derived from a small amount of labeled data are used to assign initial clusters to a large unlabeled data set, which are iteratively refined using an n-gram LM clustering procedure. The automatically-derived agreement labels can also be used for training a prosody-only predictor. The prosodic features are taken from the database developed by Shriberg and colleagues at ICSI, and we use an iterative feature selection algorithm that involved running multiple decision trees [16].

Experiments were conducted on a pilot version of the meeting corpus using both hand transcripts and ASR output. We found that significant gains are obtained with weakly supervised training over using only the small hand-labeled subset, for both prosody-only and word-based classifiers. In moving from hand transcripts to ASR output, there is a significant performance loss for the word-based system but not for the prosody-only system. However, much of the loss can be recovered by combining keyword cues (which work very well with hand transcripts) with language model scores (which appear to be more robust to ASR errors). Unfortunately, combining prosody and word cues did not give further gains in performance, perhaps because there are few cases where the spurts are lexically ambiguous but prosodically distinct. It is encouraging that on high error rate outputs (45% WER), we still obtain a 78% rate of recovery of agreements and disagreements with only a 3% confusion rate between these critical classes.

This work is primarily due to Dustin Hillard, an undergraduate student at the University of Washington, in collaboration with Mari Ostendorf from UW and Liz Shriberg from SRI. Dustin received the 2002 UW Electrical Engineering Department award for Outstanding Undergraduate Researcher for this work. Initial work was primarily due to his senior honors project, but it has also been supported in part by DARPA grant N660019928924. The work on ASR transcriptions and automation of the keyword selection was supported this past year by an REU.

5 Dialog Acts [ICSI and SRI]

5.1 Meetings labeled

To date we have labeled dialog acts (DAs)for 40 meetings in entirety, with 3 additional meetings in the works. In addition we have labeled subsets of these meetings by all three labelers for reliability (see below). Our plan is to complete 50 meetings by the end of August, when the labeling project will end. We expect to meet this goal, perhaps exceeding it.

5.2 Reliability

Since the last report we have also made very good progress on agreement among the 3 labelers. Previously this was at about .60 (Kappa) for 3-way agreement on basic 1st tier tags, ignoring differences in segmentation. After work in this area we have raised that value to between .77 and .87, depending on the mapping of tags.

This result is due to both further discussions among labelers and Dr. Shriberg, and also revising of tags to best reflect what can be quickly and reliably lableled. We have added 10 new DA tags, removed 7 tags, and changed 8 tags from the original SWBD-DAMSL tagset. These were changed to reflect the dynamic format and interaction of our meetings. Refining the labeling system helped create a more uniform understanding of the tagset among the annotators.

We have also worked on the issue of dialog act segmentation, which is a trickly issue since there are many ambiguous cases and conventions must have coverage for as much of the data as possible We are currently in the midst of assessing reliability after work on this issue by the labelers.

5.3 Research

To date our research has focused on detection of lexically ambiguous dialog acts, based on the 20-meeting subset completed earlier, and large samples counts are needed for the machine learning experiments. Since we have 40 meetings completed at this point, we can begin investigating automatic detection of a wider range of dialog acts.

We have investigated whether automatically extracted prosodic features can serve as cues to dialog acts in naturally-occurring meetings. the classification of four short DAs, all of which can be conveyed by the same words. DAs were hand-labeled based on the discourse context. Results for classifiers trained on automatically extracted prosodic features show significant associations with DAs in unseen test data. Furthermore, the specific features used depend on the classification task at hand. Results shed light on the relationship between discourse function and prosody, and could be used to aid automatic processing for natural dialog understanding.

5.4 Cross-site collaboration

We have shared our dialog acts and prosodic features with our partners at Columbia University for use in topic detection. We have also distributed this information to colleagues in Europe, particularly for partners in the Swiss National Research Network IM2 (see http://www.im2.ch). The meetings we label have set the list of meetings that UW will label for additional types of information. In all cases we look forward to further collaborations.

6 Topic Segmentation and Classification

6.1 Speaker Turn Analysis [Columbia]

We have been trying to find higher-level structure in meetings by looking at the patterns of speaker turns. We calculated speaker turn probabilities for each minute of a set of meetings, and looked at segmentation using the BIC criterion, applied both to first order (who was speaking) and second order (who followed whom) statistics. We were interested in whether these patterns would be an indicator of topic boundaries. We also tried to model the activity of each speaker within a meeting relative to a baseline activity of that speaker averaged over all the meetings in which the speaker participated. With a pool of 10 speakers variously present in 26 meetings recorded over 9 months, we were able to build relatively stable models for each speaker's innate 'talkativity' [13].

We have also worked on the development of visualization software to help with analyzing and understanding patterns of speaker turns within meetings at the scale of minutes [11].

6.2 Topic Segmentation and Clustering for Information Extraction [UW]

An important step in building a content map is developing techniques to automatically associate coherent regions of meetings with topic labels or key phrases. 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 LM clustering approaches that were applied in BN [17]. Initial experiments applying a simplified version of this technique to the meeting data (with hand transcriptions) were not successful. We attribute the problems to conversational speech phenomena – high rate of anaphora, disfluencies, and short utterances – which together lead to fewer content words per utterance on which to base clustering decisions and dimensionality problems for estimating topic-dependent LMs. To illustrate the magnitude of this problem: after removing common words that might appear in a hand-derived stop-word list, only 25% of the original tokens were left.We have so far investigated two extensions to previous work to address these problems: automatic generation of a topic-independent word list (or, stop-word list) and use of a low dimension continuous space representation of words in combination with n-gram models. Initial results, described below, are based on broadcast news data to simplify algorithm debugging.

Stop-word lists are an important part of modern information retrieval and document clustering systems. Stop-word lists are usually generated by human experts for each target domain. It is generally acknowledged that stop-word lists are domain specific although there has not been a detailed assessment of different stop-word lists to tasks such as document clustering and topic detection for different text genre (broadcast news vs. spontaneous, conversational speech). Automatically finding a stop-word list is well-defined in supervised paradigms like text classification. Discriminative feature selection methods can be applied and the least discriminative words can be derived. In tasks such as unsupervised topic learning or document clustering though, discriminative feature selection is not meaningful since there is no class information.

In this project, we looked at automatically creating a stop-word list for the task of unsupervised topic learning. First, we trained a mixture of unigrams (i.e. a naive Bayes model) with EM, initialized randomly. Several different runs of this model were performed each with a different random initialization. For each run, the entropy of the cluster posterior for each word is computed. The average of the entropies from different runs is computed and the highest N words are printed. These words are assumed to be topic-independent therefore our topic model is altered by assuming that some words are generated from topics (p(word—topic)) and some are not (p(word)). Evaluating the new model on independent data shows that treating the highest 600 words as topic independent leads to a dramatic reduction in perplexity (680 perplexity of mixture of 50 unigrams with all words assumed topic-dependent, 488 perplexity of the new model). This is an encouraging result that we may want to explore further and compare it with hand-crafted stop-word lists, particularly for the meetings corpus. Note also that in this approach we do not actually throw out the stop words, but rather implicitly ignore them by making them topic-independent, which makes the results relevant for language modeling more generally.

We have also looked at unsupervised clustering with different methods for reducing the dimensionality of the vocabulary space. In particular, we looked at latent semantic analysis (LSA) [1] and transformation based on unigram mixture component posteriors. In both cases, the premise is that a large generic source of data might be available for estimating the dimensionality reduction transformation, and then clustering could leverage this representation. Experiments on the abstract data show improvements in unigram clustering when seeded by the results of a stage of clustering in the reduced LSA space, and that the two-stage clustering outperforms the LSA approach alone. The experiments were repeated on news and meeting data, and the approach is reasonably effective on the news data but not useful on the meeting data. We hope to improve the results by incorporating language model adaptation so as to better handle the smaller amounts of data. Also, note that this work does not yet take advantage of any of the segmentation work at Columbia, and it may be that the automatically derived segments could provide better base units for clustering than utterances or series of utterances grouped by simple heuristics.

This work is primarily due to Costas Boulis, a graduate student at the University of Washington, in collaboration with Mari Ostendorf from UW.

6.3 Anaphora Annotation [UW]

In the first year's work on automatic topic extraction from meeting transcripts, we determined that the large number of pronouns posed a significant problem for automatically clustering utterances for topic identification and segmentation. We conjectured that performance could be improved by having anaphora resolution capabilities, i.e. identifying the relationship between antecedents and referring expressions in a text as in the example

> **N:** OK, what do we do with the [stuff on top]?
> **A:** We could just start by filling [it] out....

For that reason, we initiated an effort on labeling anaphora at UW, with the support of the 2002 REU award. Prof. Katrin Kirchhoff (UW) installed software tools and further developed the infrastructure and annotation guidelines (in consultation with other team members). The tools are based on the MITRE Alembic workbench, and the annotation framework is based on a system developed by Eckert & Strube (2000) [6] tailored to the types of phenomena found in the meeting data. Annotation is currently under way, and inter-transcriber agreement studies are planned for the near term. Once the labeled data is available, we plan to initiate an effort on automatic anaphora resolution and to study the impact of oracle anaphora resolution on topic clustering.

6.4 Topic Segmentation [Columbia]

A main issue in the project is to automatically identify topic changes within meetings. It is not only helpful for the purpose of indexing meetings and creating topic maps, but it is also an important pre-processing step in automatic summarization. We designed a domain-independent topic segmentation algorithm targeting multi-party conversation. It is based on decision trees, a machine learning paradigm that can induce classifiers exploiting features from various sources. We used features that we (or previous work) identified as good indicators of topic changes (see Findings section). This feature-based segmentation algorithm integrates a broad set of features and it significantly outperforms two state-of-the-art topic segmentation algorithms designed for written texts. One of the features that it incorporates is a novel topic segmentation algorithm for written texts that utilizes lexical chains; this algorithm is comparable to, or better than, previous state-of-the-art text segmenters.

7 Discourse markers of meeting content and social structure [ICSI]

Our collaborators at University of Washington and Columbia University are finding ways of identifying topics and topic shifts. Our work focuses on identifying is intended to identify points of maximal information within topics and speaker-perceived relations between one topic and the next.

Particularly relevant is ”so”. In the literature this discourse marker (DM) has been identified with introducing main points and conclusions and with relinquishing the floor.

The literature contains successful attempts at identifying prosodic profiles for specific uses of other discourse markers (e.g., Hirschberg & Litman; Local). It is our goal to derive similar profiles for discriminating the different meeting-structuring uses of ”so” and to distinguish these from non-discourse uses (such as ”and so forth”).

The profiles will use the following parameters:

- prosodic profile of ”so” and surrounding context
- position in utterance
- syntactic role to identify those with map-relevant semantic/pragmatic functions.

In this phase of our work, we have:

- reviewed the literature
- segmented topics for two meetings and compared them with the segmentations proposed at Columbia for those meetings (to ensure we are using ”topic” in the same way - which we are).
- developed our analysis grid for semantic/pragmatic function coding and applied it to two contrasting types of meetings (two of each type),
- become acquainted with existing annotation and database resources relating to the Meetings Project data:
  - Don Baron’s prosodics data base of 33 meetings
  - Speech Act coding of meetings by Liz Shriberg’s project.
- performed a preliminary analysis of function distributions with respect to speaker roles.
- reported the preliminary results at one of ICSI’s colloquium series.

8 Summarization [Columbia]

We also started to design an automatic meeting summarizer. In this preliminary work, we studied the characteristics of utterances that were labeled as salient by humans, and identified a certain number of features that are good indicators of salience. These results are discussed in the Findings section.

9 Related work on meetings

9.1 UW efforts

As part of an undergraduate directed study project, Brandon Smith is investigating information retrieval (IR) of spoken documents, particularly for IR from meeting transcripts with speaker labels. He hopes to build on prior work combining speaker information with content queries [cite to come in a later email]. In addition, a problem that he has identified for the meeting data is determining the appropriate segment to return, since it will frequently be desirable to return more than one speaker turn. The work so far has been unfunded, though we have an REU request pending that would allow us to increase the activities.

In a separate project funded in part by an IBM faculty development award and in part by a DARPA award, we have been investigating methods for using various text sources to augment the small amount of conversational speech transcripts available for language model training. This is particularly important for meetings, where the vocabulary is quite dynamic, depending on the expertise and focus of the specific group that is meeting. Through this work, we have achieved significant reduction in word error rate (40% relative) for new vocabulary items in recognizing meeting data [15]. In addition, we have developed a new word-class-dependent data combination technique that outperforms standard mixture models [5]. These advances will help improve speech recognition, which will make for more useful automatic transcripts in our meeting maps work.

As part of a DARPA project but also as part of an exchange program with the Tokyo Institute of Technology, we have been investigating speaker tracking from multiple desktop microphones. Initial efforts have aimed at identifying speaker overlap regions, and we have developed multi-microphone processing techniques as well as new features based on Hough transform fundamental frequency analysis techniques.

9.2 ICSI efforts

Under a combination of DARPA and Swiss funding (where the latter was from the Swiss research network IM2), we completed the preparation of the ICSI Meeting Corpus, pieces of which had been used for much of the research described above. This Corpus is now undergoing final checks before delivery to the Linguistic Data Consortium at the University of Pennsylvania, who will in turn provide wide distribution of the corpus via their usual channels. The Corpus will also be delivered to the Swiss consortium IM2 for inclusion in their Media file Server. This Server will permit remote access to the data by Swiss and European partners. The leading IM2 lab, IDIAP, is also a partner with ICSI and a group of other labs in the European Union project M4, which also is concerned with the analysis of data from meetings.

We are also studying the use of vector computer architectures for the core recognition algorithms that could be used in a portable Meeting Browser. Most of this effort focuses on the vectorization of the speech decoder, which is among the most irregular computations required. This effort is also currently funded under our Swiss grant, though it was originally funded as part of a larger computer architecture grant to UC Berkeley from DARPA.

Finally, ICSI and SRI have been working for some time on the core speech recognition technology required for recognizing multiparty speech from meetings. This has primarily been funded by DARPA.

References

[1] J. Bellagarda (1998), Exploiting latent semantic information in statistical language modeling, In the Proceedings of the IEEE, vol 88, 8, 1279-1296.

[2] S. Bhagat, H. Carvey and E. Shriberg (2003), Automatically Generated Prosodic Cues to Lexically Ambiguous Dialog Acts in Multiparty Meetings. To appear in Proc. International Congress of Phonetic Sciences, Barcelona.

[3] J. Bilmes and K. Kirchhoff (2003), Factored Language Models and Generalized Parallel Backoff, Proc. HLT-NAACL, Vol. Comp., pp 1-3, 2003.

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

[5] I. Bulyko and M. Ostendorf and A. Stolcke (2003), Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures, Proc. HLT-NAACL, vol. Comp., pp 7-9, 2003.

[6] M. Eckert and M. Strube, Dialogue Acts, Synchronising Units and Anaphora Resolution (2000), Journal of Semantics, 17, pp 51-89.

[7] M. Galley, K. McKeown, E. Fosler-Lussier, H. Jing (2003). Discourse Segmentation of Multi-party Conversation (2003). In the Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03). Sapporo, Japan. (to appear)

[8] D. Hillard, M. Ostendorf and E. Shriberg (2003), Detection of agreement vs. disagreement in meetings: training with unlabeled data, Proc. HLT-NAACL, vol. Comp., pp 34-36, 2003.

[9] J. Hirschberg, and D. Litman (1992). Empirical Studies on the Disambiguation of Cue Phrases. Computational Linguistics, 1992, 19, 501-530.

[10] L. Kennedy and D. Ellis (2003), ”Pitch-based Emphasis Detection for Characterization of Meeting Recordings”, ELEN E6820 Speech and Audio Processing and Recognition Final Project Report, May 2003. http://www.ee.columbia.edu/d̃pwe/e6820/newprojects/lsk20/project/sapr_final.pdf

[11] J. Liu and D. Ellis (2003), ”Transplotter - a tool for visualizing meeting recording structure”, web-based software release. http://www.ee.columbia.edu/j̃liu/trsplot/

[12] Local, J. (1992) Conversational Phonetics: Some Aspects of News Receipts in Everyday Talk. York Papers in Linguistics 15. 37-80.

[13] S. Renals and D. Ellis (2003), ”Audio Information Access from Meeting Rooms”, Proc. ICASSP-03, Hong Kong, April 2003. http://www.ee.columbia.edu/ dpwe/pubs/icassp03-mtg.pdf

[14] M.J. Reyes-Gomez, B. Raj, D. Ellis (2003), ”Multi-channel Source Separation by Factorial HMMs”, Proc. ICASSP-03, Hong Kong, April 2003. http://www.ee.columbia.edu/ dpwe/pubs/icassp03-fhmm.pdf

[15] S. Schwarm and I. Bulyko and M. Ostendorf, Adaptive language modeling with varied sources to cover new vocabulary items, internal manuscript.

[16] E. Shriberg, A. Stolcke, D. Hakkani-Tür and G. Tür (2000), Prosody-Based Automatic Segmentation of Speech into Sentences and Topics, Speech Communication, eds. T. Robinson and S. Renals, vol. 32, 1-2, Sep, 2000, 127-154

[17] S. Sista, R. Schwartz, T. R. Leek, J. Makhoul (2002), An Algorithm for Unsupervised Topic Discovery from Broadcast News Stories, in Proceedings of the Human Language Technology Conference, 2002.

[18] B. Wrede and E. Shriberg (2003), Spotting ”Hotspots” in Meetings: Human Judgments and Prosodic Cues. To appear in Proc. Eurospeech, Geneva.

1 Findings Introduction

In months 10 through 21 of our NSF-ITR on Mapping Meetings, we expanded our efforts significantly over the intial period.

Here we describe our findings in the following areas:

1. Speaker Separation - separating multiple voices from meetings
2. Multi-speaker Language Models
3. Detecting Important Regions (”hot spots”, emphasis, and agreement/disagreement)
4. Dialog Acts
5. Topic Detection, Segmentation and Classification
6. Discourse Markers
7. Summarization

The sections below will elaborate on these themes.

2 Speaker Separation [Columbia]

In our work on overlapping voice separation, we adopted an approach of successive information retrieval to move from an oracle-knowledge ’cheating’ condition (when the appropriate state sequences of the individual sources were fully known) to a more realistic condition (where each voice has certain lexical sequence constraints, but no prior timing information is given). We showed promising results in unsupervised estimation of array weights specifically optimized for recognizer input features.

In the single-channel overlapped speech condition, we developed a coupled HMM model for the speech in each subband, and showed that this factorization gave a significant improvement to the achievable SNR in signal separation through time-varying filtering driven by the inferred pair of state-sequence estimates.

3 Multi-Speaker Language Models (MSLMs) [UW]

We introduced and produced the cross-speaker dependent language modeling (MSLM) for conversational tasks. The perplexity results on Switchboard show that the words from other speakers can significantly improve the accuracy of predicting the words for the current speaker. This alone shows broad implications for language modeling in general. Specifically, what you say is influenced by what other people say. We expect this to be particularly relevant in a meeting environment, where the specific words you choose are such that they (ideally) should align with the topic and mode of the current meeting.

4 Detecting Important Regions [ICSI, SRI, UW, and Columbia]

4.1 Hot Spots [ICSI and SRI]

We conducted a study in which human subjects were asked to rate utterances with respect to involvement. We found that despite the subjective nature of the task, raters showed significant agreement in distinguishing involved from non-involved utterances. We also found a difference in ratings depending on whether raters were native or non-native speakers of the language – which may reflect language differences, cultural differences, or both.

We correlated acoustic features based on F0 and energy with the human ratings of involvement. These acoustic features were extracted and (where applicable) normalized completely automatically based on previous work at ICSI (funded by the DARPA Communicator project). We found remarkably reliable acoustic cues to involvement, based on F0 and energy values. Furthermore, it is likely that this is a general effect over all speakers (rather than a correlation between speaker prosodic values and speaker tendency for involvement), because we found that the most affected features of an individual speaker were similar to the most affected features that were computed over all speakers.

Taken together, these results suggest that hotspots, as defined via involvement level of utterances, can be fairly reliably identified by humans, and thus could be an important construct to label in our meeting maps. The automatic detection of these regions can take advantage of prosodic cues, as we have found in this study and are continuing to pursue.

4.2 Emphasized segment detection [Columbia]

The pitch-based emphasis detection algorithm achieved over 90% correct agreement with hand-marked emphasis labels over a test set of some 400 utterances, taken from the same meeting as the test set. Future work will investigate the viability of this approach for completely unlabelled data (i.e. automatic inference of thresholds), and applications of the emphasized phrase locations in recovering useful structure such as topic boundaries.

4.3 Automatic Detection of Agreement and Disagreement in Meetings [UW]

Experiments were conducted on a pilot version of the meeting corpus using both hand transcripts and ASR output. We found that significant gains are obtained with weakly supervised training over using only the small hand-labeled subset, for both prosody-only and word-based classifiers. In moving from hand transcripts to ASR output, there is a significant performance loss for the word-based system but not for the prosody-only system. However, much of the loss can be recovered by combining keyword cues (which work very well with hand transcripts) with language model scores (which appear to be more robust to ASR errors). Unfortunately, combining prosody and word cues did not give further gains in performance, perhaps because there are few cases where the spurts are lexically ambiguous but prosodically distinct. It is encouraging that on high error rate outputs (45% WER), we still obtain a 78% rate of recovery of agreements and disagreements with only a 3% confusion rate between these critical classes.

5 Dialog Acts [ICSI and SRI]

Results for classifiers trained on automatically extracted prosodic features show significant associations with dialog act labels in unseen test data. Furthermore, the specific features used depend on the classification task at hand. Results shed light on the relationship between discourse function and prosody, and could be used to aid automatic processing for natural dialog understanding.

6 Topic Segmentation and Classification

6.1 Speaker Turn Analysis [Columbia]

Boundaries indicated by changes in participating speakers were well defined in the data, but correlated only weakly with hand-marked topic changes, with about half of 36 ground-truth topic boundaries (from 6 meetings) agreeing with automatic segments, and an equal number of 'false alarms', turn-pattern boundaries that did not correspond to topic changes. For the talkativity modeling, significant hidden variations were shown between raw proportion of meeting in which participant spoke, and the normalized value accounting for innate speaker talkativity and competition from other speakers. Informal investigation showed some correspondence between meetings at which particular speakers were unusually talkative and topics on which they would be expected to contribute, but no objective ground truth has been established for more quantitative evaluation [2].

6.2 Topic Segmentation and Clustering for Information Extraction [UW]

In this project, we looked at automatically creating a stop-word list for the task of unsupervised topic learning. First, we trained a mixture of unigrams (i.e. a naive Bayes model) with EM, initialized randomly. Several different runs of this model were performed each with a different random initialization. For each run, the entropy of the cluster posterior for each word is computed. The average of the entropies from different runs is computed and the highest N words are printed. These words are assumed to be topic-independent therefore our topic model is altered by assuming that some words are generated from topics (p(word—topic)) and some are not (p(word)). Evaluating the new model on independent data shows that treating the highest 600 words as topic independent leads to a dramatic reduction in perplexity (680 perplexity of mixture of 50 unigrams with all words assumed topic-dependent, 488 perplexity of the new model). This is an encouraging result that we may want to explore further and compare it with hand-crafted stop-word lists, particularly for the meetings corpus. Note also that in this approach we do not actually throw out the stop words, but rather implicitly ignore them by making them topic-independent, which makes the results relevant for language modeling more generally.

We have also looked at unsupervised clustering with different methods for reducing the dimensionality of the vocabulary space. In particular, we looked at latent semantic analysis (LSA) [1] and transformation based on unigram mixture component posteriors. In both cases, the premise is that a large generic source of data might be available for estimating the dimensionality reduction transformation, and then clustering could leverage this representation. Experiments on the abstract data show improvements in unigram clustering when seeded by the results of a stage of clustering in the reduced LSA space, and that the two-stage clustering outperforms the LSA approach alone. The experiments were repeated on news and meeting data, and the approach is reasonably effective on the news data but not useful on the meeting data.

6.3 Topic Segmentation [Columbia]

We studied different sets of features that can be used in the context of automatic topic segmentation. In our case, we utilize features that we identified as strongly correlated with topic changes such as the presence of many speaker overlaps and broad changes in speaker activity distribution (that we capture using a information-theoric metric). Our work also shows that some features (e.g. silences and cue phrases) that have been used to segment monologue speech preserve their usefulness in multi-party speech. This is also the case for features successfully used to segment written texts (e.g. topic words). In our final evaluation, we found out that even though lexical information is most useful, a significant increase of segmentation accuracy is obtained when we incorporate information about multi-party speech into our decision-tree model.

7 Discourse markers of meeting content and social structure [ICSI]

1. Agenda setters tend to use more topic-structuring and topic-concluding ”so” markers than do other meeting participants.

2. The frequency of topic-structuring markers will differ for an individual as a function of speaker role (that is, more of them in meetings in which that speaker is the agenda-setting than in meetings when he/she is not).

3. We replicated the finding of turn-final ”so” being used to relinquish a turn, but noticed that unlike prior work, these uses of ”so” sometimes occur as much as one second into the new speaker’s turn (suggesting they may sometimes not be cues to turn change, but merely correlates of them).

4. We discoursed a new use of ”so” in turn-taking, which is to claim the floor. This use is very common in our meetings but not reported in the conversation analysis literature.

Our findings in items 3 and 4 suggest a larger point, which is that some properties of meetings (e.g., more competition for the floor, more time constraints for relevance, and more limited bandwidth for monitoring every other speakers’ nonverbal cues) may not be covered by existing theories of discourse or conversation analysis. That is, our findings on meetings may point to needed extensions of well respected theories developed for more usual types of conversation.

8 Summarization [Columbia]

In a preliminary step in the design of an automatic summarization system for meetings, we studied the correlation between the salience of utterances (judged by humans) and features that we can extract from the speech signal and the transcription. While previous work has extensively analyzed how salience in written texts correlates with diverse features such as term frequency and position in the text (e.g. a paragraph-initial sentence is deemed more important than following sentences), there is, to our knowledge, no such study with any kind of speech corpus. Our initial investigation focused on automatically-extracted lexical, acoustic, and prosodic features, in addition to features from manually-labeled dialog acts. We determined that utterances that are tagged with dialog acts like ”statement”, ”question”, and ”answer” are more likely to be judged salient than those that are tagged as ”subjective statement” and ”acknowledgment”. We also found that sentences that are uttered with higher pitch range and amplitude are statistically more likely to be judged salient by humans. This work will help us designing an extractive summarizer based on machine learning and exploiting all features we have studied.

References

[1] J. Bellagarda (1998), Exploiting latent semantic information in statistical language modeling, In the Proceedings of the IEEE, vol 88, 8, 1279-1296.

[2] S. Renals and D. Ellis (2003), ”Audio Information Access from Meeting Rooms”, Proc. ICASSP-03, Hong Kong, April 2003. http://www.ee.columbia.edu/ dpwe/pubs/icassp03-mtg.pdf

1 Contributions

The visualization software for browsing meeting recordings and speaker change patterns at very broad timescales was made available to the community via a web site [1]. This software will act as a platform for our subsequent developments in segmentation of high-level meeting structure.

As noted in our previous annual report (when this work was far less developed), we have distributed our prosodic database (including recognition output) and dialog act annotations to others both inside and beyond the ITR project. We continue to do this, though the dialog act annotations are now far more extensive.

Several contributions emerged from our effort in automatic topic segmentation. First, a major lexical sub-component of our topic segmenter is freely available for any use in research or education. Our second contribution is the reference segmentation we created for 25 meetings of the ICSI meeting corpus. Each meeting was processed by at least human judges; a statistical test indicates a reasonably good level of agreement between them. This kind of resource is unfortunately too rare, leading many researchers working in topic segmentation to automatically create artificial reference segmentations by concatenating unrelated texts, which we believe is not the right solution. We hope that other researchers in the community will be able to make good use of this resource, and that it will encourage future work in the discourse processing of meetings.

References

[1] J. Liu and D. Ellis (2003), ”Transplotter - a tool for visualizing meeting recording structure”, web-based software release. http://www.ee.columbia.edu/j̃liu/trsplot/