Combining Speech Retrieval Results with Generalized Additive Models

J. Scott Olsson[*] **and Douglas W. Oard**[†]
UMIACS Laboratory for Computational Linguistics and Information Processing
University of Maryland, College Park, MD 20742

Human Language Technology Center of Excellence
John Hopkins University, Baltimore, MD 21211
olsson@math.umd.edu, oard@umd.edu

Abstract

Rapid and inexpensive techniques for automatic transcription of speech have the potential to dramatically expand the types of content to which information retrieval techniques can be productively applied, but limitations in accuracy and robustness must be overcome before that promise can be fully realized. Combining retrieval results from systems built on various errorful representations of the same collection offers some potential to address these challenges. This paper explores that potential by applying Generalized Additive Models to optimize the combination of ranked retrieval results obtained using transcripts produced automatically for the same spoken content by substantially different recognition systems. Topic-averaged retrieval effectiveness better than any previously reported for the same collection was obtained, and even larger gains are apparent when using an alternative measure emphasizing results on the most difficult topics.

1 Introduction

Speech retrieval, like other tasks that require transforming the representation of language, suffers from both random and systematic errors that are introduced by the speech-to-text transducer. Limitations in signal processing, acoustic modeling, pronunciation, vocabulary, and language modeling can be accommodated in several ways, each of which make different trade-offs and thus induce different error characteristics. Moreover, different applications produce different types of challenges and different opportunities. As a result, optimizing a single recognition system for all transcription tasks is well beyond the reach of present technology, and even systems that are apparently similar on average can make different mistakes on different sources. A natural response to this challenge is to combine retrieval results from multiple systems, each imperfect, to achieve reasonably robust behavior over a broader range of tasks. In this paper, we compare alternative ways of combining these ranked lists. Note, we do not assume access to the internal workings of the recognition systems, or even to the transcripts produced by those systems.

System combination has a long history in information retrieval. Most often, the goal is to combine results from systems that search different content ("collection fusion") or to combine results from different systems on the same content ("data fusion"). When working with multiple *transcriptions* of the same content, we are again presented with new opportunities. In this paper we compare some well known techniques for combination of retrieval results with a new evidence combination technique based on a general framework known as *Generalized Additive Models* (GAMs). We show that this new technique significantly outperforms several well known information retrieval fusion techniques, and we present evidence that it is the ability of GAMs to combine inputs non-linearly that at least partly explains our improvements.

The remainder of this paper is organized as follows. We first review prior work on evidence com-

[*] Dept. of Mathematics/AMSC, UMD

[†] College of Information Studies, UMD

Proceedings of ACL-08: HLT, pages 461–469,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

bination in information retrieval in Section 2, and then introduce Generalized Additive Models in Section 3. Section 4 describes the design of our experiments with a 589 hour collection of conversational speech for which information retrieval queries and relevance judgments are available. Section 5 presents the results of our experiments, and we conclude in Section 6 with a brief discussion of implications of our results and the potential for future work on this important problem.

2 Previous Work

One approach for combining ranked retrieval results is to simply linearly combine the multiple system scores for each topic and document. This approach has been extensively applied in the literature (Bartell et al., 1994; Callan et al., 1995; Powell et al., 2000; Vogt and Cottrell, 1999), with varying degrees of success, owing in part to the potential difficulty of normalizing scores across retrieval systems. In this study, we partially abstract away from this potential difficulty by using the same retrieval system on both representations of the collection documents (so that we don't expect score distributions to be significantly different for the combination inputs).

Of course, many fusion techniques using more advanced score normalization methods have been proposed. Shaw and Fox (1994) proposed a number of such techniques, perhaps the most successful of which is known as CombMNZ. CombMNZ has been shown to achieve strong performance and has been used in many subsequent studies (Lee, 1997; Montague and Aslam, 2002; Beitzel et al., 2004; Lillis et al., 2006). In this study, we also use CombMNZ as a baseline for comparison, and following Lillis et al. (2006) and Lee (1997), compute it in the following way. First, we normalize each score s_i as $norm(s_i) = \frac{s_i - min(s)}{max(s) - min(s)}$, where $max(s)$ and $min(s)$ are the maximum and minimum scores seen in the input result list. After normalization, the CombMNZ score for a document d is computed as

$$CombMNZ_d = \sum_{\ell}^{\mathcal{L}} N_{s,d} \times |N_d > 0|.$$

Here, $\mathcal{L}$ is the number of ranked lists to be combined, $N_{\ell,d}$ is the normalized score of document d in ranked list ℓ, and $|N_d > 0|$ is the number of nonzero normalized scores given to d by any result set.

Manmatha et al. (2001) showed that retrieval scores from IR systems could be modeled using a Normal distribution for relevant documents and exponential distribution for non-relevant documents. However, in their study, fusion results using these comparatively complex normalization approaches achieved performance no better than the much simpler CombMNZ.

A simple rank-based fusion technique is *interleaving* (Voorhees et al., 1994). In this approach, the highest ranked document from each list is taken in turn (ignoring duplicates) and placed at the top of the new, combined list.

Many probabilistic combination approaches have also been developed, a recent example being Lillis et al. (2006). Perhaps the most closely related proposal, using logistic regression, was made first by Savoy et al. (1988). Logistic regression is one example from the broad class of models which GAMs encompass. Unlike GAMs in their full generality however, logistic regression imposes a comparatively high degree of linearity in the model structure.

2.1 Combining speech retrieval results

Previous work on single-collection result fusion has naturally focused on combining results from multiple retrieval systems. In this case, the potential for performance improvements depends critically on the uniqueness of the different input systems being combined. Accordingly, small variations in the same system often do not combine to produce results better than the best of their inputs (Beitzel et al., 2004).

Errorful document collections such as conversational speech introduce new difficulties and opportunities for data fusion. This is so, in particular, because even the same system can produce drastically different retrieval results when multiple representations of the documents (e.g., multiple transcript hypotheses) are available. Consider, for example, Figure 1 which shows, for each term in each of our title queries, the proportion of relevant documents containing that term *in only one* of our two transcript hypotheses. Critically, by plotting this proportion against the term's inverse document frequency, we observe that the most discriminative query terms are often not available in both document represen-

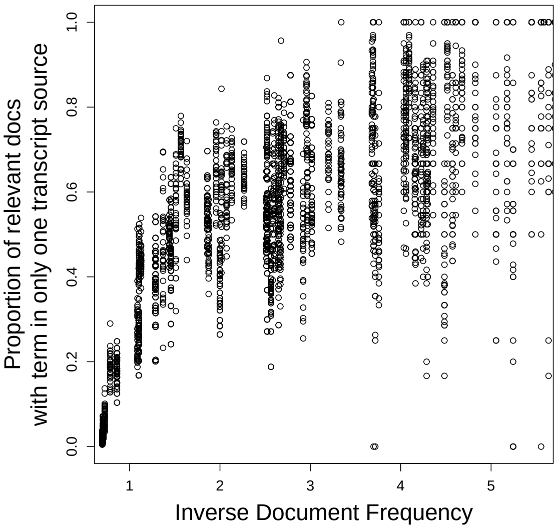

Figure 1: For each term in each query, the proportion of relevant documents containing the term vs. inverse document frequency. For increasingly discriminative terms (higher *idf*), we observe that the probability of only one transcript containing the term increases dramatically.

tations. As these high-*idf* terms make large contributions to retrieval scores, this suggests that even an identical retrieval system may return a large score using one transcript hypothesis, and yet a very low score using another. Accordingly, a linear combination of scores is unlikely to be optimal.

A second example illustrates the difficulty. Suppose recognition system A can recognize a particular high-*idf* query term, but system B never can. In the extreme case, the term may simply be out of vocabulary, although this may occur for various other reasons (e.g., poor language modeling or pronunciation dictionaries). Here again, a linear combination of scores will fail, as will rank-based interleaving. In the latter case, we will alternate between taking a plausible document from system A and an inevitably worse result from the crippled system B.

As a potential solution for these difficulties, we consider the use of generalized additive models for retrieval fusion.

3 Generalized Additive Models

Generalized Additive Models (GAMs) are a generalization of *Generalized Linear Models* (GLMs), while GLMs are a generalization of the well known linear model. In a GLM, the distribution of an observed random variable Y_i is related to the linear predictor η_i through a smooth monotonic *link function* g,

$$g(\mu_i) = \eta_i = \mathbf{X}_i\beta.$$

Here, $\mathbf{X}_i$ is the i^{th} row of the model matrix $\mathbf{X}$ (one set of observations corresponding to one observed y_i) and β is a vector of unknown parameters to be learned from the data. If we constrain our link function g to be the identity transformation, and assume Y_i is Normal, then our GLM reduces to a simple linear model.

But GLMs are considerably more versatile than linear models. First, rather than only the Normal distribution, the response Y_i is free to have any distribution belonging to the exponential family of distributions. This family includes many useful distributions such as the Binomial, Normal, Gamma, and Poisson. Secondly, by allowing non-identity link functions g, some degree of non-linearity may be incorporated in the model structure.

A well known GLM in the NLP community is *logistic regression* (which may alternatively be derived as a maximum entropy classifier). In logistic regression, the response is assumed to be Binomial and the chosen link function is the logit transformation,

$$g(\mu_i) = \texttt{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$$

.

Generalized additive models allow for additional model flexibility by allowing the linear predictor to now also contain learned smooth functions f_j of the covariates x_k. For example,

$$g(\mu_i) = \mathbf{X}_i^*\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}).$$

As in a GLM, $\mu_i \equiv \mathbb{E}(Y_i)$ and Y_i belongs to the exponential family. Strictly parametric model components are still permitted, which we represent as a row of the model matrix $\mathbf{X}_i^*$ (with associated parameters θ).

GAMs may be thought of as GLMs where one or more covariate has been transformed by a basis expansion, $f(x) = \sum_j^q b_j(x)\beta_j$. Given a set of q basis functions b_j spanning a q-dimensional space of smooth transformations, we are back to the linear problem of learning coefficients β_j which "optimally" fit the data. If we knew the appropriate transformation of our covariates (say the logarithm), we could simply apply it ourselves. GAMs allow us to learn these transformations from the data, when we expect some transformation to be useful but don't know it's form *a priori*. In practice, these smooth functions may be represented and the model parameters may be learned in various ways. In this work, we use the excellent open source package `mgcv` (Wood, 2006), which uses penalized likelihood maximization to prevent arbitrarily "wiggly" smooth functions (i.e., overfitting). Smooths (including multidimensional smooths) are represented by thin plate regression splines (Wood, 2003).

3.1 Combining speech retrieval results with GAMs

The chief difficulty introduced in combining ranked speech retrieval results is the severe disagreement introduced by differing document hypotheses. As we saw in Figure 1, it is often the case that the most discriminative query terms occur in only one transcript source.

3.1.1 GLM with factors

Our first new approach for handling differences in transcripts is an extension of the logistic regression model previously used in data fusion work, (Savoy et al., 1988). Specifically, we augment the model with the first-order interaction of scores $x_1 x_2$ and the *factor* α_i, so that

$$\texttt{logit}\{\mathbb{E}(R_i)\} = \beta_0 + \alpha_i + x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_3,$$

where the relevance $R_i \sim$ Binomial. A factor is essentially a learned intercept for different subsets of the response. In this case,

$$\alpha_i = \begin{cases} \beta_{BOTH} & \text{if both representations matched } q_i \\ \beta_{IBM} & \text{only } d_{i,IBM} \text{ matched } q_i \\ \beta_{BBN} & \text{only } d_{i,BBN} \text{ matched } q_i \end{cases}$$

where α_i corresponds to data row i, with associated document representations $d_{i,source}$ and query q_i. The intuition is simply that we'd like our model to have different biases for or against relevance based on which transcript source retrieved the document. This is a small-dimensional way of dampening the effects of significant disagreements in the document representations.

3.1.2 GAM with multidimensional smooth

If a document's score is large in both systems, we expect it to have high probability of relevance. However, as a document's score increases linearly in one source, we have no reason to expect its probability of relevance to also increase linearly. Moreover, because the most discriminative terms are likely to be found in only one transcript source, even an absent score for a document does not ensure a document is not relevant. It is clear then that the mapping from document scores to probability of relevance is in general a complex nonlinear surface. The limited degree of nonlinear structure afforded to GLMs by non-identity link functions is unlikely to sufficiently capture this intuition.

Instead, we can model this non-linearity using a generalized additive model with multidimensional smooth $f(x_{IBM}, x_{BBN})$, so that

$$\texttt{logit}\{\mathbb{E}(R_i)\} = \beta_0 + f(x_{IBM}, x_{BBN}).$$

Again, $R_i \sim$ Binomial and β_0 is a learned intercept (which, alternatively, may be absorbed by the smooth f).

Figure 2 shows the smoothing transformation f learned during our evaluation. Note the small decrease in predicted probability of relevance as the retrieval score from one system decreases, while the probability curves upward again as the disagreement increases. This captures our intuition that systems often disagree strongly because discriminative terms are often not recognized in all transcript sources.

We can think of the probability of relevance mapping learned by the factor model of Section 3.1.1 as also being a surface defined over the space of input document scores. That model, however, was constrained to be linear. It may be visualized as a collection of affine planes (with common normal vectors, but each shifted upwards by their factor level's weight and the common intercept).

4 Experiments

4.1 Dataset

Our dataset is a collection of 272 oral history interviews from the MALACH collection. The task is to retrieve short speech segments which were manually designated as being topically coherent by professional indexers. There are 8,104 such segments (corresponding to roughly 589 hours of conversational speech) and 96 assessed topics. We follow the topic partition used for the 2007 evaluation by the Cross Language Evaluation Forum's cross-language speech retrieval track (Pecina et al., 2007). This gives us 63 topics on which to train our combination systems and 33 topics for evaluation.

4.2 Evaluation

4.2.1 Geometric Mean Average Precision

Average precision (AP) is the average of the precision values obtained after each document relevant to a particular query is retrieved. To assess the effectiveness of a system across multiple queries, a commonly used measure is mean average precision (MAP). Mean average precision is defined as the arithmetic mean of per-topic average precision, $\mathrm{MAP} = \frac{1}{n}\sum_n \mathrm{AP}_n$. A consequence of the arithmetic mean is that, if a system improvement doubles AP for one topic from 0.02 to 0.04, while simultaneously decreasing AP on another from 0.4 to 0.38, the MAP will be unchanged. If we prefer to highlight performance differences on the lowest performing topics, a widely used alternative is the geometric mean of average precision (GMAP), first introduced in the TREC 2004 robust track (Voorhees, 2006).

$$\mathrm{GMAP} = \sqrt[n]{\prod_n \mathrm{AP}_n}$$

Robertson (2006) presents a justification and analysis of GMAP and notes that it may alternatively be computed as an arithmetic mean of logs,

$$\mathrm{GMAP} = \exp \frac{1}{n} \sum_n \log \mathrm{AP}_n.$$

4.2.2 Significance Testing for GMAP

A standard way of measuring the significance of system improvements in MAP is to compare average precision (AP) on each of the evaluation queries using the Wilcoxon signed-rank test. This test, while not requiring a particular distribution on the measurements, does assume that they belong to an interval scale. Similarly, the arithmetic mean of MAP assumes AP has interval scale. As Robertson (2006) has pointed out, it is in no sense clear that AP (prior to any transformation) satisfies this assumption. This becomes an argument for GMAP, since it may also be defined using an arithmetic mean of log-transformed average precisions. That is to say, the logarithm is simply one possible monotonic transformation which is arguably as good as any other, including the identify transform, in terms of whether the transformed value satisfies the interval assumption. This log transform (and hence GMAP) is useful simply because it highlights improvements on the most difficult queries.

We apply the same reasoning to test for statistical significance in GMAP improvements. That is, we test for significant improvements in GMAP by applying the Wilcoxon signed rank test to the paired, transformed average precisions, $\log \mathrm{AP}$. We handle tied pairs and compute exact p-values using the Streitberg & Röhmel Shift-Algorithm (1990). For topics with $\mathrm{AP} = 0$, we follow the Robust Track convention and add $\epsilon = 0.00001$. The authors are not aware of significance tests having been previously reported on GMAP.

4.3 Retrieval System

We use Okapi BM25 (Robertson et al., 1996) as our basic retrieval system, which defines a document D's retrieval score for query Q as

$$s(D,Q) = \sum_{i=1}^{n} idf(q_i) \frac{(\frac{(k_3+1)qf_i}{k_3+qf_i}) f(q_i,D)(k_1+1)}{f(q_i,D) + k_1(1-b+b\frac{|D|}{avgdl})},$$

where the inverse document frequency (idf) is defined as

$$idf(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

N is the size of the collection, $n(q_i)$ is the document frequency for term q_i, qf_i is the frequency of term q_i in query Q, $f(q_i, D)$ is the term frequency of query term q_i in document D, $|D|$ is the length of the matching document, and $avgdl$ is the average length of a document in the collection. We set the

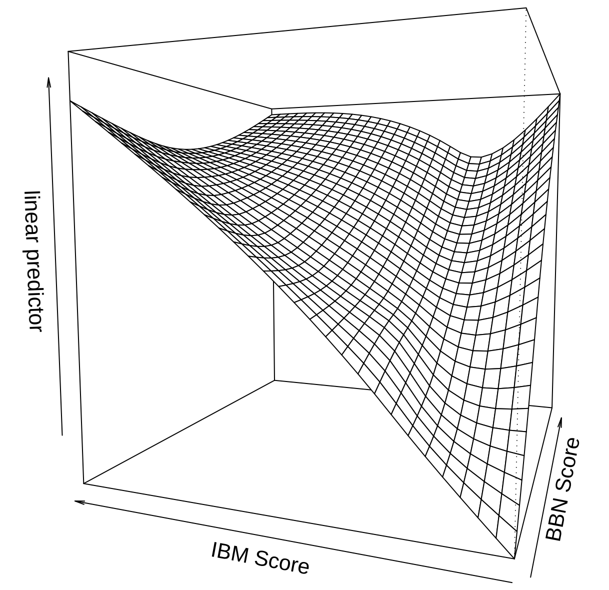

Figure 2: The two dimensional smooth $f(s_{\text{IBM}}, s_{\text{BBN}})$ learned to predict relevance given input scores from IBM and BBN transcripts.

parameters to $k_1 = 1$, $k_3 = 1$, $b = .5$, which gave good results on a single transcript.

4.4 Speech Recognition Transcripts

Our first set of speech recognition transcripts was produced by IBM for the MALACH project, and used for several years in the CLEF cross-language speech retrieval (CL-SR) track (Pecina et al., 2007). The IBM recognizer was built using a manually produced pronunciation dictionary and 200 hours of transcribed audio. The resulting interview transcripts have a reported mean word error rate (WER) of approximately 25% on held out data, which was obtained by priming the language model with metadata available from pre-interview questionnaires. This represents significant improvements over IBM transcripts used in earlier CL-SR evaluations, which had a best reported WER of 39.6% (Byrne et al., 2004). This system is reported to have run at approximately 10 times real time.

4.4.1 New Transcripts for MALACH

We were graciously permitted to use BBN Technology's speech recognition system to produce a second set of ASR transcripts for our experiments (Prasad et al., 2005; Matsoukas et al., 2005). We selected the one side of the audio having largest RMS amplitude for training and decoding. This channel was down-sampled to 8kHz and segmented using an available broadcast news segmenter. Because we did not have a pronunciation dictionary which covered the transcribed audio, we automatically generated pronunciations for roughly 14k words using a rule-based transliterator and the CMU lexicon. Using the same 200 hours of transcribed audio, we trained acoustic models as described in (Prasad et al., 2005). We use a mixture of the training transcripts and various newswire sources for our language model training. We did not attempt to prime the language model for particular interviewees or otherwise utilize any interview metadata. For decoding, we ran a fast (approximately 1 times real time) system, as described in (Matsoukas et al., 2005). Unfortunately, as we do not have the same development set used by IBM, a direct comparison of WER is not possible. Testing on a small held out set of 4.3 hours, we observed our system had a WER of 32.4%.

4.5 Combination Methods

For baseline comparisons, we ran our evaluation on each of the two transcript sources (IBM and our new transcripts), the linear combination chosen to optimize MAP (LC-MAP), the linear combination chosen to optimize GMAP (LC-GMAP), interleaving (IL), and CombMNZ. We denote our additive factor model as Factor GLM, and our multidimensional smooth GAM model as MD-GAM.

Linear combination parameters were chosen to optimize performance on the training set, sweeping the weight for each source at intervals of 0.01. For the generalized additive models, we maximized the penalized likelihood of the training examples under our model, as described in Section 3.

5 Results

Table 1 shows our complete set of results. This includes baseline scores from our new set of transcripts, each of our baseline combination approaches, and results from our proposed combination models. Although we are chiefly interested in improvements on difficult topics (i.e., GMAP), we present MAP for comparison. Results in bold indicate the largest mean value of the measure (either AP or $\log \text{AP}$), while daggers (†) indicate the

Type	Model	MAP	GMAP
T	IBM	0.0531 (-.2)	0.0134 (-11.8)
-	BBN	0.0532	0.0152
-	LC-MAP	0.0564 (+6.0)	0.0158 (+3.9)
-	LC-GMAP	0.0587 (+10.3)	0.0154 (+1.3)
-	IL	0.0592 (+11.3)	0.0165 (+8.6)
-	CombMNZ	0.0550 (+3.4)	0.0150 (-1.3)
-	Factor GLM	**0.0611 (+14.9)**†	0.0161 (+5.9)
-	MD-GAM	0.0561 (+5.5)†	**0.0180 (+18.4)**†
TD	IBM	0.0415 (-15.1)	0.0173 (-9.9)
-	BBN	0.0489	0.0192
-	LC-MAP	0.0519 (+6.1)†	0.0201 (+4.7)†
-	LC-GMAP	**0.0531 (+8.6)**†	0.0200 (+4.2)
-	IL	0.0507 (+3.7)	0.0210 (+9.4)
-	CombMNZ	0.0495 (+1.2)†	0.0196 (+2.1)
-	Factor GLM	0.0526 (+7.6)†	0.0198 (+3.1)
-	MD-GAM	0.0529 (+8.2)†	**0.0223 (+16.2)**†

Table 1: MAP and GMAP for each combination approach, using the evaluation query set from the CLEF-2007 CL-SR (MALACH) collection. Shown in parentheses is the relative improvement in score over the best single transcripts results (i.e., using our new set of transcripts). The best (mean) score for each condition is in bold.

combination is a statistically significant improvement ($\alpha = 0.05$) over our new transcript set (that is, over the best single transcript result). Tests for statistically significant improvements in GMAP are computed using our paired $\log$ AP test, as discussed in Section 4.2.2.

First, we note that the GAM model with multi-dimensional smooth gives the largest GMAP improvement for both title and title-description runs. Secondly, it is the only combination approach able to produce statistically significant relative improvements on both measures for both conditions. For GMAP, our measure of interest, these improvements are 18.4% and 16.2% respectively.

One surprising observation from Table 1 is that the mean improvement in $\log$ AP for interleaving is fairly large and yet not statistically significant (it is in fact a larger *mean* improvement than several other baseline combination approaches which *are* significant improvements. This may suggest that interleaving suffers from a large disparity between its best and worst performance on the query set.

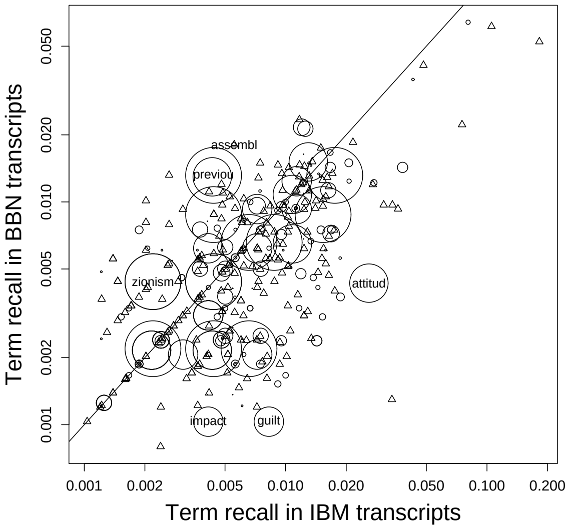

Figure 3: The proportion of relevant documents returned in IBM and BBN transcripts for discriminative title words (title words occurring in less than .01 of the collection). Point size is proportional to the improvement in average precision using (1) the best linear combination chosen to optimize GMAP (△) and (2) the combination using MD-GAM (◯).

Figure 3 examines whether our improvements come systematically from only one of the transcript sources. It shows the proportion of relevant documents in each transcript source containing the most discriminative title words (words occurring in less than .01 of the collection). Each point represents one term for one topic. The size of the point is proportional to the difference in AP observed on that topic by using MD-GAM and by using LC-GMAP. If the difference is positive (MD-GAM wins), we plot ◯, otherwise △. First, we observe that, when it wins, MD-GAM tends to increase AP much more than when LC-GMAP wins. While there are many wins also for LC-GMAP, the effects of the larger MD-GAM improvements will dominate for many of the most difficult queries. Secondly, there does not appear to be any evidence that one transcript source has much higher term-recall than the other.

5.1 Oracle linear combination

A chief advantage of our MD-GAM combination model is that it is able to map input scores non-linearly onto a probability of document relevance.

Type	Model	GMAP
T	Oracle-LC-GMAP	0.0168
-	MD-GAM	**0.0180 (+7.1)**
TD	Oracle-LC-GMAP	0.0222
-	MD-GAM	**0.0223 (+0.5)**

Table 2: GMAP results for an oracle experiment in which MD-GAM was fairly trained and LC-GMAP was unfairly optimized on the test queries.

To make an assessment of how much this capability helps the system, we performed an oracle experiment where we again constrained MD-GAM to be fairly trained but allowed LC-GMAP to cheat and choose the combination *optimizing* GMAP *on the test data*. Table 2 lists the results. While the improvement with MD-GAM is now not statistically significant (primarily because of our small query set), we found it still out-performed the oracle linear combination. For title-only queries, this improvement was surprisingly large at 7.1% relative.

6 Conclusion

While speech retrieval is one example of retrieval under errorful document representations, other similar tasks may also benefit from these combination models. This includes the task of cross-language retrieval, as well as the retrieval of documents obtained by optical character recognition.

Within speech retrieval, further work also remains to be done. For example, various other features are likely to be useful in predicting optimal system combination. These might include, for example, confidence scores, acoustic confusability, or other strong cues that one recognition system is unlikely to have properly recognized a query term. We look forward to investigating these possibilities in future work.

The question of how much a system should expose its internal workings (e.g., its document representations) to external systems is a long standing problem in meta-search. We've taken the rather narrow view that systems might only expose the list of scores they assigned to retrieved documents, a plausible scenario considering the many systems now emerging which are effectively doing this already. Some examples include *EveryZing*,[1] the MIT *Lecture Browser*,[2] and Comcast's video search.[3] This trend is likely to continue as the underlying representations of the content are themselves becoming increasingly complex (e.g., word and subword level lattices or confusion networks). The cost of exposing such a vast quantity of such complex data rapidly becomes difficult to justify.

But if the various representations of the content are available, there are almost certainly other combination approaches worth investigating. Some possible approaches include simple linear combinations of the putative term frequencies, combinations of one best transcript hypotheses (e.g., using ROVER (Fiscus, 1997)), or methods exploiting word-lattice information (Evermann and Woodland, 2000).

Our planet's 6.6 billion people speak many more words every day than even the largest Web search engines presently index. While much of this is surely not worth hearing again (or even once!), some of it is surely precious beyond measure. Separating the wheat from the chaff in this cacophony is the *raison d'etre* for information retrieval, and it is hard to conceive of an information retrieval challenge with greater scope or greater potential to impact our society than improving our access to the spoken word.

Acknowledgements

The authors are grateful to BBN Technologies, who generously provided access to their speech recognition system for this research.

References

Brian T. Bartell, Garrison W. Cottrell, and Richard K. Belew. 1994. Automatic combination of multiple ranked retrieval systems. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 173–181.

Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, Ophir Frieder, and Nazli Goharian. 2004. Fusion of effective retrieval strategies in the same information retrieval system. *J. Am. Soc. Inf. Sci. Technol.*, 55(10):859–868.

W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D.W. Oard, M. Picheny, J. Psutka, B. Ramabhadran,

[1] http://www.everyzing.com/

[2] http://web.sls.csail.mit.edu/lectures/

[3] http://videosearch.comcast.net

D. Soergel, T. Ward, and Wei-Jing Zhu. 2004. 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.

J. P. Callan, Z. Lu, and W. Bruce Croft. 1995. Searching Distributed Collections with Inference Networks . In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–28, Seattle, Washington. ACM Press.

G. Evermann and P.C. Woodland. 2000. Posterior probability decoding, confidence estimation and system combination. In *Proceedings of the Speech Transcription Workshop*, May.

Jonathan G. Fiscus. 1997. A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER). In *Proceedings of the IEEE ASRU Workshop*, pages 347–352.

Jong-Hak Lee. 1997. Analyses of multiple evidence combination. In *SIGIR Forum*, pages 267–276.

David Lillis, Fergus Toolan, Rem Collier, and John Dunnion. 2006. Probfuse: a probabilistic approach to data fusion. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 139–146, New York, NY, USA. ACM.

R. Manmatha, T. Rath, and F. Feng. 2001. Modeling score distributions for combining the outputs of search engines. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–275, New York, NY, USA. ACM.

Spyros Matsoukas, Rohit Prasad, Srinivas Laxminarayan, Bing Xiang, Long Nguyen, and Richard Schwartz. 2005. The 2004 BBN 1xRT Recognition Systems for English Broadcast News and Conversational Telephone Speech. In *Interspeech 2005*, pages 1641–1644.

Mark Montague and Javed A. Aslam. 2002. Condorcet fusion for improved retrieval. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 538–548, New York, NY, USA. ACM.

Pavel Pecina, Petra Hoffmannova, Gareth J.F. Jones, Jianqiang Wang, and Douglas W. Oard. 2007. Overview of the CLEF-2007 Cross-Language Speech Retrieval Track. In *Proceedings of the CLEF 2007 Workshop on Cross-Language Information Retrieval and Evaluation*, September.

Allison L. Powell, James C. French, James P. Callan, Margaret E. Connell, and Charles L. Viles. 2000. The impact of database selection on distributed searching. In *Research and Development in Information Retrieval*, pages 232–239.

R. Prasad, S. Matsoukas, C.L. Kao, J. Ma, D.X. Xu, T. Colthurst, O. Kimball, R. Schwartz, J.L. Gauvain, L. Lamel, H. Schwenk, G. Adda, and F. Lefevre. 2005. The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System. In *Interspeech 2005*.

S. Robertson, S. Walker, S. Jones, and M. Hancock-Beaulieu M. Gatford. 1996. Okapi at TREC-3. In *Text REtrieval Conference*, pages 21–30.

Stephen Robertson. 2006. On GMAP: and other transformations. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 78–83, New York, NY, USA. ACM.

J. Savoy, A. Le Calvé, and D. Vrajitoru. 1988. Report on the TREC-5 experiment: Data fusion and collection fusion.

Joseph A. Shaw and Edward A. Fox. 1994. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference (TREC-2)*.

Bernd Streitberg and Joachim Röhmel. 1990. On tests that are uniformly more powerful than the Wilcoxon-Mann-Whitney test. *Biometrics*, 46(2):481–484.

Christopher C. Vogt and Garrison W. Cottrell. 1999. Fusion via a linear combination of scores. *Information Retrieval*, 1(3):151–173.

Ellen M. Voorhees, Narendra Kumar Gupta, and Ben Johnson-Laird. 1994. The collection fusion problem. In D. K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 500–225. National Institute of Standards and Technology.

Ellen M. Voorhees. 2006. Overview of the TREC 2005 robust retrieval track. In Ellem M. Voorhees and L.P. Buckland, editors, *The Fourteenth Text REtrieval Conference, (TREC 2005)*, Gaithersburg, MD: NIST.

Simon N. Wood. 2003. Thin plate regression splines. *Journal Of The Royal Statistical Society Series B*, 65(1):95–114.

Simon Wood. 2006. *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC.