Vol.2009-SLP-77 No.13
2009/7/18

Inquiry Classification in a Speech-Oriented Guidance System Using Discriminative Learning

Rafael TORRES,[†1] Shota TAKEUCHI,[†1]
Hiromichi KAWANAMI,[†2] Tomoko MATSUI,[†1]
Hiroshi SARUWATARI[†1] and Kiyohiro SHIKANO[†1]

The issue of providing appropriate responses for users' inquiries in a speech-oriented guidance system can be addressed with an example-based response generation method, which compares a user's utterance with example questions stored in a question and answer database (QADB), and the answer corresponding to the most similar example question is selected. An advantage of this method is that the system can be easily expanded to cover other classes of inquiries. However, questions from different classes could present high similarity, which could hinder the selection of an appropriate answer. The selection of an appropriate answer is basically a multi-class classification problem, and to deal with it, we study the use of the bag of words (BOW) kernel to represent the features of users' inquiries and using support vector machine (SVM) for the classification, implementing a one-vs-all multi-class classification approach. Preliminary experimental results using transcribed data show that the proposed method improves the classification performance of users' inquiries, compared to a baseline method.

1. Introduction

Since one of the most important means for social interaction among humans is speech, spoken dialog systems can provide a natural and friendly interface for human-machine interaction, allowing users to have their hands free to perform other activities. Their application has been studied in call-routing services[1)], multi-modal information providers[2)], guidance in facilities, and others.

Speech-oriented guidance systems interact with users through speech to provide guidance in a specific area of knowledge or expertise. In order to have a natural interaction, the system should be able to provide appropriate answers for user's inquiries. The selection of an appropriate response is hindered if we take in consideration that spontaneous speech usually does not follow strict rules, and other forms of words that are not common in formal language can also be used. Besides, speakers' utterances can be very short, and could also contain words that do not provide relevant information for an answer selection.

Example-based response generation methods employ a question and answer database (QADB), where example questions are associated to specific answers. These methods have the advantage that the system can be easily expanded to cover other classes of inquiries. By establishing inquiry categories, we can approach the selection of an appropriate answer as a multi-class classification problem, where the variable amount of available samples for each class and the short length of the utterances in spontaneous speech are facts that should be taken in consideration when implementing a classification method.

In this paper, we study the implementation of support vector machine (SVM) in conjunction with the Bag of Words (BOW) kernel function, following a 1-vs-all multi-class classification approach for the response selection in a speech oriented guidance system. In Section 2, the speech-oriented guidance system "Takemaru-kun" is described. In Section 3, the details of the proposed method are explained. In Section 4, experimental results with the proposed method are presented. Section 5 presents the conclusions.

2. Speech-Oriented Guidance System "Takemaru-kun"

2.1 Description of the System

The "Takemaru-kun" system (Figure 1) is a real world speech-oriented guidance system placed at the entrance hall of the Ikoma-City North Community Center, located at Ikoma City, Nara Prefecture. The system has been operating daily from November 2002, to provide visitors a speech interface for information retrieval[3)]. The system offers guidance to users regarding facilities in the building, services that are offered, information about the city, sightseeing and others.

The system displays an animated agent at the front monitor, and visual information and Web pages at the monitor in the back. The interaction follows a

†1 Graduate school of Information Science, Nara Institute of Science and Technology
†2 Department of Statistical Modeling, The Institute of Statistical Mathematics

one-question and one-response principle, and when a user speaks to a microphone on the desk, the system responds with a synthesized voice and displays information if required. The speech recognition in the system is carried out with the speech recognition engine Julius[4]. The system can discriminate between adult and child's voices give answers according to the age group. The system comprises 275 answer categories for adults and 285 for children. Users can also activate a Web search feature that allows searching for Web pages over the Internet containing the uttered keywords[5].

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

2.2 Question and Answer Database

Since the Takemaru-kun system started operating, the received utterances have been recorded. A QADB containing the utterances recorded from November 2002 to October 2004 was constructed. The utterances were transcribed and classified manually by labelers. Information concerning the age group, gender and invalid inputs such as noise, level overflowed shouts and other unclear inputs were also documented.[3]

2.3 Response Generation Method

The Takemaru-kun system implements a one-question and one-answer principle, which does not require performing in-depth language understanding and dialoging control, since inquiries are not supposed to depend on the dialog history.

The response generation method implemented in the Takemaru-kun system compares an input utterance with example questions stored in the QADB, and outputs the response that corresponds to the example question that is most similar to the input.[6]

This response generation method is an extension of 1-Nearest-Neighbor (1-NN), which classifies an input based on the closest examples and do not require a training step. A similarity score is calculated between the input and example questions in the QADB. The similarity score is calculated as following:

$$\text{Similarity Score} = \frac{\sum(\#\text{ of word coincidences between input and example})}{max(\#\text{ of words in input},\#\text{ of words in example})} \quad (1)$$

This similarity score takes in consideration the number of words in common between the input and the example question, and divides it by the maximum length, normalizing the similarity score to remove the effect of the length of the samples.

On this implementation of 1-NN, the similarity score of the input is calculated for all the examples in the QADB, then the example with the highest similarity score for each category of answer is selected, constructing a response vector for the input. The input is then classified in the answer category of the example that presents highest similarity score.

3. Our Response Generation Method

To address the selection of appropriate answers in a speech-oriented guidance system, we study the use of the bag of words kernel to represent the features of users' inquiries and support vector machine for the inquiries classification, implementing a one-vs-all multi-class classification approach.

3.1 Bag of Words (BOW) Kernel Function

In BOW, a sample utterance is represented as a vector where each element indicates the frequency of appearance of a word in the sample utterance. A dictionary is constructed including every word present in the training sample set, and the length of each sample vector is the amount of words that are included in the dictionary. The polynomial function was selected to implement the nonlinear form of the BOW kernel:

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T \boldsymbol{x}_j + 1)^d \quad (2)$$

Where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represent sample vectors and d represents the dimension.

3.1.1 TF-IDF

BOW takes in consideration the frequency of words in the sample (TF). The IDF weight is applied to each word in the dictionary, as follows:

$$w_i = \ln\left(\frac{K}{t_i}\right) \tag{3}$$

Where w_i represents the weight applied to a term i, K is the total amount of classes and t_i is the number of classes where the word i appears at least once. Thus, the fewer classes the word appear in, the greater the weight. This weight acts as a measure of how important a word is for the discrimination of a specific class.

In BOW with TF-IDF, each element of the sample vector indicates the frequency of appearance of a word in the sample utterance, multiplied by the weight associated to that word.

Given the IDF weight w_i, we can define a diagonal matrix $\boldsymbol{R}$ as:

$$\boldsymbol{R}_{ii} = w_i \tag{4}$$

Then, the BOW polynomial kernel with TF-IDF can defined as defined as:

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T \boldsymbol{R}^T \boldsymbol{R} \boldsymbol{x}_j + 1)^d \tag{5}$$

3.1.2 Normalization

The length of the sample utterances is not considered as relevant for the classification. To remove the effect of the sample length, we normalize the vectors as follows:

$$\hat{\kappa}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\sqrt{\kappa(\boldsymbol{x}_i, \boldsymbol{x}_i)\kappa(\boldsymbol{x}_j, \boldsymbol{x}_j)}} \tag{6}$$

Where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represent sample vectors.

3.2 Support Vector Machine (SVM)

SVM is a supervised learning method for classification and regression. SVM tries to find optimal hyperplanes in a feature space that maximize the margin of classification of data from two different categories. The LIBSVM[7] library is used to implement SVM in the proposed method.

In the problem we are addressing, the amount of samples that are available for the classes are unbalanced. The SVM primal problem formulation in LIBSVM implementing soft margin for unbalanced amount of samples follows the form:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C_+ \sum_{y_i=1} \xi_i + C_- \sum_{y_i=-1} \xi_i \\ \text{subject to} \quad & y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, ..., l. \end{aligned} \tag{7}$$

Here $\boldsymbol{x}_i \, \epsilon \, R^n, i = 1, ..., l$ indicates a training vector and $y_i \, \epsilon \, \{1, -1\}$, $i = 1, ..., l$ a class. The hyper-parameters C_+ and C_- penalize the sum of the slack variables ξ_i for each class that allow the margin constraints to be violated. By introducing different hyper-parameters of C_+ and C_- for each class, the unbalanced data problem in which the SVM parameters are not estimated robustly due to the unbalanced numbers of training vectors for each class can be dealt with.

In our method, the values of C_+ and C_- are set according to:

$$C_+ = \frac{\text{Amount of samples in the rest of the classes}}{\text{Total amount of samples}} \times C \tag{8}$$

$$C_- = \frac{\text{Amount of samples in the class}}{\text{Total amount of samples}} \times C \tag{9}$$

where $C_+ + C_- = C$

3.2.1 Multi-class Classification Approach

SVM is originally a binary classifier. In order to classify more than two classes, we are implementing the 1-vs-all multi-class classification approach, which constructs one binary classifier for each class and each one is trained with data from one class and the rest of the classes.

SVM does not provide probability information. The LIBSVM library includes an extension of SVM that provides a pseudo probability or probability estimate[7]. In our method, the input sample is classified in the class that presents the highest pseudo probability given by LIBSVM.

4. Experiments

We evaluated the proposed method using transcribed utterances from adults collected by the speech-oriented guidance system Takemaru-kun. For these ex-

periments, we established 40 broad classes from the total amount of classes and selected the 15 broad classes with most training samples. The training dataset consisted on 14432 samples, and the test dataset on 792 samples. The amount of samples available per class is shown in Table 1.

Table 1 Amount of samples per class

Class Description	Train	Test
chat-compliments	766	49
info-services	494	35
info-news	484	37
info-local	553	32
info-facility	1795	80
info-city	504	24
info-weather	1099	62
info-time	984	53
info-sightseeing	668	10
info-access	676	33
greeting-end	912	69
greeting-start	2673	159
agent-name	1309	70
agent-likings	851	44
agent-age	664	35

In the experiments, the value of the hyper-parameter d of the BOW polynomial kernel function was varied from 2 to 6, and the value of the hyper-parameter C of SVM was varied from 0.00001 to 10000. The performance in the classification was evaluated according to a correct classification rate, which takes in consideration the amount of samples that were correctly classified in a specific class:

$$\text{Correct Classification Rate} = \frac{\text{Amount of correclty classified samples in the class}}{\text{Total amount of samples in the class}} \tag{10}$$

1-NN classification results per class using similarity score were used as baseline for the evaluation.

4.1 Experimental Results

The best classification performance was obtained by setting the hyper-parameters d in 3 and C in 0.01. Table 2 shows the correct classification rates of the proposed method and the baseline. The performance of the proposed method in the classification of the train set is shown in Figure 2 and the performance for the test set is shown in Figure 3. The performance of the baseline method in the classification of the test set is shown in Figure 4. Experimental results show that the proposed method presents better performance in comparison to the baseline in the individual classification of every class, except in the classification of the "info-service" class, where the proposed method presented lower performance, and in the "agent-age" class, where they presented equal performance. The weighted average performance of the proposed method is 91.79%, which improves the baseline performance by 6.56%. In the case of the "info-service" class, we can attribute the lower performance of the proposed method to the fact that this class is compound by samples that present a high variation. Since the baseline method classifies the samples according to the similarity to individual example questions, this condition does not significantly affect its performance. In the case of the "chat-complimented" class, it is mainly composed by very short samples, and for this case, the baseline method presents difficulties in the classification.

Table 2 Correct classification rate per class of the proposed and baseline methods

Class Description	BOW+SVM (Train set)	BOW+SVM (Test set)	Baseline (Test set)
chat-compliments	91.38%	81.63%	46.94%
info-services	84.82%	74.29%	88.57%
info-news	97.31%	89.19%	75.68%
info-local	94.76%	93.75%	75.00%
info-facility	90.25%	80.00%	78.75%
info-city	87.50%	91.67%	79.17%
info-weather	97.54%	100.00%	98.39%
info-time	98.48%	92.45%	90.57%
info-sightseeing	93.71%	100.00%	90.00%
info-access	94.53%	100.00%	90.91%
greeting-end	95.07%	92.75%	88.41%
greeting-start	96.22%	97.48%	95.60%
agent-name	94.35%	95.71%	84.29%
agent-likings	95.30%	93.18%	81.82%
agent-age	96.84%	88.57%	88.57%
Weighted average	94.30%	91.79%	85.23%

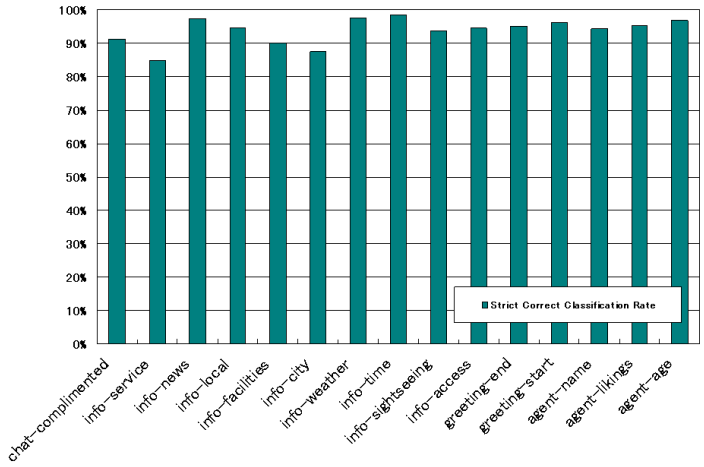

Fig. 2 Classification performance per class of the train set (BOW+SVM)

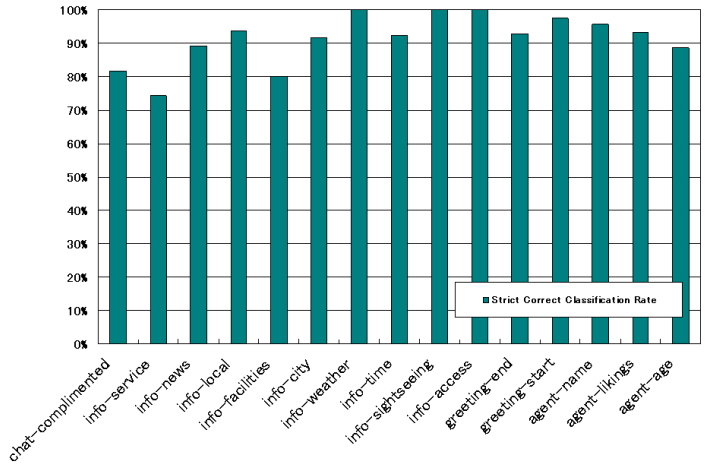

Fig. 3 Classification performance per class of the test set (BOW+SVM)

5. Conclusions

A discriminative learning method based on SVM in conjunction with BOW was considered to address the classification of inquiries in a speech-oriented guidance system. Experimental results show that the proposed method improves the average classification performance by 6.56% and performs better than the baseline in the individual classification of most of the classes. Future work will be focused on experiments with speech recognition results instead of transcribed utterances, children utterances and experiments with a larger number of classes.

References

1) A.L.Gorin, G. Riccardi, J.H.Wright: *How may I help you?*, Speech Communication, vol.23, pp.113-127, 1997.
2) Joakim Gustafson, Nikolaj Lindberg, Magnus Lundeberg: *The August Spoken Dialog System* , Proc. of EUROSPEECH'99, vol.3, pp.1151-1154, 1999.
3) Ryuichi Nisimura, Akinobu Lee, Hiroshi Saruwatari, Kiyohiro Shikano: *Public*

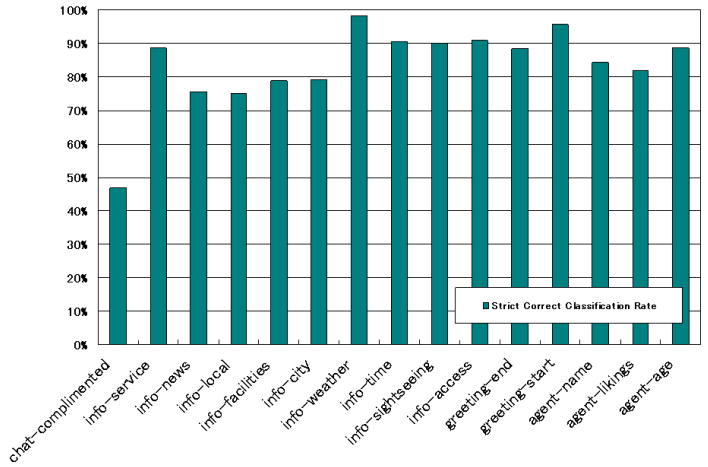

Fig. 4 Classification performance per class of the test set (Baseline method)

Speech-Oriented Guidance System with Adult and Child Discrimination Capability , Proc. of ICASSP2004, vol.1, pp.433-436, 2004.

4) Akinobu Lee, Tatsuya Kawahara, Kiyohiro Shikano: *Julius - an Open Source Real-Time Large Vocabulary Recognition Engine* , Proc. EUROSPEECH 2001, pp.1691-1694, 2001.

5) Jumpei Miyake, Shota Takeuchi, Hiromichi Kawanami, Hiroshi Saruwatari, Kiyohiro Shikano: *Language Model for the Web Search Task in a Spoken Dialogue System for Children* , Proc. of Workshop on Child, Computer and Interaction (ICMI'08 post-conference workshop), Chania, Greece, October 2008.

6) Shota Takeuchi, Tobias Cincarek, Hiromichi Kawanami, Hiroshi Saruwatari, Kiyohiro Shikano: *Question and Answer Database Optimization Using Speech Recognition Results* , INTERSPEECH 2008, pp.451-454, Sep, 2008.

7) Chih-Chung Chang and Chih-Jen Lin: *LIBSVM: a Library for Support Vector Machines* , http://www.csie.ntu.edu.tw/cjlin/libsvm