One-Sided Measures for Evaluating Ranked Retrieval Effectiveness with Spontaneous Conversational Speech

Baolong Liu and Douglas W. Oard

College of Information Studies/UMIACS, University of Maryland, College Park, MD 20742 USA

baolongliu@gmail.com; oard@glue.umd.edu

ABSTRACT

Early speech retrieval experiments focused on news broadcasts, for which adequate Automatic Speech Recognition (ASR) accuracy could be obtained. Like newspapers, news broadcasts are a manually selected and arranged set of stories. Evaluation designs reflected that, using known story boundaries as a basis for evaluation. Substantial advances in ASR accuracy now make it possible to build search systems for some types of spontaneous conversational speech, but present evaluation designs continue to rely on known topic boundaries that are no longer well matched to the nature of the materials. We propose a new class of measures for speech retrieval based on manual annotation of points at which a user with specific topical interests would wish replay to begin.

Categories and Subject Descriptors

H.3.m [**Information Storage and Retrieval**]: Miscellaneous

General Terms: Design, Experimentation

Keywords: Evaluation measures, speech retrieval, simulation

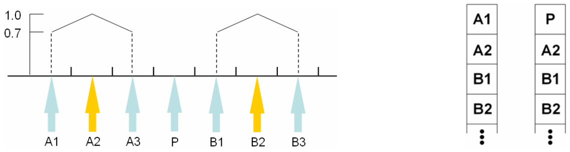
1. INTRODUCTION

Ranked retrieval test collections support insightful, explainable, repeatable and affordable evaluation of the degree to which search systems present results in best-first order. Traditionally, test collections are described as consisting of three components: topics, documents and relevance judgments [5]. Often unstated, test collection designs also reflect details of the metrics that will be computed. For example, in the unknown boundary condition of the TREC spoken document retrieval evaluations, the highest-ranked time that fell within a relevant story was scored as correct and used as a basis for computing average precision [2]. This approach can be applied when ground truth topic boundaries are known, but experience with manual topic segmentation in conversational speech suggests that is not a scalable solution [3]. In the next section we propose a "one-sided" approach that does not rely on pre-segmentation. Section 3 presents simulation results that show that our approach yields stable system rankings over a range of parameter settings; Section 4 presents next steps.

2. A ONE-SIDED MEASURE

Ranked retrieval metrics are typically applied at the scale of entire documents for written text because users with immediate access to full-text displays are typically adept at skimming to rapidly focus on what they seek. Skimming unstructured speech can be far more difficult, however. ASR errors can make automatically produced transcripts hard to interpret, so acoustic replay serves as the ultimate arbiter of utility. Interfaces designed to overcome the strict linearity of acoustic media for general-purpose browsing have been proposed [1], but the more widely adopted approach has been to identify hot spots where the user can begin replay. This suggests a "one-sided" approach to evaluation, in which we seek to characterize the accuracy with which systems are able to identify appropriate points at which replay should begin. Knowing where to stop could, of course, also be useful in some circumstances. But manually labeling appropriate start points offers some potential for minimizing the cost of the relevance assessment process. In a one-sided relevance judgment process, the result is a set of onset points at which the discussion of a topic begins. The cost of the relevance judgment process may also depend on the precision with which onset points can be defined. For the purpose of this study, we fix the granularity at 15 seconds.

An ideal one-sided search system would identify replay start points very near the ground truth onset points identified by relevance assessors and place those points near the top of a ranked list. A suitable evaluation metric must therefore reflect the "quality" (closeness in time) of a hit and the rank at which that hit is presented. A previous study generalized the notion of average precision to accommodate sharp system responses and graded relevance judgments [4]. We are faced with the dual situation: sharp relevance judgments, but with a penalty function to grade the accuracy of the system's response. Figure 1 (a) illustrates two ground truth onset points A2 and B2 and the penalty function in which A2 and B2 would receive the highest score, A1, A3, B1 and B3 lower, and P no credit at all.



(a). Ground truth and penalty function **(b). Two ranked lists**

Figure 1.

Generalized Average Precision (GAP) for the two systems in (b) can then be calculated as $GAP = (\sum_{R_k \neq 0} p_k)/N$ where N is the number of ground truth points, R_k is the score computed using the penalty function for the point at rank k, and $p_k = (\sum_{i=1}^{k} R_i)/k$ is the precision at rank k. Penalty function values are computed in ranked list order without replacement; once a relevance judgment has been used, repeated presentation of nearby points earns no further credit. Thus,

$$GAP_1 = (0.7/1 + (0.7+0.7)/3)/2 = 0.583$$

$$GAP_2 = (1.0/2 + (1.0+0.7)/3)/2 = 0.533$$

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

3. SIMULATION

The shape of the penalty function defines the desired tradeoff between temporal accuracy and ranking effectiveness. Monte Carlo simulation can offer some insight into which characteristics of the penalty function will most affect comparative system rankings, thus helping to focus the design effort on the most important issues. The basic idea is to randomly generate a set of system results based on some representative model of system behavior and then to compare the effect of alternate penalty functions on how mean GAP would rank those simulated systems.

We modeled the ground truth for 10 simulated topics as 6 to 15 onset points drawn uniformly without replacement from a set of 600 possible onset points (150 minutes at a granularity of 15 seconds). 100 ranked lists were then generated by iteratively picking points randomly from the full set of 600 and placing each in an open position on the list according to a Zipfian distribution. We used the two-state Markov model shown in Figure 2 to generate candidate points, setting p=0.5 for this study (representing a system that finds appropriate start points quite often). State 0 represents selection of points without regard to proximate ground truth onset points.

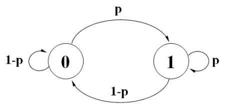


Figure 2. Markov Model as a point generator

Each time the process arrives at State 0, a point ID is emitted uniformly with the distribution $P(ptID = x) = 1/M$ where $ptID$ is the point ID, M is the number of possible points (600). Since it might happen that a randomly selected point is near a ground truth point, the penalty function is then computed for that point and stored. State 1 represents the intended selection of a ground truth onset point by the system, but with some error (modeled by a Gaussian distribution with a standard deviation of 3 points (45 seconds). Each time the process arrives at State 1, a point ID is emitted with a distribution

$$P(ptID = x) = P(x \to GTID_i)\, e^{-(x-GTID_i)^2/2\sigma^2} \Big/ \sqrt{2\pi}\sigma$$

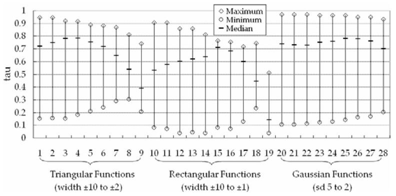
where $GTID_i$ is the *ith* ground truth point ID, $e^{-(x-GTID_i)^2/2\sigma^2} \Big/ \sqrt{2\pi}\sigma$ is the Gaussian probability mass function with a mean $GTID_i$ and a standard deviation σ, $P(x \to GTID_i)$ is the probability that x is close to $GTID_i$

$$P(x \to GTID_i) = \begin{cases} 1/N & \text{if } |x - GTID_i| \le CUTOFF \\ 0 & \text{otherwise} \end{cases}$$

where N is the number of ground truth points. No point will be emitted twice at either state. For the experiment reported here, we chose $\sigma = 3$ and $CUTOFF$=9.

Zipf's law,[1] would predict that the product of the probability of placing an (approximate) actual onset point in the list should be some (nearly) constant multiple of the rank at which that start point is placed. For M=600, the probability of finding an onset point in position k is therefore modeled as P(k) = 0.143/k, since $\sum_{k=1}^{M} P(k) = 1$. We therefore randomly select a rank under that distribution and place the modeled system output there. If that position is occupied, we discard the position and try again.

We generated 28 system rankings using three types of penalty functions. Numbering the functions in order from 1 to 28, the width of triangular functions first ranges from ±10 points down to ±2 points, rectangular function widths then range from ±10 down to ±1 points, and finally Gaussians standard deviations range from 5 down to 2 (with computational truncation at ±10). For each penalty function, we calculate the mean GAP for the same 100 simulated systems over 10 topics and rank those systems in decreasing order of mean GAP. We then compare the system rankings for different penalty functions using Kendall's , a commonly used measure of the stability of system rankings under different conditions [5]. Thus we get a 28×28 matrix, as shown in Figure 4 (darker boxes indicate the higher values). Figure 3 shows the median and range for the 28 penalty functions; triangular functions exhibit smaller variation in , and function 4 (triangular, width ±7) has the highest median (0.785). It therefore would be a good overall choice.



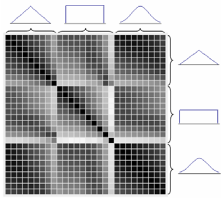


Figure 3. **Figure 4.**

tau statistics for 28 functions **tau matrix in grey scale**

4. NEXT STEPS

Using simulation to guide the design of evaluation measures has yielded results that will help focus our future work. We next need to analyze interactions between sigma and penalty function width and to try longer simulated recordings, since the minimum density of ground truth onset points here (6/600=1%) was relatively large.

5. ACKNOWLEDGMENT

This work has been supported in part by NSF IIS award 0122466.

6. REFERENCES

[1] Arons, B., *SpeechSkimmer: A System for Interactively Skimming Recorded Speech*, ACM TOCHI, 4(1)3-98, 1997.

[2] Garofolo, J. *et al*., *The TREC Spoken Document Retrieval Track: A Success Story,* in TREC-8, 2000.

[3] Gustman, S. *et al*., *Supporting Access to Large Digital Oral History Archives*, in JCDL 2002, pp.18-27.

[4] Kekäläinen, J. *et al*., *Using Graded Relevance Assessments in IR evaluation*, JASIST, 53(13), pp.1120-1129.

[5] Voorhees, E., *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, in SIGIR 1998.