Multimodal Interface for Organization Name Input Based on Combination of Isolated Word Recognition and Continuous Base-word Recognition

Norihide Kitaoka, Hironori Oshikawa, Seiichi Nakagawa

Department of Information and Computer Sciences
Toyohashi University of Technology, Toyohashi, Japan
kitaoka@slp.ics.tut.ac.jp

Abstract

We investigate a multimodal interface for organization name input to forms in WWW. The user first utters an organization name in an open vocabulary domain to the system. The system recognizes it with a combination method of isolated word recognition and continuous “base-word” recognition. Word candidates and a base-word lattice obtained by this recognition procedure are displayed on a touch panel. Then the user chooses a sequence or base-words from the candidates by pen touch to construct the desired name and the name is sent to the client. This recognition method performs well in the organization name recognition task, in which a very large vocabulary size and its successive update is needed when using isolated word recognition. Our interface design also elevates users’ input ability of organization names.

1. Introduction

Recently, we can use many information retrieval services on the WWW with easy access interfaces based on graphical user interfaces(GUI). However, for some kind of the information retrieval tasks and under certain operating conditions, such an interface may not be so efficient. For example, users with a small laptop PC, a Personal Digital Assistant (PDA) [5, 6] or a wearable computer cannot access such services efficiently or satisfactorily with such interfaces. Speech input is one of promising alternative to the keyboard and the mouse as an input modality. However, since the existing forms on the WWW are not originally intended to be filled in by speech, further investigation will be needed to apply speech to the WWW services as an alternative of the keyboard or mouse operation. So far, some researchers have proposed speech interface systems for the WWW services [1, 2, 3, 4, 5, 6].

We can find a lot of forms which expect inputs of arbitrary character sequences. So it is important to give speech-based input modality to such forms [9]. There are many opportunities to input person’s name, address, or organization name to such forms. In the speech recognition of isolated words, it is not realistic to have all person’s names or organization names in a dictionary because of an open vocabulary domain [7].

On the systems based on English speech recognition, the methods to improve recognition performance allow to accept spelling an utterances [8]. Spelling utterances indeed tend to be recognized better than word utterances, but to utter only the spell of the word is intuitively unnatural. Therefore, combination methods of word utterance and spelling utterance have been investigated [9, 10, 11, 12, 13, 14].

Spelling utterance is almost equivalent to utterance in which the syllables are separated with pauses in Japanese, but such an utterance imposes a burden to the user.

We developed a multimodal interface for person’s name input using speech and pen touch [15]. With this interface, the user first inputs a name and the system recognizes it by isolated word recognition and continuous syllable recognition. And then the system displays the 5-best name candidates, 5-best syllable sequence candidates, and a syllable lattice on a touch panel and the user selects the desired name. With this interface, users could correctly input desired last names and first names at a rate of 93% and 95%, respectively [15].

In this paper, we investigate a multimodal interface for organization name input to forms on the WWW. The vocabulary size of organization names is also large as that of person’s names and the new names are successively generated, which become unknown words, in other words, an open vocabulary domain. Because of these reasons, isolated word recognition using a dictionary with all the organization names is not realistic. On the other hand, organization names are relatively longer than person’s names. So the continuous syllable recognition of such long names is a hard task. Even if the continuous syllable recognition is correctly performed, selection of syllables from the syllable lattice is a great pain for users, especially in a mobile environment. To solve this problem, we propose continuous base-word recognition for organization name recognition. Most of the organization names are composite words and can be expressed as concatenation of short basic words (that is, base-words). Figure 1 shows the coverage of 130,000 organization names by a lexicon of N-frequent whole words and concatenation of N-frequent base-words. In this figure, the frequency of usage of each organization is considered. We can know that the base-words efficiently cover the names. But the continuous base-word recognition must be relatively more difficult than isolated word recognition. So we expect that the best performance will be achieved by the combination of these two recognition methods. The system obtains not only the N-best organization name candidates but also the base-word lattice, so we utilize the base-word lattice to the interface.

2. Recognition of organization name utterances

2.1. Continuous base-word recognition

There are too many names for Japanese organizations, and the number of them gets larger. Thus, it is not realistic to register all the names to the dictionary from the viewpoints of computational cost and the coverage. Most of the organization names are composite words and can be expressed as concatenations of short “base-words.”
So we adopted a continuous base-word recognition to an organization name recognition task. In this recognition, the statistical language knowledge of Japanese organization names can be introduced. Because the statistics of the Japanese organization names have certain characteristics in base-word arrangements, we can expect the effect to apply to a statistical language model (N-gram) of the base-words.

In this research, we made bigram language models of base-words. 130,000 organization names were split into base-words as written in Section 2.3 and the frequencies of the words were considered using the appearance frequencies in a newspaper database.

2.2. Combination of isolated whole word recognition and continuous base-word recognition

In this combination, we treated whole words for isolated word recognition and base-words for continuous base-word recognition equally. Therefore, isolated word recognition and continuous base-word recognition can work simultaneously on one recognizer. Figure 2 shows the concept of the language model. N most frequent names are contained with organization name dictionaries. Connection probability to a whole word $w$ from the start condition, $P(w|\epsilon)$, is obtained from the word frequency, i.e., unigram probability. In connection of base-word $w_1 \neq \epsilon$ to $w_2$, bigram probability $P(w_2|w_1)$ is used. This combination method may degrade individual recognition performance because beam search may prune more hypotheses which are generated by both isolated word recognition and continuous base-word recognition. However, the results of both methods can be obtained simultaneously by only one recognition process.

2.3. Splitting words into base-words

We first gathered keywords often used in organization names from a newspaper database, in which the organization names were split to general keywords, and picked up the words consisting of 2 or 3 Chinese characters (corresponding to approx. 2 to 5 syllables) as keywords. There are many homonyms in Japanese, so each keyword corresponded to a Chinese character sequence and a syllable sequence. N-frequent organization names might be used as parts of the other organization names, so we added these N-frequent organization names to the keyword set.

A user first utters an organization name to the system. The system recognizes it and displays the candidates as in Figure 4 on the touch panel. The “organization name candidates” contains the N-best (in this case, 5-best) hypotheses of isolated whole-word and continuous base-word recognition. The “base-word lattice” is created as follows: the best hypothesis generated by continuous base-word recognition is picked and the base-words contained in the hypothesis are taken. The other base-words (up to 5) ended at the same time frame as each base-word contained in the best hypothesis in HMM trellis are also picked. This procedure is illustrated in Figure 5. And then the user chooses a sequence or base-words from the candidates by pen touch to construct the desired name. The user can select a

1This procedure is a greedy algorithm and thus the optimal split may not be obtained. There is a more appropriate method for this split [16], but the accurate split was not our aim of this research. So we used this simple method.
Figure 4: The multimodal interface server for organization name input

Figure 5: Construction of a base-word lattice

name with one pen touch in the former case, whereas the user has to touch the panel several times in the latter case.

In Japanese, some forms expect not a Chinese (Kanji) character sequence but a Japanese (Kana) character sequence, which expresses the pronunciation of the word. The user can toggle between Chinese character input and Japanese character input by clicking the “Japanese/Chinese character input” button. Finally the result is sent to the client by clicking the “OK” button.

4. Experiments

In this section, we first evaluate the recognition performance of the combination of isolated name recognition and continuous base-word recognition, and then evaluate the name input ability through the interface.

4.1. Experimental conditions

100 organization names were uttered by 5 male speaker and thus the test set consisted of 500 name utterances. We controlled the coverage by the whole word dictionaries as follows: 90 % by 10000 words, 92 % by 15000 words, and 95 % by 20000 words. The coverage of the test set by the dictionaries are described in Table 1. The coverage of 130,000 names corresponding to Figure 1 are also shown in the table.

These utterances were analyzed as follows: sampling frequency of 16 kHz; Hamming window size of 25 ms; frame shift of 10 ms, and feature parameters of 16 dimensional MFCC, ∆/∆∆ MFCC, and ∆/∆∆ energy.

Left-to-right context-independent segmental unit input HMMs [17] were used as syllable acoustic models. Each model had 5 states, 4 of them had 4 Gaussian distributions with full-covariance matrices. The number of syllables was 114. The initial models were trained with syllable-segmented data from A-J sets (50 sentences each) of the ATR speech database consisting of utterances spoken by 6 male speakers. For syllable categories with a small amount of data in the database, 216 word data sets were additionally used. Afterwards, the HMMs were retrained with MAP estimation using an Acoustic Society of Japan (ASJ) database consisting of utterances by 30 male speakers (4,518 sentences) and a Japan Newspaper Article Sentences (JNAS) database with utterances by 124 male speakers (12,703 sentences).

The bigram language model was made from the most N-frequent isolated whole words and base-word sequences. Word frequencies were obtained by appearance frequencies in a news paper database. We picked the most M-frequent base-words for the lexicon and thus there appeared unknown segments in the sequences. So we used an unknown token for these segments, but such segments could not be recognized in recognition stage. We can easily find that the coverage by the combination of whole words and base-words tested to be higher than those only by whole words. For example, the coverage (99 % for test set, 98.4 % for 130,000 words) by combination of 10000 whole words and 5000 base-words was higher than that (92 % for test set, 97.4 % for 130,000 words) by 15000 whole words. Coverage by base-words were also high, so we have to compare the recognition performance.

4.2. Recognition performance evaluation

Table 2 shows the recognition rates for various lexicon settings. We adopted the 5-best candidates for the interface, so the table shows the 5-best recognition rates.

The combination method of isolated whole word recognition and continuous base-word recognition obtained better performance than isolated word recognition only and continuous base-word recognition only. For example, the result by combi-
We investigated a multimodal interface for organization name input based on a combination method of isolated whole word recognition and continuous base-word recognition. This combination method was suitable for organization names, which are often expressed by concatenation of shorter base-words. Recognition results showed that the combination method outperformed the individual use of isolated word recognition and continuous base-word recognition. This method is robust to unknown, newly generated names, in other words, an open vocabulary domain. Adoption of a base-word lattice to a multimodal interface also elevated user’s input ability of organization names.

We have implemented the interface described in Figure 4 on a PDA and we have to evaluate its usability by subjective tests as soon as possible.

6. References