A Boosting Approach to Topic Spotting on Subdialogues

Kary Myers KARY@CS.CMU.EDU
Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Michael Kearns MKEARNS@RESEARCH.ATT.COM
Satinder Singh BAVEJA@RESEARCH.ATT.COM
Marilyn A. Walker WALKER@RESEARCH.ATT.COM
AT&T Labs, 180 Park Avenue, Florham Park, NJ 07932 USA

Abstract

We report the results of a study on topic spotting in conversational speech. Using a machine learning approach, we build classifiers that accept an audio file of conversational human speech as input, and output an estimate of the topic being discussed. Our methodology makes use of a well-known corpus of transcribed and topic-labeled speech (the Switchboard corpus), and involves an interesting double use of the BOOSTEXTER learning algorithm. Our work is distinguished from previous efforts in topic spotting by our explicit study of the effects of dialogue length on classifier performance, and by our use of off-the-shelf speech recognition technology. One of our main results is the identification of a single classifier with good performance (relative to our classifier space) across all subdialogue lengths.

1. Introduction

While significant advances have been made over the last two decades in automatic speech recognition (ASR) in controlled acoustic environments, major challenges remain for ASR in noisy environments and conversational speech, such as everyday human-human dialogue. For example, while speech recognizers achieve word error rates as low as 7.8% for read speech in quiet environments, such as news broadcasts (Pallet, et al., 1999), the best recognizers for conversational speech over the telephone achieve word error rates around 36% (Martin, et al., 1998). It seems likely that highly accurate transcriptions of conversational speech will be beyond ASR technology in the near term. However, even noisy ASR transcripts may be of value for a variety of natural language problems.

This paper reports the results of a study on the problem of *conversational topic spotting*. Rather than striving for a faithful transcription of conversational spoken language, we consider the potentially easier but still useful goal of estimating the topic of the conversation from among a fixed set of possible topics. As an example of a potential application of this topic spotting task, we are building a system named EnE (for "Eyes and Ears") that will reside in a coffee room at AT&T Labs and will include a PC, microphone, video camera, and other sensors and actuators. We want the system to learn to interact in simple yet natural ways with the community of humans in its environment. As such, EnE must have an appropriately high-level internal representation of the current state of the environment. To complement other features we have designed, we would like to include features that can provide even coarse and noisy estimates of the current topic of conversation, leading us to undertake the work we describe here.

Topic spotting has been the subject of several previous studies. Like many earlier works, we use the Switchboard corpus of conversational speech (Godfrey, Holliman, & McDaniel, 1992). This corpus consists of audio files and text transcriptions, labeled by topic, for roughly 2500 spontaneous telephone conversations on topics such as music, books, movies and sports. A partial transcription of such a conversation is in Table 1. We address the problem of learning, from Switchboard corpus training data, a classifier that accepts as input only an audio file (and no text transcript), and outputs a topic label.

Our interest in eventually incorporating our learned classifier into the EnE system places some constraints on our study and solution that may be of independent interest. First, our desire for the EnE system to be able to form even crude estimates of topic from ever-shorter fragments of conversation (or *subdialogues*)[1] has led us to undertake

[1]We want the EnE system to form topic estimates from short subdialogues in order to make natural and timely conversational contributions—for instance, giving the score of a recent game if it detects that the topic is sports.

Table 1. Verbatim transcript of the first 10 turns of a Switchboard conversation on the topic Music. As described in Section 2, we remove the first five turns (marked with asterisks) from the training and test data, since they often "give away" the topic.

```
*A.1:   [Laughter].
*B.2:   Uh-huh.
*A.3:   Um, I guess we're supposed to talk about music.
*B.4:   Okay.
*A.5:   And, uh, let me go ahead and push one here.
        [tone] Uh, do you ha-, are you a musician
        yourself?
 B.6:   Uh, well, I sing.
 A.7:   Uh-huh.
 B.8:   I don't play an instrument.
 A.9:   Uh-huh.  Where, do you sing in, in a choir or
        a choral group?
B.10:   Oh, not right now.
```

a systematic study of the effects of decreasing test set conversation length on classifier accuracy. To our knowledge, all previous topic spotting studies on the Switchboard corpus have measured classifier accuracy on full conversations only. Second, since we are not speech recognition specialists, we incorporated "off-the-shelf" ASR technology rather than building specialized ASR systems for the task. Third, the desire for real-time interaction induces a bias towards faster performance, even at the possible expense of classification accuracy. Finally, we selected a subset of Switchboard topics with the coffee room target environment in mind.

Our learning algorithm involves an interesting double use of the BOOSTEXTER algorithm for document classification (Schapire & Singer, 2000), combined with the WATSON ASR system (Sharp, et al., 1997), both developed at AT&T Labs. Our results show we can identify the topic on complete Switchboard conversations with a classification accuracy of 45% (versus the majority-class baseline of 13%).[2] Furthermore, our results on the newly defined subdialogue problem establish a baseline for topic spotting on smaller segments of Switchboard conversations. We show that we can achieve topic classification accuracies ranging from 2% to 32% absolute improvement over the majority-class baseline, depending on subdialogue length, and that, as expected, the topic classification problem becomes more difficult as subdialogue length decreases. One of our main results is the identification of a single classifier whose performance is competitive (within our classifier space) across all subdialogue lengths.

The main contributions of this work are the first study of the effects of test conversation length on topic classification accuracy; a careful analysis of the effects of algorithm parameters (which include the way the Switchboard data are segmented for training, and the size of the grammar learned for the ASR) on classifier accuracy; and a new topic spotting algorithm using off-the-shelf technologies like the BOOSTEXTER classifier and the WATSON ASR system.

2. Methodology and Algorithms

This section describes both the data used to train and test our topic spotting classifiers and our learning algorithm. It begins with a summary of the Switchboard corpus.

2.1 The Switchboard Corpus

To create the Switchboard corpus (Godfrey, Holliman, & McDaniel, 1992), an automated system prompted the two participants in each telephone conversation to discuss, in unrestricted conversational English, a designated topic from a set of approximately 70 topics. For our experiments, we selected 10 topics that seemed representative of the type of "small talk" often encountered in public areas. We randomly divided the conversations for each topic into disjoint training, validation, and test sets, as shown in Table 2. (Note that this resulted in zero counts for a single category in both the validation and test sets.) The Sports topic forms the majority class (13%) over all the data, which serves as our classification accuracy baseline.

Table 2. Counts of conversations by topic in the training, validation, and test sets, obtained by random division of the full data set. The Sports topic combines conversations on Football, Basketball, Golf, and Baseball, and forms the majority class, accounting for $38/292 = 13.0\%$.

TOPIC	TRAIN	VALID	TEST
BOOKS AND LITERATURE	15	3	0
EXERCISE AND FITNESS	26	2	4
FAMILY LIFE	27	3	1
MOVIES	26	4	7
MUSIC	25	3	3
PETS	25	5	2
RECIPES, FOOD, COOKING	27	3	3
RESTAURANTS	15	1	3
WEATHER AND CLIMATE	17	0	4
SPORTS	31	5	2
Total	**234**	**29**	**29**

In addition to audio files, each five-minute conversation has a *verbatim* text transcript[3] produced by paid human listeners, as well as a label indicating which topic the participants were prompted to discuss. Thus, each conversation in the Switchboard corpus may be viewed as a triple: the audio

[2] Direct quantitative comparison of our results to earlier topic spotting studies is difficult due to the experimental differences stemming from the EnE constraints described above.

[3] The term *verbatim* distinguishes these transcripts from a noisy version discussed shortly.

of the conversation, an accurate text transcription, and the topic label. While our learning algorithm takes advantage of the verbatim transcripts, (as have previous approaches to topic spotting) the topic classifiers we learn classify conversations based only on the audio data.

Since one of our main objectives is a systematic study of the dependence of topic spotting performance on the length of the conversation being classified, we augmented the Switchboard data by dividing the audio data of the validation and test sets into *subdialogues* of decreasing lengths (full conversations, halves, quarters, and so on), thus creating six distinct subdialogue lengths for measuring classification accuracies. As we move to shorter subdialogues (say, quarters), the number of validation and test examples increases (by a factor of four), while the length of the corresponding audio decreases (by a factor of four). The subdialogues are always labeled by the same topic as the full source dialogue. Clearly the topic spotting problem becomes more difficult as we reduce conversation length. (Consider the limiting case where each "subdialogue" contains only a single word.)

2.2 A Two-Phase Learning Algorithm

Looking ahead slightly, we note that our learning algorithm has three parameters that can together assume 72 different values. We thus run our algorithm 72 times (exhaustively exploring the parameter space) to obtain 72 different classifiers. Since this set is too large to optimize directly on the test set, we use the validation set to select the "best" hypothesis for each subdialogue length. All results are then reported on the remaining independent test set.

We now describe our learning algorithm. Figure 1 shows a block diagram of the two-stage classifier constructed by our learning algorithm. At the highest level, this classifier first gives the audio input file to the ASR, producing what we call the *ASR transcript*. This transcript, which is quite different from the verbatim transcripts that form part of the training data discussed above, should be thought of as an extremely noisy text transcript of the audio. (See the sample ASR transcript in Table 3.) However, we can still hope to succeed in the topic spotting task as long as these errors are nearly invariant across topics (Seymore & Rosenfeld, 1997). The ASR transcript is then fed to a text topic classifier, which outputs the topic for the audio input (or more precisely, a weighting over the 10 topics). To instantiate such a two-stage classifier, we must learn two things: a grammar for the ASR, and a text topic classifier mapping noisy ASR transcripts to topics. We discuss these in turn.

2.3 Phase One: Learning the ASR Grammar

The WATSON ASR requires that we provide a language model or *grammar* specifying the universe of phrases or

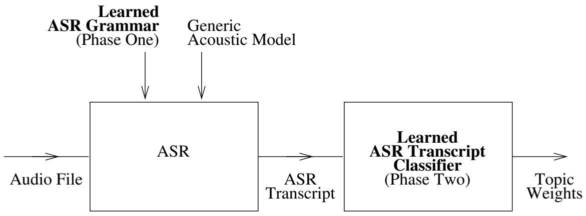

Figure 1. Simplified representation of the components of our two-stage topic spotting classifiers.

Table 3. Excerpts from an ASR transcript for the conversation in Table 1. The ASR has successfully spotted the keyword "sing", which is indicative of the Music topic, but much of the transcript consists of false recognition of unrelated words.

```
...east any uh...the deer the grass dip
...that i dip...sing ankle cook
...the fish...cat the sing yeah and um
```

sentences we expect the ASR to recognize in the audio input. One approach would be to build a grammar from a very long list of English phrases to cover as many utterances as possible (the so-called *large vocabulary recognition* (LVR) approach uses several thousand phrases). We chose to *learn* considerably smaller grammars because of our interest in the fastest possible classification, as the ASR running time grows with grammar size, which is not inconsiderable for very large grammars under real-time constraints. However, there are performance trade-offs of a precision-recall nature: smaller grammars typically result in more accurate spotting of the included phrases, at the obvious expense of missing phrases not in the grammar. For this reason, we compare performance across a number of grammar sizes, all of them considerably smaller than in LVR.

How will we learn the ASR grammar? We chose to take advantage of the verbatim transcripts, applying the BOOSTEXTER text classification algorithm to them. BOOSTEXTER (Schapire & Singer, 2000) is a document classification algorithm that works by finding a linear combination of simple rules, each of which has a (perhaps slight) advantage over random guessing on some (re)weighting of the training data. The algorithm proceeds in rounds, with each round adding a new rule to the linear combination. Our motivation in using BOOSTEXTER was an interest in its performance in this application, as well as the ease with which one can naturally extract a variable sized set of rules from its output, as described below.

The verbatim transcripts from the Switchboard corpus and their topic labels were given to BOOSTEXTER, which was run for 1000 rounds to produce a text classifier comprised of 1000 classification rules.[4] These rules test whether a particular sequence of one to five words appears in the transcript (that is, the rule *fires*), and assign weights to the 10 topics if it does.

While the usual use of such a BOOSTEXTER rule list would be to take a linear sum of the firing rules to determine the output weightings on topics, here we would like to derive an ASR grammar from the rule list by selecting the word sequences from the rules that "best" distinguish among topics. As an example, consider the two BOOSTEXTER rules and weights in Table 4. As we might hope, the presence of the word sequence "listen to" yields a strong positive weight for the topic Music and negative weights for everything else. In contrast, the word sequence "uh yeah" leads to weights comparable in magnitude across topics. Intuitively, we want to keep "listen to" in our ASR grammar, and eliminate "uh yeah". One way to quantify this is with the *variance* of the topic weights given by BOOSTEXTER. In the example here, the weights for "listen to" have a much higher variance than those for "uh yeah". We ranked the rules (word sequences) learned by BOOSTEXTER by their weight variances and selected the best ones for our grammar. In keeping with our earlier comments on investigating the performance variation over (relatively small) grammar sizes, we examined performance for ASR grammars with 20, 100, 200, and 400 word sequences. This *grammar size* is one of the three parameters of our training procedure.

Table 4. Weights assigned to each topic by two BOOSTEXTER rules when the given word sequence appears in the text.

TOPIC	'LISTEN TO'	'UH YEAH'
BOOKS AND LITERATURE	-0.183	-0.474
EXERCISE AND FITNESS	-0.972	-0.548
FAMILY LIFE	-0.815	-0.077
MOVIES	-1.604	0.530
MUSIC	2.044	-1.122
PETS	-1.022	-0.976
RECIPES, FOOD, COOKING	-1.089	-1.542
RESTAURANTS	-2.371	-0.998
WEATHER AND CLIMATE	-0.605	0.456
SPORTS	-0.639	-0.138
Variance	**1.306**	**0.470**

One possible concern with this approach to learning the ASR grammar is that by training BOOSTEXTER on the complete verbatim transcripts, we may "overfit" by choosing word sequences accurate in classifying full conversations but unlikely to appear in shorter subdialogues. For instance, a 5-minute conversation on movies may be quite likely to contain the term "movies", but very short conversations on movies may be better identified by less obvious terms like "to see it". For this reason, we varied the size of the verbatim transcripts used for ASR grammar learning by running BOOSTEXTER on verbatim transcripts divided into smaller segments (full, eighths, and sixteenths). For instance, in the case of eighths, each verbatim transcript yields eight labeled training documents for BOOSTEXTER, each containing a contiguous one-eighth segment of the original verbatim transcript and having the same label as this original transcript. This *verbatim transcript length* is another of the three parameters of our learning algorithm.

2.4 Phase Two: Mapping ASR Transcripts to Topics

The second phase of our algorithm is to learn a mapping from the noisy ASR transcripts into weightings over topics. For this we apply BOOSTEXTER a second time, this time on the ASR transcripts produced by running all the training audio files through the ASR (using the grammar learned in Phase One). Again, we face the issue that topic-identifying phrases from the full ASR transcripts of five-minute training conversations might not be useful for testing on shorter test conversations. Therefore, we created six training data sets of ASR transcripts by dividing the full ASR transcripts into pieces of different lengths, exactly as was done for the verbatim transcripts in the grammar learning.[5] This final parameter of our algorithm is the *ASR transcript length*.

In summary, our two-phase training process has three parameters: the length of the verbatim training transcripts for the first application of BOOSTEXTER (verbatim transcript length); the size of the grammar extracted from this first application of BOOSTEXTER (grammar size); and the length of the ASR transcripts for the second application of BOOSTEXTER (ASR transcript length). These three parameters and their values are summarized in Table 5. Together, we explored $3 \times 4 \times 6 = 72$ different parameter settings, resulting in 72 different classifiers mapping audio input to weightings over topics.

2.5 Testing the System

As mentioned above, exhaustive exploration of the parameter space of our two-phase learning algorithm results in 72 classifiers of the form given in Figure 1. Since this is a large number of classifiers, we used the validation set to choose a single "best" classifier for each of the six subdialogue lengths. This results in a much smaller set of six classifiers, each "tuned" to a different subdialogue length.

[4] We removed the first five conversational turns from each transcript to avoid the tendency to "give away" the topic, as in Table 1 ("Um, I guess we're supposed to talk about music."). We also stripped the remaining text of punctuation, the non-verbal tags like "[Laughter]", and the speaker/turn indicator ("A.1", "B.2").

[5] Since the beginnings of Switchboard conversations are often unnaturally rich in keywords (as noted earlier), we removed the ASR transcripts of the first three audio files for each speaker.

Table 5. Parameters varied in training the learning algorithm. The particular values explored were chosen to yield wide coverage, while obeying our constraints and keeping the total number of classifiers constructed manageable.

VERBATIM TRANSCRIPT LENGTH	GRAMMAR SIZE	ASR TRANSCRIPT LENGTH
FULL	20	FULL
EIGHTH	100	HALF
SIXTEENTH	200	FOURTH
	400	EIGHTH
		SIXTEENTH
		THIRTY-SECOND

In the next section, we present a variety of performance measures over these six classifiers, with all results being reported on the independent test set.

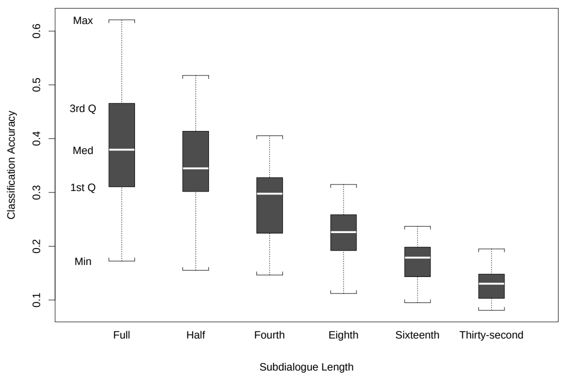

Figure 2. Boxplots of classification accuracy on the test set. Each boxplot corresponds to a single subdialogue length, and summarizes the 72 test set classification accuracies obtained by varying the learning parameters. The majority-class baseline is 13%.

3. Results and Analysis

3.1 Overview of Classification Performance

Before analyzing the test set performance of the six classifiers maximizing validation set accuracy, we first present a summary of test set accuracy across the original pool of 72. Figure 2 shows six boxplots corresponding to the six subdialogue lengths; each boxplot summarizes the test set classification accuracy of the 72 classifiers on the indicated length.[6] (Note that although for this particular analysis we are reporting test errors for a large number of classifiers, no explicit optimization over this large pool is being performed, so there is no risk of overfitting the test data.)

The boxplots clearly demonstrate the increased difficulty of the problem as the subdialogue length get smaller. In addition, the high variability in the boxplots indicate that the parameter settings do indeed impact performance, particularly for longer test subdialogues. (Note that some settings lead to classifiers that perform worse than the majority-class baseline of 13% accuracy.)

3.2 Analysis of Optimized Parameters

We now proceed to analyze the test set performance of the six classifiers obtained by optimizing validation set accuracy at each subdialogue length. The parameter settings of these six classifiers and the numeric values of their test set classification accuracies are given in Table 6 (along with many other quantities we discuss shortly). We note that the classifiers chosen for lengths eighth and sixteenth are identical, meaning we really only have a pool of five classifiers (to be evaluated on six different lengths).

The clearest conclusion we can draw from the optimized parameter settings given in Table 6 is that the longer (200- and 400-term) grammars perform best across all test subdialogue lengths. However, due to our aforementioned interest in fast classification, we explicitly examined the effect of smaller grammar sizes on classification accuracy. Figure 3 shows the test set performance of the best classifiers within each fixed grammar size on each subdialogue length. (Thus, grammar size was held fixed, and the remaining two parameters of learning were optimized for validation set classification accuracy; Figure 3 presents the test set errors of the resulting classifiers.) We see that while the 400-term grammar performs best overall in all but one case, the 200-term grammar is reasonably close for all test lengths, whereas the 20-term grammar performs considerably worse, especially on longer test dialogues. The apparent leveling off of the performance in the 200- to 400-term range suggests that acceptable performance may be obtained far short of large vocabulary methods.

The dependence of classification accuracy on grammar size is confirmed by ANOVA analyses establishing that grammar size is a significant predictor of accuracy at the 0.05 level for all six subdialogue lengths. The importance of the other two parameters (verbatim and ASR transcript lengths) is less clear, as the ANOVAs indicate significance only at some subdialogue lengths. Furthermore, at those subdialogue lengths where significance occurs, there is no clear monotonic relationship between accuracy and verbatim or ASR transcript lengths (unlike the situation for

[6] A boxplot is a graphical representation of a five-number summary: it shows the locations of the minimum, first quartile, median, third quartile, and the maximum points in the data, as labeled in the left-most boxplot in Figure 2.

Table 6. Summary of best classifiers for each subdialogue length, selected by classification accuracy on the validation set. The performance measures are over the test set.

	BEST PARAMETER SETTINGS (VALIDATION SET)			PERFORMANCE MEASURES (TEST SET)		
SUBDIALOGUE LENGTH	VERBATIM TRSCRPT LENGTH	GRAMMAR SIZE	ASR TRSCRPT LENGTH	CLASSIFICATION ACCURACY	MEAN REGRET	MAX REGRET
FULL	FULL	200	FULL	0.4483	0.0806	0.1207
HALF	SIXTEENTH	400	FULL	0.4828	0.0422	0.0862
FOURTH	EIGHTH	200	FOURTH	0.3534	0.0528	0.1034
EIGHTH	SIXTEENTH	400	THIRTY-SECOND	0.3017	0.0201	0.0862
SIXTEENTH	SIXTEENTH	400	THIRTY-SECOND	0.2371	0.0201	0.0862
THIRTY-SECOND	FULL	200	SIXTEENTH	0.1541	0.0517	0.1034

grammar size discussed above). The only significant interactions uncovered by the ANOVAs are between grammar size and verbatim transcript length for longer subdialogues.

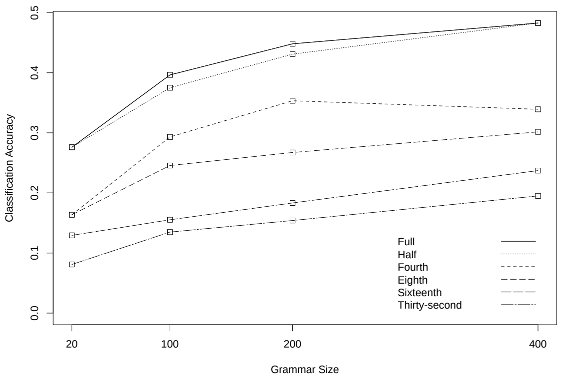

Figure 3. Performance of the best classifiers within each grammar size for each subdialogue length. The classifiers were chosen by performance on the validation set; the figure shows performance over the test set.

3.3 Anytime Classification

Given the large variation in performance across parameter settings and the increasing difficulty of the problem on shorter conversations, it is natural to ask whether there is a *single* setting of the parameters that performs well across *all* test subdialogue lengths. In other words, we would like to find a setting of parameters resulting in a single *anytime* topic spotter—able to form a reasonable estimate of the topic from very short conversations, with performance improving on longer conversations—rather than relying on a multiplicity of classifiers tuned to differing lengths.

The last two columns in Table 6 show that there is such an anytime classifier, as measured by a comparison of mean and maximum *regrets* for each classifier. For each classifier h among the five listed in Table 6, we compared the test set classification accuracy of h on a given length with the best test set classification accuracy at that length among the five. This gives us the regret of h at the given length—that is, how much we lose in performance by using classifier h rather than the best classifier for that length. Average and maximum regrets over all six lengths can then be computed. We see in the table that the classifier tuned for eighth- and sixteenth-length subdialogues jointly minimizes the mean and maximum regrets, and so is the best (and a rather good) anytime classifier.[7]

3.4 Comparison to Perfect ASR

To what extent is the difficulty of the topic spotting problem due to the spoken medium and the challenges of ASR, as opposed to the difficulty of classifying conversational text by topic? The Switchboard corpus makes a direct comparison possible. We trained new BOOSTEXTER classifiers exactly as in Phase Two, this time using the verbatim transcripts rather than the noisy ASR transcripts. This might be likened to training with "perfect" ASR performance. Table 7 compares the performance of these new classifiers on the verbatim transcripts with that of the five classifiers chosen in Table 6 on the ASR transcripts. There is again clear evidence of the increasing difficulty of the pure text classification problem as subdialogue length decreases. There are also considerable differences between pure text and pure audio performance, as much as 55% at the longest dialogue length. Thus the speech recognition component is adding considerably to the difficulty of the problem, as expected.

[7]Note that this notion of regret is weaker than regret computed over all 72 trained classifiers, instead of within the pool of five optimized on the validation set. However, we feel there is not enough test data to compute the former meaningfully, and that since each classifier in the pool of five is obtained by a "good faith" effort to optimize performance for the given length, the results indicate that a good anytime classifier has been found.

Table 7. Test set classification accuracies of BOOSTEXTER when trained and tested on verbatim transcripts vs. ASR transcripts.

SUBDIALOGUE LENGTH	VERBATIM TRANSCRIPTS	ASR TRANSCRIPTS
FULL	1.0000	0.4483
HALF	0.9483	0.4828
FOURTH	0.7845	0.3534
EIGHTH	0.7056	0.3017
SIXTEENTH	0.6442	0.2371
THIRTY-SECOND	0.5089	0.1541

3.5 Incorporating Confidence Measures

While our discussion thus far has focused on the basic and important measure of raw classification accuracy, we can also consider a richer profile of performance that incorporates some measure of confidence in our classifiers' predictions. The final vector of weights over the 10 topics that each of our classifiers produces provides a natural confidence measure: the distance γ between the largest and second-largest weights, which we might call the *margin*. We can use γ to produce the curves in Figure 4 as follows. Let A and B represent the total number of correctly and incorrectly classified test examples, respectively, for some classifier h. Let τ be a threshold such that if the margin γ in the output of h is greater than τ, we make the prediction dictated by the highest-weight topic according to h; otherwise we *abstain* from making any prediction. Each value of τ specifies values a_τ and b_τ, the number of correctly and incorrectly classified test examples with margin exceeding τ. (Note that $A = a_0$ and $B = b_0$.) Now we can calculate the *conditional* classification accuracy as $a_\tau/(a_\tau + b_\tau)$, the proportion of correctly classified examples with margin exceeding τ. This is the y-axis in Figure 4. The x-axis shows the proportion of test examples on which we abstain, which is $1 - (a_\tau + b_\tau)/(A + B)$. As we vary the threshold τ, we trace out a curve profiling the tradeoff between conditional accuracy and abstention for a given classifier.

Figure 4 shows such curves for each of the classifiers in Table 6. Note the left-most point on each curve (corresponding to $\tau = 0$) is just the raw classification accuracy for that classifier as in Table 6. We see that considerable gains in classification accuracy are possible if we abstain on larger fractions of the test examples, confirming the legitimacy of the margin as a predictive confidence measure. Such abstention may be permissible in many contexts, including the motivating EnE application. The curves again show how the problem becomes more difficult as test subdialogue length decreases.

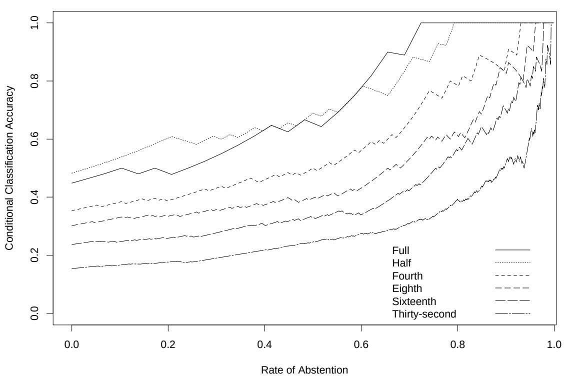

Figure 4. Conditional accuracy vs. abstention curves of the six classifiers in Table 6 over the test set.

3.6 Confusion Matrices

In addition to *overall* classification accuracy, another question is whether certain *topics* are more difficult to identify. Table 8 shows a confusion matrix generated by running the best anytime classifier on eighth-length test subdialogues. Each column of the confusion matrix represents the true topic of an eighth-length subdialogue, while each row entry indicates how the anytime classifier classified that subdialogue. The diagonal of the matrix gives classification accuracy percentages for each topic, where perfect performance would be a diagonal matrix of ones.

The table points to some confusability between the Family and Exercise topics, since the classifier predicted Exercise in 38% of the Family conversations. Likewise, the Sports topic is often predicted for Restaurant and Weather conversations, and Family is often predicted for Sports. Some of this confusability might be explained by the fact that the Sports topic combines several Switchboard topics (see Table 2), so that these conversations range quite broadly, or there may be significant content overlap between, say, the Family topic and Sports or Exercise. In general, though, the table columns show that the classifier does reasonably well in distinguishing topics, always predicting the correct topic most or second-most often.

4. Discussion

While we have introduced a new subdialogue topic spotting task in this paper, aspects of our work—such as the use of a statistical learning method for learning an ASR grammar, and the use of BOOSTEXTER to classify ASR transcripts—are related to previous work. For instance, Wright, Gorin, and Riccardi (1997) automated the selection of salient phrases to form grammars for an automatic call

Table 8. Confusion matrix for best anytime classifier on eighth-length subdialogues, with cells giving column percentages. COUNTS gives the number of subdialogues in the test set for each topic; % PRED indicates how many times the classifier predicted each topic.

TOPICS	BKS	EXER	FAM	MOV	MUS	PETS	REC	REST	WTHR	SPRT	% PRED
BOOKS	—	0.062	0.120	0.110	0.083	0.062	0	0.042	0.120	0.062	0.078
EXERCISE	—	**0.310**	0.380	0.036	0.083	0.062	0.170	0	0.062	0.120	0.110
FAMILY	—	0.190	**0.120**	0.140	0.120	0.062	0.083	0.120	0.031	0.310	0.130
MOVIES	—	0.031	0.120	**0.250**	0.083	0	0	0.120	0.062	0.062	0.100
MUSIC	—	0.031	0	0.071	**0.290**	0.062	0	0	0	0	0.056
PETS	—	0.220	0.120	0.110	0.250	**0.560**	0	0	0.094	0.120	0.150
RECIPES	—	0.031	0.120	0.089	0.042	0	**0.540**	0.120	0.094	0	0.120
RESTAURANTS	—	0.031	0	0.054	0.042	0.120	0.042	**0.210**	0.062	0.062	0.069
WEATHER	—	0.031	0	0.036	0	0	0.083	0	**0.220**	0	0.052
SPORTS	—	0.062	0	0.110	0	0.062	0.083	0.380	0.250	**0.250**	0.140
COUNTS	0	32	8	56	24	16	24	24	32	16	232

routing task known as *How May I Help You?* (HMIHY) by using simple statistical tests on phrases from verbatim transcripts. Schapire & Singer (2000) applied BOOSTEXTER to classify ASR transcripts from the HMIHY task but did not learn the language model for the speech recognizer. McDonough, et al. (1994), working with the Switchboard corpus, explored many approaches, and recommended using ASR transcripts instead of verbatim transcripts for selecting phrases for topic detection, since words that discriminate well on text might be difficult to recognize reliably. Using 4100-word grammars and specialized gender-dependent acoustic models trained on Switchboard audio data, they report classification accuracy of up to 88.3% on full conversations within their set of ten topics. This suggests that, not surprisingly, there is room to improve performance with more sophisticated language and acoustic models rather than off-the-shelf technology.

Our results differ from these and other previous efforts by setting the baseline for topic spotting performance when decreasing the length of Switchboard subdialogues, by using an off-the-shelf speech recognizer, and by demonstrating the existence of a good anytime topic spotting system within our parameter space. As future work, we intend to integrate our anytime classifier in the motivating "Eyes and Ears" application, and to explore the use of richer stochastic language models with the WATSON speech recognizer.

Acknowledgements

We thank Ralph Knag, Rob Schapire, Yoram Singer, Andrej Ljolje, and Rick Rose for many helpful conversations. Kary Myers was supported by an AT&T Labs Fellowship.

References

Godfrey, G., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research development. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 517-520).

Martin, A., et al., (1998). 1998 Hub-5 Workshop: The evaluation. *Notebook for the Ninth Hub-5 Conversational Speech Recognition Workshop*.

McDonough, J., Ng, K., Jeanrenaud, P., Gish, H., & Rohlicek, J.R. (1994). Approaches to topic identification on the Switchboard corpus. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 385-388).

Pallett, D., Fiscus, J., Garofolo, J., Martin, A., Przybocki, M. (1999). 1998 Broadcast News benchmark test results. *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. San Francisco: Morgan Kaufmann.

Schapire, R. & Singer, Y. (2000). BoosTexter: a boosting-based system for text categorization. *Machine Learning, 39*, 135-168.

Seymore, K.,& Rosenfeld, R. (1997). *Large-scale topic detection and language model adaptation* (Technical Report CMU-CS-97-152). Carnegie Mellon University, Pittsburgh, PA.

Sharp, R.D., Bocchieri, E., Castillo, C., Parthasarathy, S., Rath, C., Riley, M., & Rowland, J. (1997). The Watson speech recognition engine. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 4065-4068).

Wright, J.H., Gorin, A.L., & Riccardi, G. (1997). Automatic acquisition of salient grammar fragments for call-type classification. *Proceedings of the Fifth European Conference on Speech Communication and Technology* (pp. 1419-1422).