Elimination of Person Names in Spoken Documents for Privacy Protection

Ryo Kawaguchi*, Masatoshi Tsuchiya†, Seiichi Nakagawa*
*Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi, Aichi, Japan
E-mail: {kawaguchi, nakagawa}@slp.cs.tut.ac.jp
†Information and Media Center, Toyohashi University of Technology, Toyohashi, Aichi, Japan
E-mail: tsuchiya@slp.cs.tut.ac.jp

Abstract—There is an increasing use of sensor networks capable of sensing multimedia data including audio data. Unfortunately, public use of these is not allowed because they contain crucial privacy information such as person and location names.

Person name extraction (PNE), which is a widely investigated research topic, is an effective technique to resolve this problem. However, there is an important difference between traditional PNE and PNE for privacy protection: traditional PNE often misses out-of-vocabulary (OOV) person names that do not occur in a training corpus, and PNE for privacy protection must cover OOV person names because of the demand for privacy protection.

To resolve the issue of PNE for privacy protection, this study proposes a method consisting of two stages: the first stage is speech recognition using a language model modified to over-extract person names including OOV person names, and the second stage is filtering over-extracted person names using an SVM (Support Vector Machine). The experiments show that our method is effective in detecting / eliminating person names, and listening tests also show that the performance of our method in removing person names is promising.

I. INTRODUCTION

For person names detection from text, there are many comparative researches. For example, Tsuchiya et al. compared CRF and SVM for the performance of named entity extraction[1][2]. Iwakura proposed a method using rules acquired from unlabeled data[3]. Konkol compared the classification methods of Decision Tree, SVM, Maximum Entropy and CRF (Conditional Random Fields) for Czech, Spanish, Dutch and English, respectively. They obtained F values of about 85% for person name detection based on CRF[4].

Recently, video and audio data obtained as observation in newspapers[8]. The extremely large vocabulary speech recognition method for the detection of named entities, enlarging the vocabulary size of names in a language model, has been proposed[9][10][11]. A method using contextual feature was also proposed[12][13]. Senay et al. proposed a method of using the Poneme Confusion Networks for the extraction of the person name, and rescoring by Latent Dirichlet Allocation[14]. In addition, they also proposed a method for rescoring by context model[15]. Further, the methods for coping with the OOV word are also proposed[16][17][18].

In this study, for privacy protection we proposed a person name detection / elimination method from a spoken document, using a language model that emphasized the probability of the occurrence of person names and filtering of person name candidates using SVM. Our aim is to detect / eliminate person names in spoken documents with the high recall rate for privacy protection. We conducted experiments using the corpus of broadcast news to confirm the usefulness of the proposed method.

II. PNE FOR PRIVACY PROTECTION

A. PNE based on speech recognition

Speech recognition is formulized by the following equation, which gives a word sequence \( W = \omega_1^M \) that maximizes the probability \( P(W|O) \) for a feature vector sequence \( O = o_1^T \).

\[
W = \arg\max_W P(W|O) \quad (1)
\]

\[
W = \arg\max_W P(O|W)P(W) \quad (2)
\]

where \( P(O|W) \) is an acoustic model, and \( P(W) \) is a language model. In the language model \( P(W) \), we commonly use an \( N \)-gram model expressing the occurring probability of a word \( \omega_i \). In the \( N \)-gram model, the occurring probability of word sequence \( W \) is determined by the following equation:

\[
P(W) = \prod_{i=1}^{M} P(\omega_i|\omega_{i-N+1}^{i-1}) \quad (3)
\]

When detecting / eliminating a person name using speech recognition formulation, the method, in which “the person name appears in position \( i \) when a word \( \omega_i \) belongs to the person name class \( C_p \)”, is the simplest. However there are two problems in this method. The first is that a speech recognition error occurs depending on the position in which the person name occurs. The second is that it cannot detect OOV person names that are not included in the prepared person name class \( C_p \).

Failure to detect is a serious problem, because we need to develop a person name detection method for the purpose of privacy protection. In other words, we want to obtain a higher recall rate, although we acknowledge the low precision rate.

B. LM for PNE for privacy protection

We propose a method that emphasizes the probability of the occurrence of the person name in comparison with the
probability of occurrence of the other words to reduce the case that the recognition procedure produces a non-person word at the position where the person name appears. Thus, person names are likely to be detected (a higher recall rate). On the other hand, false alarm for person names increases (a lower precision rate). When a word $w_i$ belongs to the person name class $C_p$, the occurrence probability of the word $w_i$ is modified by the following equation:

$$p_n(w_i|w_{i-N+1}^{i-1}) = \begin{cases} \alpha \cdot P(w_i|w_{i-N+1}^{i-1}) & \text{if } w_i \in C_p \\ P(w_i|w_{i-N+1}^{i-1}) & \text{otherwise} \end{cases}$$

(4)

where $\alpha \gg 1.0$.

### C. Expansion of a PN and class language model

As a method for detecting of a person name, since OOV is not included in the vocabulary of the speech recognition language model, we adopt the class language model of a person name and expand the registration words in the person name class. We regard the mis-recognized person name as a correct detection even if the recognition result is another person name.

Instead of a full person name, we present a person name using a combination of the given name and surname to cover the registration. If an acoustically similar person name exists for the OOV person name in the class, the OOV name would be classified correctly into a person name. We divide the person class $C_p$ into the given name class $C_s$ and surname class $C_g$. We call this model a given name and surname class model.

$$p_s(w_i|w_{i-N+1}^{i-1}) = \begin{cases} P(C_s|w_{i-N+1}^{i-1})P(w_i|C_s) & \text{if } w_i \in C_s \\ P(C_g|w_{i-N+1}^{i-1})P(w_i|C_g) & \text{if } w_i \in C_g \\ P(w_i|w_{i-N+1}^{i-1}) & \text{otherwise} \end{cases}$$

(5)

Using the person name dictionary which we prepared from the name corpus, we add the frequently occurring names to a given name class $C_s$ and surname class $C_g$, respectively. Thus we perform speech recognition using the expanded given name and surname class language model and perform person name detection based on the speech recognition result. The probability $P(w_i|C_s)$, the actual occurring probability in the given name class $C_s$, and the probability $P(w_i|C_g)$, the actual occurring probability in the surname class $C_g$, are both set to a uniform distribution probability. This is because we just detect that it is person name. However, we do not change the occurrence probability of a given name and surname class, even if the number of added words increases.

Thus, even if a person name for the recognition is mis-recognized, the possibility of being mis-recognized as an acoustically similar person name increases, that is, recall rate increases.

### D. Filtering using SVM

In addition to the above two methods, we considered a method for improving person name classification that uses an SVM for the recognition result to improve precision rate in post-processing.

### TABLE II: Acoustic Analysis Condition

<table>
<thead>
<tr>
<th>Sampling frequency</th>
<th>16kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preemphasis</td>
<td>0dB</td>
</tr>
<tr>
<td>Analysis window</td>
<td>Hamming window</td>
</tr>
<tr>
<td>Analysis window length</td>
<td>25ms</td>
</tr>
<tr>
<td>Analysis frame shift</td>
<td>10ms</td>
</tr>
<tr>
<td>Feature Parameters</td>
<td>MFCC + ΔMFCC + ΔΔMFCC + ΔPow + ΔΔPow (38 dimensions)</td>
</tr>
</tbody>
</table>

As general features of the named entity detection, the target word itself and 2 words before and after the target word, surface and character type, and part of the speech have been used[19]. However, we considered the following in the speech recognition result, that is, the surface and character types are not accurate information because of recognition errors. In addition, we decided to use only a part of the speech for the features because adequate training data were not prepared. For part of the speech, 68 types were used for subclassification.

For the training data of the SVM, although there are two ways to obtain transcription such as manual transcription and speech recognition results, in this study we used as training data the speech recognition results with manually annotated labels.

On the other hand, CRF is often used in named entity detection tasks[20], but CRF is a method of solving a sequence labeling problem. We want to judge if a person name detected by a speech recognition result is correct or not. We compare SVM with CRF.

### III. Experiment

#### A. Experimental setup

In the evaluation experiment, we experimented using the NHK (Nippon Housou Kyoukai) corpus. The NHK corpus is a set of transcriptions of 30 broadcast news programs that were broadcasted from June 1st to 12th 1996. The NHK corpus has annotated person names. As a training corpus for the SVM, we removed the data that we used as a test corpus. We show the details of the corpus in Table I. For a decoder in speech recognition, we used the in-house large vocabulary continuous speech recognition system, SPOJUS++ (SPOken Japanese Understanding System)[21], and its acoustic analysis condition is shown in Table II. The Japanese context-dependent syllable-based acoustic models with 8 left contexts (in total 928 models) were trained from academic presentation speech data in the CSJ (Corpus of Spontaneous Japanese)[22]. Each continuous density HMM (Hidden Markov Mode) has 5 states, and 4 have pdf files of output probability. Each pdf file consisted of 4 Gaussians with full covariance matrices. The morphological analysis was performed using ipadic-2.7.0 and Chasen\(^1\) for the training corpus in Table I, to obtain a 3-gram word-based language model trained using Palmkit\(^2\). YamCha\(^3\) was used for training the SVM, and we used 2nd degree of polynomial kernel as SVM kernel. The evaluation experiment was intended to evaluate the correct rate for person name detection. Even if a certain person name is replaced with another person name, when it is recognized as one of person names, we judge it was detected correctly.

#### B. Objective experiment

Fig.1 shows the experimental result using the method described in Section II-B which emphasizes only the occurrence

\(^1\)http://chasen.naist.jp/
\(^2\)http://palmkit.sourceforge.net/
\(^3\)http://chasen.org/~taku/software/yamcha/
TABLE I: NHK broadcast news corpus training corpus of LM

<table>
<thead>
<tr>
<th>date</th>
<th>June 13, 1996<del>July 17, April 13</del>July 13, 1996</th>
</tr>
</thead>
<tbody>
<tr>
<td>program name</td>
<td>Ohayo NipponNews7</td>
</tr>
<tr>
<td># words</td>
<td>1564848</td>
</tr>
<tr>
<td># type of words</td>
<td>32975</td>
</tr>
<tr>
<td># person names</td>
<td>21044</td>
</tr>
<tr>
<td># type of person name</td>
<td>2637</td>
</tr>
</tbody>
</table>

Training corpus of SVM

<table>
<thead>
<tr>
<th>date</th>
<th>June 1, 1996~June 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>program name</td>
<td>Ohayo NipponNews7</td>
</tr>
<tr>
<td># words</td>
<td>64302</td>
</tr>
<tr>
<td># type of words</td>
<td>14582</td>
</tr>
<tr>
<td># person names</td>
<td>1231</td>
</tr>
<tr>
<td># type of person name</td>
<td>1231</td>
</tr>
</tbody>
</table>

Test corpus

<table>
<thead>
<tr>
<th>date</th>
<th>June 1, 2, 4, 5, 1996</th>
</tr>
</thead>
<tbody>
<tr>
<td>program name</td>
<td>News7</td>
</tr>
<tr>
<td># words</td>
<td>6620</td>
</tr>
<tr>
<td># type of words</td>
<td>1527</td>
</tr>
<tr>
<td># person names</td>
<td>67 (IV 40, OOV 27)</td>
</tr>
<tr>
<td># type of person name</td>
<td>34 (IV 20, OOV 14)</td>
</tr>
</tbody>
</table>

Fig. 1: Emphasis of the occurrence probability of person name probability of the person name. The vertical axis represents the recall and precision rates, and the horizontal axis represents the parameter \( \alpha \). The leftmost bars of Fig.1 which represent the recall and precision when \( \alpha = 1 \) is the baseline result which means the case for the occurrence probability of a person name is not emphasized. The baseline method achieved 80% recall and 89% precision against In-Vocabulary (IV) person names as shown in the left sub-figure of Fig.1, but 14% and 4% for OOV person names, respectively. On the contrary, our method when \( \alpha = 50 \) achieved 48% recall and 10% precision against OOV person names.

We added surname and given names appeared with a high frequency for the class language model as described in Section II-C. The added numbers of given names and surnames were 100 words, 200 words, 500 words and 1000 words. The person name dictionary used was a DCS-person dictionary of Nichigai Associates\(^4\). Fig.2 shows the results when 500 given names and 500 surnames were added. In this evaluation experiment, rather than using 1000 words, the result was better when we added only the top 500 words to the class. If we register person names too frequently, a similar word to non-person names is registered acoustically, which causes recognition errors. We could improve the recall and precision by introducing a class LM and adding registered names.

C. Evaluation of filtering using SVM

We conducted the person name verification experiment using SVM for the speech recognition results described in Section III using the same language model as Fig.2. We show the experimental result for IV person name detection in Table III(a) and the OOV person name in Table III(b), respectively. The result was for the recognition result as described in Section II-B, which we increased the occurrence probability of the given name and surname class 50 times. Using SVM, we exercised a judgment on whether a person name detection in the speech recognition result was correct or not. Moreover, we experimented using CRF for comparison. For the OOV person name, although the recall rate decreased a little, the precision was improved by about 10%, from 25.4% to 33.7% for IV and from 12.7% to 22.0% for OOV after using SVM based post-processing.

IV. LISTENING EXPERIMENT

We conducted the person name detection / elimination method taking into consideration in particular the recall rate; we obtained about 90% recall and 25% precision with respect to IV person names, and about 85% recall and 13% precision rate with respect to OOV person names. Improvement in the precision was necessary because the precision was still poor, while preserving the high recall rate (it is desirable that the recall rate is 100%). Therefore we actually removed the all person names detected from throughout the speech document and conduct a subject experiment to measure the degree of understanding of the content of the speech document with speech removed.

---

\(^4\)http://www.nichigai.co.jp/cgi-bin/nga_search_dcs.cgi?ID=J0588
Through the listening test, we established how much precision was necessary and how many random errors could be permitted for the person name detection. A part of a person name was transposed to a silent section, and inserted by a beep sound.

We created five test contents from the test corpus of the news speech document. As a test of the degree of understanding, we prepared the contents and questions that we could answer correctly, even if proper nouns were perfectly deleted. For each , we prepared the following three types of material.

i) correct person names: deleted only the correct person names (recall = 100%, precision = 100%)

ii) speech recognition result: deleted the person names that were detected in speech recognition including false alarm (recall = 100%, precision = 20%)

iii) less false alarm: deleted the person names that were detected in speech recognition including fewer false alarms (recall = 100%, precision = 45%)

The five different contents were selected so that the precision of the person name detection using speech recognition was about 20% for both for the IV person name and OOV person name, respectively (case ii). In addition, to investigate the influence of the precision rate, we also prepared content in which the precision was increased artificially up to 40~50% (case iii).

There were 15 subjects. Furthermore, apart from the test for degree of understanding, the subjective evaluation was also conducted. Each subject listened only to the content of one of the three ways for every content. Fig.1 shows the experimental results. For the experimental results, the rates of understanding were i) 92%, ii)72% and iii)88%, respectively.

We could conclude when the precision was improved to about 40~50%, we correctly understood the content of the speech documents from the view points of the results of objective / subjective evaluation.

V. CONCLUSION

In this study, we proposed a name detection / elimination method in spoken documents, which enhanced the occurrence probabilities of person names in an N-gram language model and used a name-class language model with the registration of many names for dealing with OOV names. We also adopted a post-processing method provided by our proposed method using SVM. It improved the precision of name detection for the speech recognition result. For OOV person names, we obtained a recall rate of 74% from 14%, and a precision rate of 22% from 4%, respectively. For the overall situation including IV person names, the recall rate and precision rate were 85% and 27%, respectively. Finally, it was established that we could understand the content at a rate of 90% if the precision was improved to 40~50%. In future work, when performing SVM, we will increase the training data and determine better features to improve the precision rate.

REFERENCES


