Journal of Information Processing Vol.21 No.2 157–167 (Apr. 2013)

[DOI: 10.2197/ipsjjip.21.157]

Regular Paper

Comparison of Methods for Topic Classification of Spoken Inquiries

Rafael Torres[1,a)] Hiromichi Kawanami[1,b)] Tomoko Matsui[2,c)]
Hiroshi Saruwatari[1,d)] Kiyohiro Shikano[1,e)]

Received: June 1, 2012, Accepted: November 2, 2012

Abstract: In this work, we address the topic classification of spoken inquiries in Japanese that are received by a speech-oriented guidance system operating in a real environment. The classification of spoken inquiries is often hindered by automatic speech recognition (ASR) errors, the sparseness of features and the shortness of spontaneous speech utterances. Here, we compare the performances of a support vector machine (SVM) with a radial basis function (RBF) kernel, PrefixSpan boosting (pboost) and the maximum entropy (ME) method, which are supervised learning methods. We also combine their predictions using a stacked generalization (SG) scheme. We also perform an evaluation using words or characters as features for the classifiers. Using characters as features is possible in Japanese owing to the presence of kanji, ideograms originating from Chinese characters that represent not only sounds but also meanings. We performed analyses on the performance of the above methods and their combination in dealing with the indicated problems. Experimental results show an F-measure of 86.87% for the classification of ASR results from children's inquiries with an average performance improvement of 2.81% compared with the performance of individual classifiers, and an F-measure of 93.96% with an average improvement of 1.89% for adults' inquiries when using the SG scheme and character features.

Keywords: topic classification, support vector machine, PrefixSpan boosting, maximum entropy, stacked generalization

1. Introduction

Improvements in automatic speech recognition (ASR) technologies have made feasible the implementation of systems that interact with users through speech. In this work, we address the topic classification of spoken inquiries in Japanese that are received by a speech-oriented guidance system operating in a real environment. The guidance system is the *Takemaru-kun* system [1], which operates in a public facility and receives daily user requests for information and collects real data.

The *Takemaru-kun* system is an open domain system, which means that the task domain was not set before its operation started, and users are free to ask the system the information they want to obtain. Since the system started collecting user's inquiries, they have been analyzed and manually labeled, to define its task domain. Therefore, we expect the results of the analysis we present in this work to be applicable to other task domains for this type of system.

Topic classification has been studied in the field of telephone call classification for the optimization of call routing [2] and to determine the reason for calling [3], [4]. These studies are similar to ours since they also deal with speech. In our work, we study topic classification in the context of an information guidance system, where inquiries tend to be short and the task domain is wider.

In text-based information retrieval, there have been studies on the determination of question type in the context of answer selection [5], [6] and also on topic detection and estimation [7], [8]. Since we are using an ASR engine to translate spoken inquiries into text, these studies share similarities with our work; however, the classification of spoken inquiries is often hindered by ASR errors.

In this work, we selected three different types of classification methods, (1) a support vector machine (SVM) with a radial basis function (RBF) kernel, (2) PrefixSpan boosting (pboost) and (3) the maximum entropy (ME) method, which are supervised learning methods, and compared their performance. In the SVM method, the estimation of a robust boundary known as the maximum-margin hyperplane is crucial. SVM has successfully been applied to a wide variety of classification tasks including speech [3], [5], [6], [9], [10]. The pboost method is for classification of sequential data, and it extracts and utilizes discriminative and sequential patterns in the data [11]. Although the method has been developed for classification of actions in videos, we introduce pboost for the classification of spoken inquiries into topics. The ME method is a probabilistic approach based on data distribution. ME has been widely used in natural language processing

[1] Nara Institute of Science and Technology, Ikoma, Nara 630–0192, Japan
[2] The Institute of Statistical Mathematics, Tachikawa, Tokyo 190–8562, Japan
[a)] rafael-t@is.naist.jp
[b)] kawanami@is.naist.jp
[c)] tmatsui@ism.ac.jp
[d)] sawatari@is.naist.jp
[e)] shikano@is.naist.jp

(NLP) tasks [8] as well as in speech classification [4], [5].

Moreover, we combined the predictions from the above different types of methods by using a stacked generalization (SG) [12] scheme and examined the complementary effect. The SG scheme and similar schemes have also been studied as a means of combining classifier predictions in other classification tasks [13], [14], [15].

We also perform an evaluation using words or characters as features for the classifiers. Using characters as features is possible in Japanese owing to the presence of kanji, ideograms originating from Chinese characters that represent not only sounds but also meanings. The use of words or characters has also been investigated for spoken document retrieval [16], [17], and better performance was obtained when using words than when using characters. However, the spoken inquiries in our topic classification task are much shorter than spoken documents; hence we are also interested in evaluating spoken inquiries.

The remainder of the paper is structured as follows: Section 2 describes the *Takemaru-kun* datasets, Section 3 explains the topic classification methods compared in this paper and their combination, Section 4 presents the experiments conducted and an analysis of their results, and Section 5 concludes this work.

2. *Takemaru-kun* Datasets

We compared the performances of the methods and their combination using the *Takemaru-kun* datasets.

2.1 Overview of the *Takemaru-kun* System

The *Takemaru-kun* system [1], shown in **Fig. 1**, is a real-environment speech-oriented guidance system placed inside the entrance hall of the Ikoma City North Community Center in Nara, Japan. The system has been operating daily since November 2002, providing information to visitors, including information on the center facilities and services, local sightseeing, the weather forecast, news, and about the system agent itself. The system uses an example-based one-question-to-one-response strategy for interaction, which fits the purpose of responding to simple questions from a large number of users. Users can also activate a Web search feature to search for Web pages over the Internet that contain the uttered keywords.

2.2 Specifications of Datasets

Utterances received by the *Takemaru-kun* system have been recorded since it first started operating. Utterances from Nov. 2002 to Oct. 2004 and from Dec. 2004 to Mar. 2005 were manually transcribed and labeled with their answers along with information concerning the age group and gender of users. Invalid inputs such as noise, coughs, laughter and unclear inputs were also documented. Some examples of inquiries received by the system are shown in **Table 1**. The signal-to-noise ratio (SNR) of the utterances recorded in this period is 38.31 dB.

Fig. 1 Speech-oriented guidance system *Takemaru-kun*.

The *Takemaru-kun* datasets consist of valid utterances from children and adults collected in the period indicated above. Acoustic models (AMs) and language models (LMs) were separately prepared for children and adults. The AMs were trained using the utterances collected by the system from Nov. 2002 to Oct. 2004, and the LMs were constructed using the transcriptions of the utterances in the same period. Details of the setup for the AMs, LMs and ASR for children and adults are shown in **Table 2**.

Spoken inquiries received by the *Takemaru-kun* system are usually short, with only a few words per utterance, as shown in **Fig. 2**. Because of this and the vocabulary sizes, shown in **Table 3**, features in the utterances tend to be sparse.

The test datasets contain utterances for Aug. 2003 and from Dec. 2004 to Mar. 2005, and the training datasets include the rest of the utterances. ASR word correct rates for children's utterances are considerably lower than those for adults, as it is shown in **Table 4**. The frequency of utterances in the datasets for the 15 most frequent topics is shown in **Table 5**. As can be observed, the frequency of utterances for each topic is variable, as some topics are more popular than others.

Table 1 Examples of utterances received by the *Takemaru-kun* system.

Utterance in Japanese	Translation to English	Topic
エレベーターはどこ？	Where is the elevator?	info-facility
生駒市の地図を見せて	Show me Ikoma city's map	info-city
さようなら	Goodbye	greeting-end
お名前は	What's your name?	agent-name

Table 2 Setup for acoustic models (AMs), language models (LMs) and ASR for children and adults.

AM training tool	HTK 3.2 [18]
Acoustic model	PTM [19], 2,781 HMMs, 1,965 states, 8,256 mixtures
Acoustic features	12 MFCC, 12 Δ MFCC, Δ E
AM training	Baum-Welch, 3 iterations
LM training tool	SRILM 1.5.0 [20]
Language model	3-gram, Kneser-Ney smoothing
LM perplexity	Children: 16.5, Adults: 9.9
ASR engine	Children: Julius 4.0, Adults: Julius 3.5.3 [21]

Table 3 Vocabulary sizes.

Inquiries	Feature	Children	Adults
Transcriptions	Word 1-grams	3,610	1,691
Transcriptions	Word 2-grams	14,096	4,221
Transcriptions	Word 3-grams	19,648	5,375
Transcriptions	Character 1-grams	858	709
Transcriptions	Character 2-grams	8,998	4,303
Transcriptions	Character 3-grams	22,252	7,469
ASR 10-best results	Word 1-grams	6,095	3,589
ASR 10-best results	Word 2-grams	68,180	22,768
ASR 10-best results	Word 3-grams	121,951	31,817
ASR 10-best results	Character 1-grams	1,228	994
ASR 10-best results	Character 2-grams	26,869	12,865
ASR 10-best results	Character 3-grams	97,337	32,126

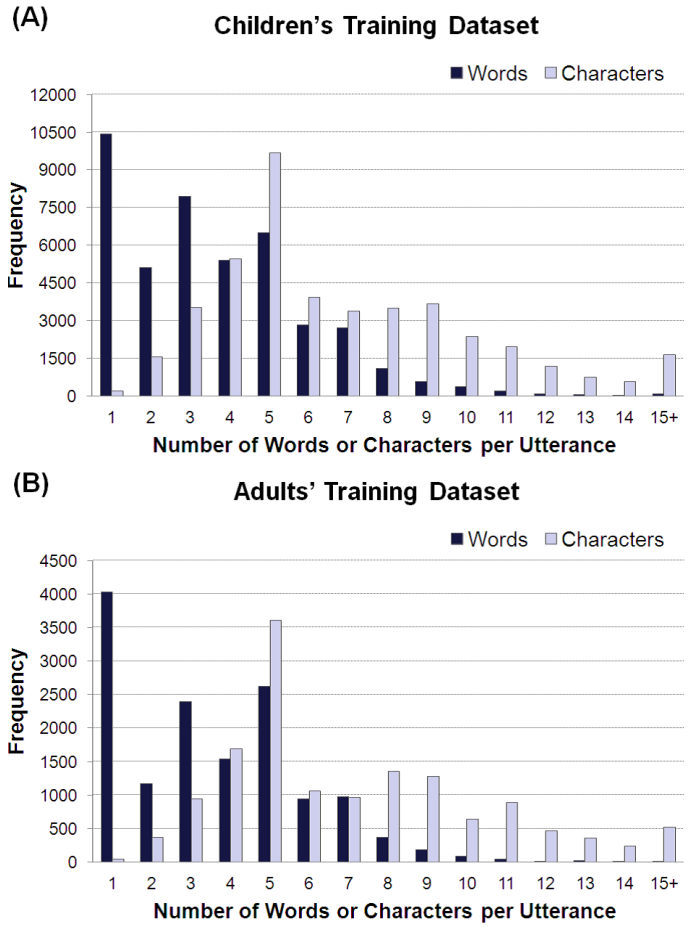


Fig. 2 Frequency of utterances by number of words and characters per utterance in (A) children's training and (B) adults' training datasets (ASR 1-best results).

Table 4 ASR word correct rate of the utterances in the datasets.

Children		Adults	
Training	Test	Training	Test
77.73%	71.59%	91.36%	85.53%

Table 5 Frequency of utterances in the datasets for each topic.

	Children		Adults	
Topic	Training	Test	Training	Test
chat-compliments	2,548	1,066	766	194
info-services	884	206	494	89
info-news	529	144	484	137
info-local	709	187	553	70
info-facility	5,007	1,653	1,795	299
info-city	1,006	317	504	93
info-weather	2,947	1,073	1,099	257
info-time	3,911	898	984	187
info-sightseeing	647	142	668	79
info-access	681	142	676	83
greeting-end	4,535	2,125	912	269
greeting-start	6,845	2,629	2,672	723
agent-name	5,381	1,574	1,309	254
agent-likings	4,418	2,260	851	194
agent-age	3,446	1,108	664	157
Total	43,494	15,524	14,431	3,085

3. Topic Classification

The classification of spoken inquiries into topics can be used to manage the interaction with users and help select appropriate answers [2]. It can also be used to improve the ASR performance by applying topic-dependent language models, as was shown by Lane et al. [22].

3.1 Compared Methods

SVM, pboost and ME have different characteristics. SVM and pboost are discriminative classifiers which means that they learn a direct map from inputs to classes without caring about underlying probability distributions. SVM can deal with nonlinearity owing to the use of kernel functions, meaning that it can robustly find boundaries among classes even when data are not linearly separable, whereas pboost performs feature selection and classifies by checking for the presence of optimal discriminative subsequence patterns in the input. On the other hand, ME is a classifier that estimates probability distributions from data, allowing multiclass classification. It also has the advantage that it is not sensitive to hyperparameter settings, in contrast to the other two classifiers.

3.1.1 Support Vector Machine

SVM maximizes the margin of classification of two different classes of data, robustly detecting boundaries between them. SVM can deal with nonlinearities by using kernels and is appropriate for sparse high-dimensional feature vectors. SVM has successfully been applied to a wide variety of classification tasks including speech [3], [5], [6], [9], [10].

In our classification task, the number of utterances for each topic is unbalanced. We use C-support vector classification (C-SVC) with a soft margin for unbalanced data. Details of the

method are described in Ref. [23]. The hyperparameters C_+ and C_- are cost parameters that control the importance given to classification errors in order to implement a soft margin. When the training data is unbalanced, SVM parameters are not estimated robustly. By introducing the different hyperparameters C_+ and C_-, this problem can be dealt with.

SVM is originally a binary classifier. We used a one-vs-rest approach for multiclass classification, which constructs one binary classifier for each topic. Each classifier is trained with data from a topic that is regarded as positive, and the rest of the topics are regarded as negative. Although SVM can only predict the topic label and not probability information, the method described in Ref. [24] can be used to obtain probability estimates or pseudo-probabilities for each topic. We used this method and classified new data in the topic with highest pseudo-probability. We selected this approach because in preliminary experiments it had better performance than the one-vs-one approach for this task.

We used a bag of words (BOW) to represent utterances as vectors, where each component of the vector indicates the frequency of appearance of a feature. The length of a vector corresponds to the size of the dictionary that includes every feature in the training dataset. We used an RBF kernel because in preliminary experiments it exhibited slightly better performance than a polynomial kernel for this task. The RBF kernel is defined as

$$\kappa(\vec{x}_i, \vec{x}_j) = \exp(-\gamma \|\vec{x}_i - \vec{x}_j\|^2), \quad \gamma > 0 \tag{1}$$

where $\vec{x}_i$ and $\vec{x}_j$ represent utterance vectors and γ is a hyperparameter of the function.

3.1.2 PrefixSpan Boosting

Pboost is a method proposed by Nozowin et al. [11] for the classification of actions in videos. In this work we introduce pboost for the classification of spoken inquiries into topics. Pboost implements a generalization of the PrefixSpan algorithm by Pei et al. [25] to find optimal discriminative subsequence patterns, and in combination with the Linear Programming boosting (LPboost) classifier, it optimizes the classifier and performs feature selection simultaneously. Boosting methods form a weighted majority prediction rule by combining the decisions of several weak learners, and have also been used for speech classification [4], [9].

Details of this method are described in Ref. [11]. Pboost uses the PrefixSpan algorithm [25] to find optimal subsequence patterns that characterize utterances from a specific topic. For example, in the topic *info-facility* we found the following utterances: "*Where* can I *find* the toilet?" and "*Where* can I *find* the library?." From these utterances, pboost can determine that an optimal pattern is the subsequence "*where find*." As can be seen from this example, subsequences can also include gaps.

The presence of a single subsequence pattern in an utterance is called a weak hypothesis and has the form $h(\vec{x}; \vec{s}, \omega)$. Here, $\vec{x} \in \{\vec{x}_i\}$, $\vec{x}_i \in \mathbb{R}^n$, $i = 1, \ldots, l$ is a training vector, $\vec{s}$ is a subsequence pattern and $\omega \in \Omega$, $\Omega = \{-1, 1\}$ is a variable that indicates if the sequence is relevant to the positive or negative class.

The classification function has the form

$$f(\vec{x}) = \sum_{(\vec{s},\omega)\in\vec{S}\times\Omega} \alpha_{\vec{s},\omega} h(\vec{x}; \vec{s}, \omega) \tag{2}$$

where $\alpha_{\vec{s},\omega}$ is the weight for feature sequence $\vec{s}$ and parameter ω such that $\sum_{(\vec{s},\omega)\in\vec{S}\times\Omega} \alpha_{\vec{s},\omega} = 1$ and $\alpha_{\vec{s},\omega} \geq 0$. $\alpha_{\vec{s},\omega}$ indicates the discriminative importance of a feature sequence.

In pboost, the primal problem formulation implementing a soft margin for an unbalanced number of samples follows the form

$$\begin{aligned} \min_{\rho,\vec{\alpha},\vec{\xi}} \; & -\rho + D_+ \sum_{\{i:y_i=+1\}} \xi_i + D_- \sum_{\{i:y_i=-1\}} \xi_i \\ \text{sb.t.} \; & \sum_{(\vec{s},\omega)\in\vec{S}\times\Omega} y_i \alpha_{\vec{s},\omega} h(\vec{x}_i; \vec{s}, \omega) + \xi_i \geq \rho, \; i = 1, \ldots, l \\ & \sum_{(\vec{s},\omega)\in\vec{S}\times\Omega} \alpha_{\vec{s},\omega} = 1, \; \vec{\alpha} \geq 0, \; \vec{\xi} \geq 0 \end{aligned} \tag{3}$$

where $\vec{x}_i \in \mathbb{R}^n$, $i = 1, \ldots, l$ indicates a training vector, $y_i \in \{1, -1\}$ is a class, ρ is the soft margin separating negative from positive samples, and D_+ and D_- are hyperparameters controlling the cost of misclassification by penalizing the sums of the slack variables ξ_i for the soft margin.

Here, we also used a one-vs-rest approach for multiclass classification, and we classified new data according to the highest value of the classification function in Eq. (2).

3.1.3 Maximum Entropy

ME is a supervised learning method that estimates probability distributions from data [26], by selecting the distribution that maximizes the entropy. Among the methods we compared in this work this is the only one that provides probability information, and is a multiclass classifier by nature. ME has been widely used in natural language processing (NLP) tasks [8] as well as in speech classification [4], [5].

Given an utterance consisting of the feature sequence c_1^N, where the suffix 1 indicates the first feature of the sequence (word or character) and N indicates the last feature of the sequence, the objective of the classifier is to provide the most likely class label $\hat{k}$ from a set of labels K, such that

$$\hat{k} = \operatorname*{argmax}_{k\in K} p(k|c_1^N), \tag{4}$$

where the ME paradigm expresses the probability $p(k|c_1^N)$ as

$$p(k|c_1^N) = \frac{\exp\Big[\sum_c N(c) \log \alpha(k|c)\Big]}{\sum_{k'} \exp\Big[\sum_c N(c) \log \alpha(k'|c)\Big]}. \tag{5}$$

Ignoring the terms that are constant with respect to k yields

$$\hat{k} = \operatorname*{argmax}_{k\in K} \sum_c N(c) \log \alpha(k|c), \tag{6}$$

where $N(c)$ is the frequency of a feature in a class, and $\alpha(k|c)$ with $\alpha(k|c) \geq 0$ and $\sum_k \alpha(k|c) = 1$ is a parameter that depends on the class k and feature c, and is calculated using methods such as L-BFGS-B [27] which is a limited-memory algorithm for solving large nonlinear optimization problems.

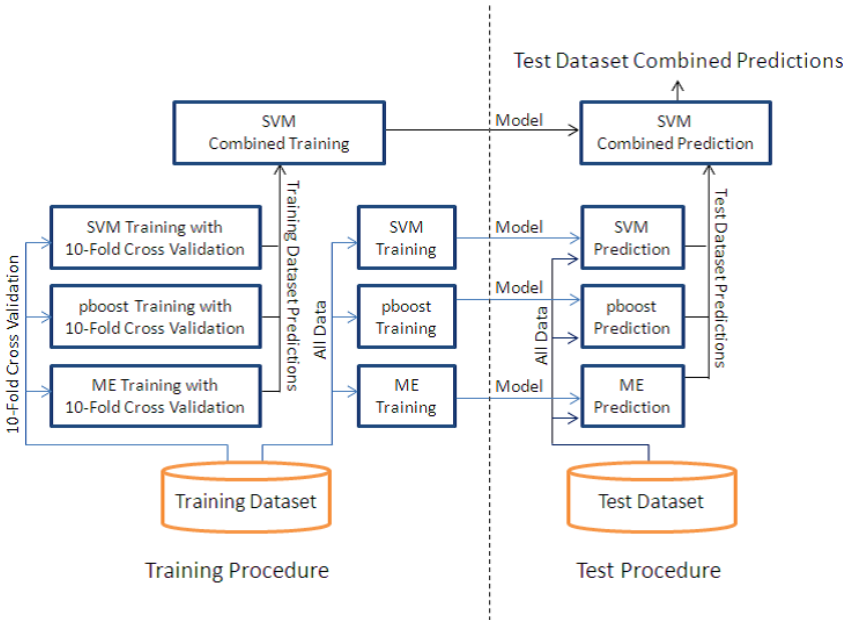


Fig. 3 Training and test procedures in the SG scheme.

3.2 Combination of Methods

Because of the differences in the classifiers we selected for comparison, we can expect them to compensate each other to improve prediction performance. Because of this, we also compare them with an SG scheme that combines their predictions.

SG, proposed by Wolpert [12], is a method that combines the outputs of multiple classifiers using a second-level classification, minimizing the generalization error of first-level classifiers and achieving greater predictive accuracy. Its success arises from its ability to exploit the diversity in the predictions of first-level classifiers in a particular classification task.

The training and test procedures in the SG scheme are illustrated in **Fig. 3**. In the first step of the training, the predictions of each of the first-level classifiers for each of the training utterances are collected to create a new dataset. Cross-validation training is used for the first-level models to avoid bias when obtaining the predictions. Each first-level method is trained with 90% of the data, and the model is used to predict the remaining 10%, until we have obtained predictions for each utterance in the training dataset.

In the second step, predictions of the first-level classifiers for each utterance in the training dataset are used as new data for training the second-level model. The feature vectors used to train the second-level model contain predictions of each of the first-level classifiers for each of the topics. For SVM and pboost, a position in the feature vector is 1 if an utterance is classified as positive in the topic represented by that position, and 0 otherwise, whereas for ME a position contains the probability for the topic represented by that position.

The test procedure is performed in a similar fashion, but in this case cross-validation is not needed, since we can obtain predictions for utterances in the test dataset by using models trained with all the training data.

As a second-level classifier, we selected SVM with an RBF kernel. We also performed preliminary experiments with SVM with a linear kernel and with ME and noticed that the results were not sensitive to the kernel or method. The classification problem at the second level is much simpler than that at the first level, since its feature vectors have very low dimensionality. Hence, the decision to use SVM with an RBF kernel was made for simplicity.

4. Experiments

We compared the performances of the methods in the topic classification of spoken inquiries.

4.1 Experimental Conditions

In our experiments, we used the *Takemaru-kun* datasets as described in Section 2.2. The experimental conditions for the first and second-level classifiers are given in detail in **Tables 6** and **7** respectively. For experiments with the SG scheme, we followed the procedure described in Section 3.2. We used a one-vs-rest approach for multiclass classification with SVM and pboost, and the "# of positive" and "# of negative" variables indicated in the experimental conditions refer to the number of utterances in the topic and in the rest of the topics respectively, for each classifier. Optimal hyperparameter values for SVM and pboost were obtained experimentally using a grid search strategy and were set a posteriori.

Owing to the considerable amount of computational time required for the PrefixSpan search-based feature selection in pboost, we used ASR 1-best results instead of ASR 10-best results. As explained in Section 3.1.2, pboost can include gaps in between optimal subsequences. In preliminary experiments, we found out that this increases the performance when using words as features; however, when characters are used as features the performance decreased when gaps are allowed.

The classification performance of the methods was evaluated using the F-measure, as defined by

$$F\text{-}measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \quad (7)$$

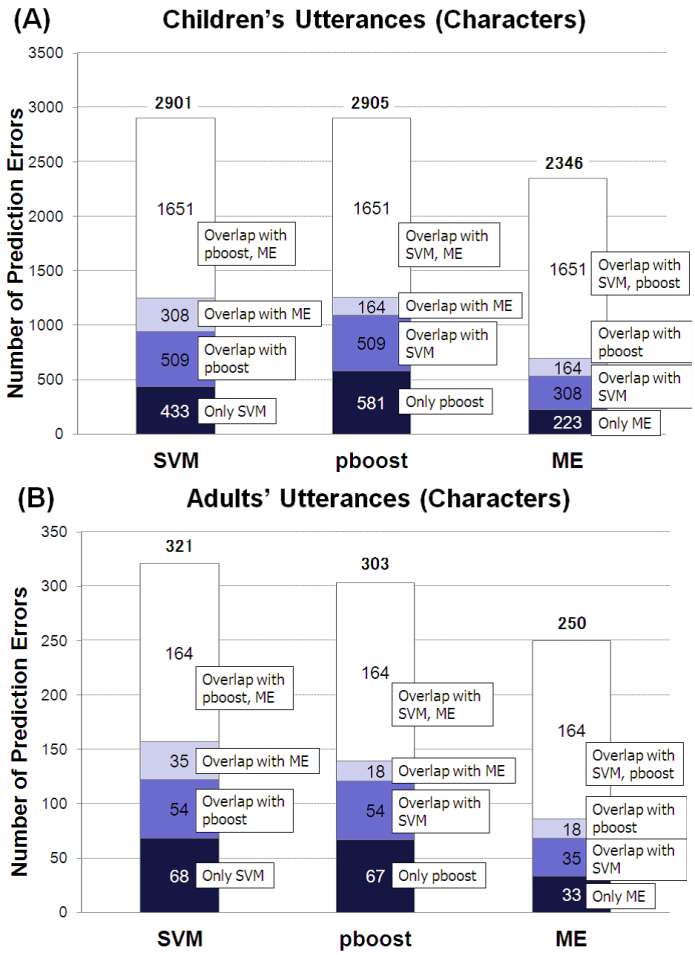


Fig. 4 Prediction error overlap by method for (A) children's and (B) adults' utterances using character features (open test). The number of prediction errors for each method is indicated above the bars in bold, and the numbers of prediction error overlaps among the methods are indicated inside the bars.

Table 6 Experimental conditions for the first-level classifiers.

SVM tool	LIBSVM 2.9 [23]
Hyperparameters C_+ and C_- for each SVM classifier	C_+ = (# of negative/total) $\times C$ C_- = (# of positive/total) $\times C$ where $C_+ + C_- = C$, and C from 1×10^{-3} to 1×10^{3} (powers of 10)
Kernel function	RBF kernel
Hyperparameter γ	1×10^{-3} to 1×10^{3} (powers of 10), and 0.5
Features	Word 1+2+3-grams, Character 1+2+3-grams
Datasets	Transcriptions and ASR 10-best results
Pboost tool	pboost 1.0 [11]
Hyperparameters D_+ and D_- for each pboost classifier	D_+ = (# of negative/total) $\times D$ D_- = (# of positive/total) $\times D$ where $D_+ + D_- = D$, and $D = 1/\nu\ell$, for ν from 0.001 to 0.100 and ℓ = number of training utterances
Max. subsequence length	3
Gaps	Allowed for word subsequences Not allowed for character subsequences
Features	Word 1-grams, Character 1-grams
Datasets	Transcriptions and ASR 1-best results
ME tool	maxent 2.11 [8]
ME model	Inequality constraints [28]
Features	Word 1+2+3-grams, Character 1+2+3-grams
Datasets	Transcriptions and ASR 10-best results

The F-measure was calculated individually for each topic and averaged over the frequency of utterances in the topics.

Table 7 Experimental conditions for the second-level classifier.

SVM tool	LIBSVM 2.9 [23]
Hyperparameters C_+ and C_- for each SVM classifier	C_+ = (# of negative/total) $\times C$ C_- = (# of positive/total) $\times C$ where $C_+ + C_- = C$, and C from 1×10^{-3} to 1×10^{3} (powers of 10)
Kernel function	RBF kernel
Hyperparameter γ	1×10^{-3} to 1×10^{3} (powers of 10), and 0.5
Features	Predictions of the first-level classifiers
Datasets	Transcriptions and ASR results

4.2 Performance Comparison

An analysis of overlaps in the prediction error among individual methods is presented in **Fig. 4**. The analysis indicates that the three methods produce some prediction errors that do not overlap with those of the other methods. Combining the methods makes it possible to correct some of these errors. On the other hand, we can observe that SVM and pboost have a higher prediction error overlap which is understandable since both are discriminative methods.

We evaluated the classification performance of the individual methods and their combination and performed a statistical significance test using a binomial proportion confidence interval of 95%. **Figures 5** and **6** present the results of each method for transcriptions and the ASR results for children's and adults' utterances respectively. The difference in the performance of the

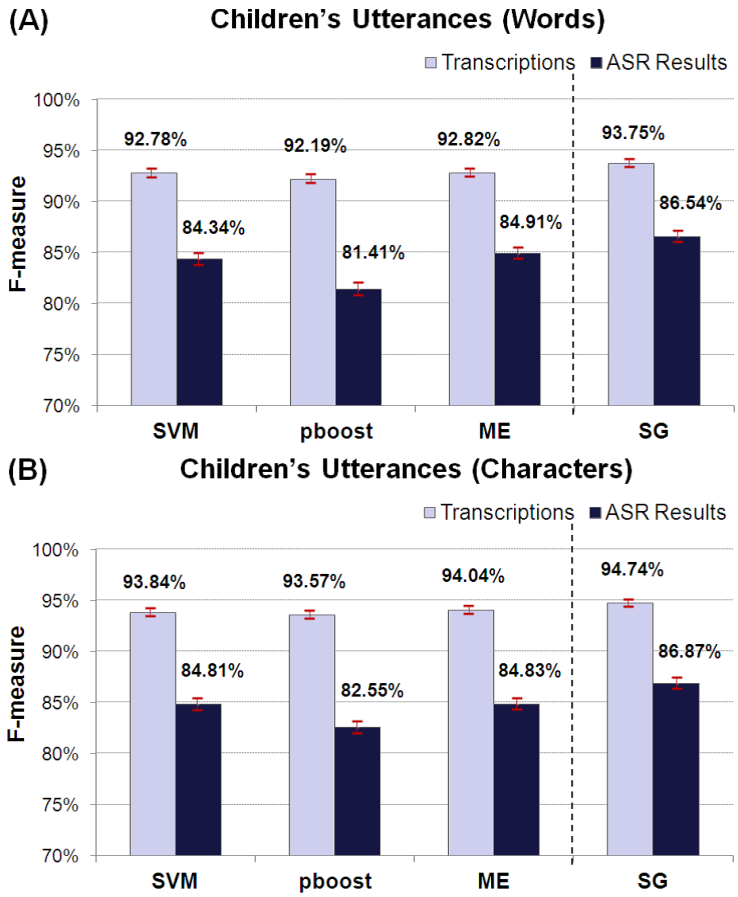


Fig. 5 F-measure for each method for transcriptions and ASR results for children's utterances using (A) word and (B) character features. The F-measure for each method is indicated above the bars in bold, and the red line segments represent 95% confidence intervals.

Table 8 Percentage of prediction errors recovered by the SG scheme by individual method (characters).

Individual Method	Children	Adults
SVM	20.13%	32.71%
pboost	20.17%	29.70%
ME	13.38%	21.60%

Table 9 Percentage of correct predictions misclassified by the SG scheme by individual method (characters).

Individual Method	Children	Adults
SVM	2.00%	0.76%
pboost	1.99%	0.86%
ME	4.08%	1.45%

individual methods was not found to be significant in most cases. However, the SG scheme performed significantly better than the individual methods. The average performance improvement was 2.81% compared with the performance of individual classifiers for the classification of ASR results of children's inquiries and 1.89% for adults' inquiries when using the SG scheme and character features. The only case in which a significant improvement could not be obtained was when classifying transcriptions of adults' inquiries using either words or characters; however, the performance was still comparatively high.

In this comparison, the performance of the methods was higher when character features were used than when words were used, although the difference was not found to be significant in the statistical test performed.

The percentage of prediction errors that the SG scheme was able to correct by an individual method using character features is presented in **Table 8**. With both children and adults the SG scheme was most beneficial for correcting SVM and pboost's prediction errors, while less benefit was seen for ME. **Table 9** presents the percentage of correct predictions by an individual method using character features that the SG scheme misclassified. Here we can observe side effects from the SG scheme which had a larger effect on ME predictions. However, these percentages are low in comparison to the prediction errors that were recovered.

Although pboost has lower classification performance than SVM and ME in many cases, experiments excluding pboost from the SG scheme yielded decreases in the classification performance. One of the advantages of pboost is that it produces results that can be interpreted. A grammatical analysis of the discriminative word subsequence patterns selected by pboost showed that the most important part of speech (POS) for the topic classification of utterances is the noun, which on average accounted for more than half of the words in the selected patterns. This is followed by the verb, which accounted on average for nearly a seventh of the words in the selected patterns. Particles, the Japanese POS that relates the preceding word to the rest of the sentence, were also selected as discriminative word subsequence patterns in some cases.

We observed that the optimal hyperparameters for SVM and pboost are highly dependent on the data. Because of this, the

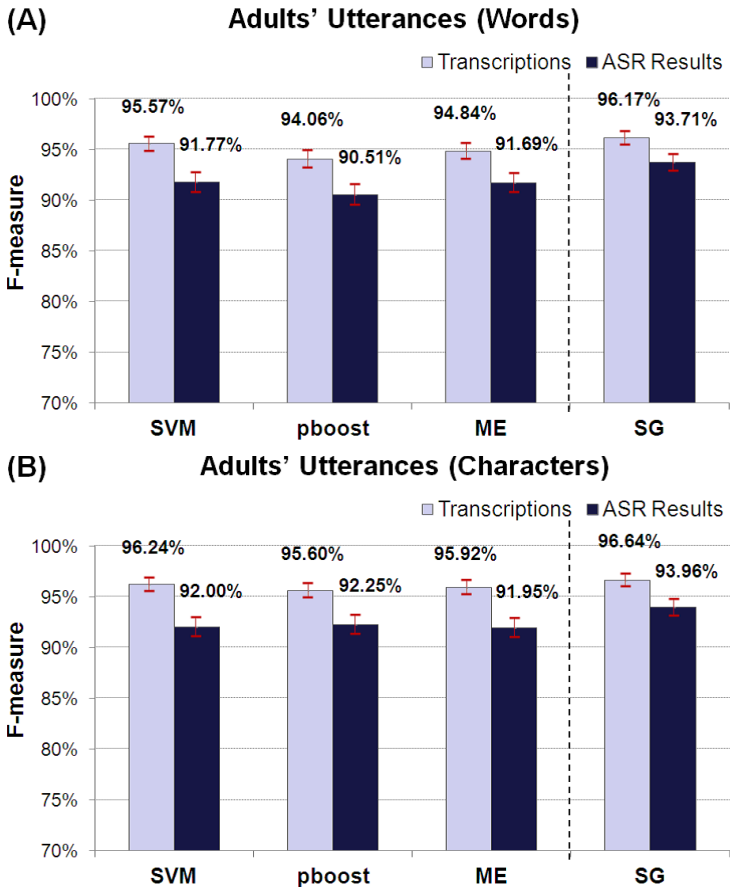


Fig. 6 F-measure for each method for transcriptions and ASR results for adults' utterances using (A) word and (B) character features. The F-measure for each method is indicated above the bars in bold, and the red line segments represent 95% confidence intervals.

same optimal hyperparameters that we found for our datasets may not be suitable for new datasets, and the hyperparameters must be tuned. ME does not have this problem since there are no hyperparameters that need to be tuned.

4.3 Effects of ASR Performance

The ASR word correct rates for children's utterances are considerably lower than those for adults which is reflected in the lower topic classification performance in the ASR results for children's inquiries. This was not evident when classifying their manual transcriptions. At the same time, the SG scheme exhibited higher performance improvements for children's utterances.

A comparison between the performance of the SG scheme and word correct rates for ASR results of children's and adults' utterances is presented in **Fig. 7**. The graphs show a tendency to obtain better classification performance as word correct rates for ASR results increase. The proportion of utterances with a word correct rate below 60% is 32.9% for children, and for adults is 15.1%; and the difference in classification performance between children and adults is evident. However, for word correct rates above 60%, the classification performances between children and adults are closer. Although some performance improvements were obtained with character features in comparison to words, this trend is not consistent.

An analysis of the performance of individual classifiers in comparison to ASR word correct rates indicated that pboost is more affected by ASR errors than SVM and ME. This is mainly because pboost uses subsequence patterns for classification, and correct recognition is important.

4.4 Word vs. Character Features

Since kanji characters also include meanings, the use of characters as features for classification of short utterances in Japanese augments the amount of available information, and hence it can help to deal with the sparseness of features present in spontaneous speech. **Figure 8** shows a comparison between the performance of the SG scheme using words or character features and the number of words per utterance. Although the use of characters yields higher classification performance in some cases, the tendency is not consistent, and the differences were not found to be significant.

5. Conclusions

In this work, we addressed the topic classification of spoken inquiries in Japanese that are received by a speech-oriented guidance system operating in a real environment. We compared the performance of SVM with an RBF kernel, pboost and ME, which are supervised learning methods, and an SG scheme to combine their predictions in a second-level classification using SVM with an RBF kernel. We evaluated the effect of using word or character features. Using characters as features yielded higher classification performances in some cases. Experimental results showed an

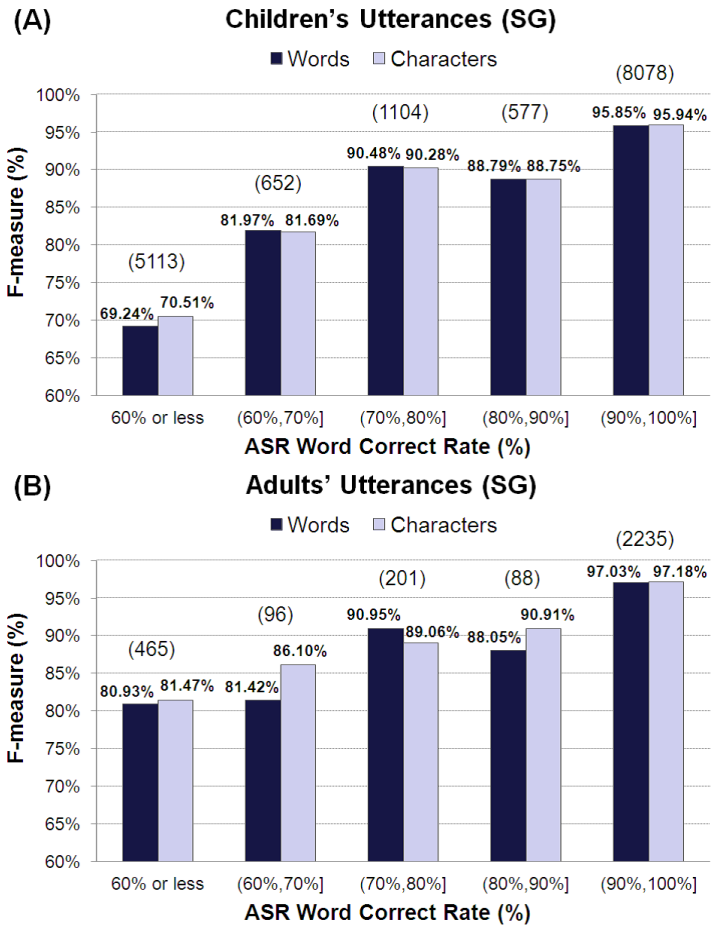


Fig. 7 F-measure of the SG scheme by showing word correct rates for ASR of (A) children's and (B) adults' utterances using word or character features (open test). Numbers of utterances are indicated above the bars inside parentheses. The F-measure for the SG scheme is also indicated above the bars in bold.

F-measure of 86.87% for the classification of ASR results from children's inquiries, with an average performance improvement of 2.81% compared with the performance of individual classifiers, and an F-measure of 93.96% with an average improvement of 1.89% for adults' inquiries when using the SG scheme and character features.

Future work will be focused on improving ASR performance of children's utterances as well as experiments combining characters and words as features in order to improve topic classification performance. Since manual data labeling, which is required for supervised learning, is a costly process and unlabeled data are usually abundant and cheap to obtain, we are also interested in the investigation of semi-supervised learning methods to improve topic classification performance.

References

[1] Nishimura, R., Lee, A., Saruwatari, H. and Shikano, K.: Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability, *Proc. ICASSP 2004*, pp.433–436 (2004).

[2] Gorin, A.L., Riccardi, G. and Wright, J.H.: How May I Help You?, *Speech Communication*, Vol.23, No.1-2, pp.113–127 (1997).

[3] Park, Y., Teiken, W. and Gates, S.: Low-Cost Call Type Classification for Contact Center Calls Using Partial Transcripts, *Proc. Interspeech 2009*, pp.2739–2742 (2009).

[4] Evanini, K., Suendermann, D. and Pieraccini, R.: Call Classification for Automated Troubleshooting on Large Corpora, *Proc. ASRU 2007*, pp.207–212 (2007).

[5] Suzuki, J., Sasaki, Y. and Maeda, E.: SVM Answer Selection for Open-Domain Question Answering, *Proc. COLING 2002*, pp.974–980 (2002).

[6] Mizuno, J., Akiba, T., Fujii, A. and Itou, K.: Non-factoid Question Answering Experiments at NTCIR-6: Towards Answer Type Detection for Realworld Questions, *Proc. NTCIR-6 Workshop Meeting*, pp.487–492 (2007).

[7] He, X., Yan, J., Ma, J., Liu, N. and Chen, Z.: Query Topic Detection for Reformulation, *Proc. WWW 2007*, pp.1187–1188 (2007).

[8] Murata, M., Uchimoto, K., Utiyama, M., Ma, Q., Nishimura, R., Watanabe, Y., Doi, K. and Torisawa, K.: Using the Maximum Entropy Method for Natural Language Processing: Category Estimation, Feature Extraction, and Error Correction, *Cognitive Computation*, Vol.2, No.4, pp.272–279 (2010). Software available at http://www.nict.go.jp/x/x161/members/mutiyama/software.html.

[9] Gupta, N., Tur, G., Hakkani-Tür, D., Bangalore, S., Riccardi, G. and Gilbert, M.: The AT&T Spoken Language Understanding System, *IEEE Trans. Audio, Speech and Language Processing*, Vol.14, No.1, pp.213–222 (2006).

[10] Lane, I., Kawahara, T., Matsui, T. and Nakamura, S.: Out-of-Domain Utterance Detection using Classification Confidences of Multiple Topics, *IEEE Trans. Speech and Audio Processing*, Vol.15, No.1, pp.150–161 (2007).

[11] Nowozin, S., Bakir, G. and Tsuda, K.: Discriminative Subsequence Mining for Action Classification, *Proc. ICCV 2007*, pp.1919–1923 (2007). Software available at http://www.nowozin.net/sebastian/pboost.

[12] Wolpert, D.: Stacked Generalization, *Neural Networks*, Vol.5, No.2, pp.241–260 (1992).

[13] Ting, K. and Witten, I.: Issues in Stacked Generalization, *Journal of Artificial Intelligence Research*, Vol.10, No.1, pp.271–289 (1999).

[14] Sigletos, G., Paliouras, G., Spyropoulos, C. and Hatzopoulos, M.: Combining Information Extraction Systems Using Voting and Stacked Generalization, *Journal of Machine Learning Research*, Vol.6, pp.1751–1782 (2005).

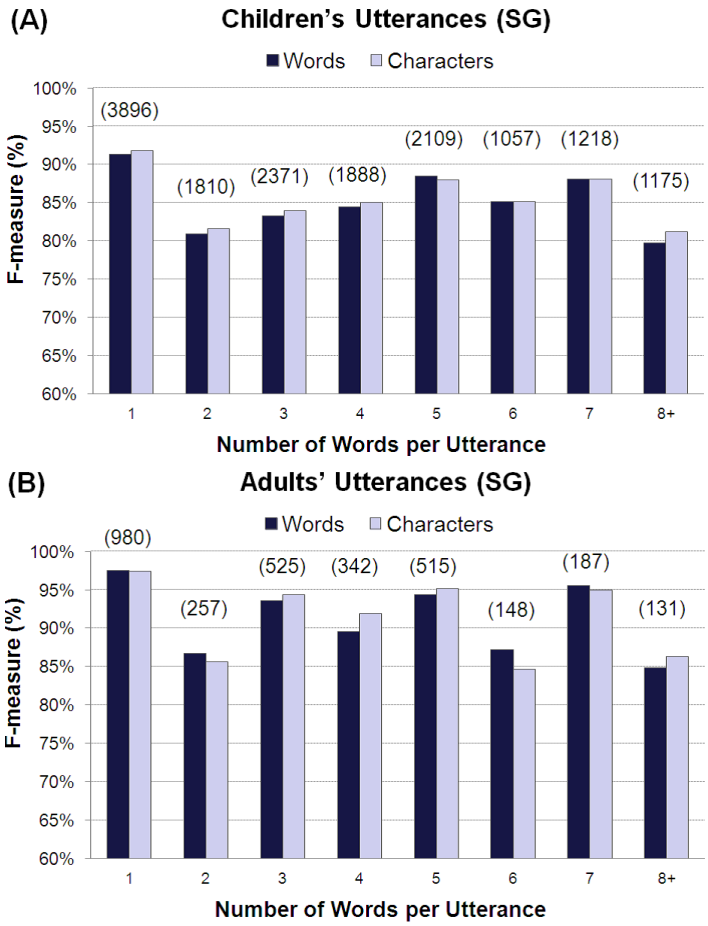


Fig. 8 F-measure of the SG scheme by showing number of words per utterance of (A) children's and (B) adults' utterances using word or character features (open test). Number of utterances are indicated above the bars inside parentheses.

[15] Halatci, I., Brooks, C. and Iagnemma, K.: Terrain Classification and Classifier Fusion for Planetary Exploration Rovers, *Proc. Aerospace Conference 2007*, pp.1–11 (2007).

[16] Akiba, T., Aikawa, K., Itoh, Y., Kawahara, T., Nanjo, H., Nishizaki, H., Yasuda, N., Yamashita, Y. and Itou, K.: Construction of a Test Collection for Spoken Document Retrieval from Lecture Audio Data, *IPSJ Journal*, Vol.50, No.2, pp.501–513 (2009).

[17] Shigeyasu, K., Nanjo, H. and Yoshimi, T.: A Study of Indexing Units for Japanese Spoken Document Retrieval, *Proc. WESPAC X* (2009).

[18] Young, S.J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. and Woodland, P.: *The HTK Book Version 3.4*, Cambridge University Press (2006). Software available at http://htk.eng.cam.ac.uk.

[19] Lee, A., Kawahara, T., Takeda, K. and Shikano, K.: A New Phonetic Tied-Mixture Model for Efficient Decoding, *Proc. ICASSP 2000*, pp.1269–1272 (2000).

[20] Stolcke, A.: SRILM — An Extensible Language Modeling Toolkit, *Proc. ICSLP 2002*, pp.901–904 (2002). Software available at http://www.speech.sri.com/projects/srilm.

[21] Lee, A., Kawahara, T. and Shikano, K.: Julius — An Open Source Real-Time Large Vocabulary Recognition Engine, *Proc. Interspeech 2001*, pp.1691–1694 (2001). Software available at http://julius.sourceforge.jp.

[22] Lane, I., Kawahara, T. and Matsui, T.: Language Model Switching Based on Topic Detection for Dialog Speech Recognition, *Proc. ICASSP 2003*, pp.I–616–I–619 (2003).

[23] Chang, C.-C. and Lin, C.-J.: LIBSVM: A Library for Support Vector Machines, *ACM Trans. Intelligent Systems and Technology*, Vol.2, No.3, pp.27:1–27:27 (2011). Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

[24] Wu, T.-F., Lin, C.-J. and Weng, R.C.: Probability Estimates for Multi-class Classification by Pairwise Coupling, *The Journal of Machine Learning Research*, Vol.5, No.1, pp.975–1005 (2004).

[25] Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U. and Hsu, M.-C.: Mining Sequential Patterns by Pattern Growth: The PrefixSpan Approach, *IEEE Trans. Knowledge and Data Engineering*, Vol.16, No.10, pp.1424–1440 (2004).

[26] Berger, A.L., Pietra, S.A.D. and Pietra, V.J.D.: A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistics*, Vol.22, No.1, pp.39–71 (1996).

[27] Zhu, C., Byrd, R.H., Lu, P. and Nocedal, J.: Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization, *ACM Trans. Mathematical Software*, Vol.23, No.4, pp.550–560 (1997).

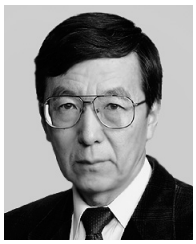
[28] Kazama, J. and Tsujii, J.: Evaluation and Extension of Maximum Entropy Models with Inequality Constraints, *Proc. EMNLP 2003*, pp.137–144 (2003).

Rafael Torres received his B.S. degree in computer systems engineering from the Technological University of Panama, Panama, in 2005; and M.E. from the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, in 2010, where he is currently a Ph.D. student sponsored through an honors scholarship by the Japan Ministry of Education, Culture, Sports, Science and Technology (MEXT). His research interests include automatic speech recognition and spoken language understanding in dialogue systems. He is a student member of ASJ and IEEE.

Hiromichi Kawanami received his B.E. degree in electrical engineering in 1994, and M.E. and Ph.D. degrees in information and communication engineering from the University of Tokyo in 1997 and 2000, respectively. During 2000–2001, he joined the Electrotechnical Laboratory (National Institute of Advanced Industrial Science and Technology). Since 2001, he has served as an Assistant Professor at the Nara Institute of Science and Technology (NAIST). His research interests are spoken dialogue systems and speech analysis. He is a member of IEICE and ASJ.

Tomoko Matsui received her Ph.D. degree from the Computer Science Department, Tokyo Institute of Technology, Tokyo, Japan, in 1997. From 1988 to 2002, she was with NTT, where she worked on speaker and speech recognition. From 1998 to 2002, she was with the Spoken Language Translation Research Laboratory, ATR, Kyoto, Japan, as a Senior Researcher and worked on speech recognition. From January to June 2001, she was an Invited Researcher at the Acoustic and Speech Research Department, Bell Laboratories, Murray Hill, NJ, working on finding effective confidence measures for verifying speech recognition results. She joined the Institute of Statistical Mathematics, Tokyo in 2003 as an Associate Professor and has been a Professor since 2008, working on statistical modeling for speech and speaker recognition applications. She received the Paper Award of IEICE in 1993.

Hiroshi Saruwatari was born in Nagoya, Japan, on July 27, 1967. He received his B.E., M.E., and Ph.D. degrees from Nagoya University, Nagoya, Japan, in 1991, 1993, and 2000, respectively. He joined the Intelligent System Laboratory, SECOM Co., Ltd., Tokyo, Japan, in 1993, where he engaged in research on the ultrasonic array system for the acoustic imaging. He is currently an Associate Professor at the Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan. His research interests include speech signal processing, array signal processing, blind source separation, and sound field reproduction. Dr. Saruwatari received Paper Awards from IEICE in 2001 and 2006, from the Telecommunications Advancement Foundation in 2004 and 2009, and at the IEEE-IROS2005 in 2006. He won the first prize at the IEEE MLSP2007 Data Analysis Competition for BSS. He is a member of IEEE, IEICE, the Virtual Reality Society of Japan, and ASJ.



Kiyohiro Shikano received his B.S., M.S., and Ph.D. degrees in electrical engineering from Nagoya University in 1970, 1972, and 1980, respectively. He is currently a Professor of Nara Institute of Science and Technology (NAIST), where he is directing the Speech and Acoustics Laboratory. From 1972 to 1993, he had been working at NTT Laboratories. During 1986–1990, he was the Head of Speech Processing Department at ATR Interpreting Telephony Research Laboratories. During 1984–1986, he was a visiting scientist at Carnegie Mellon University. He received the Yonezawa Prize from IEICE in 1975, the Signal Processing Society 1990 Senior Award from IEEE in 1991, the Technical Development Award from ASJ in 1994, IPSJ Yamashita SIG Research Award in 2000, and Paper Award from the Virtual Reality Society of Japan in 2001, IEICE Paper Award in 2005 and 2006, and Inose award in 2005. He is a fellow of IEEE, IEICE, and IPSJ, and a member of ASJ, and ISCA.