Transductive Distributional Correspondence Indexing for Cross-Domain Topic Classification

Alejandro Moreo Fernández[1], Andrea Esuli[1], and Fabrizio Sebastiani[2]

[1] Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche, 56124 Pisa, IT
alejandro.moreo@isti.cnr.it
andrea.esuli@isti.cnr.it
[2] Qatar Computing Research Institute
Qatar Foundation, PO Box 5825, Doha, QA ⋆
fsebastiani@qf.org.qa

Abstract. Obtaining high-quality annotated data for training a classifier for a new domain is often costly. Domain Adaptation (DA) aims at leveraging the annotated data available from a different but related *source* domain in order to deploy a classification model for the *target* domain of interest, thus alleviating the aforementioned costs. To that aim, the learning model is typically given access to a set of unlabelled documents collected from the target domain. These documents might consist of a representative sample of the target distribution, and they could thus be used to infer a general classification model for the domain (*inductive inference*). Alternatively, these documents could be the entire set of documents to be classified; this happens when there is *only one* set of documents we are interested in classifying (*transductive inference*). Many of the DA methods proposed so far have focused on transductive classification *by topic*, i.e., the task of assigning class labels to a specific set of documents based on the topics they are about. In this work, we report on new experiments we have conducted in transductive classification by topic using Distributional Correspondence Indexing method, a DA method we have recently developed that delivered state-of-the-art results in *inductive* classification *by sentiment*. The results we have obtained on three popular datasets show DCI to be competitive with the state of the art also in this scenario, and to be superior to all compared methods in many cases.

Keywords: Transduction, Cross-Domain Adaptation, Topic Classification, Distributional Hypothesis

1 Introduction

As a supervised task, automatic Text Classification (TC) is constrained to the availability of high-quality corpora of annotated documents to train a classifier

⋆ Fabrizio Sebastiani is on leave from Consiglio Nazionale delle Ricerche, Italy.

that will then predict the classes of new documents about a given domain of knowledge. In the absence of any such labelled collection for the domain of interest, an additional cost, economical and of time, is to be undertaken in order to collect and annotate the training examples.

Domain Adaptation (DA) is an special case of Transfer Learning (TL)[13,14] to TC, aimed at reducing, or completely avoiding, such costs, by leveraging on any different, but related, source of knowledge for which a training corpus exists already. DA thus challenges one core assumption of machine learning, usually referred to as the *iid assumption*, according to which the training and test examples are believed to be drawn from the same distribution. Traditionally, two different scenarios are considered in DA: (i) *cross-domain* adaptation [2], where the source and target domains differ in the topics they are about; and (ii) *cross-lingual* adaptation [15], where the source and target domains are expressed in different languages, although dealing with the same topics. This article focuses on cross-domain adaptation.

The transference of knowledge is typically attempted by uncovering regularities in examples that are shared across domains. To that aim, a representative (unlabeled) sample from the target distribution is collected and passed to the inference method when learning the decision function. Many of the proposed approaches to cross-domain adaptation so far though, have considered this target sampling to be, at the same time, the test set, i.e., the (only) set of documents one might be interested in classifying (see e.g., [4,10,17,18,9,1]). This fact leads us distinguish between *inductive* and *transductive* cross-domain approaches, depending on the type of inference they carry out[3]. Accordingly, inductive cross-domain approaches might be viewed as those aiming at deploying a classification model that generalizes adequately on the target domain, whereas transductive cross-domain approaches are only requested to deliver an accurate classification of the target set at one's disposal [16].

A general trend one could observe from the related literature of cross-domain adaptation is that the vast majority of the inductive approaches proposed so far have been dedicated to sentiment classification (namely, assessing positive or negative labels to opinion-laden texts), while most of the transductive approaches have instead been tested in topic classification[4]. Be that as it may, two well-differentiated folds of techniques for cross-domain adaptation exist, for which it remains unclear how much effort it entails for porting one of these methods (say, an inductive one) to the configuration of the other group (say, to the transductive setting), or to more general TL configurations (e.g., when the source and target tasks are different).

[3] This distinction is surprisingly overlooked in the related literature though. This is probably so due to the terminology Pan & Yang used in their popular survey [13], where they categorized as *transductive* all TL approaches in which the source and target task are the same, but the source and target domains are different, while term *inductive* was instead attributed the opposite meaning, i.e., when the domains are the same, but the source and target tasks differ.

[4] This seemingly deliberate partition might rather respond to the characteristics of the most popular benchmark collections available for each problem.

This paper is an extension of our former work in [5,11], where the Distributional Correspondence Indexing (DCI) method for cross-domain and cross-lingual adaptation was proposed. DCI creates words embeddings based in the *distributional hypothesis* (words with similar meanings tend to co-occur in similar contexts [6]), and delivered new state-of-the-art results for inductive classification by sentiment recently. We now put to test DCI in a different problem setting, i.e., the transductive approach, and report new experiments on a different task, i.e., cross-domain classification by topic. Results confirm that our Transductive DCI (hereafter TDCI for short), behaves robustly also in this scenario, delivering comparable classification accuracies, and even better in many cases, to state-of-the-art methods, while still being computationally cheap.

The rest of this paper is organized as follows. Section 2 offers a brief overview of related work. In Section 3 we describe our proposal. Section 4 reports the results of the experiments we have conducted, while Section 5 concludes.

2 Related work

In this section, we briefly review main related methods in the literature of domain adaptation. We will restrict our attention to transductive approaches proposed for topic classification. The interested reader can check [11] for a discussion focusing on inductive methods for sentiment classification, and [13,14] for a more general overview on transfer learning methods.

Transductive Support Vector Machines (TSVMs) for text classification where proposed in [8] as an extension of Suport Vector Machines (SVMs) aiming at minimizing the misclassification error in a concrete test set, assuming it accessible when inducing the decision function. Even though it was not particularly designed to deal with DA problems, it has often been reported as a baseline in the related literature. The Co-Clustering approach [4] uses clusters of words and documents as a bridge to propagate the class structure from the source domain to the target domain. The key idea, is to use the class labels in the source domain as a constraint for the word clusters, that are shared among both domains. The Matrix Tri-factorization [18] approach follows a somewhat similar assumption, based on the belief that associations between word clusters and classes should remain consistent between the source and target domain. The method thus performs two matrix tri-factorizations, for the source and target domains, in a joint optimization framework subject to sharing the association between word clusters and classes. Topic-bridged Probabilistic Latent Semantic Analysis [17] is an extension of Probabilistic Latent Semantic Analysis (PLSA) that models the relations between (observed) documents and terms thorough a set of (hidden) latent features, hypothesizing those latent features to be consistent across domains. Along these lines, Topic Correlation Analysis [9] establishes a distinction between latent features that could be shared between domains, and those that are rather domain specific. A joint mixture model is first used to cluster word features into shared and domain-specific topics. Then, a mapping between the domain-specific topics from both domains is induced from a correla-

tions analysis, that serves to derive a shared feature space where the transference of supervised knowledge is facilitated. Finally, the Cross Domain Spectral Classification [10] approach formulates the knowledge transference thorough spectral classification, via optimizing an objective function aimed at regularizing the supervised information contained in the source domain in order to bring to bear improved consistence with respect to the target domain structure. In [1] a probabilistic method based on Latent Dirichlet Allocation (LDA) is proposed. The method jointly optimizes the marginal and conditional distributions following a EM algorithm, while also differentiating between the domain-dependent and domain-independent latent features.

3 Transductive Distributional Correspondence Indexing

Loosely speaking, the main challenge one has to face in domain adaptation is to deal with the discrepancy of words relevance that comes about by its particular role in the source domain, and that is not generalizable to the target domain. That is to say, most important words for the source domain, on which the decision surface is likely to hinge upon, are likely not helpful enough in discriminating the positive and negative regions in the target domain.

DCI builds upon (i) the concept of *pivots* terms [3], namely, frequent and discriminant words which behave expectedly in a similar way in the source and target domains; and (ii) the *distributional hypothesis* [6], which states that terms with similar meanings tend to co-occur in similar contexts. Our idea is to model each term as a word embedding where each dimension quantifies its relative *semantic similarity* to a fixed set of pivots. The expectation is that words with equivalent role across domains might end up lying close to each other in the new embedding space, as they are expected to present similar distributions to the pivots in their respective knowledge domains. Take as an example a classifier by genre (*sci-fi*, *drama*, *horror*, *romantic*, ...), that is trained with documents from a source domain of *films*, but intended to classify documents from a target domain of *books*. Note that role equivalences between, e.g., 'director'-'writer', 'duration'-'length', or 'film'-'book' might be uncovered by inspecting their co-occurrence distribution to some pivots like 'plot', 'character', or 'story', which are expected to be approximately invariant across domains. As a result, the surface decision boundary found for the source domain will likely generalize well in the target domain. DCI is an instantiation of this model that implements a pivot selection strategy (section 3.2), and quantifies the similarity of meaning of two words thorough a Distributional Correspondence Function (DCF, section 3.3).

3.1 Preliminaries

Given a source ($\mathcal{S}$) and a target ($\mathcal{T}$) domain of documents, with different marginal distributions, for which a training set of annotated documents $Tr_{\mathcal{S}}$ exists exclusively for $\mathcal{S}$, cross-domain classification by topic might be formalized as the task of assigning class labels $C = \{c_1, \ldots, c_{|C|}\}$ to target documents in a test set $Te_{\mathcal{T}}$

by means of a classifier Φ trained on $Tr_{\mathcal{S}}$ which is also given access to a sample of (non annotated) documents $U_{\mathcal{T}}$ from $\mathcal{T}$ (and, optionally, to a sample $U_{\mathcal{S}}$ from $\mathcal{S}$), where classes in C represent predefined topics of discussion, such as e.g., "politics", "economics", or "computers".

We will here restrict our attention to the binary case $C = \{c, \overline{c}\}$, that is, deciding whether a document discusses a given topic c, or not. We will also adhere to the aforementioned "transductive setting", in which the sample of target documents given to Φ corresponds also to the unique set of documents we might be interested in classifying, i.e., $U_{\mathcal{T}} = Te_{\mathcal{T}}$, and there is not any sample $U_{\mathcal{S}}$ from the source collection other than the training set $Tr_{\mathcal{S}}$.

3.2 Pivot Selection

According to [3], pivots are frequent and discriminant terms that behave similarly in both the source and target domain. Regarding frequency, and as was done in [15], we restrict the set of pivot candidates to those which occur in at least $\phi = 30$ document in the source and target corpora. Following [2,15], we use the mutual information between the term and the classes $\{c, \overline{c}\}$ to asses the degree of discrimination of a given feature in the training set (i.e., exclusively in the source domain). Finally, we apply the *cross-consistency* heuristic defined in [11] which allows the model to be aware of the prevalence[5] drift across the source and target domains.

3.3 Distributional Correspondence Functions

DCFs are a family of real-valued functions that quantify the deviation of *correspondence* between two terms with respect to the expected correspondence due to chance. Different interpretations of correspondence could be plugged into the definition, leading to different implementations of DCF. In this work, we will restrict our attention to the cases in which correspondence is measured as the *cosine similarity* (Eq. 1), the *Asymmetric Mutual Information* (AMI – Eq. 2), the *Pointwise Mutual Information* (PMI – Eq. 3), and *linear* (Eq. 4), as discussed in [11].

Correspondence between two terms f^i and f^j in a given domain is measured by comparing their *context distribution vectors* $\mathbf{f}^i$ and $\mathbf{f}^j$. Context distribution vectors are extracted from the co-occurrence matrix of the domain, and model how a term relates to a set of contexts (e.g., documents).

$$Cosine(f^i, f^j) = \frac{\langle \mathbf{f}^i, \mathbf{f}^j \rangle}{\|\mathbf{f}^i\| \|\mathbf{f}^j\|} - \sqrt{p_i p_j} \tag{1}$$

$$AMI(f^i, f^j) = \rho(f^i, f^j) \sum_{x \in \{f^i, \overline{f^i}\}} \sum_{y \in \{f^j, \overline{f^j}\}} P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)} \tag{2}$$

[5] The prevalence of a term is typically defined as the proportion of documents in which a term appears in a corpus.

$$PMI(f^i, f^j) = \log_2 \frac{P(f^i, f^j)}{P(f^i)P(f^j)} \tag{3}$$

$$Linear(f^i, f^j) = P(f^i|f^j) - P(f^i|\overline{f^j}) \tag{4}$$

Where p_i denotes the prevalence (proportion of occurrences in the total number of contexts) of feature f^i, $P(x)$ denotes the probability that feature x occurs in a random context, $P(\overline{x})$ is the probability that x does not occur in a random context, and $\rho(x, y)$ is a function that changes the sign when x and y are negatively correlated [6].

3.4 Word Embeddings and Document Representation

The feature representations of DCI might be though as a generalization of *Co-Occurrence vectors* (see, e.g., [12]), where the co-occurrence metric is any of the DCF, and the context window is set to the document length. Once a set of m pivots $P = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m\}$ and a DCF η have been selected, each term f in the source and target domains is modeled as an m-dimensional vector

$$\overrightarrow{f} = (\eta(\mathbf{f}, \mathbf{p}_1), \eta(\mathbf{f}, \mathbf{p}_2), \ldots, \eta(\mathbf{f}, \mathbf{p}_m)) \tag{5}$$

where $\mathbf{f}$ and $\mathbf{p}_i$ are the context distribution vectors of the term f and the i^{th} pivot, respectively. Note that, because we are operating in the transductive regime, the context distribution vectors $\mathbf{f}$ and $\mathbf{p}_i$ are taken from the co-occurrence matrix of the training set when modeling the source terms, and from the co-occurrence matrix of the test set when modeling the target terms [7].

Finally, train and test documents are indexed in the embedding space via a weighted sum of all word embeddings of the terms composing the documents. That is, document d_i is represented as the m-dimensional vector

$$\overrightarrow{d}_i = \sum_{f_j \in d_i} w_{ij} \cdot \overrightarrow{f_j} \tag{6}$$

where w_{ij} is the weight of term f_j in document d_i (we used the standard cosine-normalized $tfidf$), and $\overrightarrow{f_j}$ is the word embedding for term f_j.

Once the training and test matrices have been represented in the embedding space, the classifier is learned. As the classifier we adopt the Transductive SVM [8], that also takes into account the structure of the test data while modeling the decision function. We used the linear-kernel which have consistently delivered good accuracy in text classification so far [7].

[6] That is, when the *true positive rate* plus the *true negative rate* as obtained from the 4-cell contingency table of x and y is lower than 0.

[7] In this case, and differently from [5,11] we do not apply *unification* to the common features, because during preliminary tests we observed most of the features to appear simultaneously in the source and target domains, causing thus most of the words in the vocabulary to get unified. This contradicts the rationale behind the unification process, originally proposed to consolidate the representations of shared words across languages, such as proper nouns in cross-lingual adaptation.

4 Experiments

In this section, we report on the experiments we run to test the effectiveness of our TDCI method in cross-domain topic classification.

As the evaluation measure we adopt standard *accuracy*, i.e., the ratio between the number of correctly labeled documents over the total number of documents sued to the classifier, i.e.,

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{7}$$

where TP, TN, FP, and FN stand for the numbers of true positives, true negatives, false positives, and false negatives, respectively. Note this choice is perfectly valid given that all datasets are approximately balanced with respect to the positive and negative classes,

In order to gain in reproducibility and to facilitate a comparison of performance with other methods, we consider most commonly used benchmarks in the related literature, including the Reuters-21578, SRAA, and 20 Newsgroups collections. Aside from being well-known benchmarks collections in the reign of topic classification, their class codes are organized hierarchically, and some representative subsets could thus be taken in order to generate new benchmarks that are well-suited for domain adaptation as well[8].

Reuters-21578: is one of the most used collections in TC research. Reuters-21578 is a set of 21,578 news stories appeared in the Reuters newswire in 1987. Documents in the collection are assigned to 5 top classes, among which, *orgs*, *people*, and *places* classes have commonly been selected in other works for experimenting in domain adaptation, leading to three datasets, *orgs vs people*, *orgs vs places*, and *people vs places*; a preprocessed version could be found in[9].

SRAA: consists of 73,218 Usenet posts about simulated autos, simulated aviation, real autos, and real aviation, accessible in[10]. In this dataset, the pairs of classes *real vs simulated*, and *auto vs aviation* have been used to instantiate two different domain adaptation problems. For example, in *real vs simulated*, documents about aviation were used as the source domain, while documents about autos constitute the target domain; the binary decision problem consists thus in discerning between *real* and *simulated* topics. In a similar vein, *auto vs aviation* is created, where documents about simulated vehicles act as source domain examples, and documents about real vehicles as the target ones.

[8] This procedure consists in taking two top classes, say, A and B, with subclasses $A.1 \dots A.x$ and $B.1 \dots B.y$, respectively. Then, two disjoint folds are taken for the source ($\mathcal{S}$) and target ($\mathcal{T}$) sides in each class; e.g., $A_{\mathcal{S}} = \cup_{1 \leq i < l} A.i$ and $A_{\mathcal{T}} = \cup_{l \leq i \leq x} A.i$, represent the source and target splits for the class A. Finally, the training and test sets are defined as $Tr_{\mathcal{S}} = A_{\mathcal{S}} \cup B_{\mathcal{S}}$ and $Te_{\mathcal{T}} = A_{\mathcal{T}} \cup B_{\mathcal{T}}$, where documents in A are labeled as positives, and documents in B are labeled as negatives.

[9] `http://www.cse.ust.hk/TL/dataset/Reuters.zip`

[10] `http://people.cs.umass.edu/~mccallum/data/sraa.tar.gz`

Dataset	ISVM	TSVM	CoCC	TPLSA	CDSC	MTrick	TCA	PSCC	Linear	AMI	PMI	Cosine
orgs vs places	0.721	0.740	0.680	0.653	0.682	0.768	0.730	0.742	0.792	0.787	**0.797**	0.793
orgs vs people	0.737	0.793	0.764	0.763	0.768	0.808	0.792	0.807	0.782	**0.810**	0.799	0.805
people vs places	0.595	0.614	**0.826**	0.805	0.798	0.690	0.626	0.690	0.700	0.703	0.676	0.700
real vs simulated	0.737	0.920	0.880	0.889	0.812	-	-	-	0.962	0.966	**0.967**	0.958
auto vs aviation	0.799	0.949	0.932	0.947	0.880	-	-	-	0.974	0.972	**0.978**	0.976
comp vs sci	0.699	0.842	0.870	**0.989**	0.902	-	0.891	0.900	0.906	0.879	0.910	0.858
rec vs talk	0.722	0.971	0.965	0.977	0.908	0.950	0.962	0.962	0.978	**0.981**	0.978	0.959
rec vs sci	0.803	0.945	0.945	0.951	0.876	0.955	0.879	0.955	**0.974**	0.969	**0.974**	0.959
sci vs talk	0.783	0.913	0.946	**0.962**	0.956	0.937	0.940	0.947	0.946	0.943	0.943	0.931
comp vs rec	0.782	0.903	**0.958**	0.951	**0.958**	-	0.940	**0.958**	0.913	0.909	0.914	0.909
comp vs talk	0.955	0.909	**0.980**	0.977	0.976	-	0.967	0.967	0.932	0.935	0.944	0.913
Reuters (ave.)	0.684	0.716	0.757	0.740	0.749	0.755	0.716	0.746	0.758	**0.767**	0.757	0.766
SRAA (ave.)	0.768	0.935	0.906	0.918	0.846	-	-	-	0.968	0.969	**0.973**	0.967
20News (ave.)	0.799	0.914	0.944	**0.968**	0.929	-	0.930	0.948	0.942	0.936	0.944	0.922

Table 1. Results of TDCI with different DCFs on Reuters-21578, SRAA, and 20 Newsgroups datasets. Shadowed cells highlight the configurations of TDCI that obtained higher scores than any other baseline. Values in bold highlight the best score for each dataset.

20 Newsgroups: is a publicly available[11] text collection of approximately 20,000 Usenet discussion groups, which are nearly evenly partitioned across 20 different newsgroups. Following the common practice in the related literature, we restrict our attention to the 4 most frequent top classes in the dataset (*comp*, *sci*, *rec*, and *talk*). The data is then split by their sub-classes for certain pairs of top-classes. We generated 6 different domain adaptation problems by following the same classes split as defined in [17].

We compare the performance of TDCI[12] with the following baselines (discussed in section 2): Co-Clustering (CoCC–[4]), Topic-bridged PLSA (TPLSA-[17]), Cross Domain Spectral Classification (CDSC–[10]), Matrix Trifactorization (MTrick–[18]), Topic Correlation Analysis (TCA–[9]), and Partially Supervised Cross-Collection LDA (PSCC–[1]). Additionally, we report experiments on a lower bound baseline that simply classifies the target documents using an Inductive SVM[13] trained on the source domain without carrying out any sort of adaptation (ISVM–[7]), and its transductive version (TSVM–[8]).

Table 1 reports the results of our experiments in terms of accuracy for the Reuters-21578, SRAA, and 20 Newsgroup datasets. Reported score values for CoCC, TPLSA, CDSC, MTrick, TCA, and PSCC where taken from the original papers. Columns Linear, AMI, PMI, and Cosine correspond to our TDCI with different DCFs (see Section 3.3); in this experiment, we set the number of pivots to 100, following [5,11].

[11] `http://qwone.com/~jason/20Newsgroups/`

[12] The code implementing or method is integrated in JaTeCS, and available in `https://github.com/jatecs/jatecs`. A stand-alone version could also be accessed in `http://hlt.isti.cnr.it/dciext/`

[13] We used the popular Joachims' implementation in `http://svmlight.joachims.org/`

Overall, the results of these experiments indicate that TDCI is competitive in transductive cross-domain classification by topic. In average, all configurations of TDCI outperform all baselines in terms of accuracy in Reuters-21578 and SRAA datasets. In 20 Newsgroup TDCI performs comparably in average, without surpassing though the best averaged score obtained by TPLSA. When AMI is used as the DCF, TDCI beats all baselines in 6 out of 11 cases, obtaining two best global results (*orgs vs people*, and *rec vs talk*), and the best average in Reuters-21578. The PMI function also obtained promising results, surpassing all baselines in 5 out of 11 cases, with four best global results, and the best average in SRAA. It is also noticeably that TDCI outperforms all comparison methods in SRAA, in all cases irrespective of the DCF, and by a significant margin.

When TDCI did not achieve the best score, its performance could still be considered aligned with respect to the baselines, with the sole exceptions of *comp vs rec* and *comp vs talk* cases, where TDCI performed comparatively worse. This could be explained by observing the relative performance of ISVM and TSVM. In both cases the improvements in accuracy brought about by transduction represent the lowest ones in the entire 20 Newsgroup benchmark; note it even degrades significantly in *comp vs talk*. Such observation prompted us to confront TDCI with its inductive version (noted IDCI for consistency). IDCI outperforms TDCI only in these two cases (and performed significantly worse in the rest of cases, that we omit for the sake of brevity). More precisely, accuracy scores delivered by IDCI ranged from 0.943 (Cosine) to 0.961 (Linear) in *comp vs rec* (which is now comparable to the baselines performances), and from 0.983 (PMI) to 0.992 (Linear) in *comp vs talk* (which surpasses the best accuracy score of 0.980 attributed to CoCC). This sensible variation in performance suggests further investigations are needed in order to shade light on deciding whether it is advisable to maintain an inductive strategy even when the structure of the test set is observable.

We also investigated the impact in performance due to variations in the number of pivots, i.e., the dimensionality of the embedding space. Table 1 shows two representative plots that summarizes well the casuistic we found in our experiments.

The plot for *rec vs talk* exemplifies the most frequent case in our experiments, in which accuracy improves smoothly as more pivots are selected. The case of *comp vs talk* exemplifies a less frequent case in which the performance is somehow unstable. These fluctuations seem to depend on the number of pivots, and on the DCF at hand. For example, the performance trend is increasing for AMI and decreasing for the Cosine DCFs; while PMI and Linear delivered competitive results even with only 30 pivots. This result seem to indicate the order in which pivots should be selected could, in some cases, depend also on the DCF under consideration, something we plan to clarify in future research.

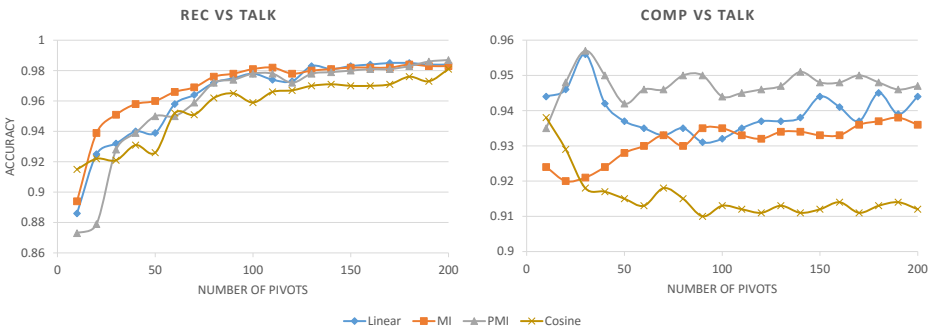

Fig. 1. Variations in accuracy at the variation of the number of pivots.

5 Conclusions

In this article, we have explored the performance efficiency of Distributional Correspondence Indexing, a method originally proposed for cross-domain and cross-lingual inductive classification by sentiment, in a different problem setting, i.e., by considering the topic classification task and the transductive inference. Results show our transductive version, dubbed TDCI, to be comparable and even better in many cases to the state of the art on three extensively used datasets. Our experiments also revealed more investigations are still required in order to automatically determine the optimal number of pivots to select, so as to find more stable distributional correspondence functions.

References

1. Bao, Y., Collier, N., Datta, A.: A partially supervised cross-collection topic model for cross-domain text classification. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. pp. 239–248. ACM (2013)
2. Blitzer, J., Dredze, M., Pereira, F.: Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007). pp. 440–447. Prague, CZ (2007)
3. Blitzer, J., McDonald, R., Pereira, F.: Domain adaptation with structural correspondence learning. In: Proceedings of the 4th Conference on Empirical Methods in Natural Language Processing (EMNLP 2006). pp. 120–128. Sydney, AU (2006)
4. Dai, W., Xue, G.R., Yang, Q., Yu, Y.: Co-clustering based classification for out-of-domain documents. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 210–219. ACM (2007)
5. Esuli, A., Moreo Fernández, A.: Distributional correspondence indexing for cross-language text categorization. In: Proceedings of the 37th European Conference on Information Retrieval (ECIR 2015). pp. 104–109. Wien, AT (2015)
6. Harris, Z.S.: Distributional structure. Word 10(23), 146–162 (1954)

7. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning (ECML 1998). pp. 137–142. Chemnitz, DE (1998)
8. Joachims, T.: Transductive inference for text classification using support vector machines. In: Proceedings of the 16th International Conference on Machine Learning (ICML 1999). pp. 200–209. Bled, SL (1999)
9. Li, L., Jin, X., Long, M.: Topic correlation analysis for cross-domain text classification. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 2012). pp. 998–1004. Toronto, CA (2012)
10. Ling, X., Dai, W., Xue, G.R., Yang, Q., Yu, Y.: Spectral-domain transfer learning. In: Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2008). pp. 488–496. Las Vegas, US (2008)
11. Moreo Fernández, A., Esuli, A., Sebastiani, F.: Distributional correspondence indexing for cross-lingual and cross-domain sentiment classification. Journal of Artificial Intelligence Research 55, 131–163 (2016)
12. Niwa, Y., Nitta, Y.: Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In: Proceedings of the 15th conference on Computational linguistics-Volume 1. pp. 304–309. Association for Computational Linguistics (1994)
13. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10), 1345–1359 (2010)
14. Pan, W., Zhong, E., Yang, Q.: Transfer learning for text mining. In: Aggarwal, C.C., Zhai, C. (eds.) Mining Text Data, pp. 223–258. Springer, Heidelberg, DE (2012)
15. Prettenhofer, P., Stein, B.: Cross-lingual adaptation using structural correspondence learning. ACM Transactions on Intelligent Systems and Technology 3(1), Article 13 (2011)
16. Vapnik, V.: Statistical Learning Theory. Wiley, New York, US (1998)
17. Xue, G.R., Dai, W., Yang, Q., Yu, Y.: Topic-bridged PLSA for cross-domain text classification. In: Proceedings of the 31st ACM International Conference on Research and Development in Information Retrieval (SIGIR 2008). pp. 627–634. Singapore, SN (2008)
18. Zhuang, F., Luo, P., Xiong, H., He, Q., Xiong, Y., Shi, Z.: Exploiting associations between word clusters and document classes for cross-domain text categorization. Statistical Analysis and Data Mining 4(1), 100–114 (2011)