Spoken Document Confidence Estimation Using Contextual Coherence

Taichi Asami, Narichika Nomoto, Satoshi Kobashikawa, Yoshikazu Yamaguchi,
Hirokazu Masataki, Satoshi Takahashi

NTT Cyber Space Laboratories, NTT Corporation,
1-1 Hikari-no-Oka, Yokosuka-shi, 239-0847 Kanagawa, Japan
asami.taichi@lab.ntt.co.jp

Abstract

Selecting well-recognized transcripts is critical if information retrieval systems are to extract business intelligence from massive spoken document databases. To achieve this goal, we target spoken document confidence measures that represent the recognition rates of each document. We focus on the incoherent word occurrences over several utterances in ill-recognized transcripts of spoken documents. The proposed method uses contextual coherence as a measure of spoken document confidence. The contextual coherence is formulated as the mean of pointwise mutual information (PMI). We also propose a smoothing method of PMI, which deals with the data sparseness problem. Compared to the conventional method, our smoothing technique offers improved correlation coefficients between spoken document confidence scores and recognition rates from 0.573 to 0.672. Moreover, an even higher correlation coefficient, 0.710, is achieved by combining the contextual-based and decoder-based confidence measures.

Index Terms: speech recognition, confidence measures, spoken documents, contextual coherence

1. Introduction

Speech mining systems that utilize massive sets of spoken documents are entering into practical use. Several studies have applied automatic speech recognition (ASR) and information retrieval techniques to large databases of call center recordings in order to extract business intelligence [1, 2].

Such systems have great difficulty in obtaining high recognition rates for all spoken documents since the speakers, noise, and channels of each spoken document vary widely and utterances are spontaneous. Therefore, the databases contain spoken documents that have low recognition rate transcripts. Such poor transcripts cause information retrieval error, since they include a lot of false information, i.e., ASR error words that have no relation to the actual content of the spoken document [1]. In order to reduce the information retrieval error, selecting reliable transcripts from the target database is an important issue for speech mining.

If the recognition rates of each spoken document can be accurately estimated, the system can select reliable transcripts easily. The goal of this paper is to make a confidence measure that accurately predicts the recognition rates of each document.

Word confidence measures have been widely studied [3, 4]. Several studies have reported that the contextual coherence of content words in recognized hypotheses is effective in estimating word confidence scores [5, 6, 7, 8]. These techniques are based on the idea that a word that appears to be incoherent in a recognized hypothesis is likely to be wrong. The contextual coherence of an utterance hypothesis can be calculated by using the relatedness measures of word-pairs.

In our task, estimating “spoken document” confidence, contextual coherence is an especially powerful source of information, since spoken documents usually consist of a few dozen to hundreds of utterances. In other words, we can use recognition hypotheses with long ranges of up to several utterances to obtain contextual coherence. The transcripts of spoken documents that offer high recognition rates exhibit strong coherence over several consecutive utterances. In contrast, low recognition rate transcripts exhibit strong incoherence over several utterances, i.e., recognized words are not making any sense at all. The difference in contextual coherence between high and low recognition rate transcripts is marked.

Based on this understanding, we propose a method that uses contextual coherence over several utterances for estimating spoken document confidence. The proposed method sets windows covering several utterances in the transcript of a spoken document, and calculates confidence scores from the contextual coherence of content words in each window.

The contextual coherence is formulated as the arithmetic mean of pointwise mutual information (PMI) in the window. However, data sparseness triggers two problems in PMI. The first one is that the PMI of word-pairs that are not present in the training sets cannot be calculated. The second one is that PMI values become too large when the occurrence frequencies of the words of a word-pair are quite low. This paper also proposes a PMI smoothing method that overcomes both problems.

This paper is organized as follows. Section 2 describes the proposed framework for estimating spoken document confidence. Section 3 describes the proposed PMI smoothing method. Experimental conditions and results are presented in Section 4, and Section 5 concludes this paper.

2. Spoken document confidence estimation

The entire flow of the proposed spoken document confidence estimation is shown in Fig. 1.

First, the windowing procedure is applied to the automatically generated transcript of spoken document $D$. The purpose of windowing is to obtain the contextual coherence from the range that includes one topic. We assume that topics can be switched within a spoken document, as in phone conversations. Strong contextual coherence is not guaranteed if multiple topics are present. Therefore, division of a document into segments that cover single topics by windowing is required.

Window length is $N$ content words and the amount of window shift is $M$ content words ($N > 1$ and $1 \leq M \leq N$). Each window includes $N$ content words and other words. The content words in each window are extracted and taken to be the bag
of words for that window. A spoken document is taken to be the set of all bags of content words, \( D = \{B_1, B_2, \ldots, B_L\} \), where \( L \) denotes the number of extracted bags of words.

The contextual coherence of each bag of words is calculated. The proposed method formulates the contextual coherence based on the idea that contextual coherence is low if the relationships between the words in a bag are weak. The contextual coherence of bag of words \( B_i \), which consists of \( N \) content words, \( w_1, w_2, \ldots, w_N \), is calculated as

\[
C_C(B_i) = \frac{1}{N} \sum_{i=1}^{N} \log \frac{P(w_i|B_i \setminus \{w_i\})}{P(w_i)}
\]

where \( B_i \setminus \{w_i\} \) is the bag of words \( B_i \) from which \( w_i \) is omitted (i.e., the set of neighbor words of \( w_i \)), \( P(w_i) \) is the probability that \( w_i \) occurs in \( B_i \), and \( P(w_i|B_i \setminus \{w_i\}) \) is the probability that \( w_i \) occurs in \( B_i \) when the neighbor words of \( w_i \) are given. \( C_C(B_i) \) is the normalized log likelihood ratio and tests whether the occurrences of each word in \( B_i \) are due to the relationship to neighbor words or due entirely to chance. When \( C_C(B_i) \) is small, the relationships of words in \( B_i \) are weak and the contextual coherence is low.

It is difficult to accurately estimate \( P(w_i|B_i \setminus \{w_i\}) \) since the number of combinations of words in \( B_i \) is enormous relative to the amount of data actually available. Consequently, the proposed method approximates \( P(w_i|B_i \setminus \{w_i\}) \) as the geometric mean of the probabilities that \( w_i \) occurs in \( B_i \) when each neighbor word of \( w_i \) is given. This is calculated as

\[
P(w_i|B_i \setminus \{w_i\}) \approx \left( \prod_{j=1, j \neq i}^{N} P(w_i|w_j) \right)^{1/(N-1)}.
\]

By substituting Eq. (2) into Eq. (1), the contextual coherence of bag of words \( B_i \) is calculated as follows.

\[
C_C(B_i) \approx \frac{1}{N} \sum_{i=1}^{N} \log \left( \prod_{j=1, j \neq i}^{N} \frac{P(w_i|w_j)}{P(w_i)} \right)^{1/(N-1)}
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} \log \left( \prod_{j=1, j \neq i}^{N} \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \right)^{1/(N-1)}
\]

\[
= \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}
\]

where \( P(w_i, w_j) \) is the probability that both \( w_i \) and \( w_j \) occur in \( B_i \).

Eq. (3) is the arithmetic mean of PMI of all word-pairs in \( B_i \). PMI indicates the strength of the relatedness of a word-pair [9]. When the contextual coherence of \( B_i \) is low (i.e., \( B_i \) includes many error words), more of the word-pairs in \( B_i \) are less likely to co-occur. Accordingly, the right side of Eq. (3) becomes small. Therefore, approximating \( C_C(B_i) \) by Eq. (3) is appropriate as a measure of contextual coherence.

Finally, the arithmetic mean of \( C_C(B_i) \) is calculated as the confidence measure of spoken document \( D \):

\[
C(D) = \frac{1}{L} \sum_{i=1}^{L} C_C(B_i).
\]

3. Smoothing pointwise mutual information

The proposed method calculates the contextual coherence of a bag of words as the mean of PMI of all word-pairs in the bag by Eq. (3). Unfortunately, if data is sparse, PMI suffers two problems. We propose a new PMI smoothing method that can handle both problems.

3.1. Two problems of PMI

The PMI of two words, \( x \) and \( y \), is expressed as follows [9].

\[
\text{PMI}(x, y) = \log \frac{P(x, y)}{P(x)P(y)} = \log \frac{f(x, y) \cdot K}{f(x)f(y)}
\]

where \( f(x) \) is the occurrence frequency of \( x \) (the number of bags of words that include \( x \)), \( f(x, y) \) is the co-occurrence frequency of \( x \) and \( y \), and \( K \) is the total frequency (the number of all bags) in the training set. Positive PMI values indicate that \( x \) and \( y \) tend to co-occur more than chance. Negative PMI values indicate that \( x \) and \( y \) tend not to co-occur more than chance, and PMI becomes 0 when \( x \) and \( y \) are independent.

The two problems are found in Eq. (5):

1. PMI cannot be calculated when \( f(x, y) = 0 \). In this case, PMI always becomes \(-\infty\) according to the above definition.

2. PMI becomes too large when \( f(x) \) and \( f(y) \) are small.

For example, when \( f(x) = f(y) = f(x, y) = 1 \), PMI(\( x, y \)) is log\( K \) (\( K \) exceeds 100,000 in practice). Meanwhile when \( f(x) = f(y) = f(x, y) = 50 \), \( x \) and \( y \) have a stronger relationship than is true in the former case. Nevertheless, PMI(\( x, y \)) is smaller, log\( K/50 \).

Guo et al. proposed a smoothing technique in order to deal with the first problem [6]. This technique corrects the co-occurrence frequencies and probabilities by adding a constant
and interpolating as follows.

\[
I(x, y) := f(x, y) + I,
\]

\[
P(x, y) := P(x, y) - \alpha P(x)P(y),
\]

where \(I\) and \(\alpha\) are parameters optimized manually on a development set. All word-pairs have more than 0 frequency by this correction. However, this smoothing method does not deal with the second problem.

### 3.2. Proposed smoothing method

The proposed PMI smoothing method deals with both problems as follows.

Against the first problem, it corrects the frequency of unobserved word-pairs using the simple Turing estimator [10]. The co-occurrence frequencies are corrected as follows.

\[
f'(x, y) = \begin{cases} 
  f(x, y) & \text{if } f(x, y) > 0 \\
  \frac{N}{N_1} & \text{else}
\end{cases}
\]

(8)

where \(N_1\) is the number of word-pairs observed once in the training set and \(N_2\) is the number of unobserved word-pairs. All word-pairs have non-zero frequency by this correction. This correction does not require manual parameter optimization.

Next, to counter the second problem, we introduce the idea that PMI should be 0 if \(f(x)\) and \(f(y)\) are too small to permit the relationship of the word-pair to be judged. The proposed method uses the t-test for examining whether \(f(x)\) and \(f(y)\) are large enough or not. The t-score, which tests whether the difference between \(P(x, y)\) and \(P(x)P(y)\) is significant or not, is calculated as follows [11].

\[
t(x, y) \approx \frac{|P(x, y) - P(x)P(y)|}{\sqrt{P(x,y)/N}}
\]

\[
= \frac{|f(x, y) - \alpha f(x) f(y)|}{\sqrt{f(x, y)}}.
\]

(9)

Finally, the smoothed PMI is obtained as follows.

\[
PMI(x, y) = \begin{cases} 
  \log \frac{f(x, y) \cdot K}{\alpha f(x) f(y)} & \text{if } t(x, y) > \theta \\
  0 & \text{else}
\end{cases}
\]

(10)

where \(\theta\) is the threshold value of the t-test, which is determined according to the significance level. Performing this t-test before PMI suppresses the second problem. For example, when the significance level is set to 5% (\(\theta = 1.65\)) and \(f(x) = f(y) = f(x, y) = 1\), \(t(x, y)\) becomes \(1 - 1/K < 1.65\). Therefore, the proposed method can let PMI \((x, y) = 0\) by Eq. (10).

### 4. Experiments

The purpose of the experiments is to evaluate how accurately the proposed spoken document confidence measure represents the recognition rates. Each spoken document in the evaluation set was recognized and given a confidence score by the method described in Section 2. The performance was evaluated by correlation coefficients between the confidence scores and recognition rates of each spoken document, which were calculated for the entire evaluation set.

Experimental conditions and results are described below.

| Table 1: Summary of evaluation task. |
|-----------------|------------------|
| **Size**        | 782 phone calls (61 hours) |
| **Utterance style** | Spontaneous conversation |
| **Speakers**    | 17 males / 31 females |
| **Recording conditions** | 16 kHz / 16 bit |
| **Acoustic model** | Triphone HMMs |
| **Language model** | Word trigram |
| **Vocabulary size** | 59,676 words |
| **ASR decoder** | VoiceRex [13] |

### 4.1. Experimental conditions

Table 1 shows the evaluation task. Each phone call was a simulated call center dialogue. Two speakers, an operator and a customer, talked to each other as in call center dialogues, and the utterances of each speaker were recorded by separate microphones. 782 phone calls (391 operator channels and 391 customer channels) were used as the evaluation set. The lengths of the phone calls ranged from 2 to 17 minutes.

We treat a phone call as a spoken document. The acoustic model was trained by a 224 hour training set, and the language model was trained by a training set including manual transcripts of call center recordings, which consisted of 1 million words. Both training sets differed from the evaluation set. The average recognition rate (character correctness) of the evaluation set was 79.56%, the minimum was 45.60% and the maximum was 92.46%.

The window length, \(N\), and the window shift amount, \(M\), in the windowing procedure described in Section 2 were optimized on a development set consisting of 212 phone calls, and were, as a result, fixed to \(N = 20\) and \(M = 10\), respectively. The development set differed from both the training and evaluation set. Nouns and verbs in the recognition vocabulary were used as content words. The word occurrence/co-occurrence frequencies used for calculating PMI were counted on the 113,079 bags of words \((K = 113079)\), which were extracted from the training set of the language model by the windowing procedure.

The following 3 conditions of PMI smoothing were compared; “no smoothing”: PMI was not smoothed at all (when \(f(x, y) = 0\), PMI was large, negative, and constant), “conventional”: PMI was smoothed by the method described in [6], “proposed”: PMI was smoothed by the proposed method described in Section 3.2. In the “conventional” condition, smoothing parameters \(I\) and \(\alpha\) (in Eq. (6) and Eq. (7)) were optimized on the development set and were, as a result, fixed to \(I = 0.10\) and \(\alpha = 0.15\), respectively. In the “proposed” condition, the significance level of the t-test was set to 5% (\(\theta = 1.65\)).

We also investigated the combination of the proposed “context-based” confidence measure and the confidence measure based on acoustic and language scores output from the ASR decoder. The acoustic log likelihood ratio between the N-best hypotheses [12], which was normalized according to the length of the spoken document, was used as the acoustic score. The averaged log 3-gram probabilities was used as the language score. The acoustic and language scores were averaged into the “decoder-based” confidence score.

The context-based and decoder-based confidence measures (ContextCM and DecoderCM) were combined by linear interpolation, \((1 - \lambda) \cdot \text{ContextCM} + \lambda \cdot \text{DecoderCM}\), in the “context-decoder” condition. The weight \(\lambda\) of the interpolation was optimized on the development set and was fixed to \(\lambda = 0.2\).
4.2. Results

Table 2 shows the correlation coefficients between recognition rates and spoken document confidence scores calculated under each condition described in the previous section. It is considered that the improvement from “no smoothing” to “conventional” represents the effect of suppressing the first problem described in Section 3.1, and the improvement from “conventional” to “proposed” is the effect of suppressing the second problem.

The highest correlation coefficient in all conditions, 0.710, was achieved in the “context+decoder” condition with “proposed” PMI smoothing. It was confirmed that the proposed PMI smoothing method was effective also in the “context+decoder” condition.

The combination of context-based and decoder-based confidence measures offered a further improvement in each condition. The ASR decoder uses short-range linguistic information, i.e., the word trigram. In contrast, our context-based confidence estimation uses longer-range linguistic information over more words (20 content words in this experiment), but does not use word order information which the word trigram contains. Moreover, the acoustic information obtained from the ASR decoder is independent of the linguistic information. Therefore, they complement each other in estimating the spoken document confidence.

The scatter plot of all spoken documents in the best condition is illustrated in Fig. 2. It is confirmed that the points have a linear arrangement and our confidence scores can predict the recognition rates of each spoken document even in the evaluation set that includes both operator’s and customer’s speech.

5. Conclusions

In this paper we presented a method that can well estimate the spoken document confidence that represents the recognition rates of each document. The proposed method uses word contextual coherence over several utterances for spoken document confidence estimation. We also proposed a new smoothing method that deals with the two problems of pointwise mutual information that are triggered by data sparseness.

Experiments were conducted to evaluate how accurately our spoken document confidence represented the recognition rate. The results showed that combining the context-based and decoder-based confidence measures is very effective in estimating the spoken document confidence. It is considered that the acoustic information, the short-range word order information (word trigram) and the long-range linguistic information (contextual coherence) complement each other. Moreover, experimental results showed that our PMI smoothing method is more effective than the conventional technique.

6. References