Diagnostic Techniques for Spoken Keyword Discovery

Peter Schulam¹, Murat Akbacak²

¹Johns Hopkins University ²Microsoft

pschulam@jhu.edu, murat.akbacak@ieee.org

Abstract

Keyword discovery is an unsupervised technology that can help to process collections of speech and capture repeated patterns. This technology becomes useful and provides solution for unsupervised content analysis tasks, especially when the acoustic and lexical characteristics are not known in advance or there is little or no data to model these characteristics via statistical models. In these situations, keyword discovery can find potentially important words for further analysis using minimal resources. Unfortunately, keyword discovery performance heavily depends on the quality of the features used to characterize the raw signal and the alignment algorithm used to find similar feature subsequences. It is not yet fully understood which features and alignment algorithms work well in different scenarios and for different tasks, and there are very few diagnostic techniques for improving our understanding. In this paper, we present two diagnostic measurements that can be used to directly assess the quality of alignments between sequences of features independently of the intended use of the alignments downstream. We argue that such diagnostic techniques are valuable for intrinsically assessing speech features and alignment algorithms for keyword detection.

Index Terms: keyword discovery, diagnostics

1. Introduction

Unsupervised spoken content analysis is a task broadly defined by the goal of extracting information from the raw speech signal. Examples include language identification, speaker identification, speaker diarization, and keyword discovery and detection. Such technologies are becoming useful as automatic speech recognition technology is more broadly used on mobiles and personal computers. As designers of such systems, it is increasingly difficult to anticipate the characteristics of the speech that will be processed, and it is therefore crucial to develop robust techniques for extracting useful information from the signal using minimal resources.

One important technology for unsupervised spoken content analysis, and the focus of this work, is keyword discovery, which is the task of identifying frequently repeated patterns of speech (e.g. words or phrases) using minimal resources. Keyword discovery can be used for a variety of downstream tasks. For example, detecting frequently repeated words in a large collection of spoken documents can help to index the documents, and, if the detections are sufficiently reliable, the indices can be used to cluster the documents without an ASR system.

Unfortunately, keyword discovery is still a developing technology, and the research community is still actively iterating and improving upon feature representations and algorithms for extracting keywords from the raw signal. It is not clear, however, how to develop improvements without a fixed downstream application to use for evaluation. Such extrinsic evaluations are difficult to use to drive research into generalizable techniques. In this paper, our aim is to propose a pair of intrinsic evaluation techniques for scoring the alignments of a keyword discovery pipeline. Because alignments are typically fed to clustering systems that produce the actual corpus-wide sets of subsequences of audio that we can consider to be instances of the same keyword, we believe that understanding and diagnosing the problems with alignment will lead to improvements in clustering and, ultimately, downstream applications.

2. Keyword Discovery Systems

A keyword discovery system can be characterized abstractly using the pipeline shown in Figure 1. State of the art algorithms are effectively providing improvements on the three steps:

Step 1 (Features) Extract features to represent the raw signal.
Step 2 (Alignment) Compare pairs of audio clips (usually at the utterance level) to find common subsequences.
Step 3 (Clustering) Cluster pair-wise alignments to find corpus-wide alignments.

Much of the work on unsupervised keyword discovery has addressed each of these steps jointly with the goal of discovering keywords for a particular corpus or collection of corpora with similar types of speech (see, for example, [1, 2, 3]). Although these papers have demonstrated impressive performance, it is still extremely difficult to develop a method that is robust to the type of speech that will be processed. Some of the ways in which the “type” of speech can vary are:

- Number of speakers (speaker heterogeneity)
- Vocabulary size (domain heterogeneity)
2.1. Related Work

Speaker Heterogeneity. Recognizing the same word spoken by different speakers is extremely challenging (see the results from [4], for example). One of the original papers on unsupervised keyword discovery ([1]) discovered keywords from lecture data. According to the types of heterogeneity described in the previous section, lecture data is speaker homogeneous, has relatively restricted vocabulary (the lecture is typically highly focused on a particular topic), and the recording conditions are typically high quality (it is relatively quiet, the microphone is usually directly in front of the speaker, etc.).

One of the first ways that we may wish to extend a keyword discovery system is by applying it to a collection of speech with multiple speakers. In [3], the authors tackle this problem on the TIMIT corpus. The TIMIT data is speaker heterogeneous, but has a limited vocabulary (the sentences are chosen to cover the articulatory spectrum and are repeated by each speaker) and the recording conditions are benign. We can therefore view TIMIT as an ideal testbed for understanding the effects that multiple speakers can have on keyword discovery results. The authors in [3] focus on one piece of the pipeline shown in Figure 1 in particular: feature extraction. They show that Gaussian mixture model posteriorgrams yield better keyword discovery results than MFCCs when there are multiple speakers.

In a more recent study [4], Carlin et al. performed an extensive evaluation of various candidate features and distance metrics on an isolated keyword similarity detection task. To evaluate the features and metrics, they identified a number of key-words in the Switchboard corpus and extracted several instances of each keyword where each instance was from a different speaker. They study keyword similarity detection in four different scenarios: the same word from the same speaker (SWSP), the same word from different speakers (SWDP), different words from the same speaker (DWSP), and different words from different speakers (DWDSP). They provide two results of interest to the current paper. First, they show that the distances obtained using dynamic time warping with the Euclidean distance over MFCC feature vectors yields a surprising amount of overlap between the four scenarios described above. In other words, the distances between the same word from the same speaker were often higher than distances between different words from different speakers. The second result that they demonstrate is that phone posteriorgrams dramatically improve over MFCCs. This is not a surprising result (since the phone recognizer should compensate for speaker and channel variation if it is properly trained), but it is still a useful result in that it shows the importance of designing features that are robust to speaker and channel heterogeneity.

Domain (Vocabulary) Heterogeneity. Carlin et al. showed that different words from the same speakers can yield surprisingly high similarities [4] (in some cases the similarities can even be greater than between the same word spoken by the same speaker). Therefore, it seems natural that broad domain speech with a potentially large number of distinct words can cause significant problems for keyword discovery systems.

Switchboard is one such corpus with a large domain. The switchboard corpus contains telephone conversations between pairs of volunteers who were prompted to discuss one of 70 topics. Because volunteers typically participate in several conversations, the domain for a particular speaker is relatively broad (i.e., a participant may talk about healthcare in one conversation and space exploration in another). Therefore, the switchboard corpus is a strong candidate for studying the effects of domain heterogeneity on keyword discovery performance. Unfortunately, the recording conditions are not ideal (telephone-quality audio with varying speaking styles for each speaker), and so there is a confounding effect.

Dredze et al. [5] perform document clustering on the switchboard conversations using keyword discovery to create features for each conversation. In Figure 1, we see that the final result of a keyword discovery system is a collection of subsequences that have been clustered into candidate keywords. Each of the subsequences (nodes in the cartoon representation in Figure 1) belongs to a document, and so we can count the number of times each candidate keyword occurs within a document to get a pseudo-bag-of-words [5]. The bag of words representations can be used to cluster the spoken documents. Dredze et al. get promising results, but rely on a phone recognizer to produce phone posteriorgrams as features in the keyword discovery process. Although this is a supervised solution, we believe it nicely cancels the confounding effect that channel heterogeneity of the Switchboard corpus has on studying broad domain scenarios.

Channel Heterogeneity: As mentioned above, the switchboard corpus can be used to study channel heterogeneity. Previous work that has done unsupervised processing of the speech to perform document clustering (as was done in [5]) but that has not relied on a supervised phone recognizer can be seen as representative state of the art solutions to keyword discovery in scenarios with high levels of channel heterogeneity. Examples of document clustering approaches on switchboard data include [6, 7].

3. Diagnosing Keyword Discovery Systems

Each of the related works that we described have made contributions to keyword discovery methodology in different scenarios that we have characterized using the tree types of heterogeneity outlined above (speaker, domain, and channel). Although the results are useful in considering the features and alignment algorithms that one can use when designing a keyword discovery system in a new domain, each of the keyword discovery systems were used for different purposes with varying levels of success. It is not clear which design decisions contributed to improvements for each scenario, and it is not clear whether similar improvements can be expected if the techniques are used for a different downstream application.

To enable translational results to be presented within the keyword discovery community, a set of intrinsic evaluation metrics are needed for assessing the quality of alignments before they are clustered and used for a particular application (e.g., document clustering). In this section, we propose an application-independent diagnostic technique for keyword discovery systems that uses two metrics that score the quality of alignments. In our experiments, we use a segmental dynamic time warping approach for extracting alignments [1], but we emphasize that our techniques can be applied to any set of alignments regardless of the chosen features and alignment algorithm used.
3.1. Metrics for Diagnosing Keyword Discovery Systems

We first introduce the pair of metrics that we use to intrinsically evaluate the quality of the segmental dynamic time warping (SDTW) alignment output. We are not aware of any previous work that has studied the quality of alignments themselves. Instead, emphasis is typically placed on the end result (usually under the assumption that alignments behave as expected). We are interested in understanding the potential for unsupervised keyword discovery in downstream applications, but also how well we can expect SDTW to perform in various conditions (e.g., multiple speakers, clean channel, telephone channel, etc.). To this end, we propose to use normalized discounted cumulative gain (NDCG) and “recall @ T” to assess the quality of alignments.

We use NDCG as a single number summary of alignment quality between each pair of utterances in our experiments. NDCG is typically used in information retrieval, where it summarizes the quality of a ranked list of documents. Given a list of documents that have been ranked by some retrieval system, the DCG (note that we have temporarily dropped the “N”) is defined to be

\[
DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i + 1)}
\]

Where \( p \) is the length of the list of documents and \( rel_i \) is the relevance of document \( i \) (according to some ground truth). It is clear that rankings that place the most relevant documents towards the top of the list will achieve higher DCG scores as they will not be penalized by the log of their position in the list. We adapt this metric for evaluating alignments by treating the collection of alignments returned by SDTW for each pair of utterances as a list of documents. The alignments for each pair of utterances are sorted in order of increasing distance, which places the alignments that we believe to be most similar at the head of the list. The relevance of each alignment is a binary value. Given the transcription of each utterance, we can determine from a particular alignment the pair of word sequences that have been connected by the alignment. We mark an alignment as good (i.e. relevant) if there is no silence on either side of the alignment and if at least 50% of the words on both sides that are not stop words are shared.

To compute NDCG, we simply normalize DCG by ideal DCG (IDCG), where IDCG is the DCG computed for the optimal ranking (i.e. if no bad alignment comes before a good alignment in the distance based ranking):

\[
NDCG = \frac{DCG}{IDCG}
\]

Since NDCG measures the quality of a set of alignments for a single utterance pair, we report the mean NDCG across all utterance pairs. Note that utterances with no shared content words will not have any good alignments and so will have an NDCG of 0. To get a better idea of how well SDTW performs when there are keywords that we are interested in discovering, we omit these bad utterance pairs from the mean.

The second measurement that we use is recall @ T, which is simply the percentage of good alignments between a pair of utterances that we retain after having thrown out all but the best (closest in distance) \( T \% \) alignments. SDTW can produce an overwhelming number of alignments (almost 14,000 for just 15 utterances in the DeathPenalty data described below), and so it may be necessary on very large data sets to prune the alignments for each utterance pair. The recall @ T measurement gives us an idea of how much information we are losing when we prune out the alignments with relatively high distances, which is something that is frequently done to reduce the number of candidate alignments that must be processed downstream.

4. Experiments

4.1. Datasets

We ran our experiments on three different datasets. Each of these datasets were chosen to reflect some combination of the three types of heterogeneity listed above. The DeathPenalty dataset is a collection of 15 utterances taken from the Switchboard corpus. Each utterance in the dataset contains the spoken phrase “death penalty”. There are two speakers, and the channel quality is poor. The 3644A dataset is a set of utterances from one side of conversation 3644 in the Switchboard corpus (speaker A). There are multiple repeated keywords in this dataset, but there is a single speaker. Because this is switchboard, the channel quality is still low. Finally, the TIMIT dataset is a collection of utterances from 70 speakers from TIMIT. The purpose of this dataset is to evaluate performance in conditions where there are many speakers and many keywords being spoken, but where the acoustic conditions are clean.

Table 1: Qualities of datasets. MKW: multiple keywords. MS: multiple speakers. LQC: low-quality channel.

<table>
<thead>
<tr>
<th>Name</th>
<th>MKW</th>
<th>MS</th>
<th>LQC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeathPenalty</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>3644A</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>TIMIT</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

In light of these results, we ran a similar set of experiments (isolated word similarity detection) to choose features for our task. We used two sets of features: MFCCs and Gaussian mixture model posteriograms (GMMPs). We did not want to use phone posteriograms because we want to assume that limited annotated resources are available. In addition to the actual feature representation, we experimented with two distance metrics for the GMMPs: cosine distance and symmetric Kullback-Leibler divergence. Our experiments yielded results similar to those in [4] and motivated our use of MFCCs with Euclidean distance, GMMPs with cosine distance, and GMMPs with symmetric KL divergence.
4.3. Alignment

Following feature extraction, the next stage of our system is to extract a set of alignments between all pairs of utterances in the corpus. We follow the segmental dynamic time warping (SDTW) approach of Park and Glass [1]. SDTW is similar in nature to the well-known dynamic time warping algorithm. The difference is in the number and location of the warp paths that are computed within the distance matrix between two utterances. In standard dynamic time warping, a single, optimal warp path is computed that aligns the frames of each utterance to one another. This approach is useful, for example, when comparing instances of isolated words and can be used to implement isolated word recognition. By computing warp paths between a query word and a set of templates (for which we know the transcription), we can “decode” the query by choosing the word associated with the template that yields the warp path with the lower distortion score.

In keyword discovery, we often do not have isolated word instances, but rather entire utterances that must be compared against one another. While silence can be used to help segment long audio clips into shorter subsequences, it is a difficult problem to properly segment a sentence or phrase into its constituent words. SDTW addresses this problem by choosing an array of starting points for the DTW computation. In practice, this translates into dividing the distance matrix into diagonal strips, and computing warp paths within each strip (i.e. each path must lie entirely within the strip in which it started). This is the most computationally intensive stage of the system, requiring $O(n^2m^2)$ time, where $n$ is the number of utterances in the corpus and $m$ is an upper bound on the length of any particular utterance (in frames).

5. Results and Discussion

For each of the datasets listed in Table 3.1, we extracted MFCCs and GMMPs. We computed alignments using SDTW [1] using Euclidean distance for MFCCs and cosine distance and symmetric KL divergence for GMMPs. We then used the computed alignments and groundtruth transcriptions for our datasets to compute the mean and median NDCG and recall @ $T$ for $T = \{0.25, 0.5, 0.75\}$.

There are some interesting points to note from our results. First, note that on the DeathPenalty dataset the mean NDCG is consistently higher than the median, suggesting that the distribution of NDCG over utterances is positively skewed. On the other hand, on 3644A and TIMIT the mean and median NDCGs are more closely matched suggesting less skewed distributions. This result suggests that the DeathPenalty dataset is more difficult on average.

We also found that MFCCs with Euclidean distance gave the best mean and median NDCG for all three datasets. We found this surprising in the case of the DeathPenalty and TIMIT datasets because we expected that the GMMPs would be more robust to speaker variation.

For recall @ $T$, we found that MFCCs yielded the highest scores for the TIMIT data, but the GMMPs improved results for DeathPenalty and 3644A. We believe that this is due to the fact that the GMMPs may be more robust to channel variation (the TIMIT data is clean read speech).

6. Conclusion

We have described a novel framework for understanding the types of heterogeneity that are encountered in the development of keyword discovery systems. We identified previous work that has built keyword discovery systems for keyword extraction or for other downstream applications and discussed how the contributions of each of the previous contributions have provided a solution for a particular type of heterogeneity in our framework. We argued for the importance of intrinsic evaluation techniques, and presented two such techniques that assess subsequence alignments.

Our aim in presenting this work is twofold. First, we believe that designing shared tasks on datasets with different types and amounts of the heterogeneity described in this paper will lead to a better understanding of the weaknesses of existing and proposed algorithms. Second, we hope that the proposed diagnostic techniques will be extended and used in future work to compare and evaluate possible improvements to the keyword discovery pipeline in an application-independent manner.
7. References


