RC23745 (W0510-070) October 11, 2005
Computer Science

IBM Research Report

Discriminative Training and Support Vector Machine for Natural Language Call Routing

Imed Zitouni
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Hui Jiang
York University
Toronto, Ontario
M3J 1P3 Canada

Qiru Zhou
Bell Labs
Lucent Technologies
Murray Hill, NJ 07974

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

DISCRIMINATIVE TRAINING AND SUPPORT VECTOR MACHINE FOR NATURAL LANGUAGE CALL ROUTING

Imed Zitouni[1]*, Hui Jiang*[2]*, Qiru Zhou*[3]

[1] IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
[2] York University, Toronto, Ont. M3J 1P3, CANADA
[3] Bell Labs, Lucent Technologies, Murray Hill, NJ 07974, USA
izitouni@us.ibm.com, hj@cs.yorku.ca, qzhou@lucent.com

ABSTRACT

In natural language call routing, callers are routed to desired departments based on natural spoken responses to an open-ended "How may I direct your call?" prompt. Natural language call classification can be performed using support vector machines (SVMs) or the popular vector-based model used in information retrieval. We recently demonstrate how discriminative training is powerful to improve any parameterized vector-based classifier to achieve minimum classification error. Discriminative training minimizes the classification error by increasing the score separation of the correct from competing documents. It makes the classifier robust to feature selection, enabling fully automated training without the injection of human expert knowledge. Support vector machines received also a lot of attention in the machine learning community. They have often achieved better performance than customized neuronal network and state-of-the-art baseline classifiers. We investigate in this paper the classification power of SVMs and discriminative training approaches on natural language call routing. Experiments are reported for a banking call routing and for Switchboard topic identification task. Results show that the application of discriminative training on vector-based model outperforms SVMs by 7% on spoken data.

1. INTRODUCTION

Call centers currently direct calls using human operators or touch-tone based interactive voice response systems. In the latter case, callers are often frustrated because the option list may be long and what they need may not appear to be related to the given choices. We investigate in this paper the application of natural language call routing for call centers, where the caller may say what he/she wants and is automatically routed to the right department or directed to a human operator when the system is unable to determine the caller's intent with certainty. Our goal is to achieve a spoken dialogue system that is an alternative to the tiresome navigation via touch-tone menus while enabling significant automation and cost savings.

Natural language call routing approaches are part of the automatic text categorization domain, where the goal is to assign the topic label to a request. This request may be either a text stream or a spoken message. In the latter case, we can apply a text-based classifier by passing the spoken message through an automatic speech to text system (STT). Recently, we presented new techniques to improve routing accuracy and robustness [1, 2]. Instead of classifiers trained by simple counting using conventional maximum likelihood training, we studied algorithms that improves the performance of individual classifiers as well as combining multiple classifiers to achieve better performance than any individual classifiers. As an example, we proposed the use of discriminative training on the vector-based model to improve classification accuracy and robustness of single classifiers [3, 1]. In this paper, we investigate the effectiveness of (1) support vector machines and (2) discriminative training (DT) on the vector-based model. We investigate here the effectiveness of both classifiers on real application with multi-class task and read data. Our goal is to find the robust classifier to be included in our call center. Experiments are conducted on a banking call routing task with USAA database. Despite the fact that our interest is for natural language call routing, we also conduct experiments on switchboard topic identification task, which belong to the automatic text categorization domain as well. Our goal is to better understand the behavior of both techniques on different databases. Experimental results are presented to demonstrate the power of discriminative training compared to SVM, and baseline vector-based classifier.

2. SUPPORT VECTOR MACHINES

Support vector machines (SVMs) have received a lot of attention in the machine learning community. SVMs have already been used for text categorization, where they showed to achieve substantial improvement over state-of-the-art methods [4]. In this paper, we will compare the performance of SVMs and discriminative training on vector-based model.

SVMs are based on the Structural Risk Minimization principle [5]. As such, it is firmly grounded in the framework of statistical learning theory, or Vapnik-Chervonenkis (VC) theory [6]. The idea of structural risk minimization (on which SVMs are based) is to find a hypothesis h for which we can guarantee the lowest true error ϵ. The true error of h is the probability that h will make an error on a randomly selected unseen test document. An upper bound can be used to connect the true error of a hypothesis h with the error of h on the training set and the complexity of h [6]:

$$\Re(h) \leq \Re_{tr}(h) + 2\sqrt{\frac{d\left(ln\frac{2N}{d}+1\right) - ln\frac{N}{4}}{N}} \qquad (1)$$

where N denotes the number of training example. The term d is the VC-dimension [5], which is a property of the hypothesis space and indicates its expressiveness. The term $\Re(h)$ denotes the true error (risk) of h and the term $\Re_{tr}(h)$ denotes the training

error of the learning machine using h. SVMs find the hypothesis h which minimizes this bound on the true error by effectively and efficiently controlling the VC dimension of hypothesis space.

An interesting property of SVMs is that their ability to learn can be independent of the dimensionality of the feature space. SVMs measure the complexity of hypotheses based on the margin with which they separate the data, not the number of features. This means that we can generalize even in the presence of huge number of features, if our data is separable with a wide margin using functions from the hypothesis space.

3. BASELINE VECTOR-BASED CLASSIFIERS

In the following we describe the two baseline vector-based classifiers we are using. They have previously shown to give very competitive results [7, 8]. In this paper, we investigate the effectiveness of discriminative training on these two baseline vector-based classifiers.

3.1. Classifier using the cosine similarity metric

The classifier using the cosine similarity metric that we denote cosine classifier is a popular vector-based classifier used in information retrieval. This model has also been adopted for natural language call routing [7]. The training process involves constructing a routing matrix R ($m \times n$). A list of ignore words are eliminated and a list of stop words are replaced with placeholders. The rows of R represent the m terms (e.g., words) and the columns the n destinations. The routing matrix R is the transpose of the term-document matrix, where r_{vw} is the frequency with which term w occurs in calls to destination v. Each term is weighted according to term frequency inverse document frequency (TFIDF) and is also normalized to unit length. New user requests are represented as feature vectors and are routed based on the cosine similarity score.

Let $\vec{x}$ be the m-dimensional observation vector representing the weighted terms which have been extracted from the user's utterance. One possible routing decision is to route to the destination with the highest cosine similarity score:

$$\text{destination } \hat{j} = \arg\max_j \cos\phi_j = \arg\max_j \frac{\vec{r_j} \cdot \vec{x}}{\|\vec{r_j}\|\|\vec{x}\|}. \quad (2)$$

3.2. Beta classifier

The beta classifier is a probabilistic method, which has previously been shown to give the best results in a study on e-mail routing [8]. Each topic is represented by a word vocabulary, which includes all words that occur at least a given number of times and excludes those which belong to a list of ignore or stop words. For each word in the vocabulary we compute its probability in the topic and its weight [8]. This weight is assigned according to a function inversely proportional to the number of topic-vocabularies in which this word is present.

A document $q = w_1, w_2, \ldots, w_M$ of M words w_k is routed to the destination j with the highest similarity measure:

$$destination\ \hat{j} = arg\max_j \left(\beta_j^{\delta_1} \times \sum_{k=1}^{M} P(w_k|j)(\eta(w_k))^{\delta_2}\right), \quad (3)$$

where $P(w_k|j)$ is the probability of w_k in topic j, and $\eta(w_k)$ the weight assigned to w_k. Parameters δ_1 and δ_2 are estimated on a development corpus to boost the accuracy. In our experiments, we obtain a value of 0.3 for δ_1, a value of 2 for δ_2 and we take into account words that occur at least three times in the corpus. We also use the same list of stop and ignore words as the cosine classifier. The term β_j is the weight assigned to topic T_j:

$$\beta_j = \frac{\sum_{t=1}^{N_j} \eta(w_t)}{\sum_{k=1}^{J}\sum_{t=1}^{N_k} \eta(w_t)}, \quad (4)$$

where N_k represents the number of words in the k^{th} topic-vocabulary.

4. DISCRIMINATIVE TRAINING TECHNIQUE

According to the way the routing matrix is constructed, there is no guarantee that the classification error rate will be minimized. The routing matrix can be improved by adjusting the entries to achieve minimum classification error (at least locally, and in the probabilistic sense). To solve this nonlinear optimization problem, we adopted the generalized probabilistic descent (GPD) algorithm for natural language call routing [3]. Intuitively, this algorithm looks at each training example and adjusts the model parameters of the correct and competing classes in order to improve the scores of the correct class relative to the other classes.

Specifically, let $\vec{x}$ be the observation vector and $\vec{r}_j$ be the model document vector for destination j. We define the *discriminant function* for class j and observation vector $\vec{x}$ to be the dot product of the model vector and the observation vector:

$$g_j(\vec{x}, R) = \vec{r}_j \cdot \vec{x} = \sum_{i=1}^{m} r_{ji} x_i. \quad (5)$$

Note that this function is identical to the cosine score if the two vectors have been normalized to unit length.

Given that the correct target destination for $\vec{x}$ is k, we define the *misclassification function* as

$$d_k(\vec{x}, R) = -g_k(\vec{x}, R) + G_k(\vec{x}, R), \quad (6)$$

where

$$G_k(\vec{x}, R) = \left[\frac{1}{n-1} \sum_{j \neq k, 1 \leq j \leq n} g_j(\vec{x}, R)^{\eta}\right]^{\frac{1}{\eta}} \quad (7)$$

is the *anti-discriminant function* of the input $\vec{x}$ in class k and $n-1$ is the number of competing classes. Notice also that $d_k(\vec{x}, R) > 0$ implies misclassification, i.e. the discriminant function for the correct class is less than the anti-discriminant function. Equation 6 essentially converts a multi-dimensional decision function into a one-dimensional metric. The GPD algorithm then iteratively optimizes a non-decreasing function of this misclassification metric. For details, please refer to earlier papers [3].

5. EXPERIMENTS

Experiments were performed on two topic identification tasks, a banking call routing task with USAA and a DARPA Switchboard text categorization task. Switchboard is a publicly available database of transcribed telephone conversations of two people talking about assigned topics. We used the database which was initially released, consisting of a total of 2284 transcribed conversations and 67 topics [9]. We divided the database into a training set consisting of about 80% of the database, and the remaining 20% was used as the test set. The vocabulary used for the switchboard task contains the 5000 most frequent words. For the banking call routing task, we

used the same training and test sets as reported in [7], consisting of a total of about 4000 calls, routed to 23 destinations. The vocabulary used for the bank call routing task contains 1232 words; each one of them appears more than three times in the training data and do not belong to the list of stop and ignore words. Only simple features were used, consisting of the most common words.

Experimental results on the banking call routing task are reported on both human transcriptions (Banking-HT) and STT recognized strings (Banking-STT). We used real-time speech recognition system, which have a word error rate of about 20% [1]. Some results are not the same as previously reported in [3] because a different set of unigram features is used in this paper. Note that the results on switchboard task are not directly comparable with published results for many reasons, one of which is that previous experiments used only 10 topics [10]. SVM experiments were conducted using software from Royal Holloway and Bedford New College, University of London [11]. We used the 1-vs-1 approach to multi-class classification [12]. We also limit ourselves to the use of dot product in $\mathbb{R}^n$ as a kernel [5].

5.1. Effects of Discriminative Training on Parameterized Classifier

In the original study on the banking task [7], very good results were already obtained. Although the algorithm was intended to be entirely automatic, some amount of manual tuning to achieve the best performance was inevitable, including choosing the stop word list, the threshold for unigrams, and the heuristic re-weighting schemes such. We showed in [3, 1, 13] that discriminative training is effective even in cases where manual optimization has already been performed. In this section, we want to investigate how much improvement can be achieved using discriminative training on classifiers that contain only simple features, consisting of the most common words. We also investigate here the effectiveness of discriminative training on switchboard database.

Table 1 shows the classification error rate (CER) of the beta classifier as well as the classifier using the cosine similarity metric with and without the use of discriminative training (DT). Results indicate that DT improves classifier accuracy of the baseline classifier using the cosine similarity metric by 35% on the human transcribed banking task. For the errorful strings obtained from speech recognition on the banking task, the relative error rate reduction is about 30%. Similar improvement rate is obtained when discriminative training is employed on beta classifier for the banking task. An important improvement is also obtained using discriminative training on switchboard task: 69% relative improvement on baseline classifier using the cosine similarity metric (19.1% vs. 5.9%) and 67% improvement on beta classifier (18.0% vs. 5.9%). These results confirm how discriminative training is effective even in cases where tasks or data to process (i.e., errorful strings from STT system) are different.

5.2. Discriminative Training and Support Vector Machines

Discriminative training showed to be able to give better classification results than other techniques such us, boosting and automatic relevance feedback [1]. Since the classification accuracy is greatly improved by using DT, we will show in this section a comparison between DT and SVM techniques. As a reminder, DT adjusts the models to increase the separation of the correct class from competitors. SVMs are based on the *structural risk minimization principle* [6] from computational learning theory. The idea of SVMs is to find a hypothesis h for which we can guarantee the lowest probability that h will make an error on a randomly selected unseen event.

	Baseline	after DT	% Change
Classifier Using the Cosine Similarity Metric			
Switchboard	19.1%	5.9%	69%
Banking-HT	9.4%	6.1%	35%
Banking-STT	12.0%	8.4%	30%
Beta Classifier			
Switchboard	18.0%	5.9%	67%
Banking-HT	12.0%	5.5%	54%
Banking-STT	14.9%	7.8%	47%

Table 1. Effects of discriminative training technique on classification error.

	SVMs	Cosine+DT	Beta+DT
Switchboard	6.3%	5.9%	5.9%
Banking-HT	6.5%	6.1%	5.5%
Banking-STT	8.4%	8.4%	7.8%

Table 2. Classification error rate of SVMs, classifier using the cosine similarity metric with DT, and beta classifier with DT.

Table 2 shows the classification error rate (CER) of SVMs, beta classifier with DT, and the classifier using the cosine similarity metric with DT. On the banking call routing task and also on switchboard task, experimental results show that DT applied on vector-based model outperforms SVMs. Compared to SVMs, 6% improvement in CER is obtained when we use baseline classifiers discriminatively trained on switchboard database (6.3% vs. 5.9%), 15% improvement in CER is obtained when we use the beta classifier discriminatively trained on the banking human transcribed database (6.5% vs. 5.5%), and 7% improvement in CER is obtained when we use the beta classifier discriminatively trained on the banking errorful strings from STT system. Because of the small improvement and also because of the relatively little training data, we cannot confirm the hypotheses that DT works better than SVMs. However, we believe that these results again testify to the effectiveness of discriminative training to achieve good performance compared to the state-of-the-art classifiers. We showed in [3, 1] that the use of DT allow the parameterized classifier to achieve minimum classification error and consequently better performance compared to other approaches, including boosting and relevance feedback [1]. In this paper, we again confirm that DT is able to outperform other state-of-the-art classifiers, including SVMs.

5.3. Multiple Classifier Combination

We investigate two different methods to combine the classifiers we have. First, linear interpolation (LI) of the different classifiers is investigated. Then, the constrained minimization technique as introduced in [2] is employed: we consider C_1 and C_2 two classifiers where their errors are uncorrelated. When both classifiers agree, the topic result is the one agreed upon. When they disagree, a third classifier C_3 is invoked as an arbiter. This classifier C_3 disambiguates among only a subset of topics chosen according to a confusion measure [2].

Let classifiers C_1 and C_2 represent SVMs and classifier using cosine metric discriminatively trained on the entire training corpus, respectively. Then, let C_3 represent the beta classifier with and without DT. One motivation of the choice of cosine classifier discriminatively trained and SVMs for C_1 and C_2 respectively is the fact that their errors are not correlated. We denote CM^* the combined classifier using only beta classifier as C_3, and CM^*_{DT} the combined classifier using beta classifier discriminatively trained as C_3. We present in Table 3 the classification error rate of this combination as well as a linear interpolation between these three classifiers (C_1, C_2, C_3).

	CM^*	CM^*_{DT}	LI
Switchboard			
Combined Classifier (SVM + Cosine w' DT)+Beta	5.4%	5.0%	5.9%
Banking Human Transcription			
Combined Classifier (SVM + Cosine w' DT)+Beta	5.5%	5.2%	5.8%
Banking STT Recognized Strings			
Combined Classifier (SVM + Cosine w' DT)+Beta	7.4%	7.1%	7.8%

Table 3. CER of three classifiers using linear interpolation (LI) and constraint minimization (CM^* and CM^*_{DT}).

Results show that the combination of these classifiers allows further improvement in term of classification error rate. The constrained minimization technique (CM^* and CM^*_{DT}) showed to be able to outperform the accuracy of individual classifiers. On the switchboard database, CM^*_{DT} (5.0%), improves the baseline version discriminatively trained (5.9%) by 15%. For the errorful strings obtained from speech recognition on the banking task, CM^*_{DT} (7.1%) outperforms the beta classifier discriminatively trained (7.8%) by 9%. A comparative improvement is obtained using the human transcribed banking data. Results also show how CM^*_{DT} is effective compared to the linear interpolation of the different individual classifiers: CM^*_{DT} outperforms the linear interpolation approach by 15% (5.0% vs. 5.9%) and 10% (5.2% vs. 5.8%) on switchboard and human transcription banking databases, respectively. Other experiments are also done using constrained minimization technique, were we changed the position of the three classifiers $C1$, $C2$ and C_3; results are quite similar. We notice that no improvement is reported when the linear interpolation is used on individual classifiers. It is important to note that DT is again able to boost the classification accuracy: CM^*_{DT} outperforms CM^* by an average of 8%.

6. CONCLUSION

We investigate in this paper the performance of support vector machines and discriminative training for natural language call routing systems. We have demonstrated the advantages of using discriminative training. We have shown that discriminative training can improves any parameterized classifier to achieve minimum classification error. SVMs have the ability to generalize well in high dimensional feature space. They eliminate the need for feature selection, making the application of natural language call routing easier. Experimental results show that SVMs achieve good performance, outperforming existing methods, such us the vector-based classifier using the cosine similarity measure. However, SVMs was not able to outperform the performance achieved by discriminative training when applied to vector-based classifiers. Results on the USAA banking and DARPA switchboard databases show that discriminative training when applied to vector-based classifiers outperforms SVMs by up to 15% in term of classification error rate. Further improvement is achieved when individual classifiers, including SVMs and discriminative training, are combined using the constrained minimization technique.

7. REFERENCES

[1] I. Zitouni, H.-K. J. Kuo, and C.-H. Lee, "Boosting and combination of classifiers for natural language call routing systems," *Speech Communication*, vol. 41, pp. 647–661, December 2003.

[2] I. Zitouni, M. Lee, and H. Jiang, "Constrained minimization technique for topic identification using discriminative training and support vector machines," in *Proceeding of the International Conference on Speech and Language Processing*, 2004.

[3] H.-K. J. Kuo and C.-H. Lee, "Discriminative training of natural language call routers," *IEEE Transaction on Speech and Audio Processing*, vol. 11, no. 1, pp. 24–35, January 2003.

[4] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European Conference on Machine Learning*, Berlin, 1998, pp. 137–142.

[5] Christopher J.C. Burges, "A tutorial on support vector machines for pattern recognition," in *Data Mining and Knowledge Discovery*, 1998, pp. 121–167.

[6] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[7] J. Chu-Carroll and B. Carpenter, "Vector-based natural language call routing," *Computational Linguistics*, vol. 25, no. 3, pp. 361–388, 1999.

[8] B. Bigi, A. Brun, J.P. Haton, K. Smaili, and I. Zitouni, "A comparative study of topic identification on newspaper and e-mail," in *String Processing and Information Retrieval-SPIRE, IEEE Computer Society*, 2001.

[9] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP*, 1992, pp. 517–520.

[10] J.W. McDonough, K. Ng, P. Jeanrenaud, H. Gish, and J.R. Rohlicek, "Approaches to topic identification on the switchboard corpus," in *Proc. ICASSP*, 1994, pp. 385–388.

[11] C. Saunders, M.O. Stitson, and J. Weston, *Support Vector Machine Reference Manual*, Department of Computer Science, Royal Holloway, University of London, 1998.

[12] Tor A. Myrvoll, "An application of support vector machines to phoneme classification," in *NORSIG-99*, Asker, Norge, September 1999.

[13] P. Liu, H. Jiang, and I. Zitouni, "Discriminative training of naive bayes classifiers for natural language call routing," in *Proceeding of the International Conference on Speech and Language Processing*, 2004.