Topic Classification of Spoken Inquiries Using Transductive Support Vector Machine

Rafael Torres, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari and Kiyohiro Shikano

Abstract In this work, we address the topic classification of spoken inquiries in Japanese that are received by a guidance system operating in a real environment, with a semi-supervised learning approach based on a transductive support vector machine (TSVM). Manual data labeling, which is required for supervised learning, is a costly process, and unlabeled data are usually abundant and cheap to obtain. TSVM allows to treat partially labeled data for semi-supervised learning, including labeled and unlabeled samples in the training set. We are interested in evaluating the influence of including unlabeled samples in the training of the topic classification models, as well as the amount of them that could be necessary for improving performance. Experimental results show that this approach can be useful for taking advantage of unlabeled samples, especially when using larger unlabeled datasets. In particular, we found gains in classification performance for specific topics, such as city information, with a 6.30% F-measure improvement in the case of children's inquiries, and 7.63% for access information in the case of adults' inquiries.

1 Introduction

The interest of this work is to improve topic classification performance of spoken inquiries in Japanese, received by a speech-oriented guidance system operating in a real environment. In previous work, we evaluated the classification performance of three supervised methods: a support vector machine (SVM) with a radial basis function (RBF) kernel, PrefixSpan boosting (pboost) and maximum entropy (ME)[5].

Rafael Torres, Hiromichi Kawanami, Hiroshi Saruwatari and Kiyohiro Shikano
Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan
e-mail: {rafael-t, kawanami, sawatari, shikano}@is.naist.jp

Tomoko Matsui
Department of Statistical Modeling, The Institute of Statistical Mathematics, Tokyo, Japan
e-mail: tmatsui@ism.ac.jp

We have also evaluated a stacked generalization scheme to combine the predictions of these three classifiers, improving predictive accuracy compared with the performance of individual classifiers[4].

Manual data labeling, which is required for supervised learning, is a costly process, and unlabeled data are usually abundant and cheap to obtain. Because of this, it is desirable to be able to use unlabeled samples to improve topic classification performance and minimize the generalization error of the classifiers. In the present work, we address the topic classification of spoken inquiries in Japanese received by a guidance system, with a semi-supervised learning approach based on a TSVM, which extends a regular SVM to treat partially labeled data, including labeled and unlabeled samples in the training set.

TSVMs were proposed by Vapnik in 1998, and were introduced by Joachims[2] 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.

Our task, topic classification of spoken inquiries, shares some similarities with text classification; however, classification of spontaneous speech includes automatic speech recognition (ASR) errors. In this work we evaluate the viability of using a TSVM for semi-supervised learning in this task, as well as the amount of unlabeled data that would be necessary for improving classification performance.

2 Speech-Oriented Guidance System *Takemaru-kun*

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 in Nara, Japan, and it has been operating daily since November 2002.

The system uses a one-question-to-one-response strategy for interaction, which fits the purpose of responding simple questions to a large number of users. It provides information about the center facilities and services, local sightseeing, weather forecast and news, among other.

Since the *Takemaru-kun* system started operating, the received utterances have been recorded. Utterances from Nov. 2002 to Oct. 2004 and from Dec. 2004 to Mar.

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

2005 have been manually transcribed and labeled. However, because this is a very costly process, there is still a vast amount of data that remains unlabeled.

3 Transductive Support Vector Machine

A transductive support vector machine (TSVM) extends a regular SVM to treat partially labeled data for semi-supervised learning, including labeled and unlabeled samples in the training set. In this work we use SVMLight[1] to implement it.

In TSVM, the primal optimization problem follows the form:

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

where $\mathbf{x}_i$ represents a labeled training sample and $\mathbf{x}_j$ 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[2] 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.

In our approach, we use labeled and unlabeled samples to train a model using a TSVM, and use the resultant model to classify test data; which differs from the approach of Joachims[2], where unlabeled samples are the test data.

We use bag-of-words (BOW) to represent utterances as vectors, and use character unigrams, bigrams and trigrams as features, as it was previously shown to improve

classification performance in comparison to words[5]. We also use a radial basis function (RBF) kernel and follow a one-vs-rest approach, constructing one binary classifier for each topic, as it showed better performance in preliminary experiments.

4 Experiments

We evaluated the performance of a TSVM against a regular SVM used as baseline. For this, we classified ASR results of inquiries in Japanese received by the speech-oriented guidance system *Takemaru-kun* in topics. We performed experiments with separate datasets for children and adults. 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.

4.1 Characteristics of the Datasets

The labeled data correspond to the utterances collected by *Takemaru-kun* in the period from Nov. 2002 to Oct. 2004 and from Dec. 2004 to Mar. 2005. Julius was used as ASR engine. Acoustic models (AMs) and language models (LMs) were separately prepared for children and adults. The AMs were trained using the samples collected by the system from Nov. 2002 to Oct. 2004, and the LMs were constructed using the transcriptions of the samples of the same period. Samples corresponding to the months of Aug. 2003 and from Dec. 2004 to Mar. 2005 were used for testing and were not included in the training sets. For these experiments we selected the 15 topics with most training samples. Table 1 shows the amount of samples and word recognition accuracy of the ASR engine 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, and we used the same AMs and LMs that were used to recognize the labeled data. We created three datasets, incrementing the size of them. Table 2 shows the amount of samples in the unlabeled datasets.

Table 1 Amount of samples and ASR word recognition accuracy in the labeled datasets

(Labeled datasets)	Children Training	Children Test	Adults Training	Adults Test
Amount of samples	43494	15524	14431	3085
ASR word recognition acc.	72.95%	66.77%	88.42%	81.60%

Table 2 Amount of samples in the unlabeled datasets

(Unlabeled datasets)	Children Training	Adults Training
Unlabeled dataset #1 (2005.04 to 2005.12)	119322	110537
Unlabeled dataset #2 (2005.04 to 2006.12)	271744	252428
Unlabeled dataset #3 (2005.04 to 2007.12)	413144	385165

4.2 Experiment Results

Table 3 presents the averaged topic classification performance per training dataset combination in the open test, for children and adults. In the case of the children's datasets, the TSVM outperformed the baseline when using the two largest unlabeled datasets, with an improvement of 0.74% when using the largest one. However, in the case of the adults' datasets, the averaged performance of the TSVM was not better than the baseline.

We can observe that the topic classification performance with children's datasets is lower in comparison to adults', which leaves more room for improvement. The main reason for this is the lower ASR accuracy for children. The results obtained with the TSVM suggest that the inclusion of unlabeled samples in the training of the topic classification models can help to deal with the influence of ASR errors. We can also observe a tendency to obtain better performance with the TSVM when using larger unlabeled datasets.

Table 4 presents the classification performance per topic. We can observe that most of the topics presented improvements in the case of children's data, while more than half of the topics were improved for adults' data. The best gain in performance for children's inquiries was presented by the *info-city* topic, with 6.30% F-measure improvement, and 7.63% by *info-access* for adults' inquiries.

Table 3 Averaged F-measure results per training dataset combination (open test)

Training Dataset Combination	Children	Adults
Labeled only (SVM)	83.54%	**93.03%**
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%

Table 4 F-measure results per topic (open test)

Topic	Children SVM	Children TSVM	Adults SVM	Adults TSVM
chat-compliments	64.24%	**66.91%**	**86.35%**	81.64%
info-services	58.06%	**59.04%**	**87.65%**	87.12%
info-news	88.89%	**92.47%**	95.52%	**96.30%**
info-local	56.71%	**59.50%**	83.08%	**84.44%**
info-facility	**82.70%**	82.15%	**89.36%**	**89.36%**
info-city	67.06%	**73.37%**	84.34%	**88.27%**
info-weather	83.89%	**85.40%**	95.46%	**95.74%**
info-time	89.67%	**90.83%**	95.56%	**96.83%**
info-sightseeing	74.74%	**75.34%**	91.39%	**92.21%**
info-access	44.58%	**44.89%**	84.39%	**92.02%**
greeting-end	**84.87%**	83.81%	**93.48%**	92.17%
greeting-start	91.64%	**91.88%**	97.76%	**97.83%**
agent-name	79.15%	**80.84%**	**92.15%**	89.50%
agent-likings	89.75%	**90.62%**	**93.30%**	91.64%
agent-age	89.26%	**89.69%**	**95.71%**	95.68%
Averaged	83.54%	**84.28%**	**93.03%**	92.81%

5 Conclusions

This work evaluated the topic classification of spoken inquiries received by a guidance system with a semi-supervised learning approach based on a TSVM. Experimental results with children's data show an overall improvement of 0.74% with the TSVM in comparison to a regular SVM. In particular, we found gains in classification performance for specific topics, such as city information, with 6.30% F-measure improvement for children's inquiries, and 7.63% for access information in the case of adults' inquiries. A tendency to obtain better performance when using larger unlabeled datasets was observed. Future work will focus on the evaluation and improvement of other semi-supervised learning approaches.

References

1. Joachims, T.: Advances in kernel methods. chap. Making large-scale support vector machine learning practical, pp. 169–184. MIT Press (1999). Software available at http://svmlight.joachims.org/
2. Joachims, T.: Transductive inference for text classification using support vector machines. In: ICML 1999 Proceedings, pp. 200–209 (1999)
3. Nisimura, R., Lee, A., Saruwatari, H., Shikano, K.: Public speech-oriented guidance system with adult and child discrimination capability. In: ICASSP 2004 Proc., pp. 433-436 (2004)
4. Torres, R., Kawanami, H., Matsui, T., Saruwatari, H., Shikano, K.: Topic classification of spoken inquiries based on stacked generalization. In: APSIPA 2011 Proceedings (2011)
5. Torres, R., Takeuchi, S., Kawanami, H., Matsui, T., Saruwatari, H., Shikano, K.: Comparison of methods for topic classification in a speech-oriented guidance system. In: Interspeech 2010 Proceedings, pp. 1261–1264 (2010)