Comparison of Methods for Topic Classification in a Speech-Oriented Guidance System

Rafael Torres¹, Shota Takeuchi¹, Hiromichi Kawanami¹, Tomoko Matsui², Hiroshi Saruwatari¹, Kiyohiro Shikano¹

¹Graduate School of Information Science, Nara Institute of Science and Technology, Japan
²Department of Statistical Modeling, The Institute of Statistical Mathematics, Japan
{rafael-t, shota-t, kawanami, sawatari, shikano}@is.naist.jp, tmatsui@ism.ac.jp

Abstract
This work addresses the classification in topics of utterances in Japanese, received by a speech-oriented guidance system operating in a real environment. For this, we compare the performance of Support Vector Machine and PrefixSpan Boosting, against a conventional Maximum Entropy classification method. We are interested in evaluating their strength against automatic speech recognition (ASR) errors and the sparseness of the features present in spontaneous speech. To deal with the shortness of the utterances, we also proposed to use characters as features instead of 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. Experimental results show a classification performance improvement from 92.2% to 94.4%, with Support Vector Machine using character unigrams and bigrams as features, in comparison to the conventional method.

Index Terms: Speech-Oriented Guidance System, Topic Classification, Support Vector Machine, PrefixSpan Boosting, Maximum Entropy

1. 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 of utterances in this kind of systems is useful to identify which are user’s main information needs and to ease the selection of proper answers.

For this task, we compare the performance of Support Vector Machine (SVM) and PrefixSpan Boosting (pboost), against a conventional Maximum Entropy (ME) classification method. We selected ME [11] using word n-grams features as baseline, as in the work of Evanini et al. [8] it 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 was not included in the comparison, as it is a discriminative classification method.

SVM has successfully been applied in speech classification [2, 5, 6, 7], because it is appropriate for sparse high-dimensional feature vectors, and it is also robust against speech recognition errors. SVM performs classification based on a large number of relevant features rather than relying on a limited set of keywords, and its training is based on margin maximization. Improving robustness even when specific keywords are erroneously recognized [6].

Boosting techniques have also been used for speech classification [2, 5, 8]; however, in this work we are introducing pboost for this task. Pboost is a method proposed by Nozowin et al. [9] for action classification in videos. Pboost implements a generalization of the PrefixSpan algorithm by Pei et al. [10] to find optimal discriminative patterns, and in combination with the Linear Programming Boosting (LPBoost) classifier, it optimizes the classifier and performs feature selection simultaneously.

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 2, the speech-oriented guidance system “Takemaru-kun” is described. In Section 3, the classification methods are briefly explained. Section 4 presents the conducted experiments and their results. Finally, Section 5 presents the conclusions.

2. Speech-Oriented Guidance System Takemaru-kun

2.1. 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 over 100K 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. Valid utterances showed to be relatively short, with an average length of 3.65 words per utterance. 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.

3. Classification Methods Overview

This sections briefly explains the topic classification methods we are comparing in this work.

3.1. Maximum Entropy

Maximum Entropy (ME) [11] is a technique for estimating probability distributions from data, and it has been widely used in natural language tasks, including speech classification, where the ME paradigm expresses the probability $p(k|\omega)$ as:

$$p(k|\omega) = \frac{\exp \left( \sum_w N(w) \log \alpha(k|w) \right)}{\sum_{\omega'} \exp \left( \sum_w N(w) \log \alpha(k|w) \right)}.$$  

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 $\sum_w \alpha(k|w) = 1$ are parameters that depend on a class $k$ and a word $w$.

We applied ME using the package maxent Ver.2.11 [12], which uses the L-BFGS-B algorithm to estimate the parameters. We also selected the ME model with inequality constraints [13], because in preliminary experiments it presented better performance.

3.2. 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 [14] 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:

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

subject to $y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i$, $\xi_i \geq 0, i = 1, \ldots, l$.

where $x_i \in \mathbb{R}^n$, $i = 1, \ldots, l$ 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.

3.3. PrefixSpan Boosting

In this work we are introducing PrefixSpan Boosting (pboost) for the classification of utterances in topics. Pboost is a method proposed by Nozowin et al. [9] for action classification in videos. Pboost implements a generalization of the PrefixSpan algorithm by Pei et al. [10] to find optimal discriminative patterns, and in combination with the Linear Programming Boosting (LPBoost) classifier, it optimizes the classifier and performs feature selection simultaneously.

In pboost, the presence of a single discriminative pattern in a sample, in our case a word sequence that could include gaps, is
checked by weak hypotheses, which have the form $h(x, s, \omega)$, where $x \in \{x_i \}, x_i \in \mathbb{R}^n, i = 1, \ldots, l$ is a training vector, $s$ is a word sequence and $\omega \in \Omega, \Omega = \{-1, 1\}$ is a variable that allows the sequence to decide for either class.

The classification function has the form:

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

(5)

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

To deal with the unbalance between positive and negative samples, we are using an extended version of the method that allows to implement soft-margin for unbalanced amount of samples. The pboost primal problem then takes this form:

$$\min_{(s, \omega) \in S \times \Omega} -\rho + D_+ \sum_{y_i = 1} \xi_i + D_- \sum_{y_i = -1} \xi_i$$

(6)

subject to:

$$\sum_{(s, \omega) \in S \times \Omega} y_i \alpha_{s, \omega} h(x_i; s, \omega) + \xi_i \geq \rho, i = 1, \ldots, l$$

where $x_i \in \mathbb{R}^n, i = 1, \ldots, l$ indicates a training vector, $y_i \in \{1, -1\}$ a class, $\rho$ is the soft margin separating negative from positive samples, and $D_+ = \frac{1}{\rho}$ and $\nu \in (0, 1)$ is a hyper-parameter controlling the cost of misclassification, which in this case is separated into $D_+$ and $D_-$, penalizing the sum of the slack variable $\xi_i$ for each class, that allows the margin constraints to be slightly violated. As in SVM, by introducing different hyperparameters $\nu_+$ and $\nu_-$ we can deal with the unbalanced amount of data problem.

We also implemented the one-vs-rest approach for multi-class classification, as in preliminary experiments it presented better performance than one-vs-one for this task.

4. Experiments

We compared the performance of the methods in the classification in topics of utterances in Japanese, received by a speech-oriented guidance system operating in a real environment. Optimal hyper-parameter values for SVM and pboost were obtained experimentally using a grid search strategy, and were set a posteriori. The experiments and obtained results are detailed below.

4.1. Characteristics of the Data

The data used in the experiments was collected by the speech-oriented guidance system Takemaru-kun from November 2002 to October 2004. For these experiments we only considered valid utterances from adults.

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. The word recognition accuracy of the ASR engine was 85.66% for the training set and 85.10% for the test set.

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

4.2. Evaluation Criteria

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

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

(7)

The classification performance of ME using word unigrams as features was used as baseline.

4.3. Experiment Results

The purpose of these experiments was to compare the classification performance of the described methods, and evaluate their strength against ASR errors. Due to the sparseness of the features and the shortness of the utterances, we also proposed to use characters as features instead of words, which is possible with the Japanese language due to the presence of kanji characters. We also conducted experiments including bigrams and unigrams.
trigrams as features.

The F-measure was calculated individually for the classification of each topic and it was averaged by frequency of samples in the topics. Table 2 and 3 present a summary of the obtained results. Figure 2 presents the best performance obtained with each method in the classification of ASR 1-best results, using words and characters as features.

We can observe that by using characters as features instead of words, we could achieve improvements in the classification performance of the three methods. 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.

We could also corroborate that the inclusion of bigrams and trigrams as features can improve classification performance; however, in some cases no further improvement was achieved by including trigrams, and there were even some cases where the performance decreased by including them.

The difference in classification performance between transcriptions and ASR 1-best results was around 2% in average, which indicates that the methods that were compared can be robust against ASR errors.

The method that presented the best performance in the classification of ASR 1-best results was SVM using character unigrams and bigrams as features, with an f-measure of 94.4%. This represents a difference of 2.2% in comparison to the baseline; followed almost without a significant difference in performance by pboost and ME using character unigrams, bigrams and trigrams.

Pboost can find discriminative patterns that include gaps, as in the pattern [where][library]. When words were used as features, this yielded to improvements in classification performance; however, when using characters as features, the classification performance decreased. Because of that, for these experiments gaps were only allowed when using words, but not with characters.

A grammatical analysis of the optimal discriminative words selected by pboost indicated that the most important part of speech (POS) for topic discrimination is the noun, which accounted in average for more than a half of the selected patterns in the different classifiers. Following, the verb, which accounted in average for nearly a seventh of the selected patterns. Particles, Japanese parts of speech that relate the preceding word to the rest of the sentence, were also selected in some cases as optimal discriminative words. Nouns proved to be important for topic discrimination, since some nouns are characteristic of specific topics, as opposed to other POS that can be found in broader numbers of topics.

5. Conclusions

This work compared the performance of Support Vector Machine, PrefixSpan Boosting and Maximum Entropy for the classification in topics of utterances in Japanese received by a speech-oriented guidance system. Support Vector Machine using character unigrams and bigrams as features presented the best classification performance. Using characters as features, instead of words, yielded to improvements in the classification performance of all the methods that were compared.

6. References