Combining Evidence for Improved Speech Retrieval

J. Scott Olsson
Department of Mathematics
University of Maryland
College Park, MD 20742
olsson@math.umd.edu

Abstract

The goal of my dissertation research is to investigate the combination of new evidence sources for improving information retrieval on speech collections. The utility of these evidence sources is expected to vary depending on how well they are matched to a collection's domain. I outline several new evidence sources for speech retrieval, situate them in the context of this domain dependency, and detail several methods for their combination with speech recognition output. Secondly, I highlight completed and proposed work for the production of this evidence.

1 Introduction and Goal

Early research in spoken document retrieval (SDR) was spurred by a new way to overcome the high cost of producing metadata (e.g., human assigned topic labels) or manual transcripts for spoken documents: large vocabulary continuous speech recognition. In this sense, SDR research has always been about making do with the available evidence. With the advent of automatic speech recognition (ASR), this available evidence simply grew from being only expensive human annotations to comparatively low-cost machine producible transcripts.

But today even more evidence is available for retrieving speech: (1) Using ASR text as input features, text classification can be applied to spoken document collections to automatically produce topic labels; (2) vocabulary independent spoken term detection (STD) systems have been developed which can search for query words falling outside of an ASR system's fixed vocabulary. These evidence sources can be thought of as two bookends to the spectrum of domain dependence and independence. On one end, topic labels can significantly improve retrieval performance but require the creation of a (presumably domain-dependent) topic thesaurus and training data. Furthermore, classification accuracy will be poor if the ASR system's vocabulary is badly matched to the collection's speech (e.g., we shouldn't expect a classifier to sensibly hypothesize automotive topics if the ASR system can not output words about *cars* or *driving*). On the other end, STD systems offer the most promise precisely when the ASR system's vocabulary is poorly matched to the domain. If the ASR system's vocabulary already includes every word in the domain, after all, STD can hardly be expected to help.

The primary goal of this dissertation is (1) to explore the combination of these new evidence sources with the features available in ASR transcripts or word lattices for SDR and (2) to determine their suitability in various domain-matching conditions. Secondarily, I'll explore improving the production of these new resources themselves (e.g., by classifying with temporal domain knowledge or more robust term detection methods).

Research in SDR has been inhibited by the absence of suitable test collections. The recently available MALACH collection of oral history data will, in large part, make this dissertation research possible (Oard et al., 2004). The MALACH test collection

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 33–36,
Rochester, April 2007.

contains about 1,000 hours of conversational speech from 400 interviews with survivors of the Holocaust[1]. The interviews are segmented into 8,104 documents with topic labels manually assigned from a thesaurus of roughly 40,000 descriptors. The collection includes relevance assessments for more than 100 topics and has been used for several years in CLEF's cross-language speech retrieval (CLSR) track (Oard et al., 2006).

Participants in the CLEF CLSR evaluations have already begun investigating evidence combination for SDR, through the use of automatic topic labels—although label texts are presently only used as an additional field for indexing. In monolingual English trials, this topic classification represents a significant effort both in time and money (i.e., to produce training data), so that these evidence combination studies have so far been rather domain dependent. Participants have also been using what are probably unnaturally good ASR transcripts. The speech is emotional, disfluent, heavily accented, and focused on a somewhat rare topic, such that the ASR system required extensive tuning and adaptation to produce the current word error rate of approximately 25%. In this setting, we'd expect STD output and topic labels to have low and high utility, respectively. To investigate the domain mismatch case, I will apply an off-the-shelf ASR system to produce new, comparatively poor, transcripts of the collection. In this setting, we'd expect STD output and topic labels to instead have high and low utility, respectively.

2 Proposed Combination Solutions

I will investigate improving SDR performance in both the poorly and well matched domain conditions through: (1) multiple approaches for utilizing automatically produced topic labels and (2) the utilization of STD output.

Throughout this paper, completed work will be denoted with a '$\star$', while proposed (non-complete, future) work will be denoted with a '†'.

[1]This is only a small subset of the entire MALACH collection, which contains roughly 116,000 hours of speech from 52,000 interviews in 32 languages. This additional data also provides training examples for classifi cation.

2.1 Speech Classification for SDR

I outline three methods of incorporating evidence from automatic classification for speech retrieval.

Creating Additional Indexable Text*

The simplest way to combine classification and speech retrieval is to use the topic labels associated with the classes as indexable text. As a participant on the MALACH project, I produced these automatic topic labels ("keywords") for the collection's speech segments. These keywords were used in this way in both years of the CLEF CLSR track. For a top system in the track, using solely automatically produced data (e.g., ASR transcripts and keyword text), indexing keyword text gave a relative improvement in mean average precision of 40.6% over an identical run without keywords (Alzghool and Inkpen, 2007).

Runtime Query Classification for SDR†

Simply using keyword text as an indexing field is probably suboptimal because information seekers don't necessarily speak the same language as the thesaurus constructors. An alternative is to classify the queries themselves at search time and to use these label assignments to rank the documents. We might expect this to be superior, insofar as information seekers use language more like interviewees (from which classification features are drawn) than like thesaurus builders.

Class Guided Document Expansion†

A third option for using classification output is as seed text for document expansion. The intuition here is that ASR text may be a strong predictor for a particular class label even if the ASR contains few terms which a user might consider for a query. In this sense, the class label text may represent a more semantically dense representation of the segment's topical content. This denser representation may then be a superior starting source for document centered term expansion.

2.2 Unconstrained Term Detection for SDR†

It is not yet clear how best to combine a STD and topical relevance IR system. One difficulty is that IR systems count words (or putative occurrences of words from an ASR system), while STD systems

report a score proportional to the confidence that a word occurs in the audio. As a solution, I propose normalizing the STD system's score for OOV query terms by a function of the STD system's score on putative occurrences of in-vocabulary terms. The intuition here is that the ASR transcript is roughly a ground truth representation of in-vocabulary term occurrences and the score on OOV query terms ought to reflect the STD system's confidence in prediction (which can be modeled from the STD system's score on "ground truth" in-vocabulary term occurrences). In this way, the presence or absence of in-vocabulary terms and their associated STD confidence scores can be used to learn a normalizer for the STD system's scores.

3 Producing the Evidence

In this section, I highlight both completed and proposed work to improve the production of evidence for combination.

3.1 Classifying with Temporal Evidence*

In spoken document collections, features beyond merely the automatically transcribed words may exist. Consider, for example, the oral history data contained in the MALACH collection. Each interview in this collection can be thought of as a time ordered set of spoken documents, produced by the guided interview process. These documents naturally arise in this context, and this temporal information can be used to improve classification accuracy.

This work has so far focused on MALACH data, although we expect the methods to be generally applicable to speech collections. For example, the topical content of a television episode may often be a good predictor of the subsequent episode's topic. Likewise, topics in radio, television, and podcasts may tend to be seasonally dependent (based on Holidays, recurring political or sporting events, etc.).

Time-shifted classification* One source of temporal information in the MALACH data is the features associated with temporally adjacent segments. Terms may be class-predictive for not only their own segment, but for the subsequent segments as well. This intuition may be easily captured by a *time shifted classification* (TSC) scheme. In TSC, each training segment is labeled with the *subsequent* segment's labels. During classification, each test segment is used to assign labels to its subsequent segment.

Temporal label weighting* We can also benefit from non-local temporal information about a segment. For example, because interviewees were instructed to relate their story in chronological order, we are more likely to find a discussion of childhood at an interview's beginning than at its end. We can estimate the joint probability of labels and segment times on held-out data and use this to bias new label assignments. We call this approach *temporal label weighting* (TLW).

In Olsson and Oard (2007), we showed that a combined TSC and TLW approach on MALACH data yields significant improvements on two separate label assignment tasks: conceptual and geographic thesaurus terms, with relative improvements in mean average precision of 8.0% and 14.2% respectively.

3.2 Classifying across languages*

In multilingual collections, training data for metadata creation may not be available for a particular language—a good example of domain mismatch. If however, training examples are available in a second language, the metadata may still be produced through *cross-language* text classification. In Olsson (2005), we used a probabilistic Czech-English dictionary to transform Czech document vectors into an English vector space before classifying them with k-Nearest Neighbors and English training examples. In this study, the cross-language performance achieved 73% of the monolingual English baseline on conceptual topic assignment.

3.3 Vocabulary Independent Spoken Utterance Retrieval*

In Olsson (2007), we examined a low resource approach to utterance retrieval using the expected posterior count of n-grams in phonetic lattices as indexing units. A query's phone subsequences are then extracted and matched against the index to produce a ranking on the lattices. Against a 1-best phone sequence baseline, the approach was shown to significantly improve the mean average precision of retrieved utterances on five human languages.

3.4 Improving Spoken Term Detection†

Phonetic lattices improve spoken term detection performance by more accurately encoding the recognizer's uncertainty in prediction. Even so, a correct lattice may not always contain a path with the query's entire phone sequence. This is so not only because of practical constraints on the size (i.e., depth) of the lattice, but also because speakers don't always pronounce words with dictionary precision. We'd like to allow approximate matching of a query's phone sequence with the phonetic lattices, and to do this as quickly as possible. This time requirement will prevent us from linearly scanning through lattices for near matches. I am currently investigating two solutions to this problem: phonetic query degradation and query expansion.

Phonetic query degradation† The idea in phonetic query degradation is to build an error model for the phone recognition system and to then degrade the query phone sequence such that it, hopefully, will more closely resemble recognized sequences. This approach incurs only a very slight cost in time and is query independent (in the sense that any term can be pushed through the degradation model—not, for example, only terms for which we can find recognized examples).

Phonetic query expansion† The idea of phonetic query expansion is, again, to transform the clean phone sequence of the query into the degraded form hypothesized by a recognizer. Instead of using a degradation model however, we simply run a first pass at STD with the non-degraded query term and use the putative occurrences to learn new, alternative, degraded forms for a second search pass. This can be thought of as blind relevance feedback or query by (putative) example.

The advantage of this approach is that we are not required to explicitly model the degradation process. Disadvantages are that we (1) require examples which may not be available and (2) assume that the degradation process is well represented by only a few examples.

4 Contributions

This dissertation will significantly contribute to speech retrieval research in several ways.

Can we improve SDR by evidence combination? By exploring evidence combination, this dissertation will advance the state of the art in speech retrieval systems and their applicability to diverse domains. I will investigate multiple methods for combining the evidence presented by both STD and classification systems with conventional ASR output (transcripts or word lattices). This work will develop upon previous research which studied, in depth, the use of only one evidence source, e.g., (Ng, 2000).

Can evidence combination decrease domain dependency? I will investigate how combining evidence sources can increase their applicability to new content domains. This will include, for example, understanding how (vocabulary independent) STD systems can be paired with fixed vocabulary ASR.

How can these evidence sources be improved? Lastly, I will explore how these new evidence sources may themselves be improved. This will include utilizing temporal domain knowledge for classification and improving the robustness of phone-based STD systems.

References

M. Alzghool and D. Inkpen. Experiments for the Cross Language Spoken Retrieval Task at CLEF 2006. In *Not yet published*.

K. Ng. 2000. Subword-based approaches for spoken document retrieval. MIT dissertation.

D.W. Oard, et al. Building an Information Retrieval Test Collection for Spontaneous Conversational Speech. In *Proceedings of SIGIR'04*.

D.W. Oard, et al. 2006. Evaluation of Multilingual and Multi-modal Information Retrieval. In *Seventh Workshop of the Cross-Language Evaluation Forum*, Alicante, Spain. Selected Papers Series: Lecture Notes in Computer Science.

J.S. Olsson and D.W. Oard. 2007. Improving Text Classifi cation for Oral History Archives with Temporal Domain Knowledge. In *Not yet published*.

J.S. Olsson, et al. Cross-language text classifi cation. In *Proceedings of SIGIR'05*.

J.S. Olsson, et al. Fast Unconstrained Audio Search in Numerous Human Languages. In *ICASSP'07*.