Domain Adaptation for Text Categorization by Feature Labeling

Cristina Kadar and José Iria

IBM Research Zurich,
Säumerstrasse 4, CH-8804 Rüschlikon, Switzerland
{cka,jir}@zurich.ibm.com

Abstract. We present a novel approach to domain adaptation for text categorization, which merely requires that the source domain data are weakly annotated in the form of labeled features. The main advantage of our approach resides in the fact that labeling words is less expensive than labeling documents. We propose two methods, the first of which seeks to minimize the divergence between the distributions of the source domain, which contains labeled features, and the target domain, which contains only unlabeled data. The second method augments the labeled features set in an unsupervised way, via the discovery of a shared latent concept space between source and target. We empirically show that our approach outperforms standard supervised and semi-supervised methods, and obtains results competitive to those reported by state-of-the-art domain adaptation methods, while requiring considerably less supervision.

Keywords: Domain Adaptation, Generalized Expectation Criteria, Weakly-Supervised Latent Dirichlet Allocation

1 Introduction

The task of domain adaptation is fundamental to real-world text categorization problems, because the simplifying assumption, often made, that documents in the training set are drawn from the same underlying distribution as documents in the test set rarely holds in practice. As a consequence, statistical models derived from training data drawn from the "source" domain typically do not perform well on test data drawn from the "target" domain. For example, [18] report that a text classification model trained on a Yahoo! directory performed poorly on a Weblog classification problem, since the distribution of terms differed significantly.

At the heart of the difficulty in applying machine learning to new domains lies the fact that labeling problem examples is expensive. In particular, annotations of documents in the target domain are usually unavailable and expensive to acquire. Recently, a new labeling paradigm was introduced that enables learning from labeled features instead of labeled instances [6]. This provides two advantages: it reduces the amount of time spent in annotating, and therefore the cost,

and it allows experts to more naturally, and thus more accurately, express their knowledge about the domain.

The feature labeling paradigm is particularly appealing for the domain adaptation task because it is often possible for domain experts to tell which features from the source domain are expected to apply robustly also in the target domain. This is easier and less time consuming than labeling documents. Unfortunately, approaches to domain adaptation have not considered the use of the feature labeling paradigm so far.

In this paper, we present a novel approach to domain adaptation for text categorization, which merely requires that the source data are weakly annotated in the form of labeled features. We propose two domain adaptation methods under this approach. The first method seeks to minimize the divergence between the distributions of the source domain, which contains labeled features, and the target domain, which contains only unlabeled data. The second method is similar to the first one, but can additionally make use of the labeled features to guide the discovery of a latent concept space, which is then used to augment the original labeled features set.

The contributions of the paper are fourfold: (i) we present, to the best of our knowledge, the first approach to domain adaptation for text categorization that relies on labeled words instead of labeled documents; (ii) we propose two different methods in order to analyse the merits of the approach (iii) we study the effect of the number of labeled features on the experimental results and verify that competitive results can be achieved even with a low number of labeled features; (iv) and we empirically show that our approach, despite only using a weak form of supervision, outperforms standard supervised and semi-supervised methods, and obtains results competitive with those previously reported by state-of-the-art methods that require the classic, more expensive, form of supervision – that of labeling documents.

The remainder of the paper is structured as follows. A brief review of related work on domain adaption is given in the next section. In Section 3 we introduce the proposed domain adaptation methods. A complete description of the experimental setting is given in Section 4, and in Section 5 a comparative evaluation of the methods is presented, followed by a discussion on the results obtained. We conclude with a mention to our plans for future work.

2 Related Work

There are roughly two variants of the domain adaptation problem, which have been addressed in the literature: the supervised case and the semi-supervised case. In the former, we have at our disposal labeled documents from the source domain, and also a small amount of labeled documents from the target domain. The goal is to take advantage of both labeled datasets to obtain a model that performs well on the target domain. For example, [5, 7] work under this setting. The semi-supervised case differs in that no labeled documents in target exist,

therefore the goal is to take advantage of an unannotated target corpus, see, e.g., [3, 9, 19, 4]. In this paper, we address the semi-supervised problem.

The problem of domain adaptation can be seen as that of finding a shared latent concept space that captures the relation between the two domains [16]. Therefore, several recent approaches sought an appropriate feature representation that is able to encode such shared concept space. [5] uses standard machine learning methods to train classifiers over data projected from both source and target domains into a high-dimensional feature space, via a simple heuristic nonlinear mapping function. In [14], the authors approach the problem from dimensionality reduction viewpoint. The method finds a low-dimensional latent feature space where the distributions between the source domain data and the target domain data are as close to each other as possible, and project onto this latent feature space the data from both domains. Standard learning algorithms can then be applied over the new space. A probabilistic approach in the same vein can be found in [19], where the authors propose an extension to the traditional probabilistic latent semantic analysis (PLSA) algorithm. The proposed algorithm is able to integrate the labeled source data and the unlabeled target data under a joint probabilistic model which aims at exploiting the common latent topics between two domains, and thus transfer knowledge across them through a topic-bridge to aid text classification in the target domain. Other relevant approaches following the same underlying principle include the feature extraction method described in [15], the method based on latent semantic association presented in [8] and the linear transformation method in [4] that takes into account the empirical loss on the source domain and the embedded distribution gap between the source and target domains.

Our approach may also be considered to belong to the above family of approaches in that we model a shared latent space between the domains, but with two major differences. First, it requires only labeled features instead of labeled instances. Second, the modeling of the latent space is not unsupervised, but partially supervised instead – by taking advantage of the availability of labeled features.

3 Domain Adaptation using Labeled Features

Rather than requiring documents in the source and target domains to be examined and labeled, our approach to the domain adaptation problem leverages a small set of words that domain experts indicate to be positively correlated with each class – the labeled features. We adopt the *generalized expectation criteria* method [13, 6] to translate this kind of domain knowledge into constraints on model expectations for certain word-class combinations. In what follows we briefly introduce this method, using the notation in [13], and then show how it can be used for domain adaptation.

A generalized expectation (GE) criterion is a term in a parameter estimation objective function that assigns scores to values of a model expectation. Let x be the input, y the output, and θ the parameters for a given model. Given a

set of unlabeled data $\mathcal{U} = \{x\}$ and a conditional model $p(y|x;\theta)$, a GE criterion $G(\theta;\mathcal{U})$ is defined by a score function V and a constraint function $G(x,y)$:

$$G(\theta;\mathcal{U}) = V(E_{\mathcal{U}}[E_{p(y|x;\theta)}[G(x,y)]]).$$

The GE formulation is generic enough to enable exploring many different choices of score functions and constraint functions. In this paper, we maximize the GE term together with an entropy regularization term in the objective function, although this can be easily combined with an empirical loss term to form a composite objective function that takes into account labeled instances as well. Moreover, we use *label regularization*, that is, the constraints are expectations of model marginal distributions on the expected output labels. As such, we use estimated label marginal distributions $\tilde{g}_{x,y} = \tilde{p}(y)$ and consider constraints of the form $G(x,y) = \mathbf{1}(y)$. Model divergence from these constraints can be computed by using, for example, KL-divergence [11]:

$$G(\theta;\mathcal{U}) = -D(\tilde{p}(y)||E_{\mathcal{U}}[\mathbf{1}(y)p(y|x;\theta)]).$$

In order to use GE for domain adaptation, we derive criteria that encourage agreement between the source and target expectations. Let $\mathcal{S}$ be source domain data and $\mathcal{T}$ be target domain data, both unlabeled. We compute the model divergence for the task of domain adaptation by:

$$G(\theta;\mathcal{S},\mathcal{T}) = -\sum_{i\in F(\mathcal{S}\cup\mathcal{T})} D\left(\hat{p}(y|x_i>0)||\tilde{p}_\theta(y|x_i>0)\right), \tag{1}$$

where F is a function that returns the set of features in the input data, $p(y|x_i>0) = \frac{1}{C_i}\mathbf{1}(y)\mathbf{1}(x_i>0)$ is an indicator of the presence of feature i in x times an indicator vector with 1 at the index corresponding to label y and zero elsewhere, and $C_i = \sum_x \mathbf{1}(x_i>0)$ is a normalizing constant; $\tilde{p}_\theta$ denotes the predicted label distribution on the set of instances that contain feature i and $\hat{p}$ are reference distributions derived from the labeled features. We estimate these reference distributions using the method proposed by [17]: let there be n classes associated with a given feature out of L total classes; then each associated class will have probability q_{maj}/n and each non-associated class has probability $(1-q_{maj})/(L-n)$, where q_{maj} is set by the domain experts to indicate the correlation between the feature and the class.

To encourage the model to have non-zero values on parameters for unlabeled features that co-occur often with a labeled feature, we select as regularizer the Gaussian prior on parameters, which prefers parameter settings with many small values over settings with a few large values. The combined objective function is finally:

$$\mathcal{O} = -\sum_{i\in F(\mathcal{S}\cup\mathcal{T})} D\left(\hat{p}(y|x_i>0)||\tilde{p}_\theta(y|x_i>0)\right) - \sum_j \frac{\theta_j^2}{2\sigma^2}, \tag{2}$$

consisting of a GE term for each for each labeled feature i, and a zero-mean σ^2-variance Gaussian prior on parameters.

We designed two methods that follow the proposed feature labeling approach to text categorization and the GE formulation above. As per equation (1), both methods are multi-class and semi-supervised (in that they make use of the unlabeled target domain data). The first method, which we will designate as *TransferLF*, directly uses the input labeled features to derive the reference distributions $\hat{p}$ (in the way described earlier). Then, given the latter and unlabeled source and target domain datasets, it estimates the classification model parameters by using an optimization algorithm, taking equation (2) as the objective function.

The second method, which we will designate as *TransferzLDALF*, is similar to the first one, but additionally aims at augmenting the set of input labeled features with new labeled features derived from the target domain data. In the same vein as related work in section 2, to discover and label new features our idea is to find a shared latent concept space that captures the relation between the two domains and bridges source and target features. This can be achieved in an unsupervised manner by using latent topic models such as Latent Dirichlet Allocation (LDA) [2]; however, we are interested in encouraging the recovery of topics that are more relevant to the domain expert's modeling goals, as expressed by the labeled features provided, than the topics which would otherwise be recovered in an unsupervised way. Weak supervision in LDA was recently introduced in works such as [1, 20]. With this goal in mind, we rehash the approach in [1], which adds supervision to LDA in the form of so-called *z-labels*, i.e., knowledge that the topic assignment for a given word position is within a subset of topics. Thus, in addition to their role in GE, we use the input labeled features as *z-labels*, in order to obtain feature clusters (containing both source and target features) where each cluster respects to one topic from the set of topics found in the labeled features. We are then able to augment the original labeled features set with the k most probable target domain features present in each cluster, in hope that the additional GE constraints lead to improved performance.

The algorithm for inducing a text categorization classifier for both methods is shown below. The first two steps only apply to *TransferzLDALF*.

Algorithm 1 TransferLF and TransferzLDALF

Input: labeled features $\mathcal{L}$, unlabeled source $\mathcal{S}$ and target $\mathcal{T}$ domain data
Output: induced classifier $\mathcal{C}$

TransferzLDALF only:
(1) $\mathcal{L}_{\mathcal{LDA}}$ = labeled features from weakly-supervised LDA using input $\mathcal{L}$, $\mathcal{S}$ and $\mathcal{T}$
(2) Augment $\mathcal{L}$ with k target domain features per topic from $\mathcal{L}_{\mathcal{LDA}}$

TransferLF and *TransferzLDALF*:
(3) Compute reference distributions $\hat{p}(y|x_i > 0)$ from $\mathcal{L}$
(4) Estimate model parameters by running optimization algorithm according to eq. (2)
(5) **return** induced classifier $\mathcal{C}$

Dataset	Source Data	Target Data	KL divergence
Cars vs Games	rec.autos rec.sport.baseball	rec.motorcycles rec.sport.hockey	0.5679
Cars vs. Hardware	rec.autos comp.sys.ibm.pc.hardware	rec.motorcycles comp.sys.mac.hardware	0.4136
Cars vs Games vs Hardware vs OS	rec.autos rec.sport.baseball comp.sys.ibm.pc.hardware comp.windows.x	rec.motorcycles rec.sport.hockey comp.sys.mac.hardware comp.os.ms-windows.misc	0.4579
Cars vs Games vs Hardware vs OS vs Politics vs Religion	rec.autos rec.sport.baseball comp.sys.ibm.pc.hardware comp.windows.x talk.politics.mideast soc.religion.christian	rec.motorcycles rec.sport.hockey comp.sys.mac.hardware comp.os.ms-windows.misc talk.politics.misc talk.religion.misc	0.3701
Comp vs Sci	comp.graphics comp.os.ms-windows.misc sci.crypt sci.electronics	comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x sci.med sci.space	0.3897
Rec vs Talk	rec.autos rec.motorcycles talk.politics.guns talk.politics.misc	rec.sport.baseball rec.sport.hockey talk.politics.mideast talk.religion.misc	0.5101
Comp vs Rec	comp.graphics comp.sys.ibm.pc.hardware comp.sys.mac.hardware rec.motorcycles rec.sport.hockey	comp.os.ms-windows.misc comp.windows.x rec.autos rec.sport.baseball	0.4741
Comp vs Talk	comp.graphics comp.sys.mac.hardware comp.windows.x talk.politics.mideast talk.religion.misc	comp.sys.ibm.pc.hardware comp.sys.mac.hardware talk.politics.guns talk.politics.misc	0.2848
Auto vs Aviation	rec.autos.simulators rec.aviation.simulators	rec.autos.misc rec.aviation.student	0.8152
Real vs Simulated	rec.autos.misc rec.autos.simulators	rec.aviation.student rec.aviation.simulators	0.6532

Table 1. Characteristics of the datasets used for evaluating the proposed approach.

4 Experiments

The first of the datasets chosen for our empirical analysis is K. Lang's original 20-newsgroups[1] dataset [12]. It contains approximately 20,000 documents

[1] http://www.cs.umass.edu/~mccallum/code-data.html

that correspond to English-language posts to 20 different newsgroups. There are roughly 1000 documents in each category. The topic hierarchy for this dataset contains four major groups: *sci* (scientific), *rec* (recreative), *talk* (discussion) and *comp* (computers), with 3 to 5 topics under each group. The second dataset used in our experiments is the SRAA[1] corpus. It contains messages about simulated auto racing, simulated aviation, real autos and real aviation from 4 discussion groups. We used the first 4,000 documents from each of the classes in this dataset.

For the purposes of evaluating domain adaptation, we gather documents drawn from related topics, having different distributions. For example, the newsgroups *rec.autos* and *rec.motorcycles* are both related to cars, whereas the newsgroups *rec.sport.baseball* and *rec.sport.hockey* both describe games. Plus, moving to the first level of the 20-newsgroups taxonomy, broader categories may also be built: recreative, talk, computers and scientific. The SRAA data set is split in a similar manner into four categories: auto, aviation, real, simulated. Table 1 summarizes the characteristics of the datasets used in the experiments, indicating the source vs. target splits, the initial number of labeled features, and the KL-divergence [11] measuring the distribution gap between the domains[2].

Minimal preprocessing was applied on the data: lowercasing the input and removing a list of English stopwords. Each document is represented as a vector of words and their frequency in the corpus. We use the MALLET[3] toolkit to solve the optimization problem using L-BFGS, a quasi-Newton optimization method that estimates the model parameters.

Class	initial seed words	top 18 words in topic
Cars	article writes car cars wheel miles toyota honda driving engine oil engines ford rear year auto autos	writes article car good **bike** time back people cars make year thing engine **ride** years **road** work front
Hardware	advance windows disk system drives computer dx software bus mode os ibm memory machine monitor dos hardware board chip card cards ram mb pc interface vlb mhz cache ide cpu controller port modem motherboard gateway scsi video isa bios floppy	system drive problem computer work **mac** card mail **apple** software mb good time pc problems disk board bit

Table 2. Initial labeled features and discovered zLDA features for *Cars vs Hardware*.

Human domain expertise is replaced in our experiments by an oracle-labeler – an experimental setup also adopted in, e.g., [6]. Making use of the true instance labels, the oracle computes the mutual information of the features within each class, and, if above a given threshold, labels the feature with the class under

[2] It may be noted that the obtained KL-divergence values are considerably larger than if we were to split randomly, which would yield values close to zero.

[3] http://www.mallet.cs.umass.edu

which it occurs most often, and also with any other class under which it occurs at least half as often. In the experiments we use as threshold the mean of the mutual information scores of the top $100L$ most predictive features, where L is the number of classes; and $q_{maj} = 0.9$ as the majority of the probability mass to be distributed among classes associated to a labeled feature. The oracle is very conservative in practice – refer to Table 3 for the actual number of labeled features for each source domain.

Dataset	*# source labeled instances*	MaxEnt	*# source labeled features*	TransferLF on source	TransferLF	*# zLDA labeled features*	TransferLF with zLDA features
Cars vs Games	*2000*	90.3	*52*	84.7	**96.1**	*29*	92.8
Cars vs. Hardware	*2000*	90.7	*57*	88.2	**94.2**	*32*	88.7
Cars vs Games vs Hardware vs OS	*4000*	76.0	*109*	72.3	**80.9**	*60*	78.8
Cars vs Games vs Hardware vs OS vs Politics vs Religion	*6000*	67.1	*167*	63.0	69	*81*	**70.2**
Comp vs Sci	*4000*	71.8	*59*	76.1	78.4	*30*	**82.2**
Rec vs Talk	*3874*	77.9	*60*	74.3	74.5	*29*	**92.8**
Comp vs Rec	*5000*	87.9	*70*	86.1	**91.3**	*32*	86.7
Comp vs Talk	*5000*	93.3	*67*	91	**94.1**	*33*	94.0
Auto vs Aviation	*8000*	77.2	*48*	78.0	86.9	*29*	**91.6**
Real vs Simulated	*8000*	63.9	*54*	60.4	59.7	*30*	**77.7**

Table 3. Classification accuracies and the amount of labeled information (either instances or features) used in different sets of experiments. Note that for the *TransferzLDALF* method, the reported results correspond to selecting a fixed number of 18 features per topic (cf. learning curves), but the features outputted by zLDA can overlap and thus the size of the feature set used is smaller when merged.

Finally, zLDA[4] was chosen as an implementation of the semi-supervised LDA method. We use the original labeled features as seeds for their latent topics and run the algorithm in its standard setup, as reported in [1]: $\alpha = .5$, $\beta = .1$, 2000 samples. Table 2 shows an example concerning the *Cars vs Hardware* experiment. The oracle identified and labeled 17 and 40 features, respectively. They all come from the source domains: *rec.autos* and *comp.sys.pc.ibm.hardware*, respectively. With these as input, zLDA identifies new associated features that are specific to the target (e.g. *bike* for *rec.motorcycles* and *apple* for *comp.sys.mac.hardware*).

[4] http://pages.cs.wisc.edu/~andrzeje/software.html

5 Results and Discussion

The results are presented using *accuracy* as evaluation metric: $Acc = (tp+tn)/d$, where tp are the true positives, tn the true negatives, and d the total number of documents in the corpus. In all comparisons, care was taken to reproduce the original authors' experimental setting with rigour.

Table 3 presents the results obtained from running the experiments on the several configurations shown in Table 1. We present results concerning two classifiers which are induced from the source domain data only: a standard supervised maximum entropy classifier as a baseline, and our proposed *TransferLF* method prevented from looking at the target domain data. The results show that our feature labeling approach to domain adaptation invariably outperforms the baseline non-domain-adaptation maximum entropy approach, while, in addition, greatly reduces the supervision requirements – compare the number of labeled features against the number of labeled instances used to induce the classifiers. It should be remarked that this is observed not only in the binary classification case, but also in the multi-class classification case. The results also suggest that the semi-supervised nature of the proposed methods is a differentiating factor, since *TransferLF* using source domain data only consistently underperforms.

Dataset	TSVM	MMD	TransferLF	TransferLF with zLDA features
Cars vs Games	87.4	94.5	96.1	92.8
Cars vs Hardware	92.5	94.6	94.2	88.7
Cars vs Games vs Hardware vs OS	75.4	82.4	80.9	78.8

Table 4. Performance comparison with [4].

Dataset	TSVM	TPLSA	TransferLF	TransferLF with zLDA features
Comp vs Sci	81.7	98.9	78.4	82.2
Rec vs Talk	96	97.7	74.5	92.8
Comp vs Rec	90.2	95.1	91.3	86.7
Comp vs Talk	90.3	97.7	94.1	94.0
Auto vs Aviation	89.8	94.7	86.9	91.6
Real vs Simulated	87	88.9	59.7	77.7

Table 5. Performance comparison with [19].

Tables 4 and 5 compare our approach with semi-supervised and latent semantic analysis-based techniques for domain adaptation in the literature. Transductive Support Vector Machines (TSVM) [10] are used as our baseline semi-supervised text classification approach. Refer to Section 2 for a brief description of MMD[4] and TPLSA[19]. It can be observed that the performance of the proposed methods is comparable with that of TSVM, which, again, is remarkable given that only a few labeled features are required to achieve that. The state-of-the-art MMD and TPLSA approaches still obtain higher accuracy in general, which is not surprising given that their supervision requirements are much greater, but it is still very interesting to see how the results obtained by the feature labeling approach remain competitive. This is important, since in many application domains the reduction of the annotation effort is an enabling factor, at the expense of a only few accuracy points.

Finally, Figures 1, 2 and 3 show the learning curves obtained by varying the number of labeled features input to the *TransferzLDALF* method. From these curves we are able to obtain a deeper insight into the supervision requirements of our proposed approach. We conclude that as little as 5 features per topic are enough to achieve performances close to the plateau of the curve, as seen in some of the experiments, and that, on average, around 18 features per topic are enough to achieve top accuracy for the majority of the experiments.

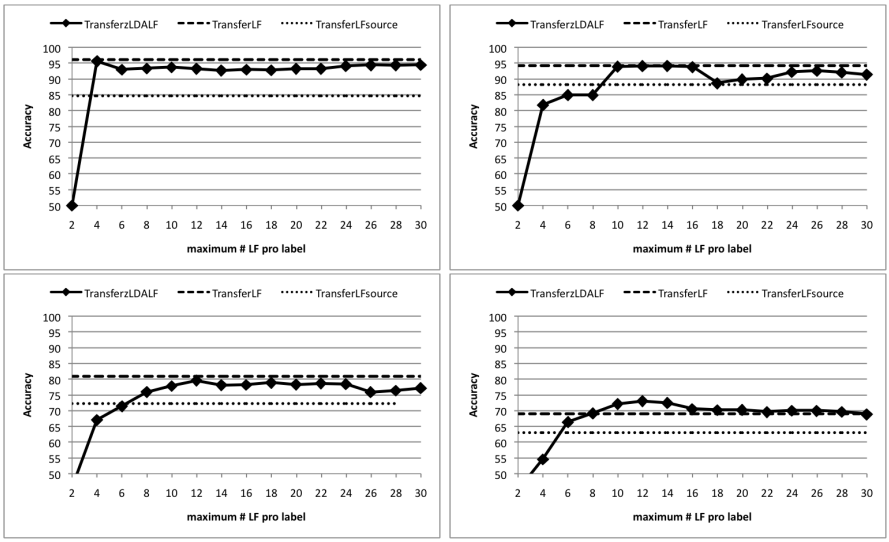

Fig. 1. Learning curves for the first dataset generated from 20-newsgroups. From left to right descending: *Cars vs Games*, *Cars vs Hardware*, *Cars vs Games vs Hardware vs OS*, and *Cars vs Games vs Hardware vs OS vs Politics vs Religion*.

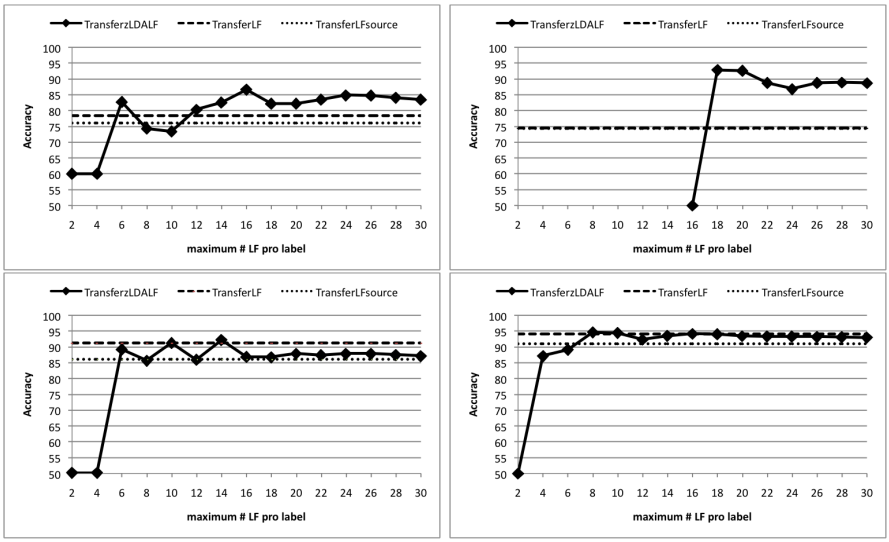

Fig. 2. Learning curves for the second dataset generated from 20-newsgroups. From left to right descending: *Comp vs Sci*, *Rec vs Talk*, *Comp vs Rec*, and *Comp vs Talk*.

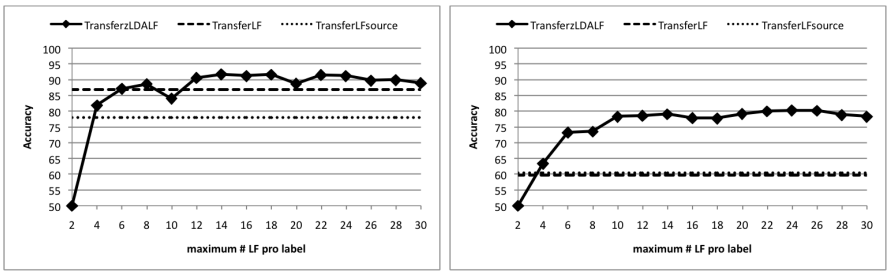

Fig. 3. Learning curves for the dataset generated from SRAA. From left to right: *Auto vs Aviation* and *Real vs Simulated*.

6 Conclusions and Future Work

In this paper, we presented a novel approach to domain adaptation for text categorization that aims at reducing the effort in porting existing statistical models induced from corpora in one domain to other related domains. Our approach is based on the new paradigm of labeling words (as opposed to labeling whole documents), which is less time consuming and more natural for domain experts. It is our expectation that the proposed approach will introduce quantifiable benefits in several information retrieval application domains.

There are several possible avenues for future work that we would like to explore. First, we will study the interplay between labeled features and labeled documents through a thorough set of experiments which will allow us to analyse

the behaviour of the induced model under varying amounts of labeled features and labeled documents in both source and target. Second, we plan to design a bootstrapping algorithm that makes use of labeled features to iteratively refine models of both source and target. Finally, we are currently developing a prototype system that implements our approach in the context of a real-world problem of classifying the textual part of tickets reporting on IT system problems.

References

1. Andrzejewski, D. and Zhu, X. Latent Dirichlet Allocation with Topic-in-Set Knowledge. In *NAACL-SSLNLP*, 2009.
2. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. In *Journal of Machine Learning*, 3:993-1022, 2003.
3. Blitzer J.,McDonald R., and Pereira F.. Domain adaptation with structural correspondence learning. In *EMNLP*, 2006.
4. Chen B., Lam W., Tsang I., and Wong T.L. Extracting discriminative concepts for domain adaptation in text mining. In *KDD*, 2009.
5. Hal Daume III. Frustratingly easy domain adaptation. In *ACL*, 2007.
6. Druck, G., Mann, G., and McCallum, A. Learning from labeled features using generalized expectation criteria. In *SIGIR*, 2008.
7. Jenny Rose Finkel and Christopher D. Manning. Hierarchical bayesian domain adaptation. In *NAACL*, 2009.
8. Guo, H., Zhu, H., Guo, Z., Zhang, X., Wu, X., and Su, Zr. Domain adaptation with latent semantic association for named entity recognition. In *NAACL*, 2009.
9. Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *ACL*, 2007.
10. Joachims, T. Transductive Inference for Text Classification using Support Vector Machines. In *ICML*, 1999.
11. Kullback, S. and Leibler, R. A. On Information and Sufficiency. In *Annals of Mathematical Statistics*, 22(1):79-86, 1951.
12. Lang, K. NewsWeeder: Learning to Filter Netnews. In *ICML*, 1995.
13. Mann, G. S. and McCallum, A. 2010. Generalized Expectation Criteria for Semi-Supervised Learning with Weakly Labeled Data. In *Journal of Machine Learning*, 11:955-984, 2010.
14. Pan, S. J., Kwok, J. T., and Yang, Q. Transfer learning via dimensionality reduction. In *AAAI*, 2008.
15. Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. Domain adaptation via transfer component analysis. In *IJCAI*, 2009.
16. Ben-David S., Blitzer J., Crammer K., and Pereira F. Analysis of representations for domain adaptation. In *NIPS*, 2006.
17. Schapire R., Rochery M., Rahim M., and Gupta N. Incorporating prior knowledge into boosting. In *ICML*, 2002.
18. Ni, X., Xue, G.-R., Ling, X., Yu, Y., Yang, Q. Exploring in the Weblog Space by Detecting Informative and Affective Articles. In *WWW*, 2007.
19. Xue, G., Dai, W., Yang, Q., and Yu, Y. Topic-bridged PLSA for cross-domain text classification. In *SIGIR*, 2008.
20. Gu Xu G., Yang S.-H., Li H. Named Entity Mining from Click-Through Data Using Weakly Supervised Latent Dirichlet Allocation. In *KDD*, 2009.