

Combining LVCSR and Vocabulary-Independent Ranked Utterance Retrieval for Robust Speech Search

J. Scott Olsson
HLT Center of Excellence
Johns Hopkins University
Baltimore, MD, USA
solsson@jhu.edu

Douglas W. Oard
University of Maryland
College Park, MD, USA
oard@umd.edu

ABSTRACT

Well tuned Large-Vocabulary Continuous Speech Recognition (LVCSR) has been shown to generally be more effective than vocabulary-independent techniques for ranked retrieval of spoken content when one or the other approach is used alone. Tuning LVCSR systems to a topic domain can be costly, however, and the experiments in this paper show that Out-Of-Vocabulary (OOV) query terms can significantly reduce retrieval effectiveness when that tuning is not performed. Further experiments demonstrate, however, that retrieval effectiveness for queries with OOV terms can be substantially improved by combining evidence from LVCSR with additional evidence from vocabulary-independent Ranked Utterance Retrieval (RUR). The combination is performed by using relevance judgments from held-out topics to learn generic (i.e., topic-independent), smooth, non-decreasing transformations from LVCSR and RUR system scores to probabilities of topical relevance. Evaluated using a CLEF collection that includes topics, spontaneous conversational speech audio, and relevance judgments, the system recovers 57% of the mean uninterpolated average precision that could have been obtained through LVCSR domain tuning for very short queries (or 41% for longer queries).

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Experimentation

Keywords: speech retrieval

1. INTRODUCTION

Speech retrieval (SR) has perhaps greater potential to revolutionize the way we store and access information than any other branch of information retrieval: The vast majority of the information we produce every day is spoken, and yet speech remains almost completely unsearchable. One key difficulty which remains is the problem of Out-Of-Vocabulary

(OOV) words. A word is said to be OOV if it is not contained within the recognition dictionary of a Large Vocabulary Continuous Speech Recognition (LVCSR) system. From the perspective of SR, we might rephrase this as, a word is OOV if it could not be *anticipated* as a potentially useful query term when the underlying recognition dictionary was constructed. Because OOV words tend to be rare, they tend to be informative, and thus of particular interest for information retrieval. And yet while OOV words occur infrequently in speech, they are comparatively common in query formulations. In our topic set, 12% of all words in short queries (titles) are OOV, and an OOV rate of 12% was also previously reported for query words in a live search engine, indexing speech audio from the Web [8]. And OOV words tend to significantly reduce retrieval effectiveness.

In order to maximize transcription accuracy without unduly increasing search complexity, LVCSR systems have over the years included increasingly larger decoding dictionaries. The words in these dictionaries must be chosen in view of a target domain to keep the OOV rate low, but of course not every potential word may be anticipated. In particular, when a new topic domain is encountered, the decoding dictionary may be quite poorly matched to the target, making it very difficult for users to find speech that is relevant to their information need.

We use this scenario of domain switching (i.e., an LVCSR system is developed for one *topic* domain but then used on another), to create a plausible distribution of OOV terms for our experiments. We investigate SR systems built with LVCSR, both when the decoding dictionary has not been adapted for the topic and when it has. We refer to a system built using a domain-adapted dictionary as being *Domain-Adapted (DA)*. When the dictionary has not been extended for the new topic domain, we refer to the system as being *Out-Of-Domain (OOD)*. While we expect an SR system built on DA LVCSR to perform best, and thus consider its performance an upper bound on retrieval utility, it will for the foreseeable future remain impossible to build one LVCSR system having good lexical coverage of all possible topic domains. We emphasize that we are considering a shift in topic domain, and that other shifts in a collection's characteristics (e.g., dialect, age, channel, or signal conditions) may also present serious difficulties that are beyond our scope.

Research in SR can broadly be divided into two camps: work focusing on indices of words produced through LVCSR and those using vocabulary-independent methods such as phoneme or subword-level indexing. Closed-vocabulary word-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

Copyright 2009 ACM 978-1-60558-483-6/09/07 ...\$10.00.

based methods have been the focus of most previous SR research [6, 28, 24, 3, 15]. This is reasonable, since for words *within* an LVCSR system’s dictionary, it is generally accepted that LVCSR systems are better than vocabulary-independent systems at detecting spoken terms [5]. Yet because OOV terms tend to be among the most informative terms in a topic’s query specification, vocabulary-independent subword representations have also been considered for *ad hoc* SR [19, 30, 39]. Subword representations are attractive primarily because they avoid the OOV problem.

Our goal, therefore, is to avoid the high costs of domain-adapting a SR system, while also maintaining retrieval effectiveness for queries containing OOV words. Naturally, we’d like to combine the strengths of OOD LVCSR and vocabulary-independent term detection for SR. We present a simple model for this evidence combination, in which we learn monotonically increasing transformations of each system’s retrieval scores which may then be easily combined for ranking pre-segmented passages. This work differs from previous SR combination efforts [9, 7, 35] in several important ways. First, we present the first results of this kind on the largest publicly available IR test collection of spontaneous, conversational, speech, the CLEF 2006/7 CL-SR collection [24]. Second, we utilize a state-of-the-art ranked utterance retrieval method recently reported in [23]. Third, we consider a new evidence combination approach which learns a transformation of our retrieval scores to predict relevance, rather than simply thresholding confidence values for augmenting an index or combining scores via arbitrary normalizations. Most importantly, we find that our combination of evidence produces a new ranking which is significantly better than either ranking alone. This combination is able to recover most of the Mean uninterpolated Average Precision (MAP) lost because of words we were not able to anticipate in the OOD recognition lexicon.

This paper is organized as follows. First, in Section 2 we introduce the collection and task we use for our experiments. In Sections 3 and 4, we introduce our LVCSR-based SR and RUR systems, respectively. Then, we discuss in Section 5.1 how these system scores may be combined. We present our experiments and results in Section 6 and conclude with remarks in Section 7.

2. SPEECH COLLECTION AND TASK

Although there has been a good deal of work on SR to date, results from reported experiments have often been difficult to interpret because of small, synthetic or proprietary test collections or because they do not incorporate human assessments of relevance in their evaluation (e.g., a passage is deemed “relevant” if it contains a specific word or, more often, if it was staged using a prompt corresponding to the topic). One notable exception was the TREC Spoken Document Retrieval (SDR) track, in which topical relevance judgments made by human assessors were used to compute standard effectiveness measures for ranked retrieval of news stories. Although a strict temporal division between LVCSR training and ranked retrieval testing did result in some OOV terms (e.g., names of people or places mentioned in breaking news), the fact that training and test materials were drawn from similar sources in a similar time frame tended to mitigate vocabulary mismatch effects.

The lack of a substantial and realistic test collection containing substantial amounts of spontaneous conversational speech has tended to focus SR research on detecting term occurrences rather than retrieving *informative* speech segments. Previous work in combining LVCSR and vocabulary-independent SR systems has focused primarily on this term detection task [12, 14]. Very little previous work has attempted to combine LVCSR and vocabulary-independent techniques for *ad hoc* SR [9]. A significant contribution of this work is that we report the first of these results using a comparatively large, publicly available test set, the CLEF 2006/7 CL-SR collection. The LDC plans to release the audio, the LVCSR training transcripts, the information retrieval topics and the relevance judgments used for these experiments in the near future. The topics and relevance judgments are, however, already available to CLEF participants through ELDA. We therefore obtained that data from ELDA, and we obtained the audio and the ASR training data from IBM Research (where the LVCSR training transcription had originally been performed).

The SR collection contains 272 interviews with survivors of the Holocaust, used previously by the Cross Language Evaluation Forum’s cross-language speech retrieval (CLEF CL-SR) track [24, 20, 34]. We present a brief overview of the collection here, while the reader is referred to [24] for further information. Note that the interviews used for training and testing the speech recognition systems are disjoint from the SR interview collection.

The test collection’s speech audio was automatically segmented into short utterances for the purpose of running speech recognition. Longer, topically coherent *segments* of the speech were also defined by professional indexers [24]. For comparison, an average utterance is 6.75 seconds (with a standard deviation of 4.16), while segments average 3.45 minutes (with a standard deviation of 137.9 seconds). There are 8,104 such segments (corresponding to roughly 589 hours of conversational speech) and 96 assessed topics. Following standard TREC conventions, the CLEF CL-SR queries are fully specified as a title, description, and narrative.

We evaluate on multiple topic sets. To allow comparison with previously published results, we run on 33 evaluation topics used in CLEF’s 2006 and 2007 CL-SR track [20, 24]. In those 33 topics however, there are only 10 and 12 topics having OOV terms in their title or title plus description fields respectively. When reporting on this topic set, we average across the complete CLEF topic set—including the topics without OOV terms. This indicates roughly how much MAP may be lost due to OOV query words in a random selection of topics. For the remainder of this paper, we refer to this topic set as the **CLEF Topics**.

We also run on the 38 topics from the complete topic set having at least one OOV word in their title and the 49 topics containing at least one OOV word in their title or description. This topic set is denoted as **OOV Topics** for the remainder of this paper.

2.1 Evaluation Measures

We evaluate our system using MAP. Given a ranked list of segments, the *precision* at position i in the list is defined as the proportion of the top i segments relevant to the corresponding query. Average Precision (AP) is the average of the precision values computed for each position contain-

ing a relevant segment. To assess the effectiveness of a system across multiple queries, Mean Average Precision is defined as the arithmetic mean of per-query average precision, $MAP = \frac{1}{n} \sum_n AP_n$.

Secondly, we report the *Fraction of Recovered Mean average precision (FRM)*, which we define as

$$FRM = \frac{MAP - MAP_{OOD}}{MAP_{DA} - MAP_{OOD}},$$

where MAP_{DA} and MAP_{OOD} are the MAPs associated with the DA and OOD word-based systems, respectively. The FRM indicates the proportion of MAP (lost because the dictionary was not adapted) which is recovered by combining the OOD word system with the vocabulary-independent system’s output. Note that, by definition, the DA SR system achieves an FRM of 100%, while the OOD system has an FRM of 0%.

Throughout this paper, when we report statistically significant improvements in MAP, we are comparing AP for paired topics using a Wilcoxon signed rank test at $\alpha = 0.05$.

3. LVCSR-BASED SR SYSTEMS

We now present our SR approach using only recognition lattices from a fixed-vocabulary LVCSR system. We construct systems using both OOD and DA LVCSR, which give lower and upper bounds respectively on the MAP attainable for each topic set. It is the scores from this OOD SR system which we combine with our vocabulary-independent results. We are thankful to BBN Technologies, who generously permitted the use of their speech recognition system *Byblos* [26, 17] for this work.

Our OOD dictionary contains about 50,000 words with manually specified pronunciations, and was previously utilized for conversational telephone and broadcast news speech transcription. To produce our DA dictionary, we added words to the OOD dictionary to cover our complete set of training transcripts, giving a dictionary of 60,378 words. For training, we use approximately 200 hours of audio transcribed in 197,220 utterances, excerpted from about 800 speakers. We use this complete set for our DA experiments. For our OOD experiments, we subset the complete set of transcriptions to exclude any utterances not covered by our OOD dictionary. This reduces the training set by 12.8% from 197,220 to 172,027 utterances. In this way, we hope to model the speaker and channel characteristics, without unfairly aiding the OOD acoustic or language models. We train separate acoustic and language models for the DA and OOD SR systems. On held-out test data, our DA system obtained a word error rate (WER) of 32.40. The OOD system’s WER on the same data was 31.63.

The output of our LVCSR system is a *lattice* of recognition hypotheses for each test speech utterance. A lattice is a directed acyclic graph that is used to compactly represent the search space for a speech recognition system. Each node represents a point in time and arcs between nodes indicates a word occurs between the connected nodes’ times. Arcs are weighted by the probability of the word occurring, so that the so-called “one-best” path through the lattice (what a system might return as a transcription) is the path through the lattice having highest probability under the acoustic and language models. From these lattices, we compute the expected

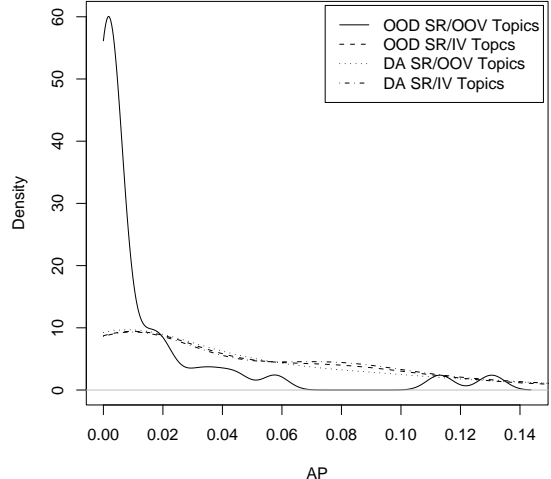


Figure 1: Density of AP for both the OOD and DA SR systems on 38 title queries with one or more OOV words and 58 title queries having only IV words.

count of each word in the corresponding utterance using a variant of the forward-backward algorithm, implemented in the SRILM toolkit [31].

To rank documents using only the expected word counts from LVCSR, we use a vector-space model with Okapi BM25 weighting [27]. The approach defines a segment d ’s retrieval score (or *retrieval status value*, **RSV**) for query q as

$$s_{d,q} = \sum_{i=1}^n idf(q_i) \frac{\binom{k_3+1}{k_3+qf_i} f(q_i, d) (k_1 + 1)}{f(q_i, d) + k_1 (1 - b + b \frac{|d|}{avgdl})},$$

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. As in previous work [22], we set the parameters to $k_1 = 1$, $k_3 = 1$, $b = 0.5$. We take as a word’s term frequency, $f(q_i, d)$, the sum of the word’s expected counts from all lattices within the segment. Because utterances can cross segments boundaries, we place word counts from an utterance in the segment containing the largest fraction of the utterance. For the purpose of computing document frequency, we define a word to be present within a segment if $f(q_i, d) \geq 0.5$.

Now that we have both OOD and DA SR systems (using BM25), we can investigate how each is affected by the presence of OOV query words. To illustrate how the systems are affected differently, we ran both the OOD and DA SR systems on the complete set of 96 CLEF CL-SR topics. This complete set includes 58 completely In-Vocabulary (IV) title queries and 38 title queries having one or more OOV words. Figure 1 shows the estimated density of AP for each system on each of the OOV and IV topics sets. First, we see that the density of AP is similar for the DA system on

both IV and OOV queries. Second, we see that the density of AP is similar for DA and OOD systems on IV queries. This is not surprising because the underlying LVCSR systems are very similar. Finally, we see that the density of AP for the OOD system on OOV queries is sharply peaked near $AP = 0$ and noticeably differs from the AP densities on the other conditions. As we would expect, this confirms that the MAP loss between the DA and OOD SR systems is primarily due to queries with OOV words. If we consider only short (title) queries having one or more OOV word, we find that a dramatic 71.1% of the DA SR system’s MAP is lost when using OOD SR. This obviously motivated special handling of queries with OOV words. To improve on these queries, we incorporate additional evidence from a vocabulary-independent RUR system.

4. RUR SYSTEMS

To rank utterances by our confidence that they contain a term, we use our RUR system reported recently in [23]. The system processes lattices from LVCSR. For our subword recognition units, we use short sequences of 1-5 phonemes called phoneme multigrams. Multigrams are learned by choosing a ML segmentation of the training phoneme transcripts (with all utterances that contain OOV words removed). The most likely segmentation defines our multigram decoding dictionary and the segmented corpus is used to retrain acoustic and language models for the multigram LVCSR system. At search time, the RUR system uses a factored phrase-based machine translation system [10] to hypothesize the 50 most probable *degradations* of an OOV word’s reference phoneme sequence and incorporates these alternate pronunciations in its term frequency estimator (ranking function). While the system had been designed to retrieve utterances, our goal now is to retrieve segments. Therefore, we sum the utterance-level term frequency estimates from each utterance in the segment to produce segment-level term frequency estimates. As in the word-based SR systems, we consider an utterance to be part of a segment if the majority of the utterance is within the segment. We refer to this system as the Ranked Utterance Retrieval (**RUR**) system.

5. COMBINATION METHODS

There are two types of approaches for combining ranked retrieval results, data-fusion and data-merging. In *data-merging*, indices are combined first, and afterwards a single RSV is computed using the combined index. The difficulty with this approach is how to transform the term frequency estimates from each system such that they are commensurate and, thus, combinable. As an example, in [9], scores above a threshold from a phonetic lattice scanner were simply added to the index as being present words. This allowed then-state-of-the-art IR methods to be used, although the combined performance was not better than either system alone.

In *data-fusion*, separate RSVs are computed from each index before the RSVs are combined. This allows us to use strong retrieval systems as inputs, but also forces us to make simplifying assumptions for their combination (e.g., that a linear combination of RSVs is sensible after normalization). In this work, we focus on data-fusion approaches. To address the RSV transformation problem, we consider a new

method which learns an appropriate normalization of the scores. First, we present several data-fusion techniques that have previously been considered.

5.1 Baselines

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 [1, 4, 25, 32] for text IR, with varying degrees of success, owing in part to the potential difficulty of normalizing scores across retrieval systems. In [9], this approach was used to combine results from a now small (20k word) LVCSR system with scores from a phone lattice scanner. Scores were normalized by the largest score for the input type. However, the combinations did not improve upon the best of the non-combined results.

More advanced score normalization methods have also been proposed for data-fusion, as in [29]. Perhaps the most successful of these is known as CombMNZ. CombMNZ has been shown to achieve strong performance and has been used in many subsequent studies [11, 18, 2, 13]. In this study, we use CombMNZ as a baseline for comparison, and following [13] and [11], compute it in the following way. First, we normalize each score $s_{d,r}$ for segment d in ranked list r as

$$N_{d,r} = \frac{s_{d,r} - \min(s_r)}{\max(s_r) - \min(s_r)},$$

where $\max(s_r)$ and $\min(s_r)$ are the maximum and minimum scores seen in the ranked list r . After normalization, the CombMNZ score for a document d is computed as

$$\text{CombMNZ}_d = \sum_{r \in \mathcal{R}} N_{d,r} \times |N_d > 0|.$$

Here, \mathcal{R} is the set of ranked lists to be combined, $N_{r,d}$ is the normalized score of segment d in ranked list r , and $|N_d > 0|$ is the number of non-zero normalized scores given to d in any ranked list.

Manmatha et al. [16] 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 this comparatively complex normalization approach achieved performance no better than the much simpler CombMNZ.

A simple rank-based fusion technique is *interleaving* [33]. 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. We use this as a second baseline for comparison.

Chiefly for the purpose of analysis, we also consider a trivial *backoff* approach as a final baseline. That is, we rank all queries using the OOD SR system unless they have at least one OOV term, in which case we backoff and rank them only by their RUR score.

5.2 Combining by Transformations of RSV

We now present our combination approach. Recall, we aim to combine an OOD SR RSV from Section 3 and an RUR RSV from Section 4 to predict a new segment’s probability of relevance. Suppose we had estimates for both the conditional probability of a segment’s relevance given its LVCSR-based score, $P(\text{rel}|W)$, and its probability of relevance given

a vocabulary-independent system’s score for an OOV title term T , $P(\text{rel}|T)$. Assuming independence between W and T , we could then compute the probability of a speech segment’s relevance given both W and T as

$$P(\text{rel}|W, T) \propto P(\text{rel}|W)P(\text{rel}|T) \quad (1)$$

where the relation is proportionality since we are only interested in ranking the segments.

Unfortunately, the RSVs obtained from our word-based SR system are not in fact probabilities of relevance. At most, we can say that, in general, a larger RSV ought to mean that a segment is more likely to be relevant. As a solution to this problem, we propose learning a smooth and monotonically increasing transformation f of the RSVs to map us from W to $P(\text{rel}|W)$. Specifically, our model is

$$\mathbb{E}(\text{rel}) = \beta_0 + f(W), \quad (2)$$

where f is constrained to be a smooth, monotonically increasing function and rel is binomial. Equation 2 is an example of a generalized additive model. Note, separate models are learned for OOD SR RSVs and RUR RSVs. We represent the smooth f using a cubic smoothing spline, and monotonicity is ensured by modifying the standard quadratic programming problem for cubic smoothing splines with a set of linear constraints, as described in [37].

In general, queries may have multiple title and description OOV terms. Accordingly, we extend Equation 1 to

$$P(\text{rel}|W, T, D) \propto \quad (3)$$

$$P(\text{rel}|W)^\lambda \left[\prod_{i=1}^t P(\text{rel}|T_i) \right]^\gamma \left[\prod_{j=1}^d P(\text{rel}|D_j) \right],$$

where λ, γ parametrize the contribution from each evidence source¹ and t, d denotes the number of OOV terms from each field type (possibly zero). We refer to this approach of Combining by Monotonic Normalizing Transformations as CMNT.

Figure 2 shows transformations $f(W)$ and $f(T)$ learned using Equation 2 on one fold of a leave-one-out cross-fold validation. On the bottom, the estimated density of normalized RSVs² for both relevant and non-relevant segments are shown, for RSVs both from OOD SR and RUR (the density estimates are heavily smoothed for visualization purposes). As we expect, the relevant and non-relevant segments are strongly mixed while relevant segments tend to have modestly larger RSVs. We note, however, that the OOD and RUR RSVs have very different RSV distributions. On top, the probability of relevance given the RSVs (e.g., $P(\text{rel}|W)$) is shown for both the OOD SR and RUR system. Probability of relevance is constrained to increase monotonically

¹It may be objected that OOV title and description terms may be weighted differently, while the OOD SR system does not know which terms are in the title and which are in the description. We evaluated this concern by running additional trials constraining $\gamma = 1$ and found no significant effect. A possible explanation is that, since OOV words tend to be good predictors of relevance, knowing which topic field an OOV word occurs in provides little additional information. Details are in [21].

²For plotting purposes, we normalize the RSVs by the largest RSV obtained by any segment.

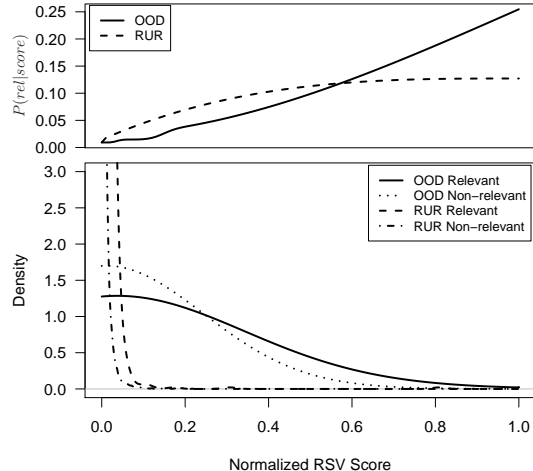


Figure 2: Bottom: the distribution of normalized RSVs for relevant and non-relevant segments from one cross-validation fold for both the OOD SR and RUR systems. Top: the smooth, monotonic transformations f learned via Equation 2.

with RSV. Note that both transformations have very different shapes. For example, the transformation learned on RUR RSVs flattens out for large RSVs (where the ratio of relevant to non-relevant RSV densities is small). For the largest normalized RSVs, we see that the probability of relevance given the OOD system’s largest RSV is about twice as large as the probability of relevance given the RUR system’s largest RSV. This is to be expected. First, because some topics still contain discriminative OOD words (in addition to their OOV words), the largest OOD RSVs are likely to be good discriminators for relevant segments. On the other hand, RUR is a harder task so that we would expect CMNT to have less confidence about the predictive strength of RUR RSVs.

To apply Equation 3 we must choose values of λ, γ . Our approach is simply to choose the parameters that give the best MAP in a leave-one-out cross-validation on the training queries. We sweep over λ, γ , on the intervals $0 \leq \lambda \leq 100$ and $0 \leq \gamma \leq 100$.

6. RESULTS

For our combined system, we consider as baselines CombMNZ, interleaving and backoff, as discussed in Section 5.1. We also consider our new approach, CMNT, defined in Section 5.2. The smooth transformations were learned using the `mgcv` package available for R [36], which fits the model using penalized likelihood maximization [38]. We use leave-one-out cross-validation (leaving out queries).

6.1 Title-Only Runs

Table 1 shows the title-only results from our experiments. We report on both the CLEF 2006/2007 set (having only 10 or 33 OOV queries) and the complete set of 38 topics having one or more OOV title word (OOV Topics). For comparison, on the CLEF Topics set, the best title-only submission at CLEF CL-SR 2006 achieved a MAP of 0.0495 using the pro-

List(s)		Combination	OOV Topics	CLEF Topics
OOD	RUR	DA	Method	MAP FRM MAP FRM
✓			no comb.	0.0158 0.0 0.0439 0.0
✓			no comb.	0.0278 30.7 — —
✓✓			CombMNZ	0.0151 -1.8 0.0454 27.3
✓✓			interleaving	0.0250 23.7 0.0464 45.5
✓✓			backoff	— — 0.0480 75.1
✓✓			CMNT	0.0382 57.5 0.0490 93.6
✓			no comb.	0.0547 100.0 0.0494 100.0

Table 1: Title run results from 38 topics having at least one OOV word and the results on the CLEF 2006/2007 test collection.

vided, DA ASR word transcripts [20].³ Our DA system, on the same topic set, achieves roughly the same MAP (0.0494).

First, we observe that neither CombMNZ nor interleaving is able to improve upon the best of the systems used alone (recall, the systems alone are the OOD SR and RUR systems). We suspect this is most likely because the RSVs from each system have very different distributions, so that more principled score normalization is necessary. This motivates our combination approach using monotonic normalizing transformations of the RSVs.

If we combine evidence using the backoff approach, we see from Table 1 that our RUR system achieves a statistically significantly higher MAP than the OOD LVCSR system alone. We also see an improvement using the same combination approach for the CLEF Topics set. We expect this simple approach works here because title queries tend to be short, so that an OOV query word often means the OOD SR RSV will not provide much information for ranking the segments.

Using CMNT to combine our OOD LVCSR and multigram RUR systems, we achieve an FRM of 57.5 on title queries with OOV terms. This improvement is statistically significant with respect to the OOD SR system alone (a state-of-the-art baseline which does not address the OOV problem). We also find that, on the OOV Topics set, CMNT significantly improves upon using RUR alone or combining evidence by simple normalizations (e.g., CombMNZ). These baselines are a sample of previous state-of-the-art methods for systems that do specifically address the OOV query word problem.

Figure 3 shows the improvements obtained for each topic. On top, we see the difference in AP between the DA and OOD SR systems for each topic, sorted by the difference in AP. In the middle, with the topics in the same sort order, the difference in AP between the CMNT system (using multigram RUR) and the OOD SR system is shown. We see that the largest improvements for the CMNT system are predominately in topics with larger differences between DA and OOD MAP. This is as we would expect.

³The same set of topics was also used in the 2007 CLEF CL-SR, although no comparable scores (i.e., using only ASR transcripts and title queries) were reported [24].

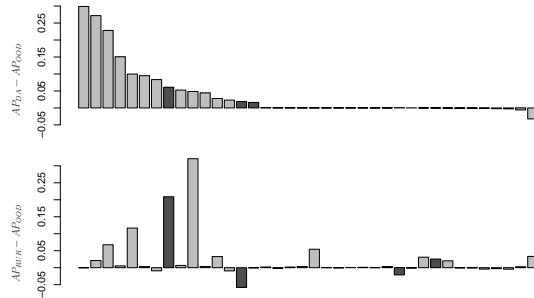


Figure 3: Per-query analysis for OOV T queries. The 10 Topics within the 33-topic CLEF Topics set are shown in darker gray. Top: the difference in AP between the DA and OOD SR systems, where topics are sorted by size of difference. Bottom: Using the same sort order, the difference in AP between the CMNT system using RUR and the OOD SR system.

mean(λ)	sd(λ)	MAP _{best}	$\frac{MAP}{MAP_{best}}$
12.8	29.6	0.0394	0.97

Table 2: Mean and standard deviation of CMNT parameter λ found in the oracle study for OOV title queries. MAP_{best} is the MAP obtained using the optimal settings of λ for each topic. The proportion of MAP_{best} obtained in the non-oracle evaluation, $\frac{MAP}{MAP_{best}}$, is also shown.

In Figure 3, we also see that, for a very few topics, the OOD system obtains a higher MAP than the DA system. In the most extreme case, OOD SR improved over DA SR by 0.0323 MAP, for the title query *The liberation of Buchenwald and Dachau*. One possible explanation for this may be that the terms *Buchenwald* and *Dachau* are rare—and therefore highly weighted by BM25, but they are not good discriminators for segments dealing specifically with the camps’ liberation.

6.1.1 Combination Parameter λ

To select our combination parameter λ for CMNT, we have used held out data in a leave-one-out cross-fold validation. We also want to know, however, how sensitive the optimal choice of λ is to different test topics. To evaluate this, we run an additional oracle experiment where we now select λ to give the best possible AP for each topic. Table 2 shows the mean and standard deviation of λ chosen for each topic. Also shown is the MAP attained by choosing the best possible values for λ for each topic, MAP_{best}, and the proportion of MAP_{best} obtained when λ was chosen fairly in the experiments reported above, $\frac{MAP}{MAP_{best}}$. First, we note that the standard deviation is large. The optimal setting of λ for most topics is zero, because OOD SR RSVs are often of little use when the title query contains an OOV word. However, a few queries contain discriminative in-vocabulary words that cause the system to benefit from the contribution from the OOD SR system (thus increasing variance in λ). Secondly, we see that when we chose λ in the fair evaluation reported

List(s)		Combination	OOV Topics	CLEF Topics	Topics		
OOD	RUR	DA	Method	MAP	FRM	MAP	FRM
✓			no comb.	0.0466	0.0	0.0374	0.0
✓			no comb.	0.0221	-70.0	—	—
✓✓			CombMNZ	0.0449	-4.9	0.0392	14.2
✓✓			interleaving	0.0365	-28.9	0.0362	-9.4
✓✓			backoff	—	—	0.0309	-51.2
✓✓			CMNT	0.0611	41.3	0.0447	57.8
✓			no comb.	0.0816	100.0	0.0501	100.0

Table 3: TD run results from 49 topics having at least one OOV word in their title or description field and the TD results on the CLEF 2006/2007 test collection.

above, we were able to obtain most (97%) of the MAP that we could have obtained if we had instead used the best possible λ for each topic. This suggests our combination approach is not particularly sensitive to choice of λ .

6.2 Title Plus Description Runs

Table 3 lists our title plus description results. Looking at CLEF Topics first, we see that our DA system achieves a MAP of 0.0501. For comparison, the best TD result from the CLEF 2006 CL-SR track (using speech recognition transcripts only) reported a MAP of 0.0381 on the same topic set [20]. For the 2007 CLEF CL-SR track, this collection was again used and the best reported TD MAP was 0.0512 [24]. We also note that MAP from TD queries on the OOV Topics set is considerably higher than the title-only counterpart. As in the title-only run, both CombMNZ and interleaving do not yield a MAP measurably higher than the best of either system alone (the apparent improvement in MAP using CombMNZ on CLEF Topics is not statistically significant).

We saw on title-only queries that a trivial backoff combination, in which we used the OOD system for all IV queries and *only* the vocabulary-independent system for OOV queries, worked better than the OOD SR system alone. Using the longer TD queries however, we see from Table 3 that this approach does not improve over OOD SR. This is not surprising because the TD queries have additional, useful IV words which are ignored when the OOD RSVs are not utilized for ranking.

Measured on OOV Topics, our CMNT approach using RUR achieved a MAP of 0.0611, with an FRM of 41.3. As before, this gain is statistically significant.

7. CONCLUSION

We introduced a new approach to combining search results from multiple ranked retrieval systems. In particular, we combined ranked lists of segments from an SR system using OOD LVCSR and from a vocabulary-independent RUR system. By learning a smooth, monotonically increasing normalization of each systems’ retrieval status values, we produced a combined ranked list that improved, with statistical significance, MAP on an established set of SR topics over either system used alone. Because DA SR systems often can

not be constructed (e.g., in open-domain speech retrieval settings or because of costs), our goal was to recover MAP lost when we are constrained to use OOD systems. On a set of topics containing OOV title words, by combining systems, we recovered 57.5% of the MAP lost. On TD queries containing OOV words, our best system recovered 41.3% of the MAP lost. This is illustrated in Figure 4, which shows MAP for each query set using short (title only) and longer (title plus description) queries, with each of the OOD, CMNT and DA SR systems. We found that the MAP obtained using our new combination scheme was significantly greater than several previously studied SR techniques, including combination by backoff, combinations using less-principled score normalizations, and of course OOD SR.

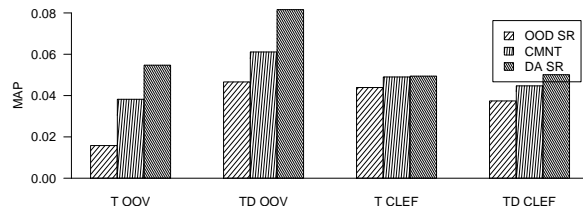


Figure 4: MAP for OOD, CMNT, and DA experiments on each test condition. Note, the leftmost two groups of bars show T and TD results using the OOV Topics set. On short (T) queries with OOV terms, we observe 71.1% of MAP is lost by using the OOD rather than DA SR system. By combining OOD SR with RUR we are able to recover 57.5% of this lost MAP.

When focusing on the OOV problem, SR researchers have often evaluated systems using collections and query sets that are not representative of real problems: many RUR papers, for example, use all present words as their test query set—while we wouldn’t expect RUR to be the system of choice for most words, i.e., words inside an OOD dictionary. In general, the problem of *creating* a reasonable set of OOV words is quite difficult, because of interaction effects with other, in-vocabulary, words (e.g., the SR system might unfairly retrieve speech using remaining words that co-occur with the artificially OOV words). We avoided these difficulties by using a real dictionary, but one that was not adapted to our topic domain. By extending this OOD dictionary to include all words found in a rather generous set of speech transcriptions (200 hours being far more than one could afford for most topic domains), we produced our DA dictionary and SR system. Together, these SR systems gave bounds on the MAP achievable without special handling of realistic OOV words. This, in turn, allowed for a simplified analysis of our combination approach’s utility (using FRM). We consider this experimental framework to be an important contribution of this work.

While we improved MAP with respect to an OOD word-based SR system, a gap remains between our combined system’s MAP and the MAP from the DA SR system. We attribute this primarily to two causes. First, vocabulary-independent spoken term frequency estimates are not as reliable as those from LVCSR. If we can anticipate a word when constructing the LVCSR system, it is best to include the word in the LVCSR dictionary and language model—

although of course anticipating the word may not be possible. Second, our combination approach does not model dependencies between the multiple retrieval status values. For example, we weighted the contribution from each OOV term equally, even though we know that different words should have different effects on the probability of a segment's relevance. We also assumed independence in our combination approach when in fact we would expect RSVs from different systems to be highly dependent. Finally, the gains we obtained required that models be trained for predicting relevance. This required costly relevance judgments and we do not yet know how sensitive CMNT is to the amount of available training data. We expect each of these difficulties will provide a fruitful venue for our future investigations.

8. REFERENCES

- [1] B. T. Bartell et al. Automatic Combination of Multiple Ranked Retrieval Systems. In *SIGIR '94*, pages 173–181, 1994.
- [2] S. M. Beitzel et al. Fusion of effective retrieval strategies in the same information retrieval system. *J. Am. Soc. Inf. Sci. Technol.*, 55(10):859–868, 2004.
- [3] 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.
- [4] J. P. Callan et al. Searching Distributed Collections with Inference Networks. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *SIGIR '95*, pages 21–28, Seattle, Washington, 1995. ACM Press.
- [5] J. Fiscus et al. English Spoken Term Detection 2006 Results. In *Presentation at NIST's 2006 STD Eval Workshop*, 2006.
- [6] J. Garofolo et al. The TREC spoken document retrieval task: A success story. Proceedings of the TREC-9 Conference, 2000.
- [7] D. A. James. A system for unrestricted topic retrieval from radio news broadcasts. In *ICASSP '96: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 279–282, 1996.
- [8] Jean-Manuel Van Thong et al. Speechbot: an experimental speech-based search engine for multimedia content on the web. *IEEE Trans. Multimedia*, 4:88–96, 2002.
- [9] G. J. F. Jones et al. Retrieving spoken documents by combining multiple index sources. In *SIGIR '96*, pages 30–38, New York, NY, USA, 1996. ACM.
- [10] P. Koehn et al. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL '07: Proceedings of the 2007 Conference of the Association for Computational Linguistics, demonstration session*, June 2007.
- [11] J.-H. Lee. Analyses of Multiple Evidence Combination. In *SIGIR Forum: Forum of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 267–276, 1997.
- [12] S. Lee et al. Combining Multiple Subword Representations for Open-Vocabulary Spoken Document Retrieval. In *ICASSP '05: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 505–508, March 2005.
- [13] D. Lillis et al. ProbFuse: a probabilistic approach to data fusion. In *SIGIR '06*, pages 139–146, New York, NY, USA, 2006. ACM.
- [14] B. Logan et al. Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio. In *HLT '02: Proceedings of the 2002 Conference on Human Language Technology*, 2002.
- [15] J. Mamou et al. Spoken document retrieval from call-center conversations. In *SIGIR '06*, pages 51–58, New York, NY, USA, 2006. ACM.
- [16] R. Manmatha et al. Modeling score distributions for combining the outputs of search engines. In *SIGIR '01*, pages 267–275, New York, NY, USA, 2001. ACM.
- [17] S. Matsoukas et al. The 2004 BBN 1xRT Recognition Systems for English Broadcast News and Conversational Telephone Speech. In *Interspeech '05: Conference of the International Speech Communication Association*, pages 1641–1644, 2005.
- [18] M. Montague and J. A. Aslam. Condorcet fusion for improved retrieval. In *CIKM '02: Proceedings of the 11th International Conference on Information and Knowledge Management*, pages 538–548, New York, NY, USA, 2002. ACM.
- [19] K. Ng and V. Zue. Subword-based approaches for spoken document retrieval. *Speech Commun.*, 32(3):157–186, 2000.
- [20] D. W. Oard et al. Overview of the CLEF-2006 Cross-Language Speech Retrieval Track. In *Proceedings of the CLEF 2006 Workshop on Cross-Language Information Retrieval and Evaluation*, September 2006.
- [21] J. S. Olsson. *Combining Evidence from Unconstrained Spoken Term Frequency Estimation for Improved Speech Retrieval*. PhD thesis, University of Maryland, College Park, MD, USA, 2008. Directed by Douglas W. Oard.
- [22] J. S. Olsson. Combining Speech Retrieval Results with Generalized Additive Models. In *ACL '08: Proceedings of the 2008 Conference of the Association for Computational Linguistics*, 2008.
- [23] J. S. Olsson and D. W. Oard. Phrase-Based Query Degradation Modeling for Vocabulary-Independent Ranked Utterance Retrieval. In *NAACL-HLT '09*, June 2009.
- [24] P. Pecina et al. 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 2007.
- [25] A. L. Powell et al. The impact of database selection on distributed searching. In *Research and Development in Information Retrieval*, pages 232–239, 2000.
- [26] R. Prasad et al. The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System. In *Interspeech '05: Conference of the International Speech Communication Association*, 2005.
- [27] S. Robertson et al. Okapi at TREC-3. In *Text REtrieval Conference*, pages 21–30, 1996.
- [28] M. Saraclar and R. Sproat. Lattice-Based Search for Spoken Utterance Retrieval. In *NAACL '04: Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2004.
- [29] J. A. Shaw and E. A. Fox. Combination of Multiple Searches. In *Proceedings of the 2nd Text REtrieval Conference (TREC-2)*, 1994.
- [30] O. Siohan and M. Bacchiani. Fast Vocabulary-Independent Audio Search Using Path-Based Graph Indexing. In *Interspeech '05: Conference of the International Speech Communication Association*, 2005.
- [31] A. Stolcke. SRILM – an extensible language modeling toolkit. In *ICSLP '02: Proceedings of 2002 International Conference on Spoken Language Processing*, 2002.
- [32] C. C. Vogt and G. W. Cottrell. Fusion Via a Linear Combination of Scores. *Information Retrieval*, 1(3):151–173, 1999.
- [33] E. M. Voorhees et al. 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, 1994.
- [34] R. W. White et al. Overview of the CLEF-2005 Cross-Language Speech Retrieval Track. In *Proceedings of the CLEF 2005 Workshop on Cross-Language Information Retrieval and Evaluation*, pages 744–759, 2005.
- [35] M. Witbrock and E. G. Hauptmann. Speech recognition and information retrieval: Experiments in retrieving spoken documents. In *In Proc. DARPA Speech Recognition Workshop '97*, 1997.
- [36] S. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC., 2006.
- [37] S. N. Wood. Monotonic smoothing splines fitted by cross validation. *SIAM Journal on Scientific Computing*, 15(5):1126–1133, 1994.
- [38] S. N. Wood. Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal Of The Royal Statistical Society Series B*, 62(2):413–428, 2000.
- [39] P. Yu and F. Seide. Fast Two-Stage Vocabulary-Independent Search In Spontaneous Speech. In *ICASSP '05: Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.