Spoken Document Clustering Using Word Confusion Networks

Shajith Ikbal, Sachindra Joshi, Ashish Verma, Om D Deshmukh
IBM Research, India.
{shajmoha, jsachind, vashish, odeshmuk}@in.ibm.com

Abstract
In this paper, we propose a word confusion network (WCN) based approach to perform clustering of the spoken documents and analyze its ability to handle the influence of speech recognition errors. WCN compactly represents multiple confidence weighted recognition hypotheses. Thus it provides scope for improving the clustering accuracy as a result of the likely presence of the correct transcription in the alternative hypotheses for those cases where 1-best transcripts are erroneous. On the other hand, several of the remaining hypotheses are incorrect and hence could pose a challenge during the clustering. In our approach, we extract TF-IDF vectors from the WCNs to perform clustering using K-Means algorithm. The components of TF-IDF vectors are further weighted with the word posterior probabilities. This is to potentially down-weight those vector components that are contributed by the incorrect hypotheses of low posterior probabilities. The experimental results obtained using switchboard data illustrate the usefulness of rich information in the WCN for clustering, showing up to 4% absolute improvement in normalized mutual information metric.

Index Terms: spoken document clustering, word confusion network, posterior weighted TF-IDF vector, k-means clustering

1. Introduction
Clustering is a fundamental statistical data analysis technique useful in many fields including data mining, information retrieval and bioinformatics. It facilitates discovering groupings among a set of data points in an unsupervised manner based on similarity among them. The groups obtained can roughly be thought of as a representation of the set of underlying topics within the data set. The focus of this paper is on clustering spoken documents, where the spoken documents are first transcribed into text using automatic speech recognition (ASR) system and then standard clustering approaches such as K-Means algorithm [1] are applied on the resulting text. We specifically consider spoken documents of conversational speech, aiming to discover the set of topics being discussed in the whole set of documents.

There is tremendous amount of literature reporting prior work on the generic text clustering exploring various algorithms, features, distance metrics and applications [2]. However, to our knowledge there is not much prior work related to the spoken document clustering. Some distinguishing aspects of the speech transcripts that could potentially influence the clustering process include recognition errors, repetitions, corrections, and non-grammatical sentences. Among these factors, in this paper, we particularly focus on the influence of speech recognition errors and explore methods of handling it. Typical word error rates (WER) for speech transcripts range from 10% to 60% depending upon the type and the source of the data. As will be reported in the later sections of this paper, we observe a moderate to large degradations in the clustering performance for various word error rates (WER) of speech transcripts. This is because, the correct recognition of certain key words spoken during the conversation is crucial for the over all clustering performance since the frequency of occurrence of those key words typically play an important role in the computation of similarity among data points. In fact, for a related problem of spoken document retrieval (SDR) whose performance also depends heavily upon the frequency of the key words, one of the conclusions of NIST SDR evaluations conducted as a part of text retrieval conference (TREC) is that the effectiveness of retrieval from spoken documents depends upon the accuracy of their ASR transcripts [3]. Issues related to the processing of ASR transcripts have been explored in the context of other problems such as information extraction [4, 5] and information retrieval [6, 7].

In this paper, we propose using a richer recognition output of the ASR system namely word confusion networks (WCN) [8] to perform the spoken document clustering. Our main motivation behind the use of WCN for clustering problem is the prior work showing the usefulness of rich ASR output in the context of problems such as multi-pass ASR systems [8], multiple ASR system output combination [10], machine translation [11], spoken document summarization [13], extraction [14] and classification [15]. In contrast to the 1-best transcript, WCN represents multiple recognition hypotheses obtained during speech decoding in a compact format. Thus it provides better scope for improving the clustering performance because of the potential presence of the correct transcriptions in the alternate hypotheses for those speech fragments whose 1-best transcriptions are erroneous. However, on the negative side, including multiple recognition hypotheses could also potentially degrade the clustering performance because of the inclusion of several incorrect words from the incorrect alternative hypotheses. In an effort to make a good trade-off between these two conflicting factors, we make use of the posterior probabilities associated with the words in the WCN to potentially down-weight the influence of uncertain incorrect words obtained during recognition. Experimental results given in the later sections of the paper show that such utilization of information from alternate hypotheses indeed leads to improved clustering performance.

The organization of this paper is as follows: In Section 2, we describe spoken document clustering using word confusion networks. In Section 3, we describe the experimental setup used for evaluation. In Section 4, we present and discuss the experimental results. In Section 5, we conclude and discuss potential future directions.

2. Clustering using WCN
2.1. Briefly about WCN
WCN is derived from word lattice which represents a portion of the decoding search space that incorporates top scoring recog-
tion hypotheses. As illustrated in Figure 1, both word lattice and WCN are acyclic graphs containing a set of nodes and arcs. Nodes correspond to the time instants and arcs correspond to the word hypotheses in the time slots between their start and end nodes. Word lattice typically represents a large number of recognition hypotheses of multiple time alignments. Whereas WCN represents hypotheses of single time alignment derived by mapping the lattice hypotheses onto a single time alignment. This is achieved by mapping all the word hypotheses in the lattice onto various slots in the single time alignment through grouping of the words based on similarity of their pronunciation and occurrence times [8]. In addition, low probability word hypotheses are pruned out. The posterior probabilities assigned to a word in the WCN is computed as the sum of probabilities of all the paths in the lattice containing that word around its time duration.

![Lattice](image1.png)

![Word confusion network](image2.png)

Figure 1: Illustration of the word lattice and word confusion network (WCN).

2.2. Why WCN for spoken document clustering?

WCN represents multiple recognition hypotheses and hence is richer in information than the 1-best transcripts. The grouping and pruning processes to derive WCN from lattice result in the removal of redundant and low probability hypotheses thus making the WCNs more compact and efficient than the word lattices. This makes WCNs more suitable for spoken language understanding tasks. Clustering being an unsupervised approach it is difficult to get it converged to an useful solution in the presence of large amount of irrelevant information. This is indeed confirmed by our early experiments showing improved clustering performance with WCNs than the word lattices. Hence, in this paper, we chose to use WCN instead of the word lattice as a representation of multiple recognition hypotheses.

2.3. Our approach

A relatively simple but an effective approach to perform clustering in the text processing domain is: preprocess the text documents to convert them into numerical vectors and apply standard K-Means algorithm to discover the groupings among them. Note that since our aim is to explore the influence of bag-of-words model used for the preprocessing step instead of plain text containing a sequence of words. The preprocessing step using WCN is explained in detail in the next subsection 2.4. The traditional preprocessing for text documents that uses bag-of-words model is explained here briefly: A vector is constructed for each document with vector components corresponding to the unique words present in the whole set of documents. The vectors components are weighed based on how important the corresponding word is to the document based on term frequency - inverse document frequency (TF-IDF) [16]. In TF-IDF representation the importance of a word is assumed to be directly proportional to the number of times it appears in the document and inversely proportional to the number of documents it appears in the whole corpus.

3) K-Means clustering: This step uses the document vectors obtained in preprocessing step to discover the groupings. K-Means is an iterative algorithm where during each iteration each data point is assigned to a cluster whose centroid vector is the closest to the document vector corresponding to that data point and the cluster centroids are updated as an average of all the data points assigned to the cluster [1].

2.4. Preprocessing using WCN

As mentioned before, our approach to cluster the spoken documents differs mainly in the preprocessing step since it should process WCN instead of the plain text. From the point of view of preprocessing, the main differences between the plain text and the WCN are: 1) Words in plain text can be strictly time ordered but in WCN more than one word can have the same start time. However, this difference does not require any special handling since the bag-of-words model used for the preprocessing step disregard the time ordering of the words. 2) Words in WCN are assigned with posterior probabilities reflecting the confidence score associated with the word. We use this as a weight for the words in the vectors in addition to the TF-IDF weighting. Note that in contrast to our claim of potential improvement in the clustering performance, the inclusion of multiple recognition hypotheses from WCN could also potentially degrade the clustering performance. This is because there could be many incorrect hypotheses in WCN that could add many incorrect words in the document vectors. The use of posterior probabilities should help in decreasing the influence of such incorrect words because the words with low probability would then be down-weighted in the document vectors. The exact preprocessing procedure is explained as below: Let \( D = \{d_1, d_2, \ldots, d_N\} \) be the set of spoken documents, and \( V = \{v_1, v_2, \ldots, v_K\} \) be the set of unique words in the vocabulary. Let WCN for document \( d_n \) be:

\[
\{[b_1^n, e_1^n, q_1^n], [b_2^n, e_2^n, q_2^n], \ldots, [b_{r_n}^n, e_{r_n}^n, q_{r_n}^n]\}
\]

where \( b_i^n \) and \( e_i^n \) respectively denote the begin and end times of \( i^{th} \) slot in the WCN, and \( q_i^n \) the word distribution at \( i^{th} \) slot as given by:

\[
q_i^n = \{[w_1^{n, i}, p_1^{n, i}], [w_2^{n, i}, p_2^{n, i}], \ldots, [w_{S_{G_n}}^{n, i}, p_{S_{G_n}}^{n, i}]\}
\]

where \( w_s^{n, i} \) denote the \( s^{th} \) word and \( p_s^{n, i} \) denote the corresponding posterior probability. The TF-IDF weight for word \( v_s \) and document \( d_n \), taking into account the word confidence score is computed as:
where $wtf(v_k, d_n)$ denote the posterior weighted term frequency computed as:

$$wtf(v_k, d_n) = \sum_{t=1}^{T_n} \sum_{s=1}^{S_n} p_s^n : w_{s,t}^n = v_k$$

and $idf(v_k)$ denote the inverse document frequency computed as

$$idf(v_k) = \log \left( \frac{N}{\sum_{n=1}^{N} 1 : v_k \in d_n} \right)$$

where $N$ is the total number of spoken documents. In the above equations, $x : y$ denote a function whose return value is $x$ if $y$ is true, 0 if $y$ is false.

The TF-IDF computation above incorporates the word confidence score as the weighing factor only in the TF component. One can think of incorporating it in several ways during the computation of the TF-IDF vector. For example, also to weigh the IDF factor for a particular word with average confidence score of that word in the document. However, based on our early experiments weighing only the TF component of the TF-IDF using word confidence score seems to give the best performance.

### 3. Experimental setup

To evaluate the performance of the WCN based spoken document clustering we used a subset of topic labeled switchboard conversations. Note that the topic labels in this data set are used only for evaluation purposes. Our main aim in this paper is to discover those topics in an unsupervised manner through clustering and the true topic labels are used to measure the quality of the clusters obtained. The data subset we took contain approximately 1100 conversations, labeled with 10 distinct and non-overlapping topics such as tourism, terrorism, middle east, etc.

We used manual transcripts, 1-best ASR transcripts and WCNs of these calls to evaluate the influence of the recognition errors and the usefulness of the rich information in the WCN. To evaluate the influence of the extent of ASR recognition accuracy on the clustering performance we generated 1-best transcripts and WCN for various word error rates of the 1-best transcripts such as 30%, 35%, 40%, 50% and 65%. In addition, we performed experiments using two different lengths of the same set of calls to evaluate the influence of the call length on the clustering performance. In one set of experiments we used the whole length of each call, approximately 10 minutes long. In the other set of experiments we used only the initial 10% of each call. The reason for doing these experiments is simulate the practically useful real-life scenarios such as call center conversations involving agents and customers where the call lengths are typically short.

The conversations in switchboard calls discuss the same topic over the entire call, thus contain enough redundant information about the topic being discussed, which potentially makes them relatively more immune to the recognition errors. On the other hand, the information of interest in call center conversations, such as problem faced by the customer, are expressed over a relatively shorter duration of time, which makes them more susceptible to the recognition errors.

To transcribe the calls for clustering experiments we used a HMM based speaker independent context-dependent speech recognition system built using IBM Attila speech recognition toolkit [17], trained with switchboard data that is not part of the data used for clustering experiments. In order to generate transcriptions at various error rates we ran the recognition with various values of the decoding beam widths and active states. We chose these values to generate transcripts of WERs similar to that typically achieved with call center conversations in practice. Prior to the processing to generate document vectors for the clustering experiments, the manual transcripts, 1-best transcripts and WCN are normalized as follows: Stop words, fillers, fog horns are removed. Words that are less frequent in the whole corpus, i.e., those occurring only three times or less in the whole corpus, are removed. Words whose posterior probabilities are less than 0.01 are removed. All the remaining words are stemmed using Porter stemmer [18].

The metric used in this paper to measure the clustering performance is normalized mutual information (NMI) [19] which is widely used in the text clustering literature. Given a set of cluster labels $C$ and true labels $T$, NMI metric is computed as follows:

$$NMI(C, T) = \frac{MI(C, T)}{\max(H(C), H(T))}$$

where $H(C)$ and $H(T)$ are respectively the entropies of $C$ and $T$, $MI(C, T)$ is the mutual information between $C$ and $T$ as defined below:

$$MI(C, T) = \sum_{c_i \in C, t_j \in T} P(c_i, t_j) \log \frac{P(c_i | t_j) P(t_j)}{P(c_i) P(t_j)}$$

where $P(c_i)$ and $P(t_j)$ respectively denote the probabilities that a document selected arbitrarily from the data belongs to the clusters $c_i$ and $t_j$ respectively, and $P(c_i, t_j)$ denote the joint probability that the arbitrarily selected document belongs to both the clusters $c_i$ and $t_j$ at the same time. The metric $NMI(C, T)$ takes values between 0 and 1, higher the value better the clustering accuracy.

In all our experiments we ran the K-Means clustering algorithm to derive a fixed number of target clusters, which in our case is 10 corresponding to the total number of true topic labels. We randomly initialized the cluster centroids during the first iteration of the K-Means algorithm. Since the clustering output is typically sensitive to the initialization, we repeated each clustering experiment 30 times with different random initializations to find and report the average NMI metric for that clustering experiment. But we also used the same set of 30 initial cluster centroids for all the clustering experiments in order to have a fair comparison between the methods.

### 4. Experimental Results and discussion

Table 1 gives a summary of all the experimental results. It compares normalized mutual information metric, $NMI(C, T)$ as obtained using (1), while performing clustering using manual transcripts, 1-best transcripts and word confusion networks for various transcription accuracies and call lengths. As can be seen from the table, in general, there is a degradation in the clustering performance while moving from the manual transcripts to the ASR outputs, i.e., 1-best transcript and WCN. This illustrates the influence of the erroneous ASR output on the clustering performance. Also there is a gradual degradation with gradual increase in the word error rate (WER) for both 1-best ASR output and WCN, illustrating the influence of the noise level in the transcripts.

From the table, the NMI metric achieved using WCN is consistently better than that obtained with the 1-best transcripts, illustrating the usefulness of the proposed usage of rich information in the WCN through posterior weighting. However, while using the whole call to perform the clustering, the improvements obtained are not significant except for higher word
error rates. On the other hand, improvements obtained are consistently higher while using the shorter version of the call, i.e., 10% of the total call length, with up to 4% absolute (12% relative) improvement in the mutual information metric (an increase of 0.04 in the absolute value for WER-50). This could be attributed to the amount of redundant information in case of the whole call in comparison to the shorter version of the call. Note that both the shorter and longer versions of each call contain features related to the same topic, except that the longer version contain relatively more topic specific features with potential repetition of the same feature many times over the entire call. Hence the chance of ASR transcription error affecting the unique feature set is less likely in case of the longer version of the calls than the shorter version of the calls. On similar lines, it can be observed that the drop in NMI metric with increase in word error rate is sharper in case of shorter version of the calls than the longer version of the calls.

One deviation from the expected performance for WCN in the table is for the case where WER is 30% while using the whole call. For this case, a closer examination showed that the particular value of decoding beam width used to generate the transcripts of 30% WER has resulted in very large sized WCNs. This in turn resulted in document vectors of relatively larger dimensions with many of their components constituted by incorrect words. This effectively lead to the marginal degradation in the clustering performance for this case.

5. Conclusions and future work

In this paper, we addressed the problem of spoken document clustering and described an approach to cluster the spoken documents using word confusion networks. The experimental results obtained illustrate the influence of the ASR transcription accuracy on the clustering performance and the usefulness of the rich information in WCN to improve the clustering accuracy of spoken documents. We observed up to 4% absolute (12% relative) improvement in the normalized mutual information metric while using WCN along with word posterior probabilities in comparison to the 1-best transcripts. The gains obtained using WCN seems to be relatively higher in the shorter version of the calls, indicating that in the absence of sufficient redundant information, i.e., when more susceptible to the transcription errors, the rich information from WCN is more useful.

One potential future direction is to explore approaches to select an useful subset of features from the whole set of features extracted from the WCNs, for example through topic summarization of the calls using WCNs. This could potentially improve the performance since the irrelevant features typically behave like noise distracting the clustering process.

### Table 1: Clustering performance for various length of calls and transcription accuracies (WER).

<table>
<thead>
<tr>
<th>Type</th>
<th>MAN</th>
<th>WER-30</th>
<th>WER-35</th>
<th>WER-40</th>
<th>WER-50</th>
<th>WER-65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual transcript</td>
<td>0.86</td>
<td>0.85</td>
<td>0.84</td>
<td>0.84</td>
<td>0.77</td>
<td>0.51</td>
</tr>
<tr>
<td>1-best ASR transcript</td>
<td>0.84</td>
<td>0.86</td>
<td>0.84</td>
<td>0.84</td>
<td>0.78</td>
<td>0.54</td>
</tr>
<tr>
<td>WCN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalized mutual information (NMI) metric while using the whole call</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual transcript</td>
<td>0.60</td>
<td>0.58</td>
<td>0.56</td>
<td>0.47</td>
<td>0.31</td>
<td>0.13</td>
</tr>
<tr>
<td>1-best ASR transcript</td>
<td>0.61</td>
<td>0.58</td>
<td>0.50</td>
<td>0.35</td>
<td>0.35</td>
<td>0.14</td>
</tr>
<tr>
<td>WCN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalized mutual information (NMI) metric while using initial 10% of the call</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### References


