Using machine learning method and subword unit representations for spoken document categorization

Qu Weidong      Katsuhiko Shirai
Department of Information and Computer Science
Waseda University, Tokyo, Japan
E-mail: qwd@shirai.info.waseda.ac.jp

ABSTRACT

In this paper, we investigate the feasibility of using machine learning method and subword units for spoken document categorization as an alternative to using words generated by word recognition or keyword spotting. An advantage of using subword acoustic unit representations to spoken document categorization is that it does not require prior knowledge about the contents of the spoken document and could attack the out of vocabulary (OOV) problem. The context-sensitive learning method is efficient on large, noisy corpora and very suitable for subword-based categorization. Given that even the best phone recognizers make a large number of mistakes, to improve phone N-gram recall, we can once again use phone lattices to obtain the bag of phone N-grams for each speech document. In this study, we examine a variety of subword unit categorization terms and measure their ability to perform effective categorization work, and also have investigated the performance when the underlying phonetic transcriptions contain different recognition errors.

1. INTRODUCTION

As the amount of accessible data continues to grow, the need for automatic methods to process, organize, and analyze this data and present it in human usable form has become more and more important. Traditionally, much work has been done on the text media, only recently there has been a considerable focus on other media such as image, video, audio, and speech. Given that speech is an attractive media and more expressive means of communication and present it in human usable form has became more and more important. Document categorization is an active research topic in the area of information retrieval, and experiments clearly indicate that automatic categorization can improve the retrieval performance compare with no categorization. The goal of the document categorization is to assign entries from a set of prespecified categories to a document. The processing can be done in batch mode on a collection or on individual documents as they arrive in a data stream. There is usually labeled data that can be used to train models to recognize the different classes [4].

One method to do the categorization work is to perform keyword spotting on the spoken documents to obtain a representation in terms of a set of keywords and then using these representations to perform the document categorization [1, 2, 12, 13]. In order to perform the keyword spotting work, the set of keywords need to be chosen a priori. This requires advanced knowledge about the content of the speech documents.

Another method is using a combination of automatic speech recognition and word based text categorization techniques to do this work. In performing spoken document categorization a speech recognition engine is applied to an audio input stream and generates a textual representation (transcription) of the speech, the word level transcription is then being categorized by using traditional text process techniques. In this method there is a major issue that the spoken word is not in the recognizer's vocabulary, thus could never be recognized. To attack the OOV problem new words need to be added, but to handle new words from growing and diverse message collection by adding the new words seems not easy due to the practical computational limit. On the other hand, sometimes the high word-recall (recognizing many of the spoken words) although comes at a cost of low word precision (recognizing many words that were not actually spoken) maybe contributes to high signal in the spoken document categorization. The combination method can not provide high word-recall because of OOV problem.

An attack on the above problems is to perform spoken document categorization on sub-word acoustic units [5, 6, and 7]. The advantages are that the recognizer is relatively independent of the vocabulary in the document and the high word-recall can be achieved [5, 6, 8]. Given that even the best phone recognizers make a large number of mistakes, to improve phone N-gram recall, we can once again use phone lattices to obtain the bag of phone N-grams for each speech document [8]. Once the recognizer outputs a phone lattice, we can use all possible sub-word units as representation units for categorization work. Since this higher word recall processing will lead to many noises in the phonetic transcriptions. A new issue is how to reduce the influence of these noises when we perform the categorization work. If we have a categorization algorithms which would benefit from a higher word-recall and are robust against poor word-precision, then the effectiveness of this kind of algorithms will give us hope that subword method can work better with spoken document for categorization. The context-sensitive learning method is efficient on large, noisy corpora and very suitable for work [3].

This paper investigates the feasibility of using machine learning method and subword units for spoken document categorization. In this study, we examine a variety of subword unit categorization terms and measure their ability to perform
effective categorization work, and also have investigated the performance when the underlying phonetic transcriptions contain different recognition errors. The outline of this paper is as follows. In section 2 we describe the subword unit representations. In section 3 we describe a context sensitive learning method for document categorization. In section 4 we present our experimental results. Finally, in section 5 we close with some conclusions and suggestions for future work.

2. SUBWORD UNIT REPRESENTATIONS

The subword units in our work are overlapping, fixed length, phonetic sequences. Of course, we can also use semi-syllables, syllables, or sequences of these as subword units. But in our work, we just examined phonetic units. Since it is difficult to obtain word and sentence boundary information from phone transcriptions, all subword units are generated by treating each story as one long phone sequence with no boundary information [5]. In our categorization work, we examined the different range of subword units from n=1 to n=6. Examples of n=1, n=2 and n=3 phone sequence subword units for the phrase “green coffee” are given in following Table 1.

<table>
<thead>
<tr>
<th>Subword Unit</th>
<th>Representation Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Green coffee</td>
</tr>
<tr>
<td>Phone (n=2)</td>
<td>/G4A/G42/G55/G4C/G5C/G51/G4E/G52/G4B/G49/G4C/G5C</td>
</tr>
<tr>
<td>Phone (n=3)</td>
<td>/G4A/G42/G55/G4C/G5C/G51/G42/G4E/G52/G4B/G49/G4C/G5C</td>
</tr>
</tbody>
</table>

Table 1: Examples of n-phone sub-word unit representations

3. CONTEXT-SENSITIVE LEARNING METHOD

3.1 Sleeping Experts Algorithms

The context-sensitive learning method we apply in this domain is sleeping experts algorithm [3]. Sleeping experts method is an inductive learning method which using n-gram phrase in classification. This method has been developed within the computational learning community over the last several years and recently applied to the text categorization work by Cohen [3]. Sleeping experts is based on a framework for combining the “advice” of difficult “experts”. In our work each subword units is kind of expert. Prediction algorithms in this framework are given a pool of fixed “experts” and build a master algorithm, which combines the opinions of the experts in some manner. Typically, the master algorithm classifies an example by using a weighted allocation algorithms are on-line algorithms.

In the context of document classification, an expert can be any lexical unit. In our work we treat each subword unit as an expert. Such an expert is “awake” and predicts if the subword unit appears in given transcriptions. In our study, we chose the pool of possible experts to be the set of all subword units from bigram to n-gram that appear in documents. In the sleeping experts framework the weight associated with each n-gram is “learned” in an on-line manner so as to minimize the classification error. The pseudo-code for the algorithm is show in following Fig. 1.

Parameters: \( \beta \in (0,1) \), \( \theta \), number of labeled Documents \( T \)

Initialize: \( \text{Pool} \leftarrow \emptyset \)

Do for \( t = 1, 2, \ldots , T \)

1. Receive a new phonetic document transcription and its classification, \( \omega_1^t \omega_2^t \ldots \omega_l^t \) and \( c^t \)

2. Define the set of active n-grams:

\[
W^t = \{ \omega \in \omega_1^t \omega_2^t \ldots \omega_l^t \mid 1 \leq i_t < i_{t+1} < \ldots < i_1 \leq l, i_t - i_{t-1} \leq n \}
\]

3. Define the weights of new mini-experts:

\[
E^t = \{ \omega_k \mid \omega \in W^t, k \in \{0,1\} \}
\]

4. Initialize the weights of new mini-experts:

\[
\forall \omega_k \in E^t \quad \text{s.t.} \quad \omega_k \notin \text{Pool} : p_{a_k}^t = 1
\]

5. Classify the document as positive if

\[
y = \frac{\sum_{\omega \in W} p_{a_k}^t}{\sum_{k=0,1} p_{a_k}^t} \geq \theta_c
\]

6. Update weights:

\[
l(\omega_k) = \begin{cases} 0 & c^t = k \\ 1 & c^t \neq k \end{cases} \Rightarrow p_{a_k}^{t+1} = p_{a_k}^t \beta^{l(\omega)} = \begin{cases} p_{a_k}^t & c^t = k \\ \beta p_{a_k}^t & c^t \neq k \end{cases}
\]

7. Renormalize weights:

\[
(a) \quad Z_i = \sum_{a_k \neq k} p_{a_k}^t \\
(b) \quad Z_{i+1} = \sum_{a_k \neq k} p_{a_k}^{t+1} \\
(c) \quad p_{a_k}^{t+1} = \frac{Z_i}{Z_{i+1}} p_{a_k}^t
\]

8. Update:

\[
\text{Pool} \leftarrow \text{Pool} \cup E^t
\]

Figure 1: The sleeping experts for n-grams algorithm
3.2 Performance Measure

The category assignments of a binary classifier can be evaluated using four contingency table values for each category:

- \( a \) = number of class members put in class
- \( b \) = number of non-class members put in class
- \( c \) = number of class members not put in class
- \( d \) = number of non-class members not put in class

Several effectiveness measures can be defined in terms of these values, for example [4, 15]:

- recall (R) = \( a / (a + c) \)
- precision (P) = \( a / (a + b) \)

To evaluate the performance of our method, we used the \( F_\beta \) measure [4, 15]; a weighted combination of recall and precision that can be defined in terms of the contingency table values:

\[
F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} = \frac{(\beta^2 + 1)a}{(\beta^2 + 1)a + b + \beta^2 c}
\]

We use \( F_1 \) with \( \beta=1 \), \( F_1 = 2a/(2a+b+c) \) if \( a, b, \) and \( c \) are 0, we define \( F_1 \) to be 1. Since there are 6 different categories in data set, we use \textit{micro-averaging} score method to evaluate the performance of our work. In micro averaging, the total number of \( a, b, c, \) and \( d \) across all categories is computed, and these totals are used to compute recall, precision and then \( F_1 \).

4. EXPERIMENTS AND RESULTS

4.1 Data Set

There are 476 news stories in our data set among 6 different categories. The 476 stories have been split into two sets, the training set and test set. The former has 331 stories and the later has 145 stories.

4.2 Baseline Performance

In order to have a baseline performance to compare with, we do the categorization work on word-level transcriptions (Word). We hope that we can compare the subword unit representations method with word (text) unit representation method under the hypothesis that performance on perfect word transcriptions (text document) can be the upper bound, and to examine if the subword units method is good enough to perform the spoken document categorization. Since sleeping experts method can also be worked on text document categorization, we also use different length word n-gram experts to get an optimal performance on word unit transcriptions.

4.3 Subword Units Performance

To examine the feasibility of different subword units in spoken document categorization, we do the phone expansions of words in news stories via a pronunciation dictionary. These kind of perfect transcriptions give us an upper bound on the performance of different subword units and also eliminate the side effect to the effectiveness of subword range when we examine the range effectiveness of different subword unit representations [5]. Performance of the different subword units using micro-averaging score is shown in Figure 2. In this experiment, we can see that when the length \( n > 2 \), the performance are almost same as to the baseline performance, even on \( n = 3 \), the performance of subword units is better than word based method. We conclude these to overlapping subword units provide more experts than the word based n-grams and some other information on phonetic level maybe contributes to the category work.

![Figure 2: Performance of different subword units with perfect underlying phonetic transcriptions](image)

4.4 Performance on Different Recognition Errors

Given that even the best phone recognizers make a large number of mistakes, we must examine the sensitivity of the subword units to erroneous phonetic transcriptions especially under different error rate. In this experiment, we simulate the phonetic recognizer to generate different error rate based on phonetic recognition error statistics derived by TIMIT corpus [9, 6]. In order to get different error rate, we use the phonetic recognition error confusion matrix to simulate the different error rate [6]. The performance of categorization under different error rate is shown in Figure 3. From this experiment we can see when the error rate is below 20%, the performance is comparable to baseline method. Since in real condition the recognition performances are often poorer than 20%, we should use some techniques to improve the performance of subword method.

![Figure 3: Performance of subword units with different erroneous phonetic transcriptions](image)
4.3 Performance on High Word Recall

Sleeping experts method is efficient on large, noisy corpora. Since if we use perfect phonetic transcriptions to train our classifier and do the categorization on errorful test transcriptions, the erroneous subword unit experts maybe have less negative effect on the prediction owing to their "sleeping state" during categorization. By using phone lattices to obtain the bag of phone N-grams for each speech document, we can obtain more "experts" than word base method and these higher word-recall may contributes to high signal in the text. By reason of these we believe that sleeping experts method can be robust against poor word precision. In experiments we simulated a high phone-recall and poor phone precision system and examine the performance under this condition. Figure 4 shows the performance of the high recall transcriptions. Experiments suggest that having higher word recall does improve the performance clearly.

![Figure 4: Performance of subword units with different erroneous phonetic transcriptions using high recall method](image)

5. CONCLUSION AND FUTURE WORK

In this paper, we have explored the use of machine learning method and subword units for spoken document categorization as an alternative to using words generated by word recognition or keyword spotting. We examine the performance of the range of subword units, and find that when n = 3, or n = 4 can achieve better performance. We also have investigated the performance when the underlying phonetic transcriptions contain different recognition errors. The experiments showed that when use phone lattices to obtain the bag of phone N-grams for each speech document, the performances are comparable to word based method under the phonetic error rate below 30% even compare with perfect transcriptions. For future research we will explore the performance of this method when the training transcriptions contain error phone. Since a phone recognizer who output error phone keeps regularly, the regular phonetic error may have less effect on performance due to the classifiers learned and remember these kind of error units. We will also study other approaches to improve the categorization performance and compare the performance with errorful word level transcriptions.

6. REFERENCES