Improving Text Classification for Oral History Archives with Temporal Domain Knowledge

J. Scott Olsson
Appl. Math. and Sci. Comp./UMIACS
University of Maryland
College Park, Maryland
olsson@math.umd.edu

Douglas W. Oard
College of Information Studies/UMIACS
University of Maryland
College Park, Maryland
oard@glue.umd.edu

ABSTRACT

This paper describes two new techniques for increasing the accuracy of topic label assignment to conversational speech from oral history interviews using supervised machine learning in conjunction with automatic speech recognition. The first, time-shifted classification, leverages local sequence information from the order in which the story is told. The second, temporal label weighting, takes the complementary perspective by using the position within an interview to bias label assignment probabilities. These methods, when used in combination, yield between 6% and 15% relative improvements in classification accuracy using a clipped R-precision measure that models the utility of label sets as segment summaries in interactive speech retrieval applications.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: Miscellaneous

General Terms: Algorithms, Measurement

Keywords: spoken document classification, automatic topic classification, classifying with domain knowledge

1. INTRODUCTION

Interactive information retrieval systems rely heavily on the user's ability to pose good queries and to recognize relevant content. Collections of conversational speech pose unique challenges for both tasks. How is the user to know which words might be correctly indexed without understanding both the way in which individuals spoke and the limitations of speech processing components? And how can we compactly summarize spoken content in ways that permit users to select useful results from large result sets? Modern Web search engines use term sequences for both purposes, accepting query terms that will be matched with terms found in the documents, and displaying document snippets containing occurrences of the query terms. That approach does not transfer well to conversational speech (e.g., recorded meetings, telephone calls, or interviews) because the best available automatic transcription yields sub-

stantial error rates. State of the art Automatic Speech Recognition (ASR) systems achieve word error rates between 15% and 50% on conversational speech [4], with that wide variation resulting from differences in the degree to which the system has been tuned (often at significant expense) to the characteristics of a particular collection. In this paper, we experiment with a 25% word error rate transcription, the best that is presently available for any collection of oral history interviews. Even so, at that error rate, many of the most selective query terms are often misrecognized, and few of the most informative snippets would be completely correct.

When a suitable thesaurus and suitable training data are available, using automatic transcription as a basis for topic classification offers a potentially useful interaction paradigm. Automatically assigned thesaurus terms can be displayed as a "bulleted list" content summary, and iterative query refinement can be done by incorporating thesaurus terms that have been seen to describe useful content. Because topic classification algorithms that leverage broad patterns of term co-occurrence are available, this approach can yield more robust summaries that are less sensitive than snippets would be to variations in the word error rate. Word error rates in large speech collections typically vary systematically by speaker, so this might also help to minimize the natural bias that has been observed from term-based systems in favor of the clearest speakers [14]. On the other hand, implementing thesaurus-based search alone can make formulation of an initial query challenging for untrained users, and search topics that were not anticipated when the thesaurus was created can be particularly difficult to express. The natural approach is therefore to use free text and thesaurus-based techniques together.

These considerations naturally raise the technical question of how accurately it is possible to assign thesaurus terms to spoken content. That is not a question that is easily answered in the abstract, so in this paper we adopt the specific context of assigning thesaurus terms to manually partitioned segments from English oral history interviews based on a one-best ASR transcript. That formulation reveals two salient characteristics of a topic classification problem that are common to many types of sequentially-told stories (e.g., television programs, or the evolution of news reporting over time): (1) the order in which the story is told provides potentially useful evidence, and (2) different aspects of a story evolve over different time scales as it is told. As a simple example, we expect to find a review of prior work early in this paper, experiment results towards the end, and a consis-

tent topical coverage throughout. In this paper we explore how those effects can be leveraged to improve classification accuracy in the context of a richly annotated oral history collection.

The remainder of this paper is organized as follows. Section 2 briefly reviews related work on topic classification for spoken content. Section 3 then describes the test collection, training data and evaluation measures that we used in our experiments. Sections 4, 5 and 6 present algorithms for our baseline kNN classifier, an enhancement using time-shifted classification, and an enhancement using temporal label weighting. Section 7 describes our approach to evidence combination, and section 8 presents the results of our experiments that show improvements in classification accuracy of between 8 and 15% for leaf nodes in the thesaurus, and improvements of between 6% and 13% for interior nodes. Section 9 concludes the paper with some observations on the broader utility of these techniques beyond the collection that we used in our experiments.

2. RELATED WORK

The BBN OnTAP system appears to have been the first to use automatically assigned topic labels to describe the content of speech in an interactive information retrieval system [10]. In their approach, topic labels are presented vertically aligned with the salient sections of the transcript during full-text display so that both can be scrolled together (along with a third vertical region depicting speaker identity).

Byrne et al. presented classification results on parts of the same collection used in this study, using the results of an early ASR system with a higher word error rate [1]. Olsson et al., also using parts of the same collection, later reported classification results where the training examples were taken from a second language [12]. Both showed that kNN was a reasonable approach, given that the problem is multi-label with many topic classes. Iurgel et al. reported classification results on spoken content using combinations of binary support vector machines, although their task contained many fewer classes [8].

A great deal of research has looked at incorporating domain knowledge to improve classification effectiveness for text documents. In [7], domain knowledge from topical hierarchies is used to enrich the document representation for search. Other work has focused largely on compensating for a shortage of available training data [2, 9, 15, 18], sometimes requiring significant modification to the learning algorithm (e.g., [18] developed a modified support vector machine classifier). Our work differs in the type of domain knowledge considered (temporal evidence as opposed to expert knowledge), in that we do not specifically consider the limited training data problem, and in that our application focus is on supporting search in speech collections. Our work also differs from [5] (which exploited temporal evidence for classification), in that we do not adapt to evidence from previously seen stories, but rather to evidence from within the same story (within the same interview).

3. EVALUATION FRAMEWORK

Exploring these questions requires a speech collection, ASR results, a thesaurus, and examples of how recognized words are used with different thesaurus labels. Fortunately, such a collection now exists. In 2005 and again in 2006, the Cross-Language Evaluation Forum (CLEF) Cross-Language Speech Retrieval (CL-SR) track distributed a collection of English oral history interviews with 246 Holocaust survivors, rescuers and witnesses with one-best ASR results, a rich thesaurus, and ground truth mappings between the ASR results and the thesaurus labels. We use those ground truth mappings as the answer key for evaluating classification accuracy, so at the end of this section we describe how those mappings were created and introduce the clipped R-precision measure that we use to characterize classification accuracy. In between, we describe the disjoint training set of mappings between text and thesaurus labels that we used to train our classifier.

3.1 Evaluation Set

The interviews from which the CLEF CL-SR collection was created were conducted by the Survivors of the Shoah Visual History Foundation (VHF)[1] late in the twentieth century and recorded on videotape. Each interview was structured by the interviewer to proceed in roughly chronological order through the interviewee's life experiences, with the first 20% or so typically addressing experiences before the Second World War, the middle 60% typically addressing experiences during the war, and the final 20% typically addressing experiences after the war. Most interviews are in the form of an extended narrative with occasional steering comments from the interviewer, but more structured question-answer formats were also sometimes used. At the end of the interview, interviewees would often hold up artifacts (e.g., photographs) for the camera to record and say a few words about them.

An initial thesaurus for indexing these materials was developed by VHF based on scholarly analysis of events during the time frame the interviewees described. Professional indexers, generally with academic training in disciplines related to the content of the collection, then manually segmented each interview into topically coherent passages that were recorded in a database as a standoff annotation to the spoken content, which at that point was still recorded on videotape. Each segment was then described by the indexer by associating several thesaurus labels with a segment. Operationally, it is useful to think of the segmentation process as having been guided in some way by the thesaurus: when a set of assigned thesaurus terms no longer described what was being discussed, insertion of a segment boundary would be appropriate. When indexers encountered concepts that were not yet present in the thesaurus, they nominated new thesaurus labels for consideration by the thesaurus maintenance team (a thesaurus extension process generally known as "literary warrant"). The resulting thesaurus thus covers the topical scope of the collection quite well. The thesaurus itself consists of two hierarchies, one a set of *part-whole* relations (the "term" hierarchy) and one a set of *is-a* relations (the "type" hierarchy). Figures 1 and 2 show some illustrative examples. These figures illustrate a distinction that we will make throughout this paper, with Figure 1 drawn from the branch of the thesaurus in which geography and time periods appear (what we call the "geography" part) and Figure 2 drawn from the remainder of the thesaurus (which we generically refer to as the "concept" part).

In parallel with the indexing process, the original video-

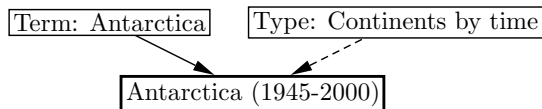[1]The successor to VHF is the USC Shoah Foundation Insti-

Figure 1: An example from the geography part of the CLEF CL-SR topic thesaurus. Solid lines denote *part-whole* **("term") relations, dashed lines denote** *is-a* **("type") relations.**
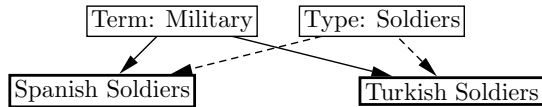

Figure 2: An example from the concept part of the CLEF CL-SR thesaurus. Solid lines denote *part-whole* **("term") relations, dashed lines denote** *is-a* **("type") relations.**

tapes were digitized by VHF and then automatically transcribed by IBM using an ASR system trained on 200 hours of manually transcribed speech from 800 held out interviewees (i.e., interviewees who do not appear in the test collection that we used) [1]. The reported mean word error rate for the one-best transcriptions that were provided by IBM is 25% for most speakers, although for logistical reasons transcriptions with an older system with a mean word error rate of 35% were used in a few cases (e.g., when glitches in the newer system that was still under development resulted in no output). The standoff annotations recorded in the database were used to automatically partition the resulting transcripts into disjoint segments (with some small automated adjustments to avoid splitting transcribed words and to align to segment boundaries with pauses where possible). The resulting segments were then associated with the unique identifiers for each thesaurus term that had been manually assigned by the indexer to that segment, and the result was stored as an XML data structure that was distributed by the Evaluation and Language Resources Distribution Agency (ELDA) to participants in the CLEF 2006 CL-SR collection, version 4.0.

The CLEF-2006 CL-SR test collection was originally intended for evaluation of ranked retrieval, and thus it contains many components (e.g., topics and relevance judgments) that we have not described here. A complete description of that collection can be found in [11]. One preprocessing step used in creating that collection affects the experiments that we report on in this paper, however. When the VHF indexers segmented the collection, they typically created one short segment for each artifact that was displayed at the conclusion of an interview. This resulted in a proliferation of very short segments, each with relatively few ASR-generated words. We elected to automatically remove all very short segments from the collection because judging topical relevance for such sections without seeing the video was often impractical. As a result, those very short segments were not used in our experiments. The remaining 8,104 segments have a unimodal segment length distribution with a median of 4 minutes (about 500 words).
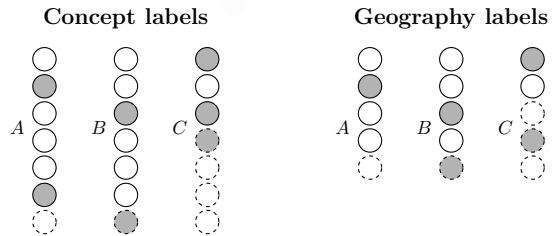
tute for Visual History and Education, or "VHI."

Figure 3: Computing clipped R**-precision for concept and geography label hypotheses on three segments,** A, B, C**. Dashed circles indicate the label falls below the clipping level** M **for the segment.**

3.2 Training Set

The traditional structure of a topic classification problem can be formulated as: given the words produced for that segment by ASR, find the set of thesaurus labels that a human indexer would have assigned. In this paper, we adopt a more general formulation: given a sequence of segments, each with ASR-generated words, find the corresponding sequence of thesaurus label sets. In order to train a classifier, we need training data in which such associations are known. As it happens, an additional set of segments, each with sets of topic labels assigned by the same indexers using the same process, are available. These segments are not distributed with the CLEF CL-SR collection, so we obtained them on a research license from VHI for use in training our system. There were initially over 186,000 segments in this collection, but after deletion of short segments near the end of an interview 168,584 training segments remained.

One important limitation of our training collection is that no ASR results are available for the words spoken in those segments. Instead, VHI provided us with three-sentence summaries written by the indexers for each segment that describe "who, what, when, where" in a fairly structured and stylized way. We therefore trained our classifiers by acting as if these summaries were representative of the words that would have been generated by ASR for those segments.

3.3 Evaluation Measure

In a content description task, we want to show the user only a small number of the best predicted labels. Supposing we could show a user N labels, we might choose as our evaluation measure precision at a cutoff of N. Unfortunately, this would unfairly penalize segments with only a few (say 3) correct labels placed at the top 3 ranks (giving a precision of $3/N$). Alternatively, we might choose a rank based measure such as R-precision (the precision at cutoff R, where R is the number of correct labels for a segment), but this may factor in label hypotheses which can never benefit the user (i.e., if $R > N$).

As a solution to these problems, we take as our measure the *clipped R-precision*. Clipped R-precision is defined as the precision at cutoff M, where

$$M = \begin{cases} R, & R \leq N \\ N, & R > N \end{cases} \quad (1)$$

Consider Figure 3. Three segments, A, B, C, have ranked lists of both concept and geography labels. We would like

to show the user 6 concept and 4 geography labels.[2] First, consider concepts ($N = 6$). Segments A, B have $R > 6$, so their clipped R-precision is $\frac{2}{6}$ and $\frac{1}{6}$ respectively. Segment C has $R = 3, R < N$, so $M = 3$ and its clipped R-precision is $\frac{2}{3}$. The calculation is the same for geography labels, now with $N = 4$. For segments A, B, $R > N$, so $M = N$ and each have clipped R-precisions of $\frac{1}{4}$. For segment C, $R = 2$, so the score is $\frac{1}{2}$. Lastly, we average over segments, so the clipped R-precisions on concepts in this example will be $(\frac{2}{6} + \frac{1}{6} + \frac{1}{3})/3 = \frac{5}{18}$. For geography, we have $\frac{1}{3}$.

Note that this evaluation measure is very severe: we give credit to our system only when the indexer assigned exactly the same content, no credit for being close enough that a savvy user could make sense of it, and no credit for being a perfectly fine assignment (i.e., one that is useful for the purpose of description) that the indexer just did not happen to make (e.g., perhaps because of strictly standardized rules of interpretation). Cumulatively, these effects may be significant because (1) there are very many labels and the segments may have multiple topics assigned (as opposed to a single-label assignment problem in which we would not expect indexers to forget to assign the one appropriate label) and (2) the thesaurus terms often have greater specificity than a user might desire. For example, in Figure 1 we see that `Antarctica (1945-2000)` is a different topic than `Antarctica`. Accordingly, the absolute value of our measure should be interpreted generously when trying to imagine the utility of the labels to the user of an interactive information retrieval system.

4. BASELINE CLASSIFIER

Our baseline is a k-Nearest Neighbors (kNN) classifier using a symmetrized variant of Okapi term weighting [6, 13],

$$w(tf, dl) = \frac{tf}{0.5 + 1.5(\frac{dl}{avdl}) + tf},$$

where $w(tf, dl)$ is the computed term weight, tf is the term frequency, dl is the length of the document in which the term occurs, and $avdl$ is the average document length. It is symmetric in the sense that both testing and training vectors use the same weighting scheme. During classification, term weights are multiplied by their inverse document frequency,

$$idf(df) = \log\left(\frac{D - df + 0.5}{df + 0.5}\right),$$

where D is the total number of segments in training. For convenience, we represent this idf weighting as a matrix vector product between $\mathbf{A}$ (a square matrix with the idf weights on the diagonal) and a document vector. For a test document with vector w_T, we first find the k nearest training vectors (neighbors) w_i, $i = 1, 2, \ldots, k$ in the document space, where our distance measure is the inner product, $(\mathbf{A}w_T)^T w_i$.

The score for class c on test document vector w_T is then computed as the sum of inner products between $\mathbf{A}w_T$ and
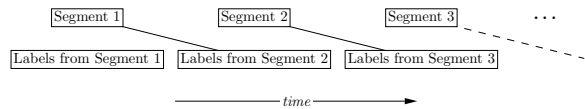

Figure 4: A schematic view of the TSC training setup. Segments are assigned labels from their temporally adjacent segment. Likewise, the classifier predicts labels for temporally adjacent (subsequent) segments.

w_j for $j \in K_c$, $K_c = \{j|$ neighbor w_j has label $c\}$. That is,

$$score(w_T, c) = \sum_{j \in K_c} (\mathbf{A}w_T)^T w_j.$$

For all experiments, we fixed the neighborhood size at $k = 100$, which was found to be roughly optimal for our baseline system.

5. TIME SHIFTED CLASSIFICATION

One new source of information in oral history data is the set of features associated with temporally adjacent segments. Features, here terms, may be class-predictive for not only their own segment, but for the subsequent segments as well. This is an example of *local* temporal evidence.

This intuition may be easily captured by a *time-shifted classification* (TSC) scheme. In TSC, each training segment is labeled with the *subsequent* segment's labels. During classification, each test segment is used to assign labels to its subsequent segment. This is illustrated in Figure 4. Because the last segment in each interview has no associated time-shifted labels, they are discarded in TSC training. Likewise, the first segment from each test interview has no preceding segment which may predict its labels, and so falls back to using only the non-shifted label hypotheses. Note, this approach may easily be extended to predict labels on segments temporally farther away.

Time shifted classification produces scores for labels on segments, just as traditional non-shifted classification. Naturally, we would like to combine these scores with those from the original, non-shifted classification problem. We used a simple linear combination of the scores for a class c and document d,

$$S_{\text{TSC.comb}}(c, d) = \lambda S_{\text{orig}}(c, d) + (1 - \lambda)S_{\text{TSC}}(c, d),$$

where S_{orig} and S_{TSC} are the original and TSC scores respectively.

We evaluated this combination approach on a set of 4,000 segments. For each setting of λ, we computed the clipped R-precision and then took 500 bootstrap resamplings of size 4,000. The mean and confidence intervals of the clipped R-precision are shown at each of several λ settings in Figure 5. Geography and concept labels are plotted separately.

We observe that optimal settings of λ occur at different positions for geography and concept labels. For the best setting on concepts, the time-shifted scores are only barely considered (i.e., λ is around 0.9), while for geography, they are strongly considered (i.e., λ is roughly 0.6). This conforms to our expectations, in that interviews were segmented by change in topic, while successive topics may naturally occur without a change in geography. On both label sets, we see the clipped R-precision varies smoothly with respect to λ.
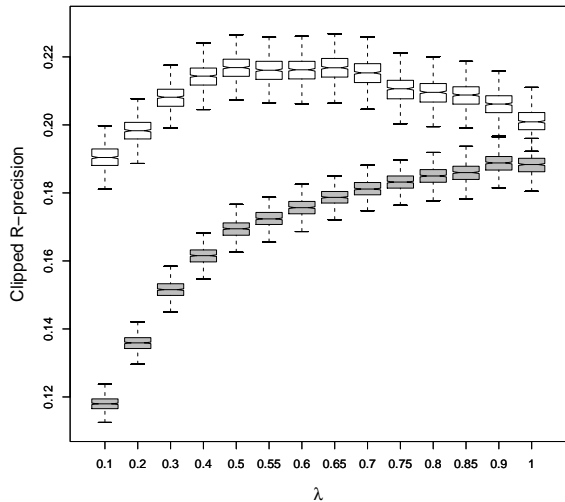
---

[2]It happens that the median number of true concept and geography labels on segments is 3 and 2 respectively. We therefore simulate showing the user twice as many of each label type (6 and 4), which gives a total number of 10 labels for presentation. The average thesaurus label contains four words, so these should easily fit on four display lines.

Figure 5: Clipped R-precision vs. mixing parameter λ for combining original and TSC classification scores. White boxes show results for geography labels, gray boxes show concept labels. Note, this is only a preliminary analysis to gauge the smoothness of the combination method.
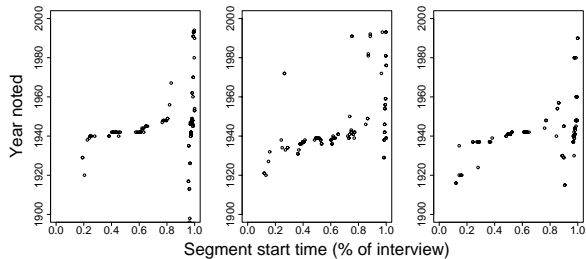

Figure 6: Years spoken in automatic speech recognition transcripts versus the corresponding segment time (as a fraction of total interview time) for three interviews.

6. TEMPORAL LABEL WEIGHTING

We can also benefit from non-local temporal information about a segment. For example, because interviewees were prompted to relate their story in chronological order, we would be less surprised to find a discussion of childhood at an interview's beginning than at its end. This chronological ordering is observed in Figure 6, which shows the years noted in the speech recognition transcripts plotted against segment time for three different interviews. The noted years ramp upwards quickly as the interviewees summarize their childhood, then progress slowly through their adult years, and finally jump about somewhat erratically as artifacts from throughout their life are introduced.

Because of this structure, topics may be more likely to occur at some times than others. For example, discussions of war crime trials are considerably more likely to occur at the end of an interview than at the beginning (simply because war crime trials tend to occur *after* a war). We can exploit this intuition by weighting our label predictions by $p(c, t)$, the probability of label c occurring during the *interval* of interview time t. We call this approach *temporal label weighting* (TLW). These label weights, $p(c, t)$ may be esti-
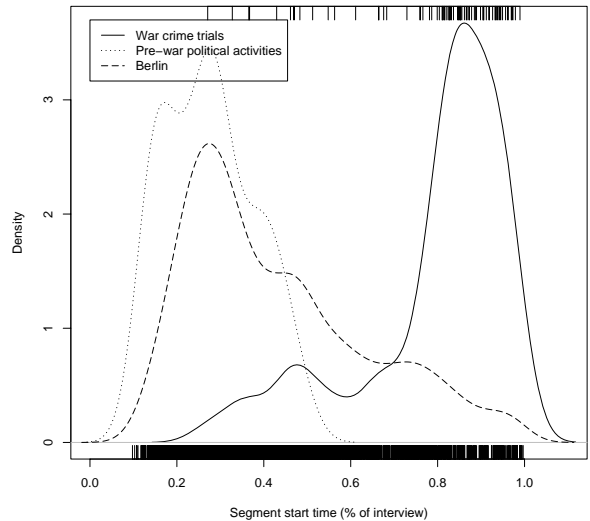

Figure 7: Time density estimates for three commonly occurring labels. The top and bottom *rugs* show where label examples occur, for `war crime trials` and `Berlin`, respectively.

mated using smoothed kernel density estimators on held out data. Figure 7 shows some example time density estimates.

Kernel density estimators are non-parametric estimators for probability density functions, similar in purpose to histograms, except that they are smooth and do not require a bin width be chosen. The intuition is that observations about a point x should contribute to the density, more so if they are nearby, less so if they are far away. This notion of distance is encoded in a *kernel* K, so that the density at a point x is estimated as

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{b}\right),$$

for observations x_i, $i = 1, 2, \ldots, n$, where the *bandwidth* b parameterizes the width of the kernel (specifically, in this case, the bandwidth is the kernel's standard deviation). An applications-oriented introduction to kernel density estimators may be found in [17].

Various kernels may be used, although they are normally chosen to be smooth, unimodal, to peak at 0, and to be a probability density function, i.e., $\int K(u)du = 1$. We produce our time density estimates using a Gaussian kernel density estimator

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right),$$

where the bandwidth is chosen such that (1) the distribution is unimodal for classes with few example and (2) the distribution may have multiple modes when they are strongly supported by available examples. Our default bandwidth is computed according to Silverman's "rule of thumb" (the default in the R statistics package) [16]. In practice, for classes with fewer than 100 examples, we iteratively increase this default smoothing bandwidth until the density function's derivative has no more than one zero crossing (i.e., the function has one maximum). This is illustrated in Figure 8 for an artificial label with only two training examples. With our
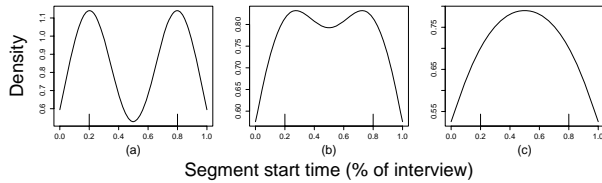
Figure 8: Three choices of bandwidth for smoothing a Gaussian kernel density estimate with very few examples (here 2). In (a), the density is bimodal with the default choice of bandwidth. At (c), a bandwidth is chosen providing a unimodal density function. Note two tick marks on the bottom edge of the graph show the position of the training examples in time.

default bandwidth, the density function is bimodal (Figure 8a), which can not be strongly supported with so few examples. In Figure 8b, the bandwidth is increased slightly, and then again in Figure 8c. We terminate at this final bandwidth, which provides a unimodal density estimate.

Note that our weighting function is a density, so that it approximately integrates to one. This is true, of course, *regardless* of the number of label examples. This is made clear in Figure 7, where `war crime trials` has a greater mode than `Berlin`, despite `Berlin` having many more examples (as seen on `Berlin`'s *rug*—the tick marks on the bottom edge showing the observations' positions). This is reasonable because the preponderance of a label's examples (i.e., its prior probability) is already modeled implicitly by kNN. Now, to estimate $p(c, t)$, we ought to integrate our estimated density function over the temporal extent of the test document. Because the segment durations have fairly low variance however, we approximate our weighting, $p(c, t)$, by the estimated density function for class c at the *start* time of interval t. This approximation will be at least roughly proportional to the integrated probability mass—and has the advantage of not requiring runtime numerical integration, provided the density function is fairly flat.[3] On the other hand, this approximation will be bad where the first derivative of the density function is large. To mitigate this effect, we dampen the values logarithmically before applying the weights to our baseline classification scores. This gives our combination formula

$$S_{\text{TLW.comb}}(c, d) = S_{\text{orig}}(c, d) \times \log\left(1 + p(c, t)\right),$$

where c is the class, d is the document, and $p(c, t)$ is the temporal label weight for label c at the start time t. We use $\log\left(1 + p(c, t)\right)$ because (1) it is positive for $p(c, t) \in (0, \infty)$ and (2) for small $p(c, t)$, $\log\left(1 + p(c, t)\right) \approx p(c, t)$.

7. COMBINING EVIDENCE

We may also combine the local evidence provided by TSC with the less localized evidence provided by TLW. Again, we use a simple linear combination of scores,

$$S_{\text{final}}(c, d) = \gamma S_{\text{TSC.comb}}(c, d) + (1 - \gamma) S_{\text{TLW.comb}}(c, d).$$

As before, we evaluated this combination approach on a set of 4,000 segments. Figure 9 shows the parameter sweep.

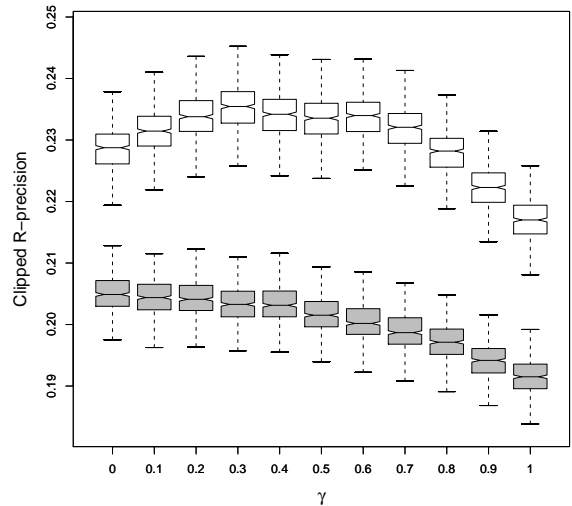[3]To see this, imagine approximating the integral over a small region by drawing a box under the density function.

Figure 9: Clipped R-precision vs. mixing parameter γ for combining TLW and TSC classification Output. White boxes show results forgeography labels, gray boxes show concept labels. Note, this is a preliminary analysis looking for smoothness. We shouldn't, for example, conclude that TSC scores are not used on concepts (we will see that they are).

For each setting of γ, we computed the clipped R-precision and then took 500 bootstrap resamplings of size 4,000. The combination parameter λ (used to produce the TSC results which are here combined with the TLW results), was taken from the similar analysis shown in Figure 5. The mean and confidence intervals of the clipped R-precision are shown at each of several γ settings in Figure 9. Again, we observe that optimal settings of γ occur at different positions for both geography and concept labels. On both label sets, we see the clipped R-precision varies smoothly with respect to γ. In the experimental section, we will determine γ from the held out portion in our cross-fold validation.

8. EXPERIMENTS

Our training set is a collection of 168,584 segments, as described in Section 3. Each segment in the training collection has one or more manually assigned thesaurus terms, from a set of 13,764 unique thesaurus labels, which in turn are drawn from a larger set of about 40,000 labels in the full thesaurus. The training features are words taken from summaries of each segment written by human indexers. The classification task is to assign thesaurus labels to a set of 8,104 new segments, where features are drawn from automatic speech recognition transcripts of the words spoken in those segments. This data is available as the `ASRTEXT2006B` field of the CLEF 2006 version 4.0 CL-SR collection. We also know every segment's position in its interview and its temporally adjacent segments.

To facilitate statistical testing and allow our combination parameters to be tuned on fair data, we use K-fold validation ($K = 10$). Our testing segments are partitioned into K folds and, for each fold, the combination parameters (λ, γ) are chosen to optimize the clipped R-precision on the remaining $K - 1$ folds.[4] We searched for optimal mixing

---

[4]We emphasize that for the experiments reported in this

	$mean(\lambda)$	$s.d.(\lambda)$	$mean(\gamma)$	$s.d.(\gamma)$
Geography	0.59	0.05	0.44	0.14
Concepts	0.93	0.0	0.18	0.01

Table 1: Mean values for the mixing parameters λ, γ (averaged over the cross-validation folds) and their standard deviation.
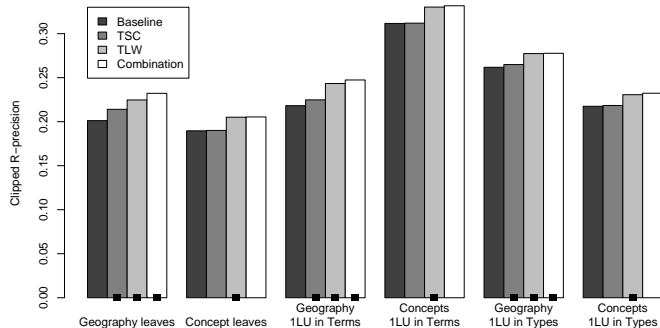

Figure 10: Clipped R-precision for each setting, averaged over the cross-validation folds. Tick marks at the base of a bar indicate that, by a paired t-test with $\alpha = 0.01$, the bar's clipped R-precision is significantly better than the left-adjacent bar.

parameters by stepping through with increment of 0.01. Table 1 shows the mean and standard deviation for the mixing parameters (averaging over the K folds).

Figure 10 shows the final results from our experiments. For each setting, the averaged clipped R-precision over the K validation folds is shown. To test for statistically significant improvement, we compare the clipped R-precision across the K validation folds using paired t-tests with $\alpha = 0.01$.[5] The results of this significance testing are shown in Figure 10: bars which have clipped R-precision significantly larger than the bar to their left are marked with a tick at their base. For example, we see that TSC significantly improves upon the baseline for geography labels (at both the leaves and one level up in each of the two thesaurus hierarchies), but not for concepts. Note that each *grouping* of bars contains at least one tick mark: accordingly, using temporal evidence improves upon our baseline for both label sets, at both levels in both thesaurus hierarchies, with statistical significance. These improvements are tabulated in Table 2.

As Table 2 shows, moving one level up ("1LU") in the "term" (i.e., *part-whole*) hierarchy to classify to the first interior node improves the overall accuracy of concept classification, but does little to benefit geography. Conversely, moving one level up in the "type" (i.e., *is-a*) hierarchy benefits geography classification more than concepts. These im-

---

section we use evidence combination parameters learned through cross-validation, not those learned on the 4,000-segment sets described in the previous sections. This distinction is important because those 4,000-segment sets are a part of the 8,104 set on which we now report results.

[5] Our training sets overlap and thus violate an independence assumption, but the probability of Type I error nevertheless tends to be acceptably small [3]. Alternatively, Fisher sign tests comparing clipped R-precision on paired segments in one fold show the same improvements are significant.

part	location	baseline	TSC&TLW	R.I. (%)
geo	leaf	0.2012	0.2322	+15.4
concept	leaf	0.1896	0.2054	+8.3
geo	1LU term	0.2182	0.2474	+13.4
concept	1LU term	0.3116	0.3317	+6.4
geo	1LU type	0.2618	0.2777	+6.1
concept	1LU type	0.2175	0.2323	+6.8

Table 2: Averaged clipped R-precision for each label set and thesaurus level, for both the baseline and combination approach. The relative improvement (R.I.) using the combined temporal evidence is also shown.

provements are not surprising by themselves—the smaller number of interior nodes simply results in an easier classification problem. In both cases, however, further statistically significant improvements of about 6% are still observed even over the stronger of the two baselines when TSC and/or TLW are applied (and mean values for the combination are never lower than either used alone). This indicates that TSC and TLW, and the combination strategy that we have employed, have utility across a range of thesaurus granularities that might be important in practical applications. This analysis also tells us something about how far the temporally informed methods are moving class hypotheses in the hierarchy to make correct class assignments. If, for example, the temporal evidence was only able to correct a class assignment having a common parent node with the correct label, we would expect classification improvements to wash away when class hypotheses were pushed up the hierarchy. As this does not occur, it appears the proposed methods are also correcting many "far misses" in the topic thesaurus.

9. CONCLUSION

The most obvious limitation to the techniques that we have described is the requirement for both a thesaurus (or some other source of appropriate topic labels) and a training set in which those labels have been associated with text in a way that is representative of how the classifier should behave. Of course, that same condition applies to any text classification problem based on supervised machine learning—all that we have really done is remove the document independence assumption by observing that in this collection, classification assignments do indeed depend on both the absolute and the relative position of segments within an interview.

This suggests several directions for future research. The most fundamental, perhaps best thought of as research in digital libraries rather than topic classification, is to identify other applications that exhibit similar properties and for which a suitable topic inventory is available or could affordably be constructed. A second research direction would be to raise our baseline by, for example, automatically transforming the human-written summaries from the training collection into something more like ASR output. This would amount to fundamental research in feature set transformation for topic classification with ASR input, and it seems likely that benefits could accrue from such an approach. Of course, we'd also hope to compare that approach to training on a complete set of ASR transcripts.
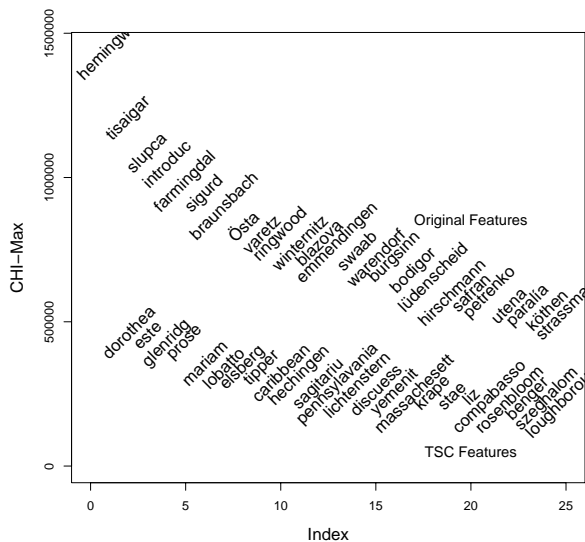
Figure 11: Features sorted by $\chi^2_{\max}$ score on both the original and time-shifted classification problem. TSC features are less informative and have a different feature ordering than the unshifted problem.

A third research direction, and the one most directly inspired by our results, is to explore other ways of leveraging position and sequence dependencies. One obvious approach to try would be a Hidden Markov Model (or some other form of sequence model) in which prior label assignments are used to bias classification decisions. Another approach to try would be to apply a decay function that decreases the contribution of individual words to a category as those words appear further back in time. Considering a more nuanced decomposition of the thesaurus than the geography vs. everything else approach that we tried might also yield additional insights. And, at the most basic level, a range of functions for combining evidence remains to be explored.

For time-shifted classification, features predictive for a segment are likely to be different than those predictive for adjacent segments. This may be especially important when feature selection is used. Consider, for example, that χ^2 feature selectors [19] are based on testing for term-class independence—and this will surely vary between the traditional and TSC case. Figure 11 shows the most predictive features, according to $\chi^2_{\max}$, for both the original and TSC case. In this study, we considered only the all-features case. We expect future work may show additional improvements by incorporating feature selection with TSC.

Ultimately, the value of topic classification is revealed in the way the results are actually used, so studying the behavior of searchers presented with a system that incorporates both text-based and topic-based speech searching will be important. Machines further down a processing pipeline can also use topic classification. For example, topic classification can serve as a source of vocabulary with which to augment an index, either by using terms from the topic labels directly, or by using the topics as pivot points in a blind feedback strategy. So extrinsic evaluations in which the utility of topic classification is assessed through its influence on ranked retrieval will also be important.

So, much remains to be done. But we should emphasize here in conclusion what we have shown—that the structure of stories told in the form of oral history interviews can be leveraged to improve topic classification effectiveness. With the substantial investments now being made in ASR for conversational speech, we can reasonably anticipate the creation of new collections for which these techniques should be directly applicable.

Acknowledgments

10. REFERENCES

[1] W. Byrne et al. Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives. *IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing*, 12(4):420–435, July 2004.

[2] A. Dayanik et al. Constructing Informative Prior Distributions from Domain Knowledge in Text Classification. In *SIGIR'06*.

[3] T. G. Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.*, 10(7), 1998.

[4] J. Fiscus et al. The Rich Transcription 2006 Evaluation Overview and Speech-To-Text Results. In *3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Recognition Workshop*, 2006.

[5] G. Forman. Tackling Concept Drift by Temporal Inductive Transfer. In *SIGIR'06*.

[6] Martin Franz. In *unpublished correspondence*.

[7] E. Gabrilovich and S. Markovitch. Feature Generation for Text Categorization Using World Knowledge. In *IJCAI'05*.

[8] U. Iurgel and G. Rigoll. Spoken Document Classification with SVMs using Linguistic Unit Weighting and Probabilistic Couplers. *Proceedings of the 17th International Conference on Pattern Recognition*, 2004.

[9] R. Jones et al. Bootstrapping for Text Learning Tasks. In *IJCAI'99 Workshop on Text Mining: Foundations, Techniques and Applications*.

[10] F. Kubala et al. Integrated Technologies for Indexing Spoken Language. *Commun. ACM*, 43(2), 2000.

[11] D. W. Oard et al. Overview of the CLEF-2006 Cross-Language Speech Retrieval Track. In *CLEF CL-SR'06*. http://clef-clsr.umiacs.umd.edu/.

[12] J. S. Olsson et al. Cross-Language Text Classification. In *SIGIR'05*.

[13] S. E. Robertson et al. Okapi at TREC-3. In *Text REtrieval Conference*, 1992.

[14] M. Sanderson and X. M. Shou. Search of Spoken Documents Retrieves Well Recognized Transcripts. In *ECIR'07*.

[15] R. Schapire et al. Incorporating Prior Knowledge into Boosting. In *Machine Learning: Proceedings of the Nineteenth International Conference*, 2002.

[16] B. W. Silverman. *Density Estimation*. Chapman and Hall, London, 1986.

[17] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag, New York, NY, USA, 2002.

[18] X. Wu and R. Srihari. Incorporating Prior Knowledge with Weighted Margin Support Vector Machines. In *KDD'04*.

[19] Y. Yang and J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *ICML'97*.