Training Data Size Requirements for Topic Classification in a Speech-Oriented Guidance System

Rafael Torres∗, Hiromichi Kawanami∗, Tomoko Matsui†, Hiroshi Saruwatari∗, Kiyohiro Shikano∗

∗Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma 630-0192 Nara Japan
E-mail: {rafael-t, kawanami, sawatari, shikano}@is.naist.jp Tel: +81-743-72-5288
†Department of Statistical Modeling, The Institute of Statistical Mathematics, Tachikawa 190-8562 Tokyo Japan
E-mail: tmatsui@ism.ac.jp Tel: +81-50-5533-8537

Abstract—In this work, we address the classification in topics of utterances in Japanese received by a speech-oriented guidance system operating in a real environment. The implementation of this kind of systems requires the collection and manual labeling of actual user’s utterances, which is a costly process. Because of this, we are interested in evaluating the influence of the amount of data for training in the context of topic classification. For this, we compared the performance of a Support Vector Machine and a Maximum Entropy classifier using training data of different sizes. We used actual data collected by the speech-oriented guidance system Takemaru-kun, from adults and children, and also evaluated the effect of automatic speech recognition (ASR) errors in the classification performance. To deal with the shortness of the utterances we proposed to use characters as features, which is possible with the Japanese language due to the presence of kanji; ideograms from Chinese characters that represent not only sound but meaning. Experimental results show an average performance decrease of 4.6% for ASR results of utterances from adults, and 2.8% for children, when reducing the amount of data for training to its 25%; and a classification performance improvement from 92.2% to 94.1% for adults and 87.2% to 88.3% for children, when using character as features instead of words.

I. INTRODUCTION

Improvements in automatic speech recognition (ASR) technologies have made feasible the implementation of systems that interact with users through speech in real environments. Their application has been studied in telephone-based services [1][2], guidance systems [3][4], and others.

In this work, we address the classification in topics of utterances in Japanese, received by a speech-oriented guidance system operating in a real environment. The system in mention operates in a public space, receiving daily user requests for information and collecting real data. Topic classification in this kind of systems is useful to identify which are user’s main information needs and to ease the selection of proper answers.

In order to perform topic classification it is necessary to collect actual user’s utterances, which are then manually labeled with its associated topic so they can be used as training data for a classifier. The performance of the classifier is expected to improve as the size of the training data increases, but the process of labeling the data is costly. Because of this, we are interested in evaluating the influence of the amount of data for training in the context of a topic classification, to verify how much data would be necessary for achieving an acceptable classification performance.

For this task, we compared the performance of a Support Vector Machine (SVM) and a Maximum Entropy (ME) classifier using training data of different sizes. We used actual data collected by the speech-oriented guidance system Takemaru-kun, from adults and children. Among the utterances received by this system, more than a half correspond to children; however, ASR accuracy tends to be lower for this age group in comparison to adults, due to irregularities in speech typical of small children.

We selected ME and SVM for this task, as both classifiers have been widely used in speech classification. In the work of Evanini et al. [5], ME was shown to outperform five other conventional statistical classifiers in the classification of calls, using user’s responses to a prompt from an automated troubleshooting dialog system. SVM has also been successfully applied in speech classification [2][6][7][8], because it is appropriate for sparse high-dimensional feature vectors and is also robust against speech recognition errors, as it performs classification based on a large number of relevant features rather than relying on a limited set of keywords, which improves robustness even when specific keywords are erroneously recognized [7].

Japanese writing is mainly composed by three scripts: kanji, hiragana and katakana; and it can occasionally include characters from the Latin alphabet. In this work, we propose to use characters as features, in comparison to words, which is possible with the Japanese language due to the presence of kanji, ideograms from Chinese characters that represent not only sound but meaning, in order to deal with the shortness of the utterances that are usually received by this kind of systems.

The remainder of the paper is structured as follows: in Section II, the speech-oriented guidance system Takemaru-kun is described. In Section III, the classification methods are briefly explained. Section IV describes the characteristics of the datasets used in the experiments. Section V presents the conducted experiments and their results. Finally, Section VI
presents the conclusions.

II. SPEECH-ORIENTED GUIDANCE SYSTEM Takemaru-kun
A. Description of the System

The Takemaru-kun system [3] (Figure 1) is a real-environment speech-oriented guidance system, placed inside the entrance hall of the Ikoma City North Community Center located in the Prefecture of Nara, Japan. The system has been operating daily from November 2002, providing guidance to visitors regarding the center facilities, services, neighboring sightseeing, weather forecast, news, and about the agent itself, among other information. Users can also activate a Web search feature that allows searching for Web pages over the Internet containing the uttered keywords. This system is also aimed at serving as field test of a speech interface, and to collect actual utterance data.

The system displays an animated agent at the front monitor, which is the mascot character of Ikoma city, Takemaru-kun. The interaction with the system follows a one-question-to-one-answer strategy, which fits the purpose of responding simple questions to a large number of users. When a user utters an inquiry, the system responds with a synthesized voice, an agent animation, and displays information or Web pages at the monitor in the back, if required.

Since the Takemaru-kun system started operating, the received utterances have been recorded. A database containing the utterances recorded from November 2002 to October 2004 was constructed. The utterances were transcribed and manually labeled, pairing them to specific answers. Information concerning the age group, gender and invalid inputs such as noise, level overflowed shouts and other unclear inputs were also documented. The answer selection in Takemaru-kun system is based on 1-nearest neighbor (1-NN), which classifies an input based on the closest example according to a similarity score. An input utterance is compared to example questions in the database, and the answer paired to the most similar example question is output.

We have heuristically defined 40 topics, grouping questions that are related, using the database constructed during the first two years of operation of the system.

III. CLASSIFICATION METHODS OVERVIEW

This section briefly explains the topic classification methods we experimented with.

A. Maximum Entropy

Maximum Entropy (ME) [9] is a technique for estimating probability distributions from data, and it has been widely used in natural language tasks, including speech classification, where it has shown to outperform other conventional statistical classifiers [5].

As it is expressed in [5], given an utterance consisting of the word sequence \( w_1^N \), the objective of the classifier is to provide the most likely class label \( k \) from a set of labels \( K \):

\[
\hat{k} = \arg\max_{k \in K} p(k | w_1^N)
\]

where the ME paradigm expresses the probability \( p(k | w_1^N) \) as:

\[
p(k | w_1^N) = \frac{\exp \left( \sum_w N(w) \log \alpha(k | w) \right)}{\sum_{k'} \exp \left( \sum_w N(w) \log \alpha(k | w) \right)}.
\]

Ignoring the terms that are constant with respect to \( k \) yields:

\[
\hat{k} = \arg\max_{k \in K} \sum_w N(w) \log \alpha(k | w).
\]

where \( N(w) \) is the frequency of a word in an utterance, and \( \alpha(k | w) \) with \( \alpha(k | w) \geq 0 \) and \( \sum_{k} \alpha(k | w) = 1 \) are parameters that depend on a class \( k \) and a word \( w \).

We applied ME using the package maxent Ver.2.11 [10], which uses the L-BFGS-B algorithm to estimate the parameters. We also selected the ME model with inequality constraints [11], because in preliminary experiments it presented better performance.

B. Support Vector Machine

Support Vector Machine (SVM) tries to find optimal hyperplanes in a feature space that maximize the margin of classification of data from two different classes. For this work, LIBSVM [12] was used to apply SVM. Specifically, we are using C-support vector classification (C-SVC), which implements soft-margin.

We used bag-of-words (BOW) to represent utterances as vectors, where each component of the vector indicates the frequency of appearance of a word. The length of a vector corresponds to the size of the dictionary that includes every word in the training sample set. We selected a Radial Basis Function (RBF) kernel, because in preliminary experiments it presented slightly better performance than a polynomial kernel for this task.

In the problem we are addressing, the amount of samples available for each topic is unbalanced. The SVM primal problem formulation implementing soft-margin for unbalanced amount of samples follows the form:

\[
\min_{w, b, \xi} \quad \frac{1}{2} w^T w + C_+ \sum_{y_i=1} \xi_i + C_- \sum_{y_i=-1} \xi_i
\]

subject to

\[
y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i, \\
\xi_i \geq 0, i = 1, \ldots, l.
\]
where $x_i \in \mathbb{R}^n, i = 1, ..., l$ indicates a training vector, $y_i \in \{1, -1\}$ a class, and $\phi$ is the function for mapping the training vectors into feature space. The hyperparameters $C_+$ and $C_-$ penalize the sum of the slack variable $\xi_i$ 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.

SVM is originally designed for binary classification. We implemented the one-vs-rest approach for multi-class classification, which constructs one binary classifier for each topic, and each one is trained with data from a topic, regarded as positive, and the rest of the topics, regarded as negative. We selected one-vs-rest as in preliminary experiments it presented better performance than one-vs-one for this task.

IV. CHARACTERISTICS OF THE DATASETS

The data used in the experiments was collected by the speech-oriented guidance system Takemaru-kun from November 2002 to October 2004, and it is composed by utterances from adults and children. Julius Ver.3.5.3 was used as ASR engine. The acoustic model was constructed using the Japanese Newspaper Article Sentences (JNAS) corpus, re-training it with valid samples collected by the system in the period indicated above. The language model was constructed using the transcriptions of the same samples. The samples corresponding to the month of August 2003 were used for the test set and were not included in the training set. The rest of the samples were used for the training set. Table I shows the word recognition accuracy of the ASR engine for the datasets.

For these experiments we selected the 15 topics with most training samples. The amount of samples available per topic is shown in Table II. We conducted experiments with transcriptions and ASR 1-best results.

We constructed additional training datasets, reducing the training data to 25% and 50% of the amount of training samples available per topic. The test datasets remained without changes.

V. EXPERIMENTS

We evaluated the influence of the training data size in the topic classification performance with SVM and ME, training the classifiers with the different training datasets. We experimented using characters and words as features, including unigrams, bigrams and trigrams.

Optimal hyper-parameter values for SVM were obtained experimentally using a grid search strategy, and were set a posteriori.

### A. Evaluation Criteria

Classification performance for each topic was evaluated using the F-measure, as defined by:

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

### B. Experiment Results

The F-measure was calculated individually for the classification of each topic and it was averaged by frequency of samples in the topics. Fig. 2 presents the best performance obtained with each training dataset and method in the classification of ASR 1-best results. Tables III and IV present a summary of the best obtained results.

We can observe that although the training dataset for children is three times larger than the one for adults, classification performance is still lower for utterances from children, due to the lower ASR accuracy. We can also see that the classification performances of transcriptions of utterances from adults and children are very close, in spite of the difference in size of the training datasets. SVM using character unigrams and bigrams
Table III: F-Measure (%) per training data size and method for adults

<table>
<thead>
<tr>
<th>Training Data Size</th>
<th>25%</th>
<th>50%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (Word 1+2gram)</td>
<td>88.7</td>
<td>91.9</td>
<td>95.5</td>
</tr>
<tr>
<td>SVM (Char 1+2gram)</td>
<td>91.0</td>
<td>93.7</td>
<td>96.4</td>
</tr>
<tr>
<td>ME (Word 1+2gram)</td>
<td>85.5</td>
<td>90.9</td>
<td>93.9</td>
</tr>
<tr>
<td>ME (Char 1+2gram)</td>
<td>90.1</td>
<td>93.7</td>
<td>95.5</td>
</tr>
<tr>
<td>ASR 1-best</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| SVM (Word 1+2gram) | 86.9 | 90.3 | 93.2 |
| SVM (Char 1+2gram) | 89.3 | 91.8 | 94.4 |
| ME (Word 1gram)    | 84.8 | 89.5 | 92.2 |
| ME (Char 1+2+3gram)| 89.2 | 92.2 | 94.1 |

As features presented the highest classification performance for the adults dataset, with an F-measure of 94.4%; while ME using character unigrams, bigrams, and trigrams presented the highest performance for the children dataset, with an f-measure of 88.3%.

There is an average performance decrease of 4.6% for ASR results of utterances from adults, and 2.8% for children, when reducing the amount of data for training to its 25%; mainly because of the smaller size of the adults dataset.

The classification performance was improved from 92.2% to 94.1% for adults and 87.2% to 88.3% for children, when using character as features instead of words. This suggests that, since kanji characters also include meaning, using characters for the classification of short utterances in Japanese can enhance the amount of information available for a proper classification.

As SVM is a binary classifier, when implementing an approach as one-vs-rest for multi-class classification, some samples can result positive for two or more topics at the same time. On these experiments, an average of 1.2% of the test samples were classified in the correct topic and some other topic at the same time. For these cases we assumed them as a correct classification. On the other hand, a sample may result positive for none of the topics, and in these cases the sample is rejected. Fig. 3 presents the rejection rate of test samples per training data size for transcriptions and ASR 1-best results. We can observe that as the training data size increases, the rejection rate decreases. These situations were not present with ME, as it is a multi-class classifier.

VI. CONCLUSIONS
This work evaluated the influence of the amount of data for training in the context of topic classification. For this, we compared the performance of Support Vector Machine and Maximum Entropy using training datasets of different sizes. We could observe that if the ASR accuracy is high enough, we can have better classification performance with smaller training datasets. We could also see that using characters as features, instead of words, can yield to improvements in the classification performance.

REFERENCES