Detecting the Noteworthiness of Utterances in Human Meetings

Satanjeev Banerjee
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
banerjee@cs.cmu.edu

Alexander I. Rudnicky
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
air@cs.cmu.edu

Abstract

Our goal is to make note-taking easier in meetings by automatically detecting noteworthy utterances in verbal exchanges and suggesting them to meeting participants for inclusion in their notes. To show feasibility of such a process we conducted a Wizard of Oz study where the Wizard picked automatically transcribed utterances that he judged as noteworthy, and suggested their contents to the participants as notes. Over 9 meetings, participants accepted 35% of these suggestions. Further, 41.5% of their notes at the end of the meeting contained Wizard-suggested text. Next, in order to perform noteworthiness detection automatically, we annotated a set of 6 meetings with a 3-level noteworthiness annotation scheme, which is a break from the binary "in summary"/ "not in summary" labeling typically used in speech summarization. We report Kappa of 0.44 for the 3-way classification, and 0.58 when two of the 3 labels are merged into one. Finally, we trained an SVM classifier on this annotated data; this classifier's performance lies between that of trivial baselines and inter-annotator agreement.

1 Introduction

We regularly exchange information verbally with others over the course of meetings. Often we need to access this information afterwards. Typically we record the information we consider important by taking notes. Note taking at meetings is a difficult task, however, because the participant must summarize and write down the information in a way such that it is comprehensible afterwards, while paying attention to and partici-

pating in the ongoing discussion. Our goal is to make note-taking easier by automatically extracting noteworthy items from spoken interactions in real time, and proposing them to the humans for inclusion in their notes.

Judging which pieces of information in a meeting are noteworthy is a very subjective task. The subjectivity of this task is likely to be more acute than even that of meeting summarization, where low inter-annotator agreement is typical e.g. (Galley, 2006), (Liu & Liu, 2008), (Penn & Zhu, 2008), etc – whether a piece of information should be included in a participant's notes depends not only on its importance, but also on factors such as the participant's need to remember, his perceived likelihood of forgetting, etc. To investigate whether it is feasible even for a human to predict what someone else might find noteworthy in a meeting, we conducted a Wizard of Oz-based user study where a human suggested notes (with restriction) to meeting participants during the meeting. We concluded from this study (presented in section 2) that this task appears to be feasible for humans.

Assuming feasibility, we then annotated 6 meetings with a 3-level noteworthiness scheme. Having 3 levels instead of the typical 2 allows us to explicitly separate utterances of middling noteworthiness from those that are definitely noteworthy or not noteworthy, and allows us to encode more human knowledge than a 2-level scheme. We describe this annotation scheme in more detail in section 3, and show high inter-annotator agreement compared to that typically reported in the summarization literature. Finally in sections 4 and 5 we use this annotated data to train and test a simple Support Vector Machine-based predictor of utterance noteworthiness.

2 Can Humans Do this Task?

As mentioned in the introduction, given the degree of subjectivity involved in identifying note-

worthy utterances, it is reasonable to ask whether the notes-suggestion task can be accomplished by humans, let alone by automatic systems. That is, we ask the question: Is it possible for a human to identify noteworthy utterances in a meeting such that

(a) For at least some fraction of the suggestions, one or more meeting participants agree that the suggested notes should indeed be included in their notes, and

(b) The fraction of suggested notes that meeting participants find noteworthy is high enough that, over a sequence of meetings, the meeting participants do not learn to simply ignore the suggestions.

Observe that this task is more restricted than that of generic note-taking. While a human who is allowed to summarize discussions and produce to-the-point notes is likely to be useful, we assume here that our system will not be able to create such abstractive summaries. Rather, our goal here is to explore the feasibility of an extractive summarization system that simply picks noteworthy utterances and suggests their contents to the participants. To answer this question, we conducted a Wizard of Oz-based pilot user study, as follows.

2.1 Wizard of Oz Study Design

We designed a user study in which a human Wizard listened to the utterances being uttered during the meeting, identified noteworthy utterances, and suggested their contents to one or more participants for inclusion in their notes. In order to minimize differences between the Wizard and the system (except for the Wizard's human-level ability to judge noteworthiness), we restricted the Wizard in the following ways:

(a) The Wizard was allowed to only suggest the contents of individual utterances to the participants, and not summarize the contents of multiple utterances.

(b) The Wizard was allowed to listen to the meeting speech, but when suggesting the contents of an utterance to the participants, he was restricted to using a real-time automatic transcription of the utterance. (He was allowed to withhold suggestions because they were too erroneously transcribed.)

(c) In order to be closer to a system that has little or no "understanding" of the meetings, we chose a human (to play the role of the Wizard) who had not participated in the meetings before, and thus had little prior knowledge of the meetings' contents.

2.2 Notes Suggestion Interface

In order to suggest notes to meeting participants during a meeting – either automatically or through a Wizard – we have modified the SmartNotes system, whose meeting recording and note-taking features have been described earlier in (Banerjee & Rudnicky, 2007). Briefly, each meeting participant comes to the meeting with a laptop running SmartNotes. At the beginning of the meeting, each participant's SmartNotes client connects to a server, authenticates the participant and starts recording and transmitting his speech to the server. In addition, SmartNotes also provides meeting participants with a note-taking interface that is split into two major panes. In the "notes" pane the participant types his notes that are then recorded for research purposes. In the "suggestions" pane, Wizard-suggested notes are displayed. If at any time during the meeting a participant double-clicks on one of the suggested notes in the "suggestions" pane, its text gets included in his notes in the "notes" pane. The Wizard uses a different application to select real-time utterance transcriptions, and insert them into each participant's "suggestions" pane. (While we also experimented with having the Wizard target his suggestions at individual participants, we do not report on those experiments here; those results were similar to the ones presented below.)

2.3 Results

We conducted the Wizard of Oz study on 9 meetings that all belonged to the same sequence. That is, these meetings featured a largely overlapping group of participants who met weekly to discuss progress on a single project. The same person played the role of the Wizard in each of these 9 meetings. The meetings were on average 33 minutes long, and there were 3 to 4 participants in each meeting. Although we have not evaluated the accuracy of the speech recognizer on these particular meetings, the typical average word error rate for these speakers is around 0.4 – i.e., 4 out of 10 words are incorrectly transcribed.

On average, the Wizard suggested the contents of 7 utterances to the meeting participants, for a total of 63 suggestions across the 9 meetings. Of these 63 suggestions, 22 (34.9%) were accepted by the participants and included in their notes. Thus on average, about 2.5 Wizard-suggested notes were accepted and included in participants' notes in each meeting. On average, meeting participants took a total of 5.9 lines of notes per

meeting; thus, 41.5% of the notes in each meeting were Wizard-suggested.

It cannot be ascertained if the meeting participants would have written the suggested notes on their own if they weren't suggested to them. However the fact that some Wizard-suggested notes *were* accepted implies that the participants probably saw some value in including those suggestions in their notes. Further, there was no drop-off in the fraction of meeting notes that was Wizard-suggested: the per-meeting average percentage of notes that was Wizard-suggested was around 41% for both the first 4 meetings, as well as the last 5. This implies that despite a seemingly low acceptance rate (35%), participants did not "give up" on the suggestions, but continued to make use of them over the course of the 9-meeting meeting sequence. We conclude that an extractive summarization system that detects noteworthy utterances and suggests them to meeting participants can be perceived as useful by the participants, if the detection of noteworthy utterances is "accurate enough".

3   Meeting Data Used in this Paper

Assuming the feasibility of an extraction-based notes suggestion system, we turn our attention to developing a system that can automatically detect the noteworthiness of an utterance. Our goal here is to learn to do this task over a sequence of related meetings. Towards this end, we have recorded sequences of natural meetings – meetings that would have taken place even if they weren't being recorded. Meetings in each sequence featured largely overlapping participant sets and topics of discussion. For each meeting, we used SmartNotes (Banerjee & Rudnicky, 2007) (described in section 2 above) to record both the audio from each participant as well as his notes. The audio recording and the notes were both time stamped, associated with the participant's identity, and uploaded to the meeting server. After the meeting was completed the audio was manually segmented into utterances and transcribed both manually and using a speech recognizer (more details in section 5.2).

In this paper we use a single sequence of 6 meetings held between April and June of 2006. (These were separate from the ones used for the Wizard of Oz study above.) The meetings were on average 28 minutes and 43 seconds long (± 3 minutes and 48 seconds standard error) counting from the beginning of the first recorded utterance to the end of the last one. On average each meeting had 28 minutes and 38 seconds of speech – this includes overlapped speech when multiple participants spoke on top of each other. Across the 6 meetings there were 5 unique participants; each meeting featured between 2 and 4 of these participants (average: 3.5 ± 0.31).

The meetings had, on average, 633.67 (± 85.60) utterances each, for a total of 3,796 utterances across the 6 meetings. (In this paper, these 3,796 utterances form the units of classification.) As expected, utterances varied widely in length. On average, utterances were 2.67 ± 0.18 seconds long and contained 7.73 (± 0.44) words.

4   Multilevel Noteworthiness Annotation

In order to develop approaches to automatically identify noteworthy utterances, we have manually annotated each utterance in the meeting data with its degree of "noteworthiness". While researchers in the related field of speech summarization typically use a binary labeling – "in summary" versus "out of summary" (e.g. (Galley, 2006), (Liu & Liu, 2008), (Penn & Zhu, 2008), etc) – we have observed that there are often many utterances that are "borderline" at best, and the decision to label them as "in summary" or "out" is arbitrary. Our approach instead has been to create three levels of noteworthiness. Doing so allows us to separate the "clearly noteworthy" utterances from the "clearly not noteworthy", and to label the rest as being between these two classes. (Of course, arbitrary choices must still be made between the edges of these three classes. However, having three levels preserves more information in the labels than having two, and it is always possible to create two labels from the three, as we do in later sections.)

These multilevel noteworthiness annotations were done by two annotators. One of them – denoted as "annotator 1" – had attended each of the meetings, while the other – "annotator 2" – had not attended any of the meetings. Although annotator 2 was given a brief overview of the general contents of the meetings, his understanding of the meeting was expected to be lower than that of the other annotator. By using such an annotator, our aim was to identify utterances that were "obviously noteworthy" even to a human being who lacks a deep understanding of the context of the meetings. (In section 5.2 we describe how we merge the two sets of annotations.)

The annotators were asked to make a 3-level judgment about the relative noteworthiness of each utterance. That is, for each utterance, the

annotators were asked to decide whether a note-suggestion system should "definitely show" the contents of the utterance to the meeting participants, or definitely not show (labeled as "don't show"). Utterances that did not quite belong to either category were asked to be labeled as "maybe show". Utterances labeled "definitely show" were thus at the highest level of noteworthiness, followed by those labeled "maybe show" and those labeled "don't show". Note that we did not ask the annotators to label utterances directly in terms of noteworthiness. Anecdotally, we have observed that asking people to label utterances with their noteworthiness leaves the task insufficiently well defined because the purpose of the labels is unclear. On the other hand, asking users to identify utterances they would have included in their notes leads to annotators taking into account the difficulty of writing particular notes, which is also not desirable for this set of labels. Instead, we asked annotators to directly perform (in some sense) the task that the eventual notes-assistance system will perform.

In order to gain a modicum of agreement in the annotations, the two annotators discussed their annotation strategies after annotating each of the first two meetings (but not after the later meetings). A few general annotation patterns emerged, as follows: Utterances labeled "definitely show" typically included:

(a) Progress on action items since the last week.
(b) Concrete plans of action for the next week.
(c) Announcements of deadlines.
(d) Announcements of bugs in software, etc.

In addition, utterances that contained the *crux* of any seemingly important discussion were labeled as "definitely show". On the other hand, utterances that contained no information worth including in the notes (by the annotators' judgment) were labeled as "don't show". Utterances that did contain some additional elaborations of the main point, but without which the main point could still be understood by future readers of the notes were typically labeled as "maybe show".

Table **1** shows the distribution of the three labels across the full set of 3,796 utterances in the dataset for both annotators. Both annotators labeled only a small percentage of utterances as "definitely show", a larger fraction as "maybe show" and most utterances as "don't show". Although the annotators were not asked to shoot for a certain distribution, observe that they both labeled a similar fraction of utterances as "definitely show". On the other hand, annotator 2, who

did not attend the meetings, labeled 50% more utterances as "maybe show" than annotator 1 who did attend the meetings. This difference is likely due to the fact that annotator 1 had a better understanding of the utterances in the meeting, and was more confident in labeling utterances as "don't show" than annotator 2 who, not having attended the meetings, was less sure of some utterances, and thus more inclined to label them as "maybe show".

Annotator #	Definitely show	Maybe show	Don't show
1	13.5%	24.4%	62.1%
2	14.9%	38.8%	46.3%

Table 1: Distribution of Labels for Each Annotator

4.1 Inter-Annotator Kappa Agreement

To gauge the level of agreement between the two annotators, we compute the Kappa score. Given labels from different annotators on the same data, this metric quantifies the difference between the observed agreement between the labels and the expected agreement, with larger values denoting stronger agreement.

For the 3-way labeling task, the two annotators achieve a Kappa agreement score of 0.44 (± 0.04). This seemingly low number is typical of agreement scores obtained in meeting summarization. (Liu & Liu, 2008) reported Kappa agreement scores between 0.11 and 0.35 across 6 annotators while (Penn & Zhu, 2008) with 3 annotators achieved Kappa of 0.383 and 0.372 on casual telephone conversations and lecture speech. (Galley, 2006) reported inter-annotator agreement of 0.323 on data similar to ours.

To further understand where the disagreements lie, we converted the 3-way labeled data into 2 different 2-way labeled datasets by merging two labels into one. First we evaluate the degree of agreement the annotators have in separating utterances labeled "definitely show" from the other two levels. We do so by re-labeling all utterances not labeled "definitely show" with the label "others". For the "definitely show" versus "others" labeling task, the annotators achieve an inter-annotator agreement of 0.46. Similarly we compute the agreement in separating utterances labeled "do not show" from the two other labels – in this case the Kappa value is 0.58. This implies that it is easier to agree on the separation between "do not show" and the other classes, than between "definitely show" and the other classes.

4.2 Inter-Annotator Accuracy, Prec/Rec/F

Another way to gauge the agreement between the two sets of annotations is to compute accuracy, precision, recall and f-measure between them. That is, we can designate one annotator's labels as the "gold standard", and use the other annotator's labels to find, for each of the 3 labels, the number of utterances that are true positives, false positives, and false negatives. Using these numbers we can compute precision as the ratio of true positives to the sum of true and false positives, recall as the ratio of true positives to the sum of true positives and false negatives, and f-measure as the harmonic mean of precision and recall. (Designating the other annotator's labels as "gold standard" simply swaps the precision and recall values, and keeps f-measure the same). Accuracy is the number of utterances that have the same label from the two annotators, divided by the total number of utterances.

Table **2** shows the evaluation over the 6-meeting dataset using annotator 1's data as "gold standard". The standard error for each cell is less than 0.08. Observe in Table **2** that while both the "definitely show" and "maybe show" classes have nearly equal f-measure, the precision and recall values for the "maybe show" class are much farther apart from each other than those for the "definitely show" class. This is due to the fact that while both annotators label a similar number of utterances as "definitely show", they label very different numbers of utterances as "maybe show". If the same accuracy, precision, recall and f-measure scores are computed for the "definitely show" vs. "others" split, the accuracy jumps to 87%, possibly because of the small size of the "definitely show" category. The accuracy remains at 78% for the "don't show" vs. "others" split.

	Definitely show	Maybe show	Don't show
Precision	0.57	0.70	0.70
Recall	0.53	0.46	0.93
F-measure	0.53	0.54	0.80
Accuracy	69%		

Table 2 Inter-Annotator Agreement using Accuracy Etc.

4.3 Inter-Annotator Rouge Scores

Annotations can also be evaluated by computing the ROUGE metric (Lin, 2004). ROUGE, a popular metric for summarization tasks, compares two summaries by computing precision, recall and f-measure over ngrams that overlap between them. Following previous work on meeting summarization (e.g. (Xie, Liu, & Lin, 2008), (Murray, Renals, & Carletta, 2005), etc), we report evaluation using ROUGE-1 F-measure, where the value "1" implies that overlapping *unigrams* are used to compute the metric. Unlike previous research that had one summary from each annotator per meeting, our 3-level annotation allows us to have 2 different summaries: (a) the text of all the utterances labeled "definitely show" and, (b) the text of all the utterances labeled either "definitely show" or "maybe show". On average (across both annotators over the 6 meetings) the "definitely show" utterance texts are 18.72% the size of the texts of all the utterances in the meetings, while the "definitely or maybe show" utterance texts are 61.6%. Thus, these two texts represent two distinct points on the compression scale. The average R1 F-measure score is 0.62 over the 6 meetings when comparing the "definitely show" texts of the two annotators. This is twice the R1 score – 0.3 – of the trivial baseline of simply labeling every utterance as "definitely show". The inter-annotator R1 F-measure for the "definitely or maybe show" texts is 0.79, marginally higher than the trivial "all utterances" baseline of 0.71. In the next section, we compare the scores achieved by the automatic system against these inter-annotator and trivial baseline scores.

5 Automatic Label Prediction

So far we have presented the annotation of the meeting data, and various analyses thereof. In this section we present our approach for the automatic prediction of these labels. We apply a classification based approach to the problem of predicting the noteworthiness level of an utterance, similar to (Banerjee & Rudnicky, 2008). We use leave-one-meeting-out cross validation: for each meeting *m*, we train the classifier on manually labeled utterances from the other 5 meetings, and test the classifier on the utterances of meeting *m*. We then average the results across the 6 meetings. Given the small amount of data, we do not test on separate data, nor do we perform any tuning.

Using the 3-level annotation described above, we train a 3-way classifier to label each utterance with one of the multilevel noteworthiness labels. In addition, we use the two 2-way merged-label annotations – "definitely show" vs. others and "don't show" vs. others – to train two more 2-way classifiers. In each of these classification

problems we use the same set of features and the same classification algorithms described below.

5.1 Features Used

Ngram features: As has been shown by (Banerjee & Rudnicky, 2008), the strongest features for noteworthiness detection are ngram features, i.e. features that capture the occurrence of ngrams (consecutive occurrences of one or more words) in utterances. Each ngram feature represents the presence or absence of a single specific ngram in an utterance. E.g., the ngram feature "action item" represents the occurrence of the bigram "action item" in a given utterance. Unlike (Banerjee & Rudnicky, 2008) where each ngram feature captured the *frequency* of a specific ngram in an utterance, in this paper we use boolean-valued ngram features to capture the presence/absence of ngrams in utterances. We do so because in tests on separate data, boolean-valued features out-performed frequency-based features, perhaps due to data sparseness. Before ngram features are extracted, utterances are normalized: partial words, non-lexicalized filler words (like "umm", "uh"), punctuations, apostrophes and hyphens are removed, and all remaining words are changed to upper case. Next, the vocabulary of ngrams is defined as the set of ngrams that occur at least 5 times in the entire dataset of meetings, for ngram sizes of 1 through 6 word tokens. Finally, the occurrences of each of these vocabulary ngrams in an utterance are recorded as the feature vector for that utterance. In the dataset used in this paper, there are 694 unique unigrams that occur at least 5 times across the 6 meetings, 1,582 bigrams, 1,065 trigrams, 1,048 4-grams, 319 5-grams and 102 6-grams. In addition to these ngram features, for each utterance we also include the number of Out of Vocabulary ngram – ngrams that occur less than 5 times across all the meetings.

Overlap-based Features: We assume that we have access to the text of the agenda of the test meeting, and also the text of the notes taken by the participants in previous meetings (but not those taken in the test meeting). Since these artifacts are likely to contain important keywords we compute two sets of overlaps features. In the first set we compute the number of ngrams that overlap between each utterance and the meeting agenda. That is, for each utterance we count the number of unigrams, bigrams, trigrams, etc that also occur in the agenda of that meeting. Similarly in the second set we compute the number of ngrams in each utterance that also

occur in the notes of previous meetings. Finally, we compute the degree of overlap between this utterance and other utterances in the meeting. The motivation for this last feature is to find utterances that are repeats (or near-repeats) of other utterances – repetition may correlate with importance.

Other features: In addition to the ngram and ngram overlap features, we also include term frequency – inverse document frequency (tf-idf) features to capture the information content of the ngrams in the utterance. Specifically we compute the TF-IDF of each ngram (of sizes 1 through 5) in the utterance, and include the maximum, minimum, average and standard deviation of these values as features of the utterance. We also include speaker-based features to capture who is speaking when. We include the identity of the speaker of the current utterance and those of the previous and next utterances as features. Lastly we include the length of the utterance (in seconds) as a feature.

5.2 Evaluation Results

In this paper we use a Support Vector Machines-based classifier, which is a popular choice for extractive meeting summarization, e.g. (Xie, Liu, & Lin, 2008); we use a linear kernel in this paper. In the results reported here we use the output of the Sphinx speech recognizer, using speaker-independent acoustic models, and language models trained on publicly available meeting data. The word error rate was around 44% – more details of the speech recognition process are in (Huggins-Daines & Rudnicky, 2007). For training purposes, we merged the annotations from the two annotators by choosing a "middle or lower ground" for all disagreements. Thus, if for an utterance the two labels are "definitely show" and "don't show", we set the merged label as the middle ground of "maybe show". On the other hand if the two labels were on adjacent levels, we chose the lower one – "maybe show" when the labels were "definitely show" and "maybe show", and "don't show" when the labels were "maybe show" and "don't show". Thus only utterances that *both* annotators labeled as "definitely show" were also labeled as "definitely show" in the merged annotation. We plan to try other merging strategies in the future. For testing, we evaluated against each annotator's labels separately, and averaged the results.

	Definitely show	Maybe show	Don't show
Precision	0.21	0.47	0.72
Recall	0.16	0.40	0.79
F-measure	0.16	0.43	0.75
Accuracy	61.4%		

Table 3 Results of the 3-Way Classification

Table **3** presents the accuracy, precision, recall and f-measure results of the 3-way classification task. (We use the Weka implementation of SVM that internally devolves the 3-way classification task into a sequence of pair-wise classifications. We use the final per-utterance classification here.) Observe that the overall accuracy of 61.4% is only 11% lower relative to the accuracy obtained by comparing the two annotators' annotations (69%, Table **2**). However, the precision, recall and f-measure values for the "definitely show" class are substantially lower for the predicted labels than the agreement between the two annotators. The numbers are closer for the "maybe show" and the "don't show" classes. This implies that it is more difficult to accurately detect utterances labeled "definitely show" than it is to detect the other classes. One reason for this difference is the size of each utterance class. Utterances labeled "definitely show" are only around 14% of all utterances, thus there is less data for this class than the others. We also ran the algorithm using manually transcribed data, and found improvement in only the "Definitely show" class with an f-measure of 0.21. This improvement is perhaps because the speech recognizer is particularly prone to getting names and other technical terms wrong, which may be important clues of noteworthiness.

Table **4** presents the ROUGE-1 F-measure scores averaged over the 6 meetings. (ROUGE is described briefly in section 4.3 and in detail in (Lin, 2004)). Similar to the inter-annotator agreement computations, we computed ROUGE between the text of the utterances labeled "definitely show" by the system against that of utterances labeled "definitely show" by the two annotators. (We computed the scores separately against each of the annotators in turn and then averaged the two values.) We did the same thing for the set of utterances labeled either "definitely show" or "maybe show". Observe that the R1-F score for the "definitely show" comparison is nearly 50% relative higher than the trivial baseline of labeling every utterance as "definitely show". However the score is 30% lower than the corresponding inter-annotator agreement. The

corresponding R1-Fmeasure score using manual transcriptions is only marginally better – 0.47. The set of utterances labeled either definitely or maybe shows (second row of table 4) does not outperform the all-utterances baseline when using automatic transcriptions, but does so with manual transcriptions, whose R1-F value is 0.74.

Comparing What	R1-Fmeasure
Definitely show	0.43
Definitely or maybe show	0.63

Table 4 ROUGE Scores for the 3-Way Classification

These results show that while the detection of definitely show utterances is better than the trivial baselines even when using automatic transcriptions, there is a lot of room for improvement, as compared to human-human agreement. Although direct comparisons to other results from the meeting summarization literature are difficult because of the difference in the datasets, numerically it appears that our results are similar to those obtained previously. (Xie, Liu, & Lin, 2008) uses Rouge-1 F-measure solely, and achieve scores between 0.6 to 0.7. (Murray, Renals, & Carletta, 2005) also achieve Rouge-1 scores in the same range with manual transcripts.

The trend in the results for the two 2-way classifications is similar to the trend for the inter annotator agreements. Just as inter-annotator accuracy increased to 87% for the "definitely show" vs. "others" classification, so does accuracy of the predicted labels increase to 88.3%. The f-measure for the "definitely show" class falls to 0.13, much lower than the inter-annotator f-measure of 0.53. For the "don't show" vs. "others" classification, the automatic system achieves an accuracy of 66.6%. For the "definitely plus maybe" class, the f-measure is 0.59, which is 22% relatively lower than the inter-annotator f-measure for that class. (As with the 3-way classification, these results are all slightly worse than those obtained using manual transcriptions.)

5.3 Useful Features

In order to understand which features contribute most to these results, we used the Chi-Squared test of association to find features that are most strongly correlated to the 3 output classes. The best features are those that measure word overlaps between the utterances and the text in the agenda labels and the notes in previous meetings. This is not a surprising finding – the occurrence of an ngram in an agenda label or in a previous note is highly indicative of its importance, and

consequently that of the utterances that contain that ngram. Max and average TF-IDF scores are also highly ranked features. These features score highly for utterances with seldom-used words, signifying the importance of those utterances. Domain independent ngrams such as "action item" are strongly correlated with noteworthiness, as are a few domain *dependent* ngrams such as "time shift problem". These latter features represent knowledge that is transferred from earlier meetings to latter ones in the same sequence. The identity of the speaker of the utterance does not seem to correlate well with the utterance's noteworthiness, although this finding could simply be an artifact of this particular dataset.

6   Related Work

Noteworthiness detection is closely related to meeting summarization. Extractive techniques are popular, e.g. (Murray, Renals, & Carletta, 2005), and many algorithms have been attempted including SVMs (Xie, Liu, & Lin, 2008), Gaussian Mixture Models and Maximal Marginal Relevance (Murray, Renals, & Carletta, 2005), and sequence labelers (Galley, 2006). Most approaches use a mixture of ngram features, and other structural and semantic features – a good evaluation of typical features can be found in (Xie, Liu, & Lin, 2008). Different evaluation techniques have also been tried, with ROUGE often being shown as at least adequate (Liu & Liu, 2008). Our work is an application and extension of the speech summarization field to the problem of assistive note-taking.

7   Conclusions and Future Work

In our work we investigated the problem of detecting the noteworthiness of utterances produced in meetings. We conducted a Wizard-of-Oz-based user study to establish the usefulness of extracting the text of utterances and suggesting these as notes to the meeting participants. We showed that participants were willing to accept about 35% of these suggestions over a sequence of 9 meetings. We then presented a 3-level noteworthiness annotation scheme that breaks with the tradition of 2-way "in/out of summary" annotation. We showed that annotators have strong agreement for separating the highest level of noteworthiness from the other levels. Finally we used these annotations as labeled data to train a Support Vector Machine-based classifier which performed better than trivial baselines but not as well as inter-annotator agreement levels.

For future work, we plan to use automatic noteworthiness predictions to suggest notes to meeting participants during meetings. We are also interested in training the noteworthiness detector directly from the notes that participants took in previous meetings, thus reducing the need for manually annotated data.

Reference

Banerjee, S, and A. I. Rudnicky. "Segmenting Meetings into Agenda Items by Extracting Implicit Supervision from Human Note-Taking." Proceedings of the International Conference on Intelligent User Interfaces. Honolulu, HI, 2007.

Banerjee, Satanjeev, and A. I. Rudnicky. "An Extractive-Summarization Baseline for the Automatic Detection of Noteworthy Utterances in Multi-Party Human-Human Dialog." IEEE Workshop on Spoken Language Technology. Goa, India, 2008.

Galley, Michel. "A Skip-Chain Conditional Random Field for Ranking Meeting Utterances by Importance." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Sydney, Australia, 2006.

Huggins-Daines, David, and A. I. Rudnicky. "Implicitly Supervised Language Model Adaptation for Meeting Transcription." Proceedings of the HLT-NAACL. Rochester, NY. 2007.

Lin, Chin-Yew. "ROUGE: A Package for Automatic Evaluation of Summaries." Proceedings of the ACL-04 Workshop: Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics, 2004. 74-81.

Liu, Feifan, and Y. Liu. "Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries." Proceedings of ACL-HLT. Columbus, OH, 2008.

Murray, Gabriel, S. Renals, and J. Carletta. "Extractive Summarization of Meeting Recordings." Proceedings of Interspeech. Lisbon, Portugal, 2005.

Penn, Gerald, and X. Zhu. "A Critical Reassessment of Evaluation Baselines for Speech Summarization." Proceedings of ACL-HLT. Columbus, OH, 2008.

Xie, Shasha, Y. Liu, and H. Lin. "Evaluating the Effectiveness of Features and Sampling in Extractive Meeting Summarization." IEEE Workshop on Spoken Language Technology. Goa, India, 2008.