NAIST-IS-DD1061031

Doctoral Dissertation

Comparison of Topic Classification Methods for Spoken Inquiries

Rafael Antonio Torres Rodriguez

March 22, 2013

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Rafael Antonio Torres Rodriguez

Thesis Committee:

Professor Kiyohiro Shikano	(Supervisor)
Professor Yuji Matsumoto	(Co-supervisor)
Professor Tomoko Matsui	(ISM, Co-supervisor)
Associate Professor Hiroshi Saruwatari	(Co-supervisor)
Assistant Professor Hiromichi Kawanami	(Co-supervisor)

Comparison of Topic Classification Methods for Spoken Inquiries*

Rafael Antonio Torres Rodriguez

Abstract

One of the most natural means for social interaction among humans is speech. Automatic speech recognition (ASR) technologies have made feasible the usage of speech as an interface for human-machine interaction. As a result it has been applied to telephone-based services, smartphone applications, guidance systems, car navigation systems, and others; aiming to provide a more natural interaction.

Topics group inquiries that are related by sharing a common subject. The classification of spoken inquiries into topics is useful to manage the interaction with users by reducing the range of possible responses and for dialog management. However, topic classification of spoken inquiries is often hindered by ASR errors, sparseness of features and phenomena peculiar to spontaneous speech.

This work addresses the topic classification of spoken inquiries in Japanese by comparing the performances of three supervised learning methods with different characteristics: support vector machine (SVM) with a radial basis function (RBF) kernel, PrefixSpan boosting (pboost) and the maximum entropy (ME) method. SVM robustly finds boundaries among topics 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 estimates probability distributions from data and allows multi-class classification.

An evaluation using words or characters as features for the classifiers is also performed. Using characters as features is possible in Japanese owing to the

*Doctoral Dissertation, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1061031, March 22, 2013.

presence of kanji, ideograms originating from Chinese characters that represent not only sound but also meaning.

The differences among the three classifiers allow them to compensate each other's performance. Because of this, the usage of a stacked generalization (SG) scheme that combines their predictions to achieve greater classification performance is proposed.

An analysis on the performance of the above methods and their combination in the topic classification of spoken inquiries from a speech-oriented guidance system operating in a real environment was carried out. Experimental results show that the three methods individually produce some prediction errors that do not overlap, and that the SG scheme improves the topic classification performance by correcting some of them. There was an F-measure of 86.87% for the classification of ASR results from children's inquiries with an average performance improvement of 2.81%, 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, spoken inquiry, support vector machine, PrefixSpan boosting, maximum entropy, stacked generalization

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General Background and Problem Definition

Speech is one of the most important and natural means for social interaction among humans. We learn to speak at an early age, and to write a little bit later in life, so we usually prefer communication through speech to other means of communication. However, speaking is a very complex act. Our speaking skills start to develop when we are children and continuously improve as we grow. We still keep learning new words and expressions through interaction with other people even when we have already reached maturity, therefore it is a very dynamic process.

Human-machine interaction has historically required devices such as control panels and keyboards, which require people to adapt to them. However, improvements in automatic speech recognition (ASR) technologies have made feasible the usage of speech as an interface for human-machine interaction. As a result it has been applied to telephone-based services [1, 2, 3], guidance systems [4, 5, 6], call center automation [7, 8], car navigation systems, smartphone applications, video games, and others; aiming to provide a more natural interaction (Fig. 1). In recent years, the wide availability of smartphones has brought these technologies closer to people in the form of personal assistant applications like Apple's Siri and applications for Voice Search. The usage of ASR technologies in video games has also opened new possibilities for interaction which enriches the experience.

Speech as an interface for human-machine interaction has many advantages, including that nearly all of us can speak without additional training, we can have our hands and eyes free to perform other tasks or still be able to operate a machine if they are impaired, and we can have a more fluent interaction since we can talk faster than we type. Nevertheless, the difficulties of human-machine interaction through speech have been observed since the first spoken language technologies started to be developed and are still matter of discussion [9, 10]. Spontaneous speech includes jargon, slangs, ungrammatical constructions, mispronounced words, and disfluencies such as filled pauses, fillers, false starts,

Figure 1. Examples of human-machine interactions through speech.

repetitions and repairs, among other issues.

Topic classification of speech is a subject of interest in spoken language processing because of its several applications. It has been studied in the field of telephone call classification to optimize call routing [1, 11] and to resolve call type or call reason in contact centers [8, 7]. The "How May I Help You?" (HMIHY) [1] system from AT&T automatically routes telephone calls to appropriate destinations in a telecommunications environment based on a user's spoken response to the prompt "How may I help you?" The system uses a dialog strategy to determine the call type, classifying the speech into one of fifteen possible categories using a statistical classifier that uses salient grammar fragments as features. A study on call type classification in the context of contact centers is presented in [8], and the particularity in this case is that they tried to classify human-to-human conversations in free format. These studies are similar to the research presented in this work since they also deal with speech. However, this research deals with topic classification of spoken inquiries in the context of an information guidance system, where utterances are shorter and their features are sparse.

Other uses of topic classification of speech include the improvement of ASR performance by detecting the topic of a user's utterance and then performing again the speech recognition, applying an appropriate topic-dependent language model [12], and the detection of out-of-domain (OOD) utterances [13].

Topic classification of spoken inquiries can also be used to ease the answer selection from a high number of possible answers in an information guidance system, where a topic would group inquiries that are related by sharing a common subject. This approach is frequently used in text-based information retrieval (IR) [14, 15, 16], classifying a text inquiry in a topic and then selecting an answer only from the possibilities included in that topic. Other studies on topic detection and estimation in text-based IR are presented in [17, 18]. The research presented in this work, however, deals with spoken inquiries instead of text, whose classification performances are often hindered by ASR errors and phenomena peculiar to spontaneous speech. Although incorporating grammatical information as features has proven to yield high classification performances in text-based IR, in this work we decided to focus on evaluating the effect of using words or characters as features due to the shortness of the utterances.

1.2 Scope of This Work

The research presented in this work aims to improve topic classification performance of spoken inquiries in Japanese received by a speech-oriented guidance system operating in a real environment. For this three different types of classification methods were selected, (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 their performances were compared.

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, 8, 11, 12, 13, 15, 16]. The pboost method is for classification of sequential data, and it extracts and utilizes discriminative and sequential patterns in the data [19]. Although the method has been developed for classification of actions in videos, in this work pboost is introduced 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 (NLP) tasks [18] as well as in speech classification [7, 15].

Moreover, the predictions from the above different types of methods were

combined in this work by using a stacked generalization (SG) [20] scheme and the complementary effect was examined. The SG scheme and similar schemes have also been studied as a means of combining classifier predictions in other classification tasks [21, 22, 23, 24].

An evaluation using words or characters as features for the classifiers was also performed. 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 [25, 26], and better performance was obtained when using words than when using characters. However, spoken inquiries in this topic classification task are much shorter than spoken documents; hence this evaluation is also of interest.

An analysis on the performance of the above methods and their combination in the topic classification of spoken inquiries was carried out. Prediction errors of each method were analyzed in order to determine their overlap, and to observe how many of the errors that did not overlap were able to be corrected by the SG scheme, as well as how many correct predictions were misclassified by it. The influence of ASR performance in the topic classification was also analyzed.

The experiments, evaluations and analysis were carried out using data obtained from a speech-oriented guidance system that operates in a real environment. The guidance system is the *Takemaru-kun* system [5], and it operates in a public facility receiving daily user requests for information and collecting 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 for the information they want to obtain. When the system started collecting user's inquiries, they were analyzed and manually labeled to define its task domain. Therefore, the results of the analysis and evaluations presented in this work are expected to be applicable to other task domains for this type of systems.

1.3 Thesis Overview

This thesis is organized as follows.

In Chapter 2, the three supervised classification methods that were selected for

comparison are explained. Section 2.1 presents the SVM-based method, Section 2.2 the pboost-based method, and Section 2.3 the ME-based method, finalizing with Section 2.4 which presents a summary of the chapter.

In Chapter 3, the proposed combination of methods using an SG scheme is explained. In this chapter the SG algorithm is detailed, including the training and test procedures and the features that are used. By the end of the chapter a summary is also provided.

In Chapter 4, the details of the datasets used in the experiments, analysis and evaluations are presented. This chapter also includes an overview of the speech-oriented guidance system *Takemaru-kun*, and also includes a summary by the end of the chapter.

In Chapter 5, the performed experiments, results, analysis and evaluations are presented. First, the experimental setup is explained. Then, an analysis of prediction error overlaps is presented. After this, the performance comparison among methods and their combinations is presented with its corresponding analysis, including an analysis of the effects of ASR performance and an evaluation of the differences in performances when using words and character features. This chapter finalizes with the conclusions that were derived from the experimental results.

In Chapter 6, the conclusion of the thesis is presented, including a summary of the thesis and the future work.

Chapter 2

Topic Classification Methods

SVM, pboost and ME were selected for comparison because of their different characteristics. We selected SVM and ME because even though they are different, they have presented very competitive performances in different classification tasks. Pboost was developed for classification of actions in videos, and in this work we introduce it for the classification of spoken inquiries into topics.

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 and pboost classify by maximizing the separation margin between two classes; however, SVM deals 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, while pboost does not. Pboost performs feature selection and classifies by checking for the presence of optimal discriminative subsequence patterns in the input, while SVM and ME do not perform feature selection. ME is a method that estimates probability distributions from data, and is a multi-class classifier by nature; while SVM and pboost do not estimate probabilities, and need to make use of approaches like one-vs-one or one-vs-rest for multi-class classification. ME also has the advantage that it is not sensitive to hyperparameter settings, in contrast to the other two classifiers.

This chapter explains the details of the methods that we compare in this thesis.

2.1 Support Vector Machine-Based Method

Support Vector Machine (SVM) maximizes the margin of classification of two different classes of data, robustly detecting boundaries between them. SVM deals 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, 8, 11, 12, 13, 15, 16].

2.1.1 C-Support Vector Classification

C-support vector classification (C-SVC) [27, 28, 29] implements soft-margin and solves the following primal problem:

$$\begin{aligned} \min_{\vec{w},b,\vec{\xi}} \quad & \frac{1}{2}\vec{w}^T\vec{w} + C\sum_{i=1}^{l}\xi_i \\ \text{sb.t.} \quad & y_i(\vec{w}^T\phi(\vec{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, ..., l, \end{aligned} \tag{1}$$

where $\vec{x}_i \in \mathbb{R}^n, i = 1, ..., l$ indicates a training vector, $y_i \in \{1, -1\}$ a class, and ϕ is the function for mapping the training vectors into feature space. The hyper-parameter C penalizes the sum of the slack variable ξ_i, that allows the margin constraints to be slightly violated to reduce the influence of outliers.

The dual form of the problem is:

$$\begin{aligned} \min_{\vec{\alpha}} \quad & \frac{1}{2}\vec{\alpha}^T Q\vec{\alpha} - \vec{e}^T\vec{\alpha} \\ \text{sb.t.} \quad & \vec{y}^T\vec{\alpha} = 0, \\ & 0 \leq \alpha_i \leq C, i = 1, ..., l, \end{aligned} \tag{2}$$

where $\vec{e}$ is the vector of all ones, $C > 0$ is the upper bound, and Q is an l by l positive semidefinite matrix, $Q_{ij} \equiv y_i y_j \kappa(\vec{x}_i, \vec{x}_j)$, where $\kappa(\vec{x}_i, \vec{x}_j)$ is the kernel.

The decision function is:

$$sgn\left(\sum_{i=1}^{l} y_i\alpha_i\kappa(\vec{x}_i, \vec{x}) + b\right). \tag{3}$$

A sample vector is classified in the positive or the negative class according to the sign, which indicates the side of the hyperplane where the sample is located.

2.1.2 Soft-Margin Approach for Unbalanced Data

In the topic classification task presented in this thesis the number of utterances for each topic is unbalanced. When the training data is unbalanced, SVM parameters are not estimated robustly. C-support vector classification (C-SVC) with soft

margin for unbalanced data is used to deal with this problem. The SVM primal problem formulation implementing soft-margin for unbalanced amount of samples follows the form:

$$\begin{aligned} \min_{\vec{w},b,\vec{\xi}} \quad & \frac{1}{2}\vec{w}^T\vec{w} + C_+ \sum_{\{i:y_i=+1\}} \xi_i + C_- \sum_{\{i:y_i=-1\}} \xi_i \\ \text{sb.t.} \quad & y_i(\vec{w}^T\phi(\vec{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, ..., l \end{aligned} \tag{4}$$

where $\vec{x}_i \in \mathbb{R}^n, i = 1, ..., l$ indicates a training vector, $y_i \in \{1, -1\}$ a class, and ϕ is the function for mapping the training vectors into feature space. The hyper-parameters C_+ and C_- penalize the sum of the slack variable ξ_i for each class, that allows the margin constraints to be slightly violated. By introducing different hyperparameters C_+ and C_-, the unbalanced amount of data problem, in which SVM parameters are not estimated robustly due to unbalanced amount of training vectors for each class, can be dealt with.

The dual form of the problem is:

$$\begin{aligned} \min_{\vec{\alpha}} \quad & \frac{1}{2}\vec{\alpha}^T Q\vec{\alpha} - \vec{e}^T\vec{\alpha} \\ \text{sb.t.} \quad & 0 \leq \alpha_i \leq C_+, \text{ if } y_i = 1, \\ & 0 \leq \alpha_i \leq C_-, \text{ if } y_i = -1, \\ & \vec{y}^T\alpha = 0, \end{aligned} \tag{5}$$

where $\vec{e}$ is the vector of all ones, $C > 0$ is the upper bound, and Q is an l by l positive semidefinite matrix, $Q_{ij} \equiv y_i y_j \kappa(\vec{x}_i, \vec{x}_j)$, where $\kappa(\vec{x}_i, \vec{x}_j)$ is the kernel.

2.1.3 Bag-of-Words Feature Representation

The bag-of-words (BOW) vector space model was used to represent utterances as vectors, where each component of the vector indicates the frequency of appearance of a feature. A bag is a set in which repeated elements are allowed, so that not only the presence of a feature but also its frequency is taken into account [30]. The length of a vector corresponds to the size of the dictionary that includes every feature in the training dataset. BOW does not take in consideration grammar or word order. If a feature has a high frequency of appearance in a sample vector in

comparison to the rest it may undesirably influence the importance of the other features, overshadowing them. Horizontal scaling of the vectors was used to deal with this problem. The feature's frequencies were scaled from 0 to 1 on each sample vector.

2.1.4 Kernel Function

A kernel function performs the computation of inner products as a direct function of the input features, without explicitly computing the mapping ϕ that aims at converting nonlinear relations into linear ones. The objective is to find a mapping such that, in the new space, problem solving is easier. They make possible the use of feature spaces with an exponential or even infinite number of dimensions [30].

As the concept of a kernel is formulated as an inner product in a feature space, if we have an algorithm formulated in such a way that the input vector $\vec{x}$ enters only in the form of scalar products, the kernel trick, also known as kernel substitution, allows to replace that scalar product with some other choice of kernel [31]. As SVM uses kernel functions as input, it also benefits from the kernel trick.

In the approach presented in this thesis the RBF kernel was used because in preliminary experiments it exhibited better performance than a linear kernel and slightly better performance than a polynomial kernel for this task. The RBF kernel is defined as

$$\kappa(\vec{x}_i, \vec{x}_j) = \exp(-\gamma||\vec{x}_i - \vec{x}_j||^2), \gamma > 0 \tag{6}$$

where $\vec{x}_i$ and $\vec{x}_j$ represent utterance vectors and $\gamma > 0$ is a hyperparameter of the function.

2.1.5 One-vs-Rest Multi-Class Classification

SVM is originally a binary classifier. For multi-class classification the one-vs-rest approach was selected, 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. This approach was selected because in

preliminary experiments it had better performance than the one-vs-one approach for this task.

Although SVM can only predict the topic label and not probability information, the method described in [32] can be used to obtain probability estimates or pseudo-probabilities for each topic. This method was used to classify new data in the topic with highest pseudo-probability.

Given k topics, for any sample $\vec{x}$ and topic label y, the goal is to estimate

$$p_i = p(y = i|\vec{x}), i = 1, ..., k. \tag{7}$$

A probability estimate p_i of a sample for a category i is calculated using the decision value $\hat{f}$ obtained in (3) without using the sign operator. A probability estimate p_i is then calculated by applying a sigmoid function to the decision value $\hat{f}$:

$$p_i \approx \frac{1}{1 + e^{A\hat{f}+B}}, \tag{8}$$

where A and B are estimated by minimizing the negative log-likelihood function using known training data and their decision values $\hat{f}$. Labels and decision values are required to be independent, so five-fold cross-validation is conducted to obtain the those decision values [29, 32].

2.2 PrefixSpan Boosting-Based Method

PrefixSpan Boosting (pboost) is a method proposed by Nozowin et al. [19] for the classification of actions in videos. In this work pboost is introduced for the classification of spoken inquiries into topics. Pboost implements a generalization of the PrefixSpan algorithm by Pei et al. [33] 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 [3, 7].

Pboost uses the PrefixSpan algorithm [33] to find optimal subsequence patterns that characterize utterances from a specific topic. For example, in the topic

info-facility we can find 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 seen from this example, subsequences can also include gaps.

2.2.1 LPBoost-Based Classifier

The idea of boosting classifiers is to combine multiple weak classifiers into a powerful composite classifier. Pboost classification is based on Linear Programing Boosting (LPBoost), which is a supervised binary classifier form the boosting family, which maximizes a margin between training samples of two different classes, and therefore it belongs to the class of margin-maximizing supervised classification algorithms, as SVM.

In pboost, 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, $\vec{x} \in \{\vec{x}_i\}$, $\vec{x}_i \in \mathbb{R}^n, i = 1, ..., l$ 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

$$f(\vec{x}) = \sum_{(\vec{s},\omega)\in\vec{S}\times\Omega} \alpha_{\vec{s},\omega} h(\vec{x}; \vec{s}, \omega) \tag{9}$$

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.

The primal form of the training problem is:

$$\begin{aligned} \min_{\vec{\alpha},\xi,\rho} \quad & -\rho + D\sum_{i=1}^{\ell} \xi_i \\ \text{sb.t.} \quad & \sum_{(\vec{s},\omega)\in\vec{S}\times\Omega} y_i \alpha_{\vec{s},\omega} h(\vec{x}_i; \vec{s}, \omega) + \xi_i \geq \rho, i = 1, ..., l \\ & \sum_{(\vec{s},\omega)\in\vec{S}\times\Omega} \alpha_{\vec{s},\omega} = 1, \vec{\alpha} \geq 0, \vec{\xi} \geq 0, \end{aligned} \tag{10}$$

where $\vec{x}_i \in \mathbb{R}^n, i = 1, ..., l$ indicates a training vector, $y_i \in \{1, -1\}$ a class, ρ is the soft margin separating negative from positive samples, $D = \frac{1}{\nu\ell}$ and $\nu \in (0, 1)$ is a hyperparameter controlling the cost of misclassification.

It is not feasible to solve this optimization problem directly, due to the large number of variables in α. Instead, the equivalent dual form of the problem is solved, which takes the form:

$$\begin{aligned} \min_{\vec{\lambda},\gamma} \quad & \gamma \\ \text{sb.t.} \quad & \sum_{i=1}^{l} \lambda_i y_i h(\vec{x}_i; \vec{s}, \omega) \leq \gamma, (\vec{s}, \omega) \in \vec{\bar{S}} \times \Omega \\ & \sum_{i=1}^{l} \lambda_i = 1, 0 \leq \lambda_i \leq D, i = 1, ..., l. \end{aligned} \tag{11}$$

The primal solution of $\vec{\alpha}$ is obtained from the Lagrange multipliers. The dual problem has a limited amount of variables, however the amount of constraints is very large. This problem is solved using a constraint generation technique, which starts with an empty hypothesis set, and adds iteratively the hypothesis whose constraint (12) is violated the most. Each time a hypothesis is added, the optimal solution is updated by solving the restricted dual problem. This method optimizes the classifier and performs feature selection simultaneously. In each iteration, the following problem is solved to find an optimal hypothesis:

$$(\vec{\hat{s}}, \hat{\omega}) = \text{argmax}_{(\vec{s},\omega) \in \vec{\bar{S}} \times \Omega} g(\vec{s}, \omega), \tag{12}$$

where the gain function is defined as

$$g(\vec{s}, \omega) = \sum_{i=1}^{l} \lambda_i y_i h(\vec{x}_i; \vec{s}, \omega). \tag{13}$$

The constraint generation algorithm terminates if there is no hypothesis violating the constraint (12).

2.2.2 Optimal Subsequence Pattern Search

This section describes how the maximum-gain search problem formulated in (12) is solved. This problem is difficult due to the size of the combinatorial space to be considered. This problem is solved by using a generalization of the PrefixSpan algorithm by Pei et al. [33], which is an algorithm to enumerate all frequent

subsequences. The problem consists on finding all the subsequences $\vec{s} \in \vec{S}$ whose occurrence is no less than a threshold:

$$\sum_{i=1}^{l} I(\vec{s} \subseteq \vec{x}_i) \geq \tau, \tag{14}$$

where τ is the threshold, called minimum support parameter.

A search tree is generated, starting from an empty root node. In the search tree, each child node contains a sequence that is an extension of its parent node's sequence. By defining an ordering in the sequences, duplicate sequences are not generated.

The gain function $g(\vec{s}, \omega)$ defined in (13) is used to calculate the gain of a subsequence.

In order to minimize the size of the explored search tree, tree pruning is essential. If a search tree is generated up to a pattern $\vec{s}$, and a gain g^* is the maximum gain among the ones observed so far, if we can guarantee that $g(\vec{s'}, \omega)$ is not larger than g^* for any extensions s' of s and any ω, we can prune the downstream nodes without losing the optimal pattern.

In the algorithm, a gain bound function $\mu(\vec{s})$ is defined as follows:

$$\mu(\vec{s}) = \max \left\{ 2 \sum_{\{i|y_i=+1, \vec{s} \subseteq \vec{x}_n\}} \lambda_i - \sum_{i=1}^{l} y_i \lambda_i,\ 2 \sum_{\{i|y_i=-1, \vec{s} \subseteq \vec{x}_n\}} \lambda_i + \sum_{i=1}^{l} y_i \lambda_i \right\}. \tag{15}$$

If the condition $g^* > \mu(\vec{s})$ is satisfied, the gain $g(\vec{s'}, \omega)$ of any downstream sequence $\vec{s'} \supset \vec{s}$ does not exceed the current best g^* for any $\omega \in \Omega$, and the downstream nodes can be pruned.

Essentially, the differences between the algorithm implemented in pboost and PrefixSpan are that, it finds the optimal patterns that maximizes a gain function instead of enumeration, and a gain bound μ is used for tree pruning. The algorithm recursively generates a subsequence search tree, and it keeps a variable g^* which contains the highest gain value observed so far, and is updated whenever a subsequence with higher gain value is observed. Tree pruning occurs if the pruning condition holds.

2.2.3 Soft-Margin Approach for Unbalanced Data

In the classification problem addressed in this work the amount of samples for each topic is unbalanced. An extended version of the method is used to deal with this. The primal problem formulation implementing a soft margin for an unbalanced number of samples follows the form:

$$\begin{aligned} \min_{\rho,\vec{\alpha},\vec{\xi}} \quad & -\rho + D_+ \sum_{\{i:y_i=+1\}} \xi_i + D_- \sum_{\{i:y_i=-1\}} \xi_i \\ \text{sb.t.} \quad & \sum_{(\vec{s},\omega)\in\vec{S}\times\Omega} y_i \alpha_{\vec{s},\omega} h(\vec{x}_i;\vec{s},\omega) + \xi_i \geq \rho, i = 1,...,l \\ & \sum_{(\vec{s},\omega)\in\vec{S}\times\Omega} \alpha_{\vec{s},\omega} = 1, \vec{\alpha} \geq 0, \vec{\xi} \geq 0 \end{aligned} \tag{16}$$

where $\vec{x}_i \in \mathbb{R}^n, i = 1,...,l$ 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.

The dual problem follows the form:

$$\begin{aligned} \min_{\vec{\lambda},\gamma} \quad & \gamma \\ \text{sb.t.} \quad & \sum_{i=1}^{l} \lambda_i y_i h(\vec{x}_i;\vec{s},\omega) \leq \gamma, (\vec{s},\omega) \in \vec{S}\times\Omega \\ & \sum_{i=1}^{l} \lambda_i = 1, \\ & 0 \leq \lambda_i \leq D_+, \text{ if } y_i = 1, \\ & 0 \leq \lambda_i \leq D_-, \text{ if } y_i = -1, \\ & i = 1,...,l. \end{aligned} \tag{17}$$

This problem is solved similarly as it was described in Section 2.2.1.

2.2.4 One-vs-Rest Multi-Class Classification

Here a one-vs-rest approach for multi-class classification is also used, 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. New data is classified in a topic according to the highest value of the classification function in (9).

2.3 Maximum Entropy-Based Method

ME is a supervised learning method that estimates probability distributions from data [34], 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 multi-class classifier by nature. ME has been widely used in natural language processing (NLP) tasks [18] as well as in speech classification [7, 15].

2.3.1 Maximum Entropy Model

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

$$\hat{k} = \underset{k \in K}{\operatorname{argmax}}\ p(k|c_1^N), \tag{18}$$

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

$$p(k|c_1^N) = \frac{\exp\left[\sum_c N(c) \log \alpha(k|c)\right]}{\sum_{k'} \exp\left[\sum_c N(c) \log \alpha(k'|c)\right]}. \tag{19}$$

Ignoring the terms that are constant with respect to k yields

$$\hat{k} = \underset{k \in K}{\operatorname{argmax}} \sum_c N(c) \log \alpha(k|c), \tag{20}$$

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 [35] which is a limited-memory algorithm for solving large nonlinear optimization problems.

2.4 Summary

In this chapter, an explanation about the supervised learning methods compared in this work was given, highlighting the characteristics and differences among them. First an explanation about the SVM-based method was presented, describing C-SVC which implements the soft-margin approach to improve generalization in spite of outliers, and the approach for dealing with unbalanced amount of training samples. A brief explanation about kernel functions and BOW representation was also given, as well as a description of the one-vs-rest multi-class classification approach. Then, an explanation about the pboost-based method was given, describing its LPBoost-based classifier and its optimal subsequence pattern search, the approach for dealing with unbalanced amounts of training samples and the one-vs-rest multi-class classification. Finally, an explanation about the ME-based method was given, indicating the model formulation.

Chapter 3

Combination of Methods

3.1 Stacked Generalization

Stacked Generalization (SG), proposed by Wolpert [20], is a method that combines the outputs of multiple classifiers using a second-level classification. In [21] SG was compared against voting, which does not use a second-level classification but takes in consideration the prediction of the majority of the classifiers, concluding that SG was consistently effective in the tested domains while voting was not. In [22], SG was effective for combining learning algorithms for the classification of datasets from the UCI repository of machine learning databases. In [23], a similar approach was used in a study of multi-sensor terrain classification for planetary rovers. In [24], SG was used in a collaborative filtering algorithm to predict user ratings for films. In this work, an SG scheme is proposed for the topic classification of spoken inquiries.

3.2 Compensation Effect

The objective of SG is to reduce the generalization error of first-level classifiers by achieving greater predictive accuracy in a second-level classification using predictions as input data. Its success arises from its ability to exploit the diversity in the predictions of first-level classifiers. Because of the differences in the classifiers we selected for comparison, we can expect them to compensate each other to improve prediction performance.

3.3 Training and Test Procedures

The training and test procedures in the SG scheme are illustrated in Fig. 2. 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

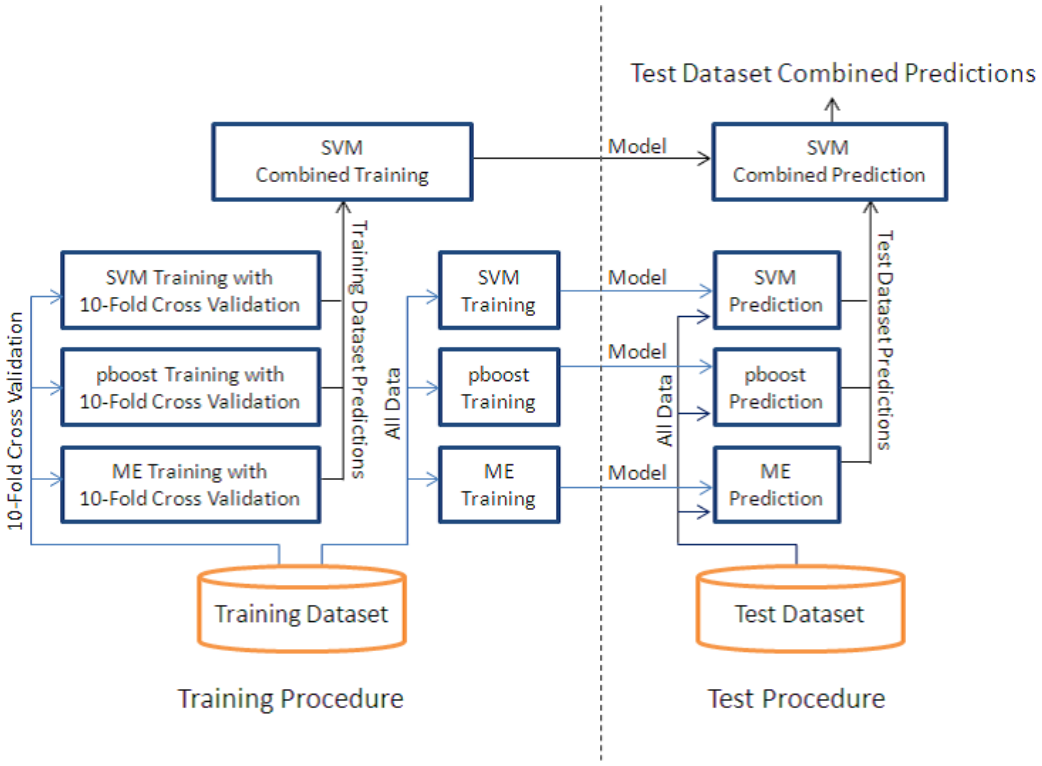

Figure 2. Training and test procedures in the SG scheme.

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.

3.4 Summary

In this chapter, an explanation of SG was given, indicating its training and testing procedures, the importance of cross-validation in the training of first-level models and the features that result from them, which are used as input data for the second-level classifier.

Chapter 4

Spoken Inquiries Datasets

The experiments, evaluations and analysis were carried out using data obtained from a speech-oriented guidance system that operates in a real environment. The guidance system is the *Takemaru-kun* system [5], and it operates in a public facility receiving daily user requests for information and collecting 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 for the information they want to obtain. When the system started collecting user's inquiries, they were analyzed and manually labeled to define its task domain. Therefore, the results of the analysis and evaluations presented in this work are expected to be applicable to other task domains for this type of systems.

4.1 Overview of the *Takemaru-kun* System

The *Takemaru-kun* system [5], shown in Fig. 3, 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

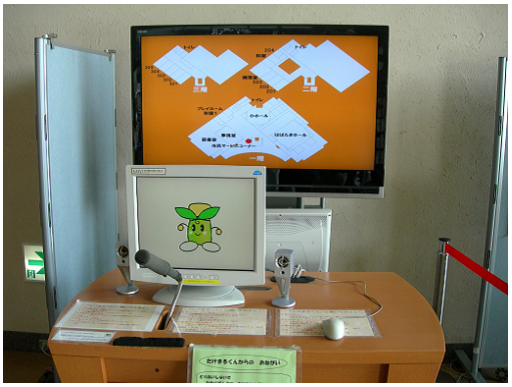

Figure 3. Speech-oriented guidance system *Takemaru-kun*.

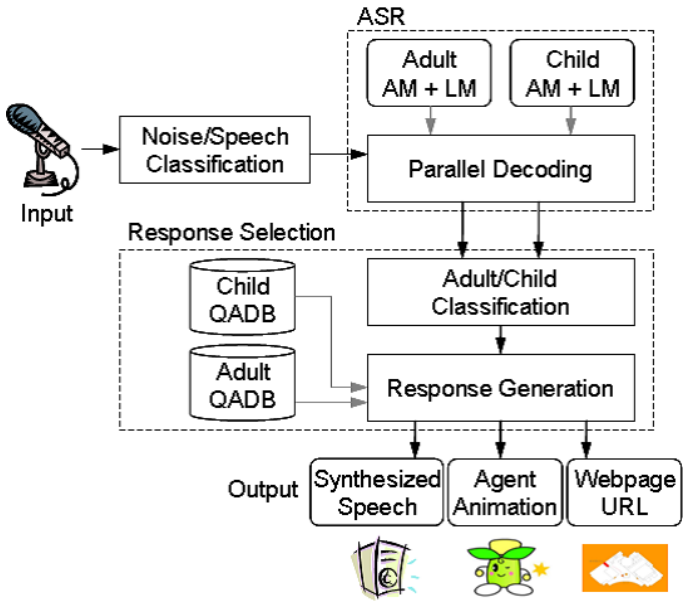

Figure 4. Block diagram of the main components of the *Takemaru-kun* system.

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 [36].

Figure 4 shows a block diagram of the main components and general process flow in the system. The microphone records an input, which is then analyzed to distinguish if it is valid speech or an invalid input, such as noise, level overflowed shouts, laughter, or coughing, among others. This is done by comparing acoustic likelihoods of the input given by Gaussian Mixture Models (GMMs) trained with valid and invalid data respectively, and selecting the highest one [37].

After that, the input is decoded using the large vocabulary continuous speech recognition (LVCSR) engine Julius [38], with two parallel speech recognition decoders, using acoustic models (AMs) and language models (LMs) with adults and children data respectively. The system determines if it is an utterance from an adult or a child on the basis of speech recognition logarithmic likelihood scores

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 [39]
Acoustic model	PTM [40], 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 [41]
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 [38]

from each recognized result [5], in order to answer accordingly.

4.2 Specifications of the 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. These utterances were also classified in heuristically defined topics, grouping inquiries that were related by sharing a common subject. Invalid inputs such as noise, coughs, laughter and unclear inputs were also documented. The signal-to-noise ratio (SNR) of the utterances recorded in this period is 38.31 dB. Some examples

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

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%

of inquiries received by the system are shown in Table 1.

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. 5. Because of this and the vocabulary sizes, shown in Table 3, features in the utterances tend to be sparse.

Figure 5. 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 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

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. These 15 topics were used in the experiments, evaluations and analysis that are presented in this work.

4.3 Summary

In this chapter, the spoken inquiries datasets that are used in this work were described. An overview of the speech-oriented guidance system *Takemaru-kun* was also provided, explaining its characteristics and functionalities. The specifications of the datasets were given in detail, including examples of utterances, setup for AMs, LMs and ASR, vocabulary sizes, utterances' lengths and frequency of samples.

Chapter 5

Experimental Evaluations and Analysis

5.1 Experimental Setup

We compared the performances of the methods in the topic classification of spoken inquiries. Additionally, we compared the classification performance of the SG scheme against a voting strategy which classifies a sample utterance in a topic selected by the majority of the three methods compared in this work.

In our experiments, we used the *Takemaru-kun* datasets as described in Section 4.2. We used the 15 most frequent topics, which were previously shown in Table 5. The experimental conditions for the first and second-level classifiers are given in detail in Table 6 and Table 7 respectively. For experiments with the SG scheme, we followed the procedure described in Section 3.1. We used a one-vs-rest approach for multi-class classification with SVM and pboost, and the "# of pos" and "# of neg" variables indicated in the experimental conditions refer to the number of utterances in the topic (positive) and in the rest of the topics (negative) 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 2.2, 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

$$F\text{-}measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \tag{21}$$

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

Table 6. Experimental conditions for the first-level classifiers

SVM tool	LIBSVM 2.9 [29]
Hyperparameters C_+ and	$C_+ =$ (# of neg / # of pos + # of neg) $\times C$
C_- for each SVM	$C_- =$ (# of pos / # of pos + # of neg) $\times C$
classifier	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 [19]
Hyperparameters D_+ and	$D_+ =$ (# of neg / # of pos + # of neg) $\times D$
D_- for each pboost	$D_- =$ (# of pos / # of pos + # of neg) $\times D$
classifier	where $D_+ + D_- = D$, and
	$D = 1/\nu\ell$, for ν from 0.001 to 0.100
	and $\ell =$ 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 [18]
ME model	Inequality constraints [42]
Features	Word 1+2+3-grams,
	Character 1+2+3-grams
Datasets	Transcriptions and ASR 10-best results

Table 7. Experimental conditions for the second-level classifier

SVM tool	LIBSVM 2.9 [29]
Hyperparameters C_+ and C_- for each SVM classifier	$C_+ =$ (# of neg / # of pos + # of neg) $\times C$ $C_- =$ (# of pos / # of pos + # of neg) $\times C$ 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

5.2 Performance Comparison

An analysis of overlaps in the prediction error among individual methods is presented in Fig. 6. 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%. Fig. 7 and Fig. 8 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 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

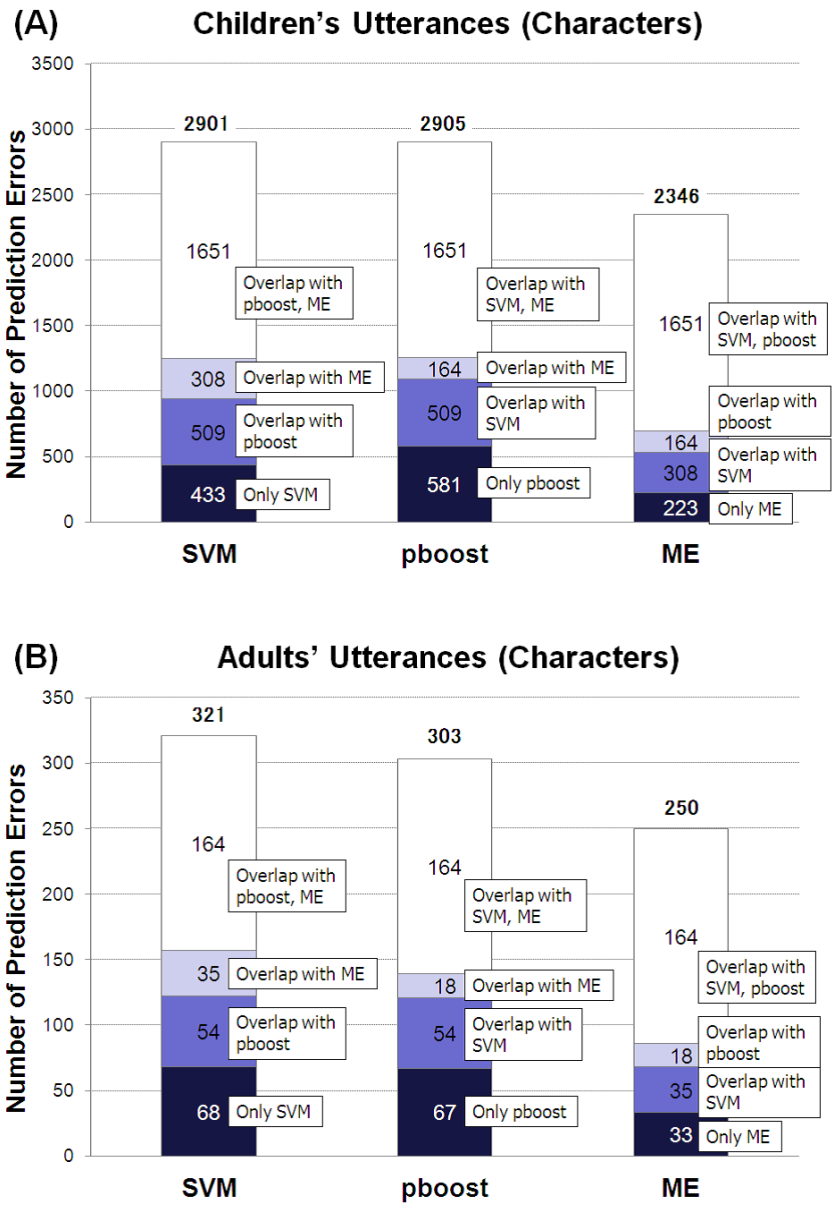

Figure 6. Prediction error overlap by method for (A) children's and (B) adults' utterances using character features (open test). The number of prediction errors for each method is indicated above the bars in bold, and the numbers of prediction error overlaps among the methods are indicated inside the bars.

Table 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%

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.

Table 10 shows the performance of the voting strategy against the SG scheme in the classification of ASR results using words or character features. The voting strategy classifies a sample utterance in a topic selected by the majority of the methods, in this case SVM, pboost and ME. The classification performance of the SG scheme was significantly higher than that of the voting strategy for both children and adults, either using words or character features. The main reason

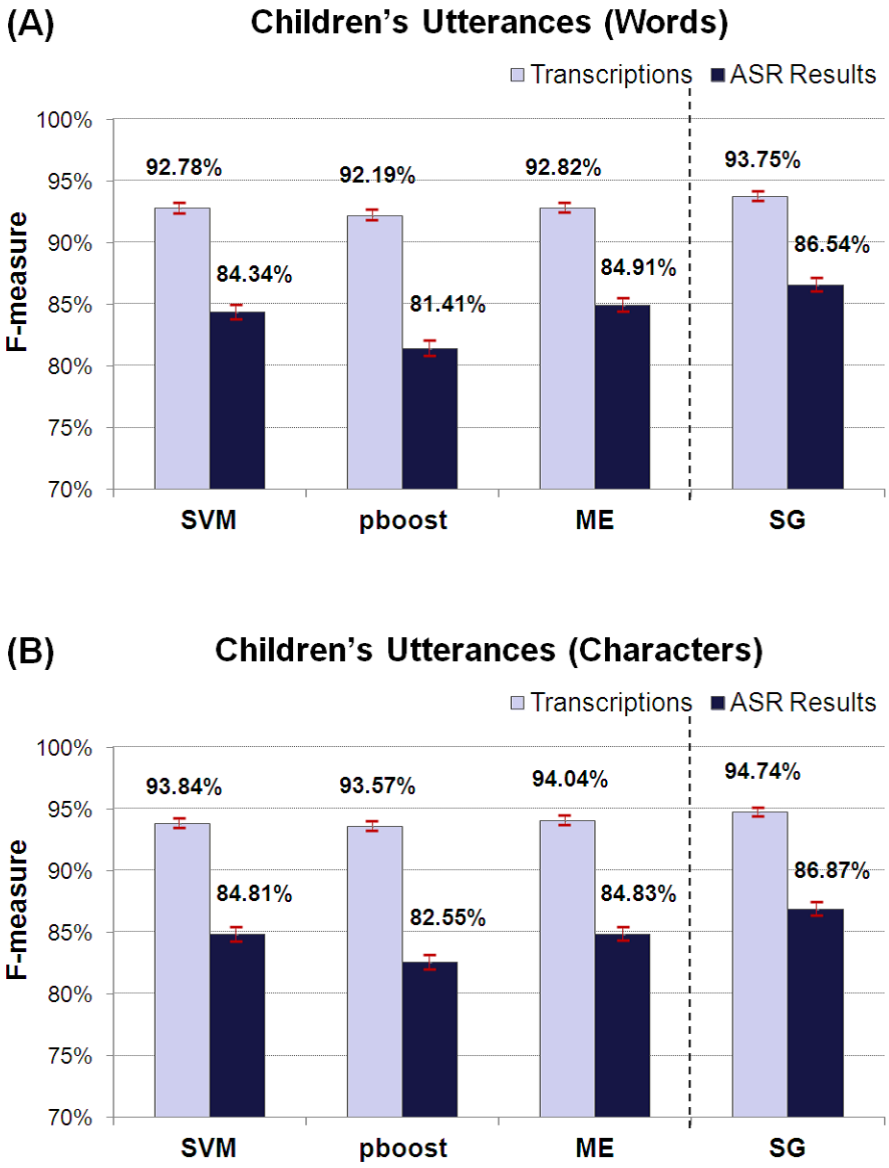

Figure 7. F-measure for each method for transcriptions and ASR results for children's utterances using (A) word and (B) character features. The F-measure for each method is indicated above the bars in bold, and the red line segments represent 95% confidence intervals.

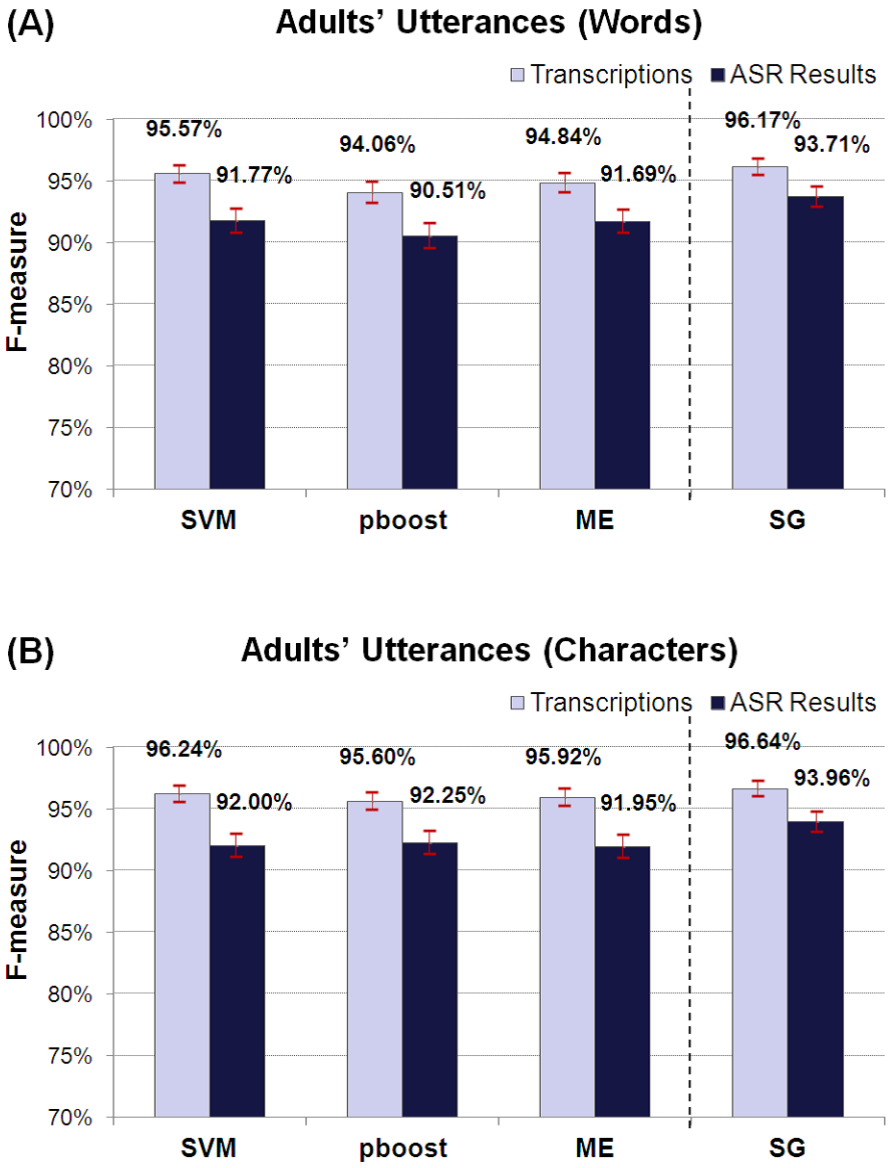

Figure 8. F-measure for each method for transcriptions and ASR results for adults' utterances using (A) word and (B) character features. The F-measure for each method is indicated above the bars in bold, and the red line segments represent 95% confidence intervals.

Table 10. F-measure of the SG scheme vs. the voting strategy using predictions of SVM, pboost and ME (ASR Results)

Method	Children	Adults
SG scheme (Words)	**86.54%**	**93.71%**
Voting strategy (Words)	85.07%	92.64%
SG scheme (Characters)	**86.87%**	**93.96%**
Voting strategy (Characters)	85.50%	92.96%

Table 11. F-measure of the SG scheme including pboost vs. excluding pboost from the combination (ASR Results)

SG Scheme	Children	Adults
Including pboost (Words)	**86.54%**	**93.71%**
Excluding pboost (Words)	85.59%	92.69%
Including pboost (Characters)	**86.87%**	**93.96%**
Excluding pboost (Characters)	85.34%	92.93%

for this is that the voting strategy gives equal importance to each classifier in the classification of every utterance. This is not the case with the SG scheme, which gives different weights to each classifier according to the utterance.

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, as is shown in Table 11. When excluding pboost from the combination in the SG scheme, the classification performance was significantly lower for both ASR results from children's and adults' utterances, either by using words or character features.

Both SVM and pboost perform classification by maximizing separation margin among data, which makes them similar; however, pboost performs a feature selection by finding optimal discriminative subsequence patterns, and in SVM there is originally no feature selection. On the other hand, SVM uses kernel functions to deal with nonlinearities and pboost does not. Because of this, even

though both classifiers have similarities, both classifiers can still compensate each other. Preliminary experiments using Recursive Feature Elimination (RFE)[43] for feature selection in SVM were also performed, however they did not present classification performance improvements in our task.

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 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.

5.3 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. 9. 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

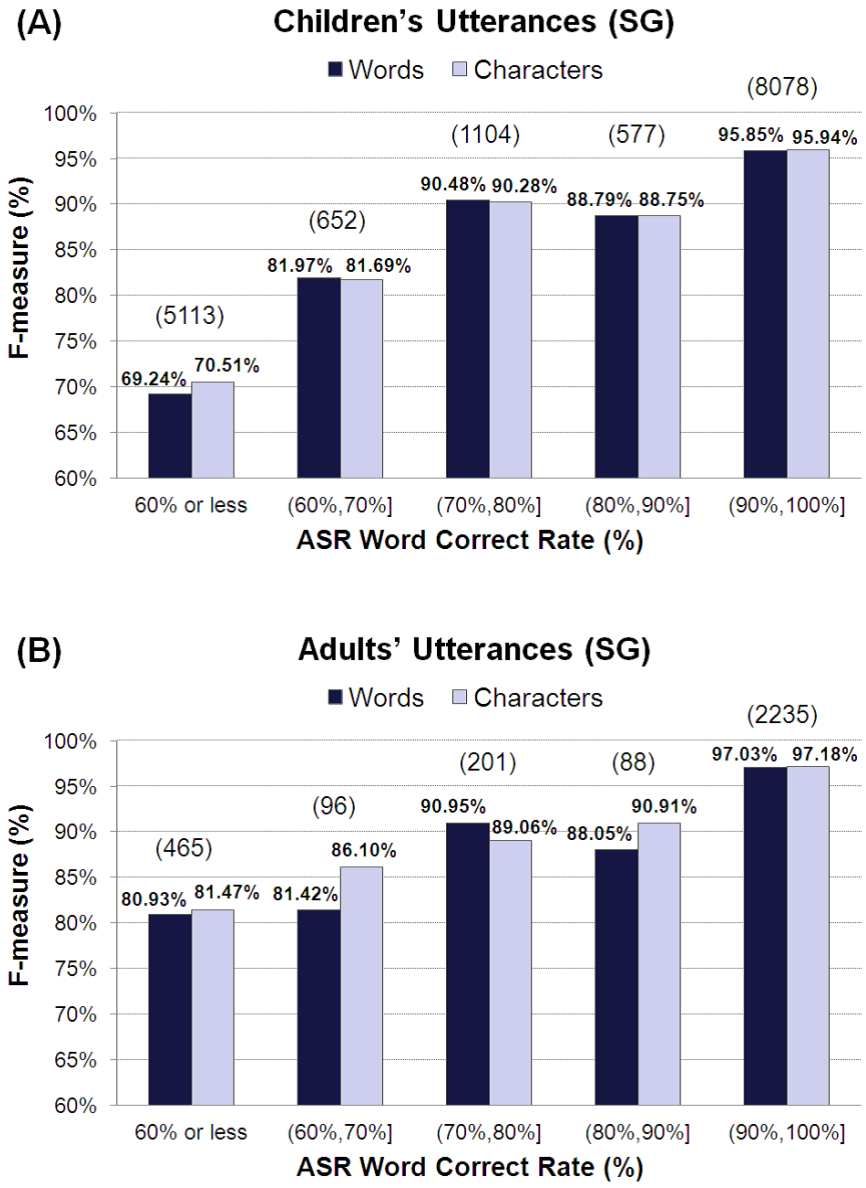

Figure 9. F-measure of the SG scheme by showing word correct rates for ASR of (A) children's and (B) adults' utterances using word or character features (open test). Numbers of utterances are indicated above the bars inside parentheses. The F-measure for the SG scheme is also indicated above the bars in bold.

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.

5.4 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. Fig. 10 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.

5.5 Conclusion

Experiments with three supervised 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, were performed for the topic classification of spoken inquiries received by a speech-oriented guidance system operating in a real environment, and their performances were compared. An analysis on prediction error overlaps indicated that the three methods produce some prediction errors that do not overlap with those of the other methods.

The predictions from the above different types of methods were combined by using a stacked generalization (SG) scheme. 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. Experimental results showed 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

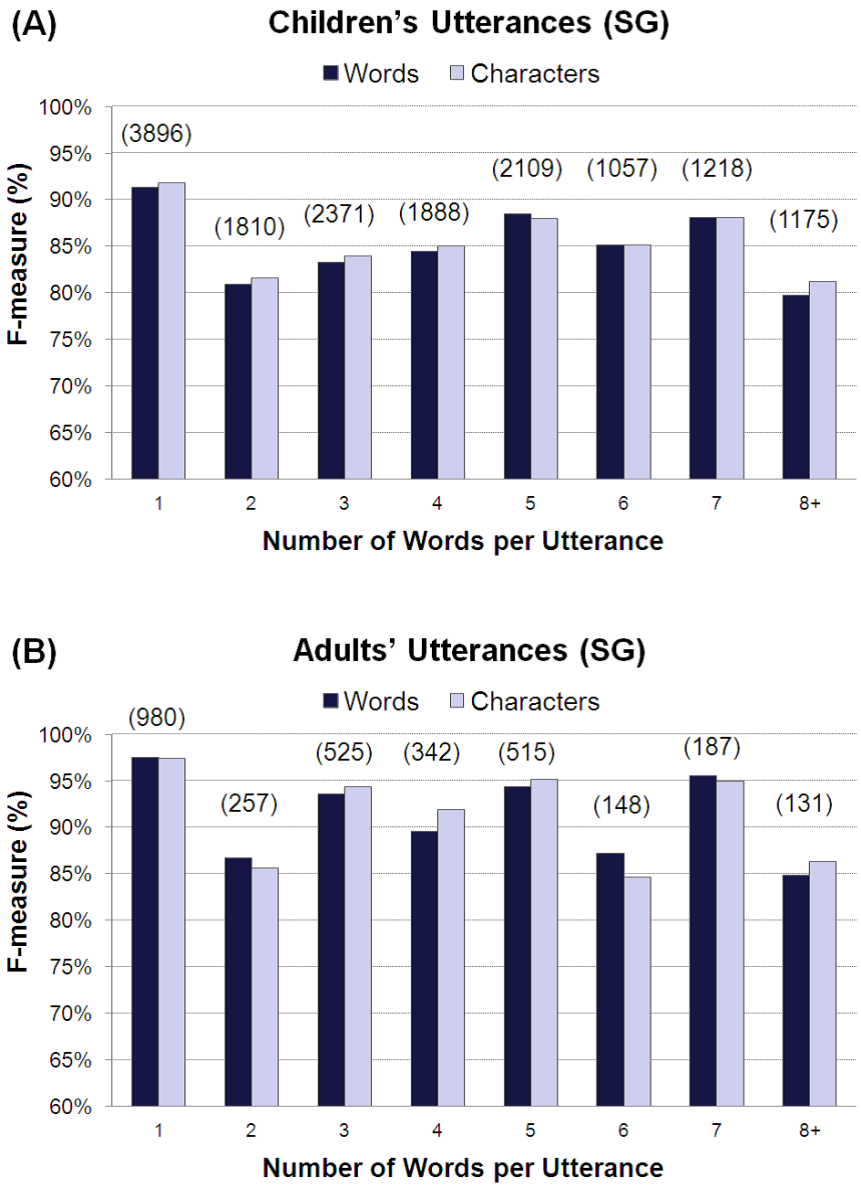

Figure 10. F-measure of the SG scheme by showing number of words per utterance of (A) children's and (B) adults' utterances using word or character features (open test). Number of utterances are indicated above the bars inside parentheses.

of 1.89% for adults' inquiries when using the SG scheme and character features.

The classification performance of the SG scheme was also compared against a voting strategy, which classifies a sample utterance in a topic selected by the majority of the methods. The SG scheme presented significantly higher classification performance than that of the voting strategy in the classification of ASR results for both children an adults, either using words or character features.

A comparison between the performance of the SG scheme and word correct rates for ASR results of children's and adults' utterances showed a tendency to obtain better classification performance as word correct rates for ASR results increase. For word correct rates above 60% the classification performances between children and adults are closer, while the differences are larger for word correct rates below that percentage.

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. 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.

An evaluation using words or characters as features for the classifiers was also performed. 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.

Chapter 6

Conclusion

6.1 Summary of the Thesis

Speech is one of the most important and natural means for social interaction among humans, thus human-machine interaction through speech presents great advantages. However, speaking is a very complex act and human-machine interaction through speech poses several difficulties.

Topic classification of speech is a subject of interest in spoken language processing because of its several applications. It has been studied for call routing, call type resolution, improvement of ASR performance, out-of-domain utterance detection, and others.

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, by comparing the performance of three supervised 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. The differences among the three classifiers allow them to compensate each other's performance. Because of this, the usage of a stacked generalization (SG) scheme that combines their predictions to achieve greater classification performance was proposed.

An analysis on the performance of the above methods and their combination in the topic classification of spoken inquiries was carried out. Prediction errors of each method were analyzed in order to determine their overlap, and to observe how many of the errors that did not overlap were able to be corrected by the SG scheme, as well as how many correct predictions were misclassified by it. The influence of ASR performance in the topic classification was analyzed. An evaluation using words or characters as features for the classifiers was also performed.

The experiments, evaluations and analysis were carried out using data obtained from a speech-oriented guidance system that operates in a real environment. The guidance system is the *Takemaru-kun* system, which is an open do-

main system whose task domain was not set before its operation started, and users are free to ask the system for the information they want to obtain. When the system started collecting user's inquiries, they were analyzed and manually labeled to define its task domain. Therefore, the results of the analysis and evaluations presented in this work are expected to be applicable to other task domains for this type of systems.

An analysis on prediction error overlaps indicated that the three methods produce some prediction errors that do not overlap with those of the other methods. 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.

Experimental results showed 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. The classification performance of the SG scheme was also compared against a voting strategy, where the SG scheme presented a significantly higher classification performance.

A comparison between the performance of the SG scheme and word correct rates for ASR results of children's and adults' utterances showed a tendency to obtain better classification performance as word correct rates for ASR results increase. For word correct rates above 60% the classification performances between children and adults are closer, while the differences are larger for word correct rates below that percentage.

The evaluation using words or characters as features for the classifiers showed that although the use of characters yielded higher classification performance in some cases, the tendency is not consistent, and the differences were not found to be significant.

6.2 Future Work

Future work will be focused on the following subjects.

Semi-Supervised Learning Methods: Manual data labeling, which is required for supervised learning, is a costly process and unlabeled data are usually abundant and cheap to obtain. Semi-supervised learning methods allow to take

advantage of unlabeled data by using them in conjunction with labeled data for training the models. It is desirable to be able to improve topic classification performance by taking advantage of unlabeled data.

Combination of different features for classification: In this work the use of words or character features was evaluated. However, their combination was not explored. Different classification models trained with different features may yield improvements. It would also be interesting to incorporate other kind of features such as acoustic and semantic features.

Improvement of ASR performance for children's utterances: When spoken dialog systems are set in public places, people from different age groups make use of them, but children represent one of the highest percentages of users. Among the utterances received by the *Takemaru-kun* system, nearly 70% correspond to children. However, ASR performance tends to be lower for children's utterances due to speech disfluencies and other issues. Topic-dependent LMs and AMs could be beneficial for improving ASR performance on these cases.

Appendix

A. Additional Experimental Results

The F-measure of the methods in the classification of children's and adults' utterances in the open test are presented in Tables 12 and 13 respectively. ASR 10-best results were used for SVM and ME, and ASR 1-best were used for pboost.

Table 12. F-measure of the methods in the classification of children's utterances (open test)

Method	Features	Transcriptions	ASR Results
SVM (RBF kernel)	Word 1-grams	92.73%	84.38%
	Word 1+2-grams	93.10%	84.00%
	Word 1+2+3-grams	92.78%	84.34%
	Char. 1-grams	92.80%	84.30%
	Char. 1+2-grams	93.51%	84.60%
	Char. 1+2+3-grams	93.84%	84.81%
pboost	Word 1-grams (max lg. 1)	91.51%	81.40%
	Word 1-grams (max lg. 2)	92.29%	81.37%
	Word 1-grams (max lg. 3)	92.19%	81.41%
	Char. 1-grams (max lg. 1)	89.92%	79.72%
	Char. 1-grams (max lg. 2)	93.34%	82.93%
	Char. 1-grams (max lg. 3)	93.57%	82.55%
ME	Word 1-grams	92.68%	83.73%
	Word 1+2-grams	92.95%	84.57%
	Word 1+2+3-grams	92.82%	84.91%
	Char. 1-grams	92.31%	81.09%
	Char. 1+2-grams	93.96%	84.54%
	Char. 1+2+3-grams	94.04%	84.83%
SG	Predictions (Word)	93.75%	86.54%
	Predictions (Char.)	94.74%	86.87%

In the pboost details, max lg. indicates the maximum subsequence pattern length that was set. The SG scheme combines the predictions of SVM and ME using 1+2+3-grams and pboost using 1-grams with max lg. of 3, using word or character features. This setting was chosen because it presented better results, although there were some exceptions.

Table 13. F-measure of the methods in the classification of adults' utterances (open test)

Method	Features	Transcriptions	ASR Results
SVM (RBF kernel)	Word 1-grams	95.30%	91.62%
	Word 1+2-grams	95.79%	91.67%
	Word 1+2+3-grams	95.57%	91.77%
	Char. 1-grams	96.18%	92.29%
	Char. 1+2-grams	96.54%	92.14%
	Char. 1+2+3-grams	96.24%	92.00%
pboost	Word 1-grams (max lg. 1)	93.83%	90.43%
	Word 1-grams (max lg. 2)	94.16%	90.54%
	Word 1-grams (max lg. 3)	94.06%	90.51%
	Char. 1-grams (max lg. 1)	94.10%	90.67%
	Char. 1-grams (max lg. 2)	95.47%	92.15%
	Char. 1-grams (max lg. 3)	95.60%	92.25%
ME	Word 1-grams	94.74%	90.76%
	Word 1+2-grams	94.84%	91.49%
	Word 1+2+3-grams	94.84%	91.69%
	Char. 1-grams	95.23%	89.80%
	Char. 1+2-grams	95.89%	91.74%
	Char. 1+2+3-grams	95.92%	91.95%
SG	Predictions (Word)	96.17%	93.71%
	Predictions (Char.)	96.64%	93.96%

The F-measure of the methods in the classification of children's and adults' utterances in the closed test are presented in Tables 14 and 15 respectively. ASR 10-best results were used for SVM and ME, and ASR 1-best were used for pboost.

In the pboost details, max lg. indicates the maximum subsequence pattern length that was set. The SG scheme combines the predictions of SVM and ME using 1+2+3-grams and pboost using 1-grams with max lg. of 3, using word or character features.

Table 14. F-measure of the methods in the classification of children's utterances (closed test)

Method	Features	Transcriptions	ASR Results
SVM (RBF kernel)	Word 1-grams	99.67%	93.07%
	Word 1+2-grams	99.80%	94.64%
	Word 1+2+3-grams	99.89%	96.53%
	Char. 1-grams	99.03%	90.64%
	Char. 1+2-grams	98.69%	96.34%
	Char. 1+2+3-grams	99.43%	95.42%
pboost	Word 1-grams (max lg. 1)	96.94%	86.65%
	Word 1-grams (max lg. 2)	98.49%	91.43%
	Word 1-grams (max lg. 3)	98.15%	91.49%
	Char. 1-grams (max lg. 1)	92.96%	82.47%
	Char. 1-grams (max lg. 2)	98.93%	93.32%
	Char. 1-grams (max lg. 3)	98.87%	94.86%
ME	Word 1-grams	97.03%	88.71%
	Word 1+2-grams	97.66%	91.57%
	Word 1+2+3-grams	97.73%	91.40%
	Char. 1-grams	95.61%	84.50%
	Char. 1+2-grams	98.05%	91.52%
	Char. 1+2+3-grams	98.34%	92.99%
SG	Predictions (Word)	97.01%	90.22%
	Predictions (Char.)	97.56%	90.91%

Table 15. F-measure of the methods in the classification of adults' utterances (closed test)

Method	Features	Transcriptions	ASR Results
SVM (RBF kernel)	Word 1-grams	99.97%	98.25%
	Word 1+2-grams	99.99%	99.64%
	Word 1+2+3-grams	99.99%	98.86%
	Char. 1-grams	99.99%	98.70%
	Char. 1+2-grams	99.97%	99.36%
	Char. 1+2+3-grams	99.82%	98.69%
pboost	Word 1-grams (max lg. 1)	97.20%	94.88%
	Word 1-grams (max lg. 2)	98.56%	97.55%
	Word 1-grams (max lg. 3)	98.17%	97.33%
	Char. 1-grams (max lg. 1)	98.58%	96.80%
	Char. 1-grams (max lg. 2)	99.88%	98.98%
	Char. 1-grams (max lg. 3)	99.71%	98.94%
ME	Word 1-grams	97.59%	95.86%
	Word 1+2-grams	97.84%	96.85%
	Word 1+2+3-grams	97.93%	96.91%
	Char. 1-grams	97.30%	94.53%
	Char. 1+2-grams	98.67%	97.63%
	Char. 1+2+3-grams	98.76%	97.98%
SG	Predictions (Word)	96.99%	95.52%
	Predictions (Char.)	97.70%	96.13%

B. Optimal Hyperparameters from Experiments

Optimal hyperparameter values for SVM and pboost were obtained experimentally using a grid search strategy and were set a posteriori. Tables 16, 17, 18 and 19 present the experimentally obtained optimal hyperparameters for children and adults, for transcriptions and ASR results respectively.

In the pboost details, max lg. indicates the maximum subsequence pattern length that was set. The SG scheme combines the predictions of SVM and ME using 1+2+3-grams and pboost using 1-grams with max lg. of 3, using word or character features. First-level SVM corresponds to SVM as individual classifier, and Second-level SVM corresponds to the classifier used in the second-level of SG. ASR 10-best results were used for First-level SVM and ASR 1-best were used for pboost.

Table 16. Optimal hyperparameters for SVM and pboost in the classification of transcriptions of children's utterances (open test)

Method	Features	Transcriptions
First-level SVM (RBF kernel)	Word 1-grams	$\gamma = 0.1$, C = 1000
	Word 1+2-grams	$\gamma = 0.1$, C = 100
	Word 1+2+3-grams	$\gamma = 0.1$, C = 100
	Char. 1-grams	$\gamma = 0.5$, C = 100
	Char. 1+2-grams	$\gamma = 0.01$, C = 100
	Char. 1+2+3-grams	$\gamma = 0.01$, C = 100
pboost	Word 1-grams (max lg. 1)	$\nu = 0.043$
	Word 1-grams (max lg. 2)	$\nu = 0.032$
	Word 1-grams (max lg. 3)	$\nu = 0.036$
	Char. 1-grams (max lg. 1)	$\nu = 0.074$
	Char. 1-grams (max lg. 2)	$\nu = 0.017$
	Char. 1-grams (max lg. 3)	$\nu = 0.019$
Second-level SVM (RBF kernel)	Predictions (Word)	$\gamma = 0.1$, C = 1
	Predictions (Char.)	$\gamma = 0.1$, C = 1

Table 17. Optimal hyperparameters for SVM and pboost in the classification of ASR results of children's utterances (open test)

Method	Features	ASR Results
First-level SVM (RBF kernel)	Word 1-grams	$\gamma = 0.1$, C = 10
	Word 1+2-grams	$\gamma = 0.0001$, C = 1000
	Word 1+2+3-grams	$\gamma = 0.01$, C = 10
	Char. 1-grams	$\gamma = 0.5$, C = 10
	Char. 1+2-grams	$\gamma = 0.001$, C = 1000
	Char. 1+2+3-grams	$\gamma = 0.0001$, C = 1000
pboost	Word 1-grams (max lg. 1)	$\nu = 0.15$
	Word 1-grams (max lg. 2)	$\nu = 0.094$
	Word 1-grams (max lg. 3)	$\nu = 0.095$
	Char. 1-grams (max lg. 1)	$\nu = 0.23$
	Char. 1-grams (max lg. 2)	$\nu = 0.071$
	Char. 1-grams (max lg. 3)	$\nu = 0.051$
Second-level SVM (RBF kernel)	Predictions (Word)	$\gamma = 0.1$, C = 1
	Predictions (Char.)	$\gamma = 0.1$, C = 1

Table 18. Optimal hyperparameters for SVM and pboost in the classification of transcriptions of adults' utterances (open test)

Method	Features	Transcriptions
First-level SVM (RBF kernel)	Word 1-grams	$\gamma = 0.1$, C = 1000
	Word 1+2-grams	$\gamma = 0.1$, C = 1000
	Word 1+2+3-grams	$\gamma = 0.1$, C = 1000
	Char. 1-grams	$\gamma = 0.5$, C = 1000
	Char. 1+2-grams	$\gamma = 0.1$, C = 100
	Char. 1+2+3-grams	$\gamma = 0.01$, C = 100
pboost	Word 1-grams (max lg. 1)	$\nu = 0.050$
	Word 1-grams (max lg. 2)	$\nu = 0.031$
	Word 1-grams (max lg. 3)	$\nu = 0.044$
	Char. 1-grams (max lg. 1)	$\nu = 0.015$
	Char. 1-grams (max lg. 2)	$\nu = 0.005$
	Char. 1-grams (max lg. 3)	$\nu = 0.008$
Second-level SVM (RBF kernel)	Predictions (Word)	$\gamma = 0.1$, C = 10
	Predictions (Char.)	$\gamma = 0.1$, C = 1

Table 19. Optimal hyperparameters for SVM and pboost in the classification of ASR results of adults' utterances (open test)

Method	Features	ASR Results
First-level SVM (RBF kernel)	Word 1-grams	$\gamma = 0.01$, C = 100
	Word 1+2-grams	$\gamma = 0.001$, C = 1000
	Word 1+2+3-grams	$\gamma = 0.0001$, C = 1000
	Char. 1-grams	$\gamma = 0.5$, C = 10
	Char. 1+2-grams	$\gamma = 0.001$, C = 1000
	Char. 1+2+3-grams	$\gamma = 0.0001$, C = 1000
pboost	Word 1-grams (max lg. 1)	$\nu = 0.1$
	Word 1-grams (max lg. 2)	$\nu = 0.032$
	Word 1-grams (max lg. 3)	$\nu = 0.035$
	Char. 1-grams (max lg. 1)	$\nu = 0.033$
	Char. 1-grams (max lg. 2)	$\nu = 0.012$
	Char. 1-grams (max lg. 3)	$\nu = 0.013$
Second-level SVM (RBF kernel)	Predictions (Word)	$\gamma = 0.1$, C = 10
	Predictions (Char.)	$\gamma = 0.1$, C = 10

C. Transductive Support Vector Machine

C.1 Introduction

Supervised learning requires manually labeled data, which are very costly, while unlabeled data are usually abundant and cheap to obtain. Because of this, it would be ideal to be able to use unlabeled data to improve the topic classification performance of spoken inquiries.

Transductive Support Vector Machine (TSVM) extends a regular SVM to treat partially labeled data for semi-supervised learning, including labeled and unlabeled data in the training set. TSVMs were proposed by Vapnik in 1998, and were introduced by Joachims [44] for text classification. TSVMs use labeled samples to find optimal hyperplanes that maximize the separation margin of two classes of data, and then use unlabeled samples to adjust that margin.

In this appendix, the viability of using a TSVM for semi-supervised learning in the topic classification of spoken inquiries is evaluated.

C.2 Method Explanation

In TSVM, the primal optimization problem follows the form:

$$\begin{aligned} \min_{\vec{w},b,\vec{\xi},\vec{\xi}^*} \quad & \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\xi_i + C_-^*\sum_{\{j:y_j^*=-1\}}\xi_j^* + C_+^*\sum_{\{j:y_j^*=+1\}}\xi_j^* \\ \text{sb.t.} \quad & \forall_{i=1}^{n} : y_i[\vec{w}\cdot\vec{x}_i + b] \geq 1 - \xi_i \\ & \forall_{j=1}^{k} : y_j^*[\vec{w}\cdot\vec{x}_j + b] \geq 1 - \xi_j^* \end{aligned} \tag{22}$$

where $\vec{x}_i$ represents a labeled training sample and $\vec{x}_j$ an unlabeled training sample, $y_i \in \{1, -1\}$ and $y_j^* \in \{1, -1\}$ a class for labeled and unlabeled samples respectively. The hyperparameters C, C_-^* and C_+^* penalize the sum of the slack variables ξ_i and ξ_j^* to allow soft-margin, where * is used to denote unlabeled samples.

The TSVM algorithm [44] begins with labeling unlabeled samples based on the classification of a regular SVM trained with only labeled samples. Then, it re-trains the model using all samples and improves the solution by switching the

labels of the newly-labeled samples so that the objective function decreases. The label switching part of the algorithm consists of two embedded loops:

- An external loop uniformly increases the influence of the newly-labeled samples by incrementing C^*_- and C^*_+, which are initialized with a very low value, up to a defined value C^*. Very low values of C^*_- and C^*_+ mean that these samples are almost ignored when finding the classification margin, because these are still considered not reliable. As the reliability of the newly-labeled samples improves, the values of C^*_- and C^*_+ are increased.

- An internal loop identifies two newly-labeled samples for which switching the labels leads to a decrease in the current objective function, and switches the labels if this condition is met. For this, it identifies two samples with opposite labels and checks if the value of ξ^*_j, which measures classification error, is greater than a predefined value, which indicates that the samples may be mislabeled, and then it switches both labels. In each iteration, the optimization problem is solved again.

C.3 Experiments

SVMLight [45] was used for the experiments with TSVM. In the approach that was followed in the experiments, labeled and unlabeled data were used to train a model using a TSVM and the resultant model was used to classify test data.

The experiment consisted in the topic classification of ASR results of inquiries in Japanese received by the speech-oriented guidance system *Takemaru-kun*. BOW was used to represent utterances as vectors, and character unigrams, bigrams and trigrams were used as features. An RBF kernel was used and one-vs-rest was used for multi-class classification.

Experiments with separate datasets for children and adults were performed. Classification performance was evaluated using the F-measure, which was calculated individually for each topic and then averaged by frequency of samples. Optimal hyperparameter values were obtained experimentally using a grid search strategy, and were set a posteriori.

Table 20. Amount of samples in the labeled datasets

(Labeled datasets)	Children Training	Children Test	Adults Training	Adults Test
Amount of samples	43,494	15,524	14,431	3,085

Table 21. Amount of samples in the unlabeled datasets

(Unlabeled datasets)	Children Training	Adults Training
Unlabeled dataset #1 (2005.04 to 2005.12)	119,322	110,537
Unlabeled dataset #2 (2005.04 to 2006.12)	271,744	252,428
Unlabeled dataset #3 (2005.04 to 2007.12)	413,144	385,165

C.3.1 Characteristics of the Datasets

The labeled data are the same data that were described in Chapter 4 and used in the experiments described in Chapter 5. Table 20 shows the amount of samples in the labeled datasets. The unlabeled data correspond to the utterances collected by *Takemaru-kun* in the period from Apr. 2005 to Dec. 2007. Julius was also used as ASR engine, using and the same AMs and LMs that were used to recognize the labeled data, as described in Chapter 4. Three unlabeled datasets were created with different sizes. Table 21 shows the amount of samples in the unlabeled datasets.

C.3.2 Experimental Results

Table 22 presents the averaged topic classification performance per training dataset combination in the open test, for children and adults. Although the topic classification performances presented by the TSVM were quite competitive, they were not better than the performances of the regular SVM that were presented in Chapter 5. There is a slight tendency to obtain better classification performances with larger unlabeled datasets.

Table 22. Averaged F-measure results per training datasets (open test)

Training Dataset Combination	Children	Adults
Labeled dataset + Unlabeled dataset #1 (TSVM)	83.02%	91.75%
Labeled dataset + Unlabeled dataset #2 (TSVM)	84.17%	**92.86%**
Labeled dataset + Unlabeled dataset #3 (TSVM)	**84.28%**	92.81%

C.4 Conclusion

An evaluation of topic classification of spoken inquiries using a semi-supervised learning approach based on a TSVM was evaluated. Although the topic classification performances of the TSVM were competitive, they were not better than the performances of a regular SVM. There was a slight tendency to obtain better classification performances with larger unlabeled datasets.

References

[1] 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).

[2] Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T. J. and Hetherington, L.: Jupiter: A Telephone-Based Conversational Interface for Weather Information, *IEEE Transactions on Speech and Audio Processing*, Vol. 8, pp. 85–96 (2000).

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

[4] Gustafson, J., Lindberg, N. and Lundeberg, M.: The August Spoken Dialogue System, *Proceedings of Eurospeech 1999*, pp. 1151–1154 (1999).

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

[6] Misu, T. and Kawahara, T.: Speech-Based Interactive Information Guidance System using Question-Answering Technique, *Proceedings of ICASSP 2007*, pp. IV–145–IV–148 (2007).

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

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

[9] Ward, W.: Understanding Spontaneous Speech, *Proceedings of the workshop on Speech and Natural Language*, pp. 137–141 (1989).

[10] Shriberg, E.: Spontaneous Speech: How People Really Talk and Why Engineers Should Care, *Proceedings of Interspeech 2005*, pp. 1781–1784 (2005).

[11] Haffner, P., Tur, G. and Wright, J. H.: Optimizing SVMs for Complex Call Classification, *Proceedings of ICASSP 2003*, pp. 632–635 (2003).

[12] Lane, I. R., Kawahara, T., Matsui, T. and Nakamura, S.: Dialogue Speech Recognition by Combining Hierarchical Topic Classification and Language Model Switching, *IEICE Transactions on Information and Systems*, Vol. E88-D, No. 3, pp. 446–454 (2005).

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

[14] Skowron, M. and Araki, K.: Effectiveness of Combined Features for Machine Learning Based Question Classification, *Information and Media Technologies*, Vol. 1, No. 1, pp. 461–481 (2006).

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

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

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

[18] 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.

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

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

[21] 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).

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

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

[24] Sill, J., Takács, G., Mackey, L. and Lin, D.: Feature-Weighted Linear Stacking, *CoRR*, Vol. abs/0911.0460, pp. 1–17 (2009).

[25] 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).

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

[27] Boser, B. E., Guyon, I. M. and Vapnik, V. N.: A Training Algorithm for Optimal Margin Classifiers, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152 (1992).

[28] Cortes, C. and Vapnik, V.: Support-Vector Networks, *Machine Learning*, pp. 273–297 (1995).

[29] Chang, C.-C. and Lin, C.-J.: LIBSVM: a Library for Support Vector Machines, *ACM Transactions on 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.

[30] Shawe-Taylor, J. and Cristianini, N.: *Kernel Methods for Pattern Analysis*, Cambridge University Press, New York, NY, USA (2004).

[31] Bishop, C. M.: *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006).

[32] 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).

[33] 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 Transactions on Knowledge and Data Engineering*, Vol. 16, No. 10, pp. 1424–1440 (2004).

[34] 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).

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

[36] Miyake, J., Takeuchi, S., Kawanami, H., Saruwatari, H. and Shikano, K.: Language Model for the Web Search Task in a Spoken Dialogue System for Children, *Proceedings of WOCCI 2008 (ICMI post-conference workshop)* (2008).

[37] Nisimura, R., Lee, A., Yamada, M. and Shikano, K.: Operating a Public Spoken Guidance System in Real Environment, *Proceedings of Interspeech 2005*, pp. 845–848 (2005).

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

[39] 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.

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

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

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

[43] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V.: Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, Vol. 46, No. 1-3, pp. 389–422 (2002).

[44] Joachims, T.: Transductive Inference for Text Classification using Support Vector Machines, *ICML 1999 Proceedings*, pp. 200–209 (1999).

[45] Joachims, T.: Making large-scale support vector machine learning practical, *Advances in Kernel Methods* (Schölkopf, B., Burges, C. J. C. and Smola, A. J., eds.), MIT Press, pp. 169–184 (1999). Software available at http://svmlight.joachims.org/.

Acknowledgements

The work presented in this thesis would not have been possible without the support of many important persons, whom hereby I would like to acknowledge.

I would like to express my sincere and deepest gratitude to Professor Kiyohiro Shikano, for giving me the opportunity to perform research under his guidance and for his advises, patience and encouragement.

I would also like to express my gratitude to Professor Yuji Matsumoto and Associate Professor Hiroshi Saruwatari, for their constructive comments and helpful suggestions.

I would like to thank Assistant Professor Hiromichi Kawanami for his kind guidance and useful comments.

I feel very much obliged to express my gratitude to Professor Tomoko Matsui, from The Institute of Statistical Mathematics, for her invaluable instruction, encouragement and continuous support.

I would also like to express my deep gratitude to the Japanese Ministry of Education, Culture, Sports, Science and Technology, for providing me the opportunity to study in Japan and to engage in to this unforgettable journey.

My greatest thanks to all the students and staff of the Speech and Acoustics Laboratory, who in a greater or lesser extent have helped me in daily situations and contributed to maintain a positive work environment.

Finally, but no less important, I would like to express my sincere and deep gratitude to my family and friends, whom I feel inexorably indebted to for supporting and encouraging me during difficult times.

List of Publications

Journal Papers

1. Rafael Torres, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Comparison of methods for topic classification of spoken inquiries. *IPSJ Journal of Information Processing*, Vol. 21, No. 2, Apr. 2013 (Accepted).

2. Haruka Majima, Yoko Fujita, Rafael Torres, Hiromichi Kawanami, Sunao Hara, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Invalid input rejection using bag-of-words for speech-oriented guidance system. *IPSJ Journal of Information Processing*, Vol. 54, No. 2, pp. 443–451, Feb. 2013 (In Japanese).

International Conferences

1. Rafael Torres, Shota Takeuchi, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Comparison of methods for topic classification in a speech-oriented guidance system. In *Proceedings of Interspeech 2010*, pp. 1261–1264, Chiba, Japan, Sept. 2010.

2. Rafael Torres, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Training data size requirements for topic classification in a speech-oriented guidance system. In *Proceedings of APSIPA ASC 2010*, pp. 486–489, Biopolis, Singapore, Dec. 2010.

3. Rafael Torres, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Topic classification of spoken inquiries based on stacked generalization. In *Proceedings of APSIPA ASC 2011*, Xi'an, China, Oct. 2011.

4. Hiromichi Kawanami, Shota Takeuchi, Rafael Torres, Hiroshi Saruwatari, and Kiyohiro Shikano. Development and operation of speech-oriented information guidance systems, Kita-chan and Kita-robo. In *Proceedings of APSIPA ASC 2011*, Xi'an, China, Oct. 2011.

5. Haruka Majima, Rafael Torres, Yoko Fujita, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Spoken inquiry discrimination using bag-of-words for speech-oriented guidance system. In *Proceedings of Interspeech 2012*, Portland, USA, Sept. 2012.

6. Rafael Torres, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Topic classification of spoken inquiries using transductive support vector machine. In *Proceedings of IWSDS 2012*, pp. 231–236, Paris, France, Nov. 2012.

7. Haruka Majima, Rafael Torres, Hiromichi Kawanami, Sunao Hara, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Evaluation of invalid input discrimination using bag-of-words for speech-oriented guidance system. In *Proceedings of IWSDS 2012*, pp. 339–347, Paris, France, Nov. 2012.

Technical Reports

1. Rafael Torres, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Stacked generalization for topic classification of spoken inquiries. *IPSJ SIG Technical Report*, 2011-SLP-85(6), pp. 1–6, Feb. 2011.

2. Haruka Majima, Yoko Fujita, Rafael Torres, Hiromichi Kawanami, Sunao Hara, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Invalid input rejection using bag-of-words for speech-oriented guidance system. *IPSJ SIG Technical Report*, 2012-SLP-92(7), pp. 1–6, July 2012 (In Japanese).

Meetings

1. Rafael Torres, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Topic classification in a speech-oriented guidance system using character based methods. *ASJ Autumn Meeting*, 1-Q-32, pp. 195–198, Sept. 2010.

2. Rafael Torres, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Topic classification of spoken inquiries with stacked generalization. *ASJ Spring Meeting*, 2-P-44(b), pp. 217–220, Mar. 2011.

3. Rafael Torres, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Evaluation of topic classification of spoken inquiries. *ASJ Autumn Meeting*, 3-P-23, pp. 205–208, Sept. 2011.

4. Hiromichi Kawanami, Keigo Kubo, Yusuke Kisaki, Rafael Torres, and Kiyohiro Shikano. Database extension of the information guidance system Takemaru-kun for implementation at an exhibition site. *ASJ Autumn Meeting*, 3-10-8, pp. 89–92, Sept. 2011 (In Japanese).

5. Rafael Torres, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Semi-supervised topic detection with transductive SVM. *ASJ Spring Meeting*, 3-P-26, pp. 281–284, Mar. 2012.

6. Rafael Torres, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Semi-supervised learning algorithms for topic classification using maximum entropy and transductive SVM. *ASJ Autumn Meeting*, 3-P-34, pp. 209–212, Sept. 2012.

7. Haruka Majima, Rafael Torres, Hiromichi Kawanami, Sunao Hara, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Evaluation of invalid input discrimination using ME for speech-oriented guidance system. *ASJ Autumn Meeting*, 3-1-8, pp. 113–116, Sept. 2012 (In Japanese).

8. Haruka Majima, Rafael Torres, Hiromichi Kawanami, Sunao Hara, Tomoko Matsui, Hiroshi Saruwatari, and Kiyohiro Shikano. Evaluation of portability of invalid input discrimination model using bag-of-words for speech-oriented guidance system. *ASJ Spring Meeting*, 3-9-5, Mar. 2013 (In Japanese).