AUTOMATIC LEXICON GENERATION AND DIALOGUE MODELING FOR SPONTANEOUS SPEECH

Hiroki Mori and Hideki Kasuya
Faculty of Engineering, Utsunomiya University
7-1-2, Yoto, Utsunomiya-shi, 321-8585 Japan

ABSTRACT
This paper describes novel framework for dialogue modeling based on a superword model, a superset of word n-gram. This has a remarkable advantage, because only transcribed text is needed to obtain the model, and no word dictionary is needed. In this paper, it is shown that the expressions specific to dialogue speech are extracted automatically from the transcriptions of spoken dialogue corpora by applying the acquisition method of the superword model. From experimental results based on a Japanese spoken dialogue database which consists of 42 sessions from 6 different tasks, it has been found that the proposed language modeling method has an ability to acquire task-independent lexical entry characteristic of dialogue speech, and many lexical entries are found to be relevant to discourse structures.

1. INTRODUCTION
To be able to realize robust and intelligent spoken dialogue systems that can handle spontaneous speech in the real world, it is essential to develop a language model that copes with various speaking styles of spoken dialogue. Such a language model is expected to be useful not only to improve recognition accuracy by predicting users’ utterance, but also to generate adequate responses from the system. In conventional spoken dialogue systems, the network grammar has often been used as a language model for speech recognition and dialogue management. Spontaneous utterance is often analyzed mainly from its content words and case markers, and many expressions specific to spoken dialogue are discarded as out-of-vocabulary words. In recent studies, however, such ignored expressions have proved to play essential roles in making dialogues fluent. It is important, therefore, to reorganize the expressions that compose spoken dialogue, and to find a way in which the expressions are related to discourse structures, in order to improve fluent voice communication with machines. However, preparation and standardization of a lexicon for speech recognition in dialogues has difficulties caused by the variety of expressions, especially with languages such as Japanese, which have numerous variations of speaking style according to a speaker’s social background or the social relationship with another speaker. This makes it difficult to incorporate such various expressions into conventional, rigid language models. In addition, we must rely on some heuristic method to distinguish expressions as discourse cue in conventional frameworks.

2. LEXICAL CHARACTERISTICS OF DIALOGUE SPEECH
In this section we will briefly sketch the difficulty of defining a lexicon for dialogue speech, especially in Japanese. The following example is an exchange from a simulated dialogue, “find the differences” task[2].

A: dene-... kuchi wa toji te te
‘and well... his mouse is closed, and’
B: un
‘uh-huh’

Long vowels are indicated by a trailing hyphen. As for Japanese, the concept of word is ambiguous, and sentences are written without any space orthographically. Word-to-word segmentation in the example above is therefore one possibility. The expression “dene-” requires special treatment because of the following reasons: (1)The expression is peculiar to dialogue speech, and never appears in written language. (2)The expression is quite informal but an intimate one, and is permissible only in some kinds of certain interpersonal relationships. (3)The expression is thought to be a kind of discourse marker—that is, the speaker A intended to keep her ‘turn’ by uttering “dene-”. The last vowel could be arbitrarily lengthened according to the speaker’s dialogue strategy. (4)The expression “dene-” is not regarded as an independent
lexical entry in most dictionaries. However, it seems to provide a somewhat characteristic usage, rather than merely composed of “de” and “ne-.”

In addition to that, there are so many non-lexical expressions characteristic of dialogue speech, and most of them have no meaning themselves[3]. To reorganize such expressions composing spoken dialogue, we employ an example-based, statistical modeling approach—namely, superword model. Such an approach is promising because it has potential ability to incorporate a variety of expressions, discourse contexts and restriction of pragmatics into a single model.

3. SUPERWORD MODEL

The superword model[1] is a language model originally used for Japanese document recognition. In some Asian languages, such as Japanese and Chinese, word boundaries are not explicitly indicated in written text, so building a word n-gram model for such languages is a non-trivial problem. The superword model is a solution to the problem, and previous studies showed its superiority to conventional word n-gram based on automatic morphological analysis. This section outlines the concept of the superword model.

3.1. Formulation

Any possible string is defined as a superword. A superword is identified based on only one principle—it must appear at least twice in the training text. In addition, we also regard every string consisting of a single character as a superword. This guarantees that every utterance can be represented as a sequence of superwords.

The superword n-gram probability is defined as a conditional probability that depends on the last \((n-1)\) superwords:

\[
P(w_i|w_{i-(n-1)} \cdots w_{i-2}w_{i-1}),
\]

where \(w_i\) denotes \(i\)-th superword which consists of several characters. Generally, character-to-superword mapping is not unique for a given utterance. The probability of an utterance \(C = C_1C_2 \cdots C_k\), where \(C_i\) is the \(i\)-th character, is given by the total sum of the probability of any possible superword sequences in the utterance:

\[
P(C) = \sum_{w_1 \cdots w_L \in C} \prod_{i=1}^{L} P(w_i|w_{i-(n-1)} \cdots w_{i-1})
\]

If \(n = 1\), that is, if we make an assumption that each superword occurs independently, the definition is identical to that of multigram[4].

Apparently, the superword n-gram model is a subclass of HMM(hidden Markov model), so its well-established learning schemes, such as the forward-backward algorithm, can be utilized for acquisition of the superword model.

3.2. Model Combination

Stochastic language models, such as the n-gram model, are often faced with the “sparse data” problem—that is, we must estimate parameters from limited samples. To cope with both reliability and accuracy, it is known to be effective to utilize some smoothing technique like interpolation[5] or back-off[6][7].

As for superword bigram model, the interpolated probability is given by weighted sum of superword bigram and superword unigram probabilities:

\[
\hat{P}(w_i|w_{i-1}) = \lambda_0 P(w_i|w_{i-1}) + (1 - \lambda_0) P(w_i) \quad (3)
\]

The weight can be learned through an iterative reestimation process, so as to maximize the probability of held-out[5] sample of the text.

The previous study also showed that the superword model gets more robust when combined with the character n-gram of Japanese. The combined superword bigram probability is defined by:

\[
P_{wc}(w_i|w_{i-1}) = \lambda_w \hat{P}(w_i|w_{i-1}) + (1 - \lambda_w) \hat{P}_c(w_i|w_{i-1}), \quad (4)
\]

where \(\hat{P}_c(w_i|w_{i-1})\) is the superword \(w_i\)'s probability given by the interpolated character trigram as follows:

\[
\hat{P}_c(w_i|w_{i-1}) = \begin{cases} 
\lambda_3 P(C^3|C_1C_2) + \lambda_2 P(C^3|C_2) + \lambda_1 P(C^3) + \lambda_0 P_0 & \text{if } w_i = C^3 \\
0 & \text{otherwise} 
\end{cases} \quad (5)
\]

where \(C^3\) denotes the last two characters of \(w_{i-1}\), and \(\lambda_3 + \lambda_2 + \lambda_1 + \lambda_0 = 1\). \(\hat{P}_c(w_i|w_{i-1})\) has a non-zero value only if the length of the superword \(w_i\) is equal to 1. When \(\lambda_w \to 0\), Equation (2) falls into the interpolated character trigram model.

4. MODEL ACQUISITION

The acquisition process of superword model can be divided into two stages: (1) collection of superword set, (2) iterative estimation of probability distribution.

4.1. Lexicon Generation

At the first stage, we must count frequency of strings in training text and store any string which appears two or more times. A general framework for counting arbitrary long strings is discussed in [8], but we can employ a more simple approach where singletons can be ignored.

It is clear that any substring of a superword appears twice or more, because the superword also appears twice or more. Therefore, the recursive procedure shown in Figure 1 finds a set of superwords consisting of up to \(L\) characters.

Now, we discuss a lexicon for spoken dialogue. In general, task-oriented dialogue speech is composed of (1) task-dependent part, and (2) task-independent part. The former usually consists of task-specific content words and case markers, and can be efficiently described by the task’s grammar. The latter does include greetings or irregular chat, but mainly concerns metacommunication
To obtain task-independent superword sets, we apply the following process of the superword model, refer to [1]. It can be summarized as:

1. Assign uniform distribution to all superword’s initial unigram probabilities.
2. Reestimate the distribution by forward-backward algorithm until it converges.
3. (optional) Prune the superword set of unlikely superwords (=ones with small probability assigned).
4. Assign the estimated distribution to higher-order n-gram’s initial probabilities (ex. initialize superword bigram probabilities by superword unigram probabilities). Reestimate it. Repeat this step if necessary.

Figure 1: Pseudocode for finding a set of superwords from the training text. “chop” is a function which truncates the last character of the given string and returns the result.

Unlike monologue speech or written text, transcription of dialogue speech has two (or more) channels. For simplification, it is assumed here that each utterance occurs independently of others. By this assumption, utterances from multiple channels can be freely merged into a single stream in order to build a superword model.

5. EVALUATION

5.1. Dialogue Corpora

Transcribed speech data used for training consists of 42 sessions from 6 different tasks (“secretary,” “scheduling,” “crossword,” “telephone ordering,” “find the differences,” “chat”). They are collected from simulated dialogue corpus[2], excluding the sessions collected with WOZ and the ones which have no separate sound tracks for individual speakers. Each transcription is canonicalized to phoneme string by hand, and then converted to mora string. Fillers and disfluent utterances are not excluded.

5.2. Extracted Expressions

Fig. 2 shows (a) top 16 superwords which have unigram probability, and (b) typical superwords that are in dialogues are picked from top 100 superwords. “Hai(yes)” is the most frequent expression because it is also uttered as a polite backchannel response. The corresponding impolite expression “un” is also extracted. Most of the other frequent expressions can be classified as case markers or fillers. The proposed method can also find typical discourse cues such as “desune,” “dewa” and “yone,” which most word dictionaries do not cover.

Fig. 3 shows superword bigrams with large joint probability, and thought to represent the structure of dialogue speeches well.

5.3. Perplexity

The proposed dialogue model is evaluated with perplexity. Lower perplexity generally means it is a better language model for recognizing speech. The goal of our model is to emulate task-
In this paper, we showed the framework of superword-based dialogue modeling and its efficiency. So far, the model captures the intra-utterance structure of dialogue speech. This should be extended to inter-utterance relationships, by modifying the superword’s formulation to enable handling of opposite utterances as contexts. For example, we believe that the expression “Qte (topic marker)” shown in Fig. 2 plays an important role in discourse, because this is often followed by a question about the topic (sometimes even omitted completely), and the answer to the topic should follow. Long-distance prediction of superwords, such as the trigger-based model[10], may be useful to model this kind of discourse structures.

Task-oriented spoken dialogue systems involve task grammar. To make full use of the superword model in the systems, integration of the grammar and superword model is needed. Dialogue management using superword-based framework is also challenging and possible future work.

### 7. REFERENCES


### Table 1: Mora perplexity for test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Perplexity</th>
<th># of Entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>mora trigram</td>
<td>5.648</td>
<td>11243</td>
</tr>
<tr>
<td>superword unigram</td>
<td>5.420</td>
<td>3331</td>
</tr>
<tr>
<td>superword bigram</td>
<td>5.265</td>
<td>70545</td>
</tr>
<tr>
<td>superword bigram (pruned)</td>
<td>5.267</td>
<td>13654</td>
</tr>
</tbody>
</table>

Figure 3: Typical superword bigrams.