Language Modeling of Chinese Personal Names Based on Character Units for Continuous Chinese Speech Recognition

Xinhui Hu[1,2], Hirofumi Yamamoto[1,2,3], Genichiro Kikui[2], Yoshinori Sagisaka[1,2,3]

[1]National Institute of Information and Communications Technology,
[2]ATR Spoken Language Translation Research Laboratories,
Hikaridai 2-2-2, Seika-cho, Soraku-gun, Kyoto 619-0228 Japan
{xinhui.hu, hirofumi.yamamoto, genichio.kikui, yoshinori.sagisaka}atr.jp
[3]Global Information and Telecommunication Institute, Waseda University
Nishi-Waseda 1-3-1, Shinjuku-ku, Tokyo, 169-0051, Japan

Abstract

In this paper, we analyze Chinese personal names to model their statistical phonotactic characteristics for continuous Chinese speech recognition. The analysis showed language-specific characteristics of Chinese personal names and strongly suggested the advantage of character-unit oriented modeling. A hierarchical language model was composed by reflecting statistical phonotactic characteristics of Chinese personal names as a lower intra-word model, and ordinary inter-word neighboring characteristics as an upper multi-class composite N-gram model. These two layers of models were trained independently using different language corpora. For the modeling of given names, the syllable without tone information was selected as the unit for training the bi-gram. The properties of either one or two characters of a given name were introduced to simplify the length constraint of the modeling process. For Chinese family names, we simply added them directly in the recognition lexicon, since their numbers are very restricted. The results from Chinese speech recognition experiments revealed that the proposed hierarchical language model greatly improved the identification accuracy of the Chinese given names compared with the conventional word-class N-gram model.

Index Terms: Personal Name Identification, Hierarchical language model, Chinese Speech Recognition

1. Introduction

It is well known that the paradigm of large vocabulary continuous speech recognition based on stochastic methods can only recognize those words that exist in the recognition lexicon. This means that if a word is to be recognized, it must be registered in the lexicon in advance. However, in real applications, it is unrealistic to register every word in the lexicon, especially proper nouns such as personal names, local names, organization names etc. In the case of conversation, information on such kinds of proper nouns is crucial for understanding. Therefore, if these words cannot be identified, the progress of a conversation will collapse.

To cope with this problem, a hierarchical structure language model has been proposed for Japanese personal names (family and given names) and city names [1]. It consists of conventional word-class N-grams and a set of the independently trained, class-specific sub-word N-grams. A class-specific sub-words, an OOV intra-word, is divided into phones (morae) by which the word's pronunciation is determined. This phone sequence is modeled by probabilistic finite-state automation (PFSA). Good performance has been obtained for identifying Japanese family names, given names and city names, and that any OOV word of one class is never misrecognized as being from other classes.

In this paper, we adopt this hierarchical language model to the Chinese speech recognition system. As Chinese word units are generally smaller than those of Japanese, even recognition of Chinese personal names of the "Han nationality (汉族)" are expected to be tougher than for Japanese. In Section 2, based on the statistics of our name corpus, we analyze the characteristics of modern Chinese personal names, and show the advantage of character unit based modeling for speech recognition. In Section 3, the structure of hierarchical language model is explained, and the experimental conditions and results are shown in Section 4. Finally, in Section 5 we summarize the advantages of the proposed model and offer concluding remarks.

2. Characteristics of Chinese personal names

Modern Chinese comprise a family name, or surname (姓), which is always placed first, and a one- or two-character given name (名). A family name can be one of the thousands of name sets that have been historically used by the Han nationality. In fact, more than 3,500 family names are reported as being used in modern China [2]. In this study, we gathered a name corpus from the Beijing district. Of the 560,000 names within it, there are about 1,900 different family names in this corpus.

Although Chinese family names consist of one or two characters, the single-character version is predominant. According to [3], more than 99.5% of the names in the two corpora, which are from Beijing and Hong Kong, are single-character family names.

In our name corpus mentioned above, only 17 two-character family names are found, such as "呼延，令狐，上官，欧阳，司马，尉迟，宇文，诸葛" etc. According to the statistics on our name corpus, the most common family names

are "王 (6.21%)，张 (5.75%)，李 (5.51%), 刘 (4.55%), 赵 (2.91)，杨 (2.56%)，and 陈(2.13%)."

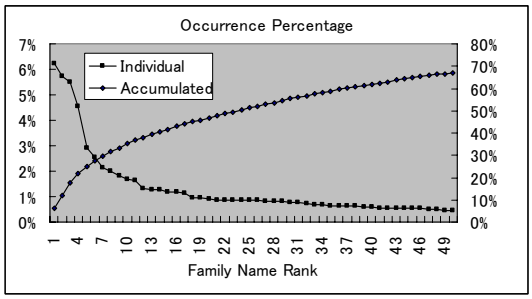

Figure 1: Individual and accumulated occurrences of the top-50 ranked family names

Figure 1 shows the individual and accumulated occurrence ratios for the top 50 family names among this corpus.

Different to given names, new family names are rarely produced, so family names can be regarded as a known fixed set. The total vocabulary of such words is not very big, so it is practical and feasible to add these words to the recognition lexicon.

For Chinese given names, only one-character or two-character names are possible, several characteristics make them difficult to recognize. Much more so than for characters used in family names, which are strictly comprise subset of a Chinese character set that forms common words, most characters in given names can be located freely in the first or second position, or used individually. Of the 567,000-name corpus, 32,400 are one-character given names, and 544,000 are two-character given names. All of these factors introduce a lot of potential ambiguities in identification. The situation becomes even more complicated due to the fact that some polysyllabic common words are also possible in names, e.g. "前进" (Go Forward) and "国庆" (Celebrate the National Day). In these cases, the algorithm may easily mistake names as common words and produce irrecoverable errors.

In this study, we added all the family names from the name corpus, a total of 1,900, to the recognition lexicon, and modeled the given names as an inter-word model of a hierarchical language model.

3. Hierarchical Language Model

The speech recognition problem is given in the following equation:

$$\hat{W} = \arg\max\, p(A \mid W)p(W), \qquad (1)$$

Here, $p(A \mid W)$ is the acoustic model, $p(W)$ is the language model. The acoustic model usually consists of phonemes as its units, and a recognition lexicon is used here to concatenate words and phonemes. In this case, the Eq. (1) can be further changed as shown in the following equation:

$$\hat{W} = \arg\max\, p(A \mid P)p(P \mid W)p(W), \qquad (2)$$

where, P is the phoneme sequence, and $p(A \mid P)$ is the acoustic model based on the phoneme sequences. Here, $p(P \mid W)$ represents the recognition lexicon.

In Eq. (1), $p(P \mid W)p(W)$ can be expanded to the following equation when bi-grams are used.

$$p(P \mid W)p(W) \approx \prod_i p(P_i \mid w_i)p(w_i \mid w_{i-1}) \qquad (3)$$

Here, P_i denotes the phoneme sequence that represents the reading of word w_i. By replacing a word bi-gram with a word–to-word class bi-gram (here, a word class corresponds to a word category in which a given word belongs), we obtain the following equation:

$$\begin{aligned} & p(P_i \mid w_i)p(w_i \mid w_{i-1}) \\ & \quad \approx p(P_i \mid w_i)p(w_i \mid c_i)p(c_i \mid w_{i-1}) \\ & \quad = p(P_i \mid c_i)p(c_i \mid w_{i-1}), \end{aligned} \qquad (4)$$

where the term $p(c_i \mid w_{i-1})$ represents the probability of the transition from word w_{i-1} to word class c_i. The term $p(P_i \mid c_i)$ denotes the constraint for reading a word belonging to the class c_i. Concretely speaking for a Chinese given name, $p(c_i \mid w_{i-1})$ represents the probability from a preceding word w_{i-1} to the class of a given name, and $p(P_i \mid c_i)$ restricts the possible reading of all the given names.

The constraint phone sequences of a word in Eq. (4) can be further expanded using an M-gram,

$$p(P_i \mid c_i) \approx \prod_{j=1}^{I} p(p_j \mid p_{j-M+1}, p_{j-M+2,\ldots} p_{j-1}, c_i), \qquad (5)$$

where I is the phoneme length of the word belonging to the word class.

To give an explicit constraint on the phone length of words, we update the right side of the Eq. (5) using

$$p(L = I)\prod_{j=1}^{I} p(p_j \mid p_{j-M+1}, p_{j-M+2,\ldots} p_{j-1}, c_i) \qquad (6)$$

where $p(L = I)$ represents the probability of a word consisting of I phones.

In [1], instead of syllables, the "mora" is selected as a phonetic unit for word identification since the mora set is efficient for the description of Japanese phonotactics. Compared with Japanese, the syllable is a basic unit for Chinese, and all Chinese characters are read in the unit of syllables. There are about 1,300 tonal syllables, and about 410 toneless syllables in standard Chinese. To efficiently describe Chinese phonotactics, we selected toneless syllables as the phonetic units for the intra-word model in the hierarchical language model. These syllables are represented by their PINYIN strings.

For the reading length in Eq. (6), only I=1 *and* I=2 are limited to suit the fact that in Chinese given names, only one- and two-character names are permitted.

The hierarchical language model was adopted for identifying Chinese given names as illustrated in Figure 2. The example utterance is "..是盖志海 (. . am Gai Zhi-Hai). " The family name "盖" (denoted as class CHFN) is identified at the inter-word level, while the given name (denoted as class CHGN) is modeled in the lower layer. The two-character given name "志海" (Zhi-Hai) is divided into the two individual characters' toneless syllables "Zhi" and "Hai."

To avoid expansion in lexicon size when an intra-word model is used, we introduced an approximation with the help of syllable network shown in Fig. 3. In that figure, syllables and syllable sequences are classified into only three categories: the head, middle, and end positions. The number of new entries was reduced from the number of syllables and syllable sequences multiplied by position variation *I* to syllables and syllable sequences multiplied by three. By considering the specialty of Chinese given names (*I <=2),* the above network can be further simplified by removing the intermediate states ($P_{BI} = P_{IE} = 0$), thus the number of new entries becomes syllables and syllable sequences multiplied by two. A given name can be directly controlled using the following equations:

$$p(L=1) = p_0 \times p_\phi \quad (7)$$

$$p(L=2) = p_0 \times p_{BE} \quad (8).$$

To prevent prohibited combinations of syllables, intersection of FSA and N-grams are used as new constraints. This FSA only accepts reasonable combinations of syllables or syllable sequences and words.

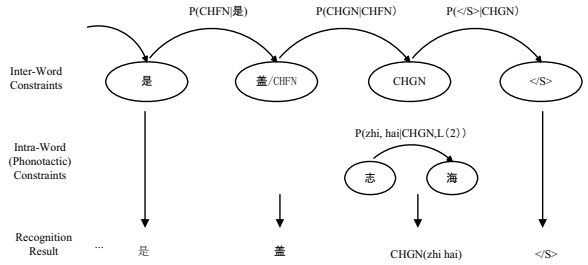

Figure 2: The hierarchical language model with phone length for Chinese name identification.

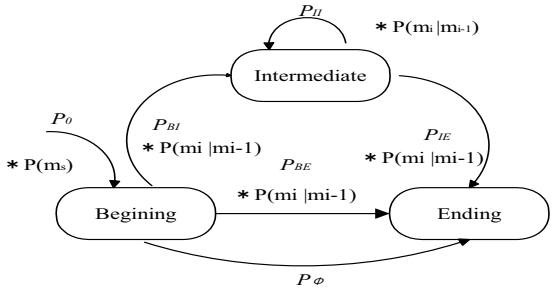

Figure 3: The reduced network for the intra-word model

4. Recognition Experiments

To evaluate the effectiveness of the proposed model, we performed speech recognition experiments for Chinese given names.

The experiments compared two language models using two test data sets. The first language model is the baseline model (denoted as the in-lexicon model) which is the multi-class composite bi-grams [4]. In this model, both all the 1,900 Chinese family names extracted from the name corpus and the 200 given names appearing in one of the test sets are added to the recognition lexicon. These words are also added to the class files, by which the multi-class composition bi-gram is built. The second language is the proposed model, in which syllable networks are used instead of lexical entries. In training the intra-word model, the Chinese given names are moved from the lexicon and corresponding class files.

4.1. Training data and OOV units

The corpora used for training the baseline model and the proposed model are shown in Table 2. Both the baseline model and the inter-word based class N-gram were trained using a travel conversation corpus. The corpus contains 43.7K word entries and a total of 3.5M words. The segmented & POS tagged corpus was annotated manually. The annotations were guided by our own specification, which is constituted by referring to the guidelines of segmentation and POS tagging of Peking University [5]. For the proper nouns, we defined family names and given names for Chinese, Japanese, and Western people, respectively.

As the texts of this corpus are mainly translated from Japanese or English and related to the travel domain, personal names within it were mostly Japanese and Western; Chinese names rarely appeared. Since the naming method is similar among Japanese and Chinese - both are in the order of "family name + given name" - we classified the Japanese and Chinese names into the same group, and let the Chinese names share the class of Japanese ones. Therefore, to slide over the scarcity of Chinese names in the training data, we first trained the baseline model using the original corpus, and then substituted the Japanese family-name class with the Chinese family-name class, and the Japanese given-name class with the Chinese given name-class.

The intra-word bi-gram was trained by using a Chinese name corpus. There were a total of 560K names, containing 1,900 family names, and 4,200 Chinese characters were found in the given names. Compared with the approximately 6,700 common Chinese characters defined in the GB2312 encoding table, a national standard in the mainland China, it was found that most of the Chinese characters were used in given names. By converting the HANZI (Chinese characters) of the given names into pronunciation strings, 83.9K unduplicated tonal PINYIN strings were obtained (from which 1,112 different individual syllables are found), and 44.1K unduplicated toneless PINYIN string combinations were obtained (from which 388 individual syllables were found). This also means that a majority of the total permitted syllables (about 410) were used in the pronunciation of given names.

Two test sets were extracted randomly from the test data of the travel conversation corpus, each containing 200 utterances. One set (denoted as NONAME) did not contain any Japanese or Chinese personal names, but the other set (denoted as NAME) had at least one Japanese name (family name, given name, or both).

Table 2: Training Data for Chinese Personal Names

	Chinese Name	Base corpus
Total Words	560K	3.5M
Num. of Entries	141.8K	43.7K

After the test utterances were selected, we extracted about 1,000 Chinese names randomly from two Chinese corpora, SINICA and the People's Daily, and randomly replaced the Japanese names in the test set with these Chinese names. The rationality of the replaced names was checked manually, such as in the fitness of genders and the order of family names and given names. The obviously irrational replacements were corrected manually and replaced again with other candidates.

These two test sets were recorded in speech by four male speakers and four female speakers from different regions of mainland China. Everyone spoke 25 utterances for each test set, a total of 50 utterances each.

To compare the performance of OOV identification using the in-lexicon model (lexical closed condition) and OOV model, we also added those given names appearing in the test set NAME to the lexicon and in-lexicon class files.

4.2. Experiment Results

The recognition performance was evaluated using two criteria:

- Whole-word Accuracy (Word Acc.): Whole words in utterances, including words registered in lexicon and OOV words (Chinese given names), are evaluated.
- OOV word accuracy (OOV Acc.): Only the Chinese given names in the utterances are evaluated (based on the DP-matching). We further divided the evaluations into two cases. One is the given name's accuracy (Acc.) in the 1-best search, another is the given names accuracy embedded in the recognition lattice (Lat. Acc.). The latter can help us to estimate the maximum possibility in identifying the given names with this method.

The data sets and their word counts are shown in Table 3. The experimental results are shown in Table 4. For the NON-NAME test set, the whole-word accuracy of the proposed model decreased by 1.7% compared with the conventional in-lexicon model. On the other hand, for the NAME test set, the whole-word accuracy of proposed model was nearly 4.5% higher than the in-lexicon model. By investigating the OOV word accuracy, we found that the proposed model's result was 30% higher than the in-lexicon. These facts imply that even without any Chinese given-name words in the lexicon, the proposed method has a great ability to identify the Chinese given names. For many applications of language processing, such as machine translation, these identifications are helpful, since the sentence meaning is correctly kept even if the pronunciation of the personal names is misrecognized. With 96% accuracy for the result in recognition lattice, it is implied that almost all the Chinese given names can be identified with the recognition lattice. So the high accuracy of the Lat. Acc left us a big space to improve the OOV's accuracy. One reason for the large decline in word recognition accuracy from non-names to names might be that the names for testing are randomly selected forcibly from other domains that are different from the baseline training corpus. Some names are in fact not so commonly used, therefore, some utterances are read unnaturally. In these cases, not only the names themselves, but also their neighboring words are influenced in a certain extent. This brings on them easily being recognized incorrectly.

Table 3: Words in Evaluation Data Sets

Test Set	NONAME	NAME	
Total word	1022	1591	
Total name	0	200	200

Table 4: Words Recognition Performance with Chinese Given-Name Hierarchical Language Model (%)

	Word Acc.		OOV Acc.	
	NONAME	NAME	Acc.	Lat.Acc.
In-lexicon	88.4	60.1	13.1	87.5
Proposed	86. 8	64.5	43.5	96.0

5. Conclusions

We proposed a hierarchical language model incorporating class-dependent word models to cope with the identification of Chinese given names in continuous Chinese speech recognition. Although the Chinese given name's identification is one of the difficult tasks in Chinese language and speech processing due to their frequent ambiguities with common words, the proposed model shows its effectiveness to these problems. While having high ability to identify the Chinese given names, the model does not significantly hamper the recognition performance for other non-name words. In addition, its ability to identify the most of given names in the recognition lattice inspires us keep improving the searching algorithm, enhancing both the OOV and whole word's recognition performance further in the future. Moreover, the basic ideas within this model are expected to be applied to other kinds of OOV words in Chinese.

6. References

[1] Koichi Tanigaki, Hirofumi Yamamoto and Yoshinori Sagisaka, "A Hierarchical Language Model Incorporating Class-dependent Word Models for OOV Words Recognition," Proc. ICSLP 2000, pp.123-126, 2000.

[2] http://www.genetics.ac.cn/expert/yuanyida.htm

[3] Lawrence Cheung, Benjamin K. Tsou, Maosong Sun, "Identification of Chinese Personal names in Unrestricted Texts," Proceedings of the 16th Pacific Asia Conference, Korea, pp28-35, 2002.

[4] Hirofumi Yamamoto, Shuntaro Isoga, Yoshinori Sagisaka, "Multi-class Composition N-gram Language Model," Speech Communication, Vol. 41-003, pp. 369-379, Oct. 2003.

[5] Shiwen Yu, Huiming Duan, Xuefeng Zhu, et al, "The Specification of Basic Processing of Contemporary Chinese Corpus," Journal of Chinese Information Processing, Issue5 & 6, 2002.