Journal of Artificial Intelligence Research 55 (2016) 131-163 Submitted 3/2015; published 1/2016

Distributional Correspondence Indexing for Cross-Lingual and Cross-Domain Sentiment Classification.

Alejandro Moreo Fernández ALEJANDRO.MOREO@ISTI.CNR.IT
Andrea Esuli ANDREA.ESULI@ISTI.CNR.IT
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
56124 Pisa, IT

Fabrizio Sebastiani FSEBASTIANI@QF.ORG.QA
Qatar Computing Research Institute
Hamad bin Khalifa University
PO Box 5825, Doha, QA

Abstract

Domain Adaptation (DA) techniques aim at enabling machine learning methods learn effective classifiers for a "target" domain when the only available training data belongs to a different "source" domain. In this paper we present the Distributional Correspondence Indexing (DCI) method for domain adaptation in sentiment classification. DCI derives term representations in a vector space common to both domains where each dimension reflects its distributional correspondence to a pivot, i.e., to a highly predictive term that behaves similarly across domains. Term correspondence is quantified by means of a distributional correspondence function (DCF). We propose a number of efficient DCFs that are motivated by the distributional hypothesis, i.e., the hypothesis according to which terms with similar meaning tend to have similar distributions in text. Experiments show that DCI obtains better performance than current state-of-the-art techniques for cross-lingual and cross-domain sentiment classification. DCI also brings about a significantly reduced computational cost, and requires a smaller amount of human intervention. As a final contribution, we discuss a more challenging formulation of the domain adaptation problem, in which both the cross-domain and cross-lingual dimensions are tackled simultaneously.

1. Introduction

Automated text classification methods usually rely on a training set of labelled examples in order to learn a classifier that will predict the classes of unlabelled documents. One important bottleneck that supervised machine learning methods have to deal with has to do with their dependence on high-quality annotated examples in order for the model to be trained. Deploying a model for a domain where these examples are not available thus entails a substantial human labelling effort.

Transfer learning (TL – see e.g., Pan & Yang, 2010; Pan, Zhong, & Yang, 2012) focuses on alleviating this problem by leveraging training examples from a different, although related, *source domain* (a.k.a. *out-domain*, or *auxiliary domain*) for which the amount of available labelled examples is higher. TL allows making use of these examples in order to train an effective classifier for the *target domain* (a.k.a. *in-domain*), thus allowing to diminish or completely do away with the cost involved in the manual generation of training documents for the target domain. As a consequence, when TL is applied one of the fundamental assumptions of "traditional" machine learning, i.e., that the training and the test data are randomly drawn from the same distribution (the so-called "iid assumption"), no longer holds.

One applicative scenario of particular interest for TL is *sentiment classification*, the task of classifying opinion-laden documents as conveying a positive or a negative sentiment towards a given entity (e.g., a product, a policy, a political candidate). Determining the users' stance towards such an entity is of the utmost importance for market research, customer relationship management, the social sciences, and political science among others, and several automated methods have been proposed for this purpose (see e.g., Liu, 2012; Pang & Lee, 2008). However, when sentiment classification needs to deal with a completely new entity, it is likely that the amount of available, pre-labelled opinion-laden documents is scarce or even null. In such cases, promptly generating a sentiment classifier might become difficult, due to the considerable cost and time involved in producing a representative corpus of training documents.

In sentiment classification, TL finds a natural application in *domain adaptation* (DA), i.e., the task of adapting a sentiment classifier to operate on a new domain. For example, we might want to use a training set of book reviews written in English to classify movie reviews written in English, or to classify book reviews written in German. The former case is typically known as *cross-domain adaptation*, while the second one is instead known as *cross-lingual adaptation* (Pan et al., 2012). In this article we will use the notation $L_sC_s \rightarrow L_tC_t$ to indicate the domain adaptation setup, where L_s and L_t are the source and target languages, and C_s and C_t are the source and target domains, respectively. Therefore, the previously discussed examples will be written as EnglishBooks→EnglishDVDs (or simply Books→DVDs for short), which is an example of cross-domain adaptation, and EnglishBooks→GermanBooks, which is an example of cross-lingual adaptation.

A common practice in text classification is to represent a dataset as a term-document matrix M according to the so-called *bag-of-words model* (BoW), where the value M_{ij} is a function of the frequency of term f_i in document d_j and in the dataset as a whole. Accordingly, rows (resp., columns) can be regarded as vectorial representations of terms (resp., documents) in a vector space generated by documents (resp., terms). Here, the expectation is that distances between vectors in this vector space model (VSM) reflect the semantic distance between terms or between documents. Since each term is a dimension of the vector space where documents lie, terms are represented by orthogonal dimensions even if they have similar meanings. For example, the term beautiful is viewed as being as dissimilar to nice as it is to awful; the base of the vector space is therefore "unaware" of any semantic nuance, which lies hidden in the joint term distributions. Statistical methods like Latent Semantic Indexing (LSA – Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990; Landauer & Dumais, 1997) and Latent Dirichlet Allocation (LDA – Blei, Ng, & Jordan, 2003) attempt to discover these hidden correlations among terms. Discovering such semantic correspondences becomes crucial when dealing with different domains; for example, when "adapting" the target domain of book reviews from a source domain of film reviews, identifying cross-domain semantic correspondences among important terms (e.g., book ≈ film, writer ≈ director, length ≈ duration) might be helpful for the task, as it is likely that the decision boundaries for the model hinge upon these terms.

With respect to this aspect, a general assumption in domain adaptation is that large sets of unlabelled documents for both the source and the target domain are available. When leveraging these unlabelled collections, various techniques can be applied in order to better explore the term distributions in the two domains, in an attempt to map the similarities between the terms in the two domains. This idea rests on the belief that the meaning of a term is somehow determined by its distribution in text and by the terms it co-occurs with, an idea that is generally referred to as the *distributional hypothesis* (Harris, 1954).

The discovery of hidden correlations between highly predictive terms, with the goal of improving document representations, was studied in the Structural Learning paradigm (Ando & Zhang, 2005). These correlations are discovered by learning predictive structures of the input data as auxiliary binary problems which consist of predicting term presence using the surrounding terms in the unlabelled data and then applying LSA to the predictors. This framework was then extended in Structural Correspondence Learning for domain adaptation (SCL – Blitzer, McDonald, & Pereira, 2006). SCL unifies in a latent space the correspondences among terms from different domains that derive from the auxiliary prediction problems of *pivot* terms – highly predictive terms expected to behave in a similar way in both domains (e.g., intriguing, annoying, or captivating in both film and book reviews). SCL was first applied to cross-domain adaptation in sentiment classification (Blitzer, Dredze, & Pereira, 2007). It was later applied to cross-lingual adaptation (Prettenhofer & Stein, 2010) by enhancing each pivot term with one of its equivalent translations in the target language (e.g., intriguing $\approx$ intrigante, annoying $\approx$ noioso, or captivating $\approx$ travolgente in EnglishBooks → ItalianBooks cross-lingual adaptation). Even though SCL was successfully validated in the cross-domain and cross-lingual scenarios, it suffers from a considerable computational cost, deriving from the intermediate optimizations of the auxiliary predictive problems and from the use of LSA.

The method we present in this paper, which we dub *Distributional Correspondence Indexing* (DCI), is inspired by SCL but follows a different, simpler approach, with a more direct application of the distributional hypothesis. We propose to mine the distributional correspondences between each term and a small set of pivot terms. We hypothesize these correspondences to be approximately invariant across both domains for terms that have equivalent roles in the two domains. For example, we expect the distributional correspondence between the source term $f_s = \textsf{book}$ and pivots $\mathbf{p}_s^1 = \textsf{intriguing}$, $\mathbf{p}_s^2 = \textsf{annoying}$, $\mathbf{p}_s^3 = \textsf{captivating}$, to be approximately similar to the distributional correspondence between the target term $f_t = \textsf{film}$ and the same pivots $\mathbf{p}_t^1$, $\mathbf{p}_t^2$, $\mathbf{p}_t^3$ (Books → DVDs cross-domain adaptation). Analogously, we expect the distributional correspondence between the source (English) term $f_s = \textsf{book}$ and pivots $\mathbf{p}_s^1 = \textsf{intriguing}$, $\mathbf{p}_s^2 = \textsf{annoying}$, $\mathbf{p}_s^3 = \textsf{captivating}$, to be approximately similar to the distributional correspondence between the target (Italian) term $f_t = \textsf{libro}$ and the pivot translations $\mathbf{p}_t^1 = \textsf{intrigante}$, $\mathbf{p}_t^2 = \textsf{irritante}$ and $\mathbf{p}_t^3 = \textsf{accattivante}$ (EnglishBooks → ItalianBooks cross-lingual adaptation). Contrarily to SCL, we define these distributional correspondences through a *distributional correspondence function* that directly mines term vectors from unlabelled collections, and that can be computed efficiently.

The present work is an extension of work by Esuli and Moreo Fernández (2015), where some preliminary intuitions underlying our method were applied to the cross-lingual case. The present improved version of DCI compares favourably with respect to the state of the art in extensive experimental comparisons we have carried out on two popular sentiment classification datasets that cover both cross-domain and cross-lingual adaptation. The same experiments also show that DCI has a substantially smaller computational cost with respect to the competition. In the cross-lingual scenario, we show DCI to require a smaller amount of human intervention in order to create the cross-lingual pivots. As a final contribution, we explore a more general and more complex formulation of the domain adaptation problem that combines the cross-domain setting with the cross-lingual setting; we present experimental results where we compare our method with some state-of-the-art methods.

The rest of this article is structured as follows. Section 2 overviews related work in domain adaptation. Section 3 formally states the problem and presents the notation we are going to use. Section 4 formally defines the distributional correspondence functions, while the method as a whole is described in Section 5. Section 6 presents the experiments and our analysis of the results, while Section 7 concludes.

2. Related Work

This section offers a brief overview of the main related methods in the literature of domain adaptation for sentiment classification. Traditionally, two different types of approaches to domain adaptation can be identified. A first group of *instance transfer* methods aim at re-weighting the relative importance of each training document, in order to compensate the disagreements between the source and the target marginal probability distributions (Dai, Xue, Yang, & Yu, 2007; Gao, Fan, Jiang, & Han, 2008). Such methods can be applied only to cross-domain (and not cross-lingual) adaptation problems, since they cannot solve the problem posed by the fact that, in the cross-lingual setting, the term sets of the source and target domains are disjoint. A second group of *feature-representation transfer* methods project documents from both domains into a common vector space in which standard classification algorithms could be applied. The DCI method we propose belongs to the feature-representation transfer class, and can thus be applied to both the cross-domain and the cross-lingual settings. In the following we review the most relevant work on the cross-domain setting (Section 2.1) and the cross-lingual setting (Section 2.2). The interested reader can check (Pan & Yang, 2010; Pan et al., 2012) for surveys on transfer learning methods.

2.1 Cross-Domain Adaptation

The Structural Correspondence Learning method (Blitzer et al., 2006), already discussed in the introduction, extends the Structural Learning paradigm of Ando and Zhang (2005) by introducing the concept of "pivot features". SCL has been applied to cross-domain adaptation (Blitzer et al., 2007) by leveraging the notion of "predictive power" of a pivot. A similar criterion has been adopted to discern between domain-specific and domain-independent terms in Spectral Feature Alignment (SFA – Pan, Ni, Sun, Yang, & Chen, 2010), a method for clustering domain-specific terms from source and target domains into a latent space by mining their relationships with domain-independent terms.

Aside from sets of unlabelled documents for each domain, some methods take advantage of external general-purpose knowledge resources in order to bridge the gap between domains. For example, Wang, Domeniconi, and Hu (2008) extended the co-clustering approach to propagate labels between the two domains by using Wikipedia to represent documents by means of concepts. More recently, the Bridging Information Gap method (Xiang, Cao, Hu, & Yang, 2010) followed a similar motivation, exploiting Wikipedia or the Open Directory Project as the general-purpose knowledge base. In sentiment classification, other related methods use a sentiment lexicon as the external resource. The Joint Sentiment-Topic Model (JSTM), proposed by He, Lin, and Alani (2011) as an extension of Latent Dirichlet Allocation, consists of augmenting terms with polarity-bearing topics using a sentiment lexicon as a repository of prior word sentiment. The JSTM was found to perform better than SCL and comparably to SFA. Denecke (2009) studied the viability of SentiWordNet, a well-known sentiment lexicon, as a lexicon for cross-domain sentiment adaptation; the main drawback of methods such as these is their dependence on the availability of suitable public resources / lexicons for the language which the application targets. Li, Pan, Jin, Yang, and Zhu (2012a) alleviated such constraint by automatically co-extracting a topic lexicon and a sentiment lexicon for the target domain, exploiting the information from the source domain. Similarly, Bollegala, Weir, and Carroll (2011) obtained a sentiment-sensitive thesaurus to augment term vectors. The main peculiarity of this approach is that the lexicon is created by mining multiple source domains. Similarly, but following a completely different approach based on deep learning architectures, the Stacked Denoising Autoencoder (SDA$_{sh}$) method (Glorot, Bordes, & Bengio, 2011) exploits the unlabelled information contained in multiple domains in order to improve the vector representations of terms in an unsupervised fashion. Glorot et al. found the method (which we use as one of the baselines in our experiments) to scale well on large multi-domain collections, outperforming SCL and SFA when using 22 different domains of unlabelled documents.

Other branches of research related to cross-domain methods for binary classification exist that were not tested on sentiment classification but on topic classification, using the popular datasets Reuters-21578, 20-Newsgroups, and SRAA. Some relevant examples include Spectral Domain Transfer Learning (Ling, Dai, Xue, Yang, & Yu, 2008), Matrix Trifactorization (Zhuang, Luo, Xiong, He, Xiong, & Shi, 2011), Topic Correlation Analysis (Li, Jin, & Long, 2012b), and Topic-Bridged Probabilistic LSA (Xue, Dai, Yang, & Yu, 2008). Aside from the fact that these methods were not explicitly designed to classify according to sentiment, these approaches also faced a different problem setting, i.e., the test set in the target domain is available – though with labels omitted – when modelling the classification hypothesis, and there is no other collection of unlabelled documents available for the source or target domain. These approaches thus fall in the domain of *transductive* learning (see e.g., Joachims, 1999), and are thus not directly related to our approach, which is completely inductive.

2.2 Cross-Lingual Adaptation

We review prior work on cross-lingual adaptation, covering three different types of approaches: (i) methods relying on automatic machine translation, (ii) methods exploiting parallel corpora, and (iii) methods exploiting unlabelled topic-specific collections.

Rigutini, Maggini, and Liu (2005) proposed a method for cross-lingual adaptation that first translates the source documents into the target language by means of an automatic machine translation service. Then, an Expectation Maximization method refines the translated representations by mining the statistical properties of large sets of unlabelled documents in the target language. Along these lines, Wan, Pan, and Li (2011) proposed a bi-weighting method to re-weight both terms and training instances in order to correct the word drift that machine translation could have introduced during the translation process. Motivated by the lack of labelled Chinese sentiment corpora, Wan (2009) proposed instead an English-Chinese co-training approach based on automatic machine translation. The method translates the labelled English (source) documents into Chinese (target), and the Chinese unlabelled documents into English. A classifier is then created for each of the languages, that is later used to classify their respective set of unlabelled documents to improve the model. Finally, each Chinese test document is attached to its translation equivalent in English and given as input to the classifier.

Even though machine translation represents a promising solution to cross-lingual problems (a solution that will presumably become more and more viable as the field of machine translation evolves), current machine translation services are not always free-to-use, nor available for all language pairs either, and are computationally expensive. All other things being equal, cross-lingual methods that do not rely on them are thus preferable.

Latent Semantic Analysis is a well-known technique which originated within monolingual text analysis (Deerwester et al., 1990) but was later extended to deal with cross-lingual retrieval (Dumais, Letsche, Littman, & Landauer, 1997) and multilingual classification (Xiao & Guo, 2013). LSA consists of mapping the original term-document matrix into a lower-dimensional latent semantic space that captures the (linear) relations among the original terms and the documents. In a cross-lingual context, this mapping is performed via a singular value decomposition of the original term-document matrix extracted from multilingual documents. The main problem with the use of LSA for cross-lingual applications is its dependence on a parallel corpus. In order to relax this constraint, Xiao and Guo (2014) proposed a method that induces cross-lingual terms via matrix completion. The method requires only a small set of parallel documents that is used to construct a dual-language co-occurrence matrix; LSA is then applied to the completed dual-language matrix in order to produce a low-dimensional interlingual representation. Cross-lingual Kernel Canonical Correlation Analysis (KCCA – Vinokourov, Shawe-Taylor, & Cristianini, 2002) produces instead a semantic cross-lingual representation by investigating correlations between aligned bilingual fragments. KCCA takes advantage of kernel functions in order to map aligned texts into a high-dimensional space in such a manner that the correlations between both mappings are mutually maximized. Finally, Oriented Principal Component Analysis (OPCA – Platt, Toutanova, & Yih, 2010) finds a discriminative projection that maximizes the document variance across languages, at the same time minimizing the distance between comparable documents, thus avoiding the use of artificially concatenated documents.

The techniques based on correlation analysis that we have discussed above are rather expensive from a computational point of view, and their use requires the availability of a parallel or comparable corpus. For this reason Moen and Marsi (2013) proposed the use of Random Indexing (Sahlgren, 2005), a computationally lighter indexing approach that approximates LSA (Kanerva, Kristofersson, & Holst, 2000), as an alternative for use in cross-lingual information retrieval. Cross-lingual Random Indexing requires only a monolingual corpus for each language, plus a bilingual dictionary.

Even though some of the approaches discussed above succeed in discovering hidden correlations between terms belonging to different languages, they are still based on the availability of a suitable parallel corpus or a bilingual dictionary. As a response, Gliozzo and Strapparava (2005, 2006) provided a means for automatically obtaining a Multilingual Domain Model (MDM) by defining soft relations between words and domain topics. Making up for the lack of a bilingual dictionary or a parallel corpus, a MDM can be automatically obtained from comparable corpora by performing LSA. In a similar vein, Rapp (1999) proposed a method for acquiring a bilingual dictionary based on the assumption that there is a correlation between word co-occurrence patterns in different languages (Rapp, 1995). The dimensions of the co-occurrence matrices are rearranged so as to make translation equivalents of the same word correspond to identical positions in the vector, by using a small bilingual dictionary. Word translation is then performed as vector similarity in the two co-occurrence matrices, ignoring all dimensions that are not aligned. The dictionary is iteratively expanded with the inclusion of newly translated terms. Koehn and Knight (2002) proposed a method for automatically constructing a word-level translation lexicon by taking a monolingual corpus for each language as input, neither requiring the corpora to be parallel or even comparable, nor requiring the availability of an initial dictionary. Roughly speaking, this was done by first taking words with identical spelling ("cognates") or similar spelling as the initial entries of the dictionary, and then by expanding the dictionary by assuming the context, frequency, and word correlations to be approximately preserved across languages. More recently, Peirsman and Padó (2010) proposed a method to induce selectional preferences for resource-poor languages which also takes advantage of cognates. A bilingual vector space is initially derived by taking cognates as the dimensions of the vector space. This space, that is subsequently bootstrapped from large (unparsed) corpora of both languages, allows direct word translations to be performed based on vector distances.

In a realistic cross-lingual setting we do not expect to have any sort of labelled corpus available for the target domain, and machine translation tools, when available, are still expensive. Therefore, Prettenhofer and Stein (2010, 2011) assume that a word translation oracle is available but only for a limited budget of calls (450, in their experiments). The resulting word translations allow Structural Correspondence Learning (SCL – see Section 2.1) to be applied to cross-lingual domain adaptation by pairing each source pivot with its equivalent translation in the target language. Even though cross-lingual SCL succeeds in relaxing the constraints discussed above thanks to the fact that it does not need any linguistic resource, it still suffers from a considerable computational cost deriving from the need to perform intermediate optimizations of structural problems, and from the need to use LSA. Following a similar strategy relying on bilingual pivots, our DCI method requires significantly fewer word translations, and avoids the use of any expensive statistical analysis technique.

Our method bears some resemblance to Explicit Semantic Analysis (ESA), a method that indexes any given text with respect to a set of explicitly given external categories (Gabrilovich & Markovitch, 2007). In the work of Sorg and Cimiano (2008, 2012) different language-specific views of Wikipedia articles were considered as the external categories on which semantic term vectors were defined. Each dimension thus models the strength of association between a given term and a given article in a cross-lingual information retrieval setting (CL-ESA). Arguably, the main difference with our method is that CL-ESA relies on high-dimensional spaces (about 10,000 dimensions) of interlingual and universal concepts, while DCI instead projects each term into a low-dimensional space (about 100 dimensions) of highly predictive concepts, i.e., bilingual pivots. Additionally, our method does not require any sort of external resource apart from a word translation oracle, and the strength of association is rather defined in terms of distributional correspondence, computed efficiently on unlabelled sets (Section 4).

3. Problem Statement

In this section we formally state our problem and set the notation we are going to use throughout this paper.

Sentiment classification may be viewed as the task of approximating the unknown target function $\Phi : X \rightarrow \mathcal{Y}$, that indicates how documents ought to be classified, by means of a function $\hat{\Phi} : X \rightarrow \mathcal{Y}$, called the classifier, where X denotes the space of documents and $\mathcal{Y} = \{+1, -1\}$ denotes the space of labels, indicating positive (+1) or negative (-1) sentiment. In domain adaptation (see e.g., Pan & Yang, 2010) it is customary to define a *domain* as a pair $\mathcal{D} = \langle F, P(X) \rangle$, where $P(X)$ is the marginal probability distribution that governs the likelihood with which documents (represented in the term space F) are drawn. Given a *source* domain $\mathcal{D}_s = \langle F_s, P_s(X) \rangle$ and a *target* domain $\mathcal{D}_t = \langle F_t, P_t(X) \rangle$, *domain adaptation* is a subtask of transfer learning that consists of improving the accuracy of the classifier $\hat{\Phi}$ on $\mathcal{D}_t$ by using knowledge from $\mathcal{D}_s$, a domain such that $\mathcal{D}_s \neq \mathcal{D}_t$.

The precondition $\mathcal{D}_s \neq \mathcal{D}_t$ leads to two different interpretations of the domain adaptation problem. On one side, *cross-domain* adaptation (e.g., DVDs → Books) is typically characterized by $F_s = F_t$ and $P_s \neq P_t$; that is, the term space is common – or is trivially made common by joining the two term spaces – but the marginal probability distributions differ. On the other side, *cross-lingual* adaptation (e.g., EnglishBooks → GermanBooks) is typically characterized by $F_s \neq F_t$ and $P_s = P_t$; that is, documents are drawn from similar distributions but are described in different term spaces.

In this paper we also define and investigate a third case of domain adaptation, in which both the term spaces and the probability distributions differ. This is the *cross-domain/cross-lingual* adaptation problem, characterized by $F_s \neq F_t$ *and* $P_s \neq P_t$. We argue that this case is of particular interest, since it enables cross-lingual adaptation to be performed even in the absence of an auxiliary dataset that acts as a “bridge” in a two-steps adaptation (e.g., EnglishBooks → GermanBooks → GermanDVDs). For example, when a sentiment classifier for a resource-poor language needs to be deployed that analyses sentiment about a new topic, a common scenario is one in which we want to leverage data from a resource-rich language (e.g., English) on a different, already known topic. This scenario is a realistic generalization of the domain adaptation problem; to the best of our knowledge, nobody has tackled it before in published work.

As a final note regarding notation, we will use subscripts s and t to indicate whether the data is drawn from the source or from the target domain, respectively. Accordingly, U_s denotes the unlabelled source dataset, while U_t refers to the unlabelled target dataset. Similarly, Tr_s and Te_t denote the training set and test set, respectively.

4. Distributional Correspondence Functions

The goal of this section is to introduce Distributional Correspondence Functions (DCFs). We first characterize the family of DCFs and then propose some specific ones.

4.1 Definition

DCFs are a family of real-valued functions that quantify the degree of correspondence between two terms f^i and f^j by comparing their *context distribution vectors* $\mathbf{f}^i$ and $\mathbf{f}^j$ from an unlabelled collection U. A context distribution vector is a unit-length n-dimensional vector that models how a term relates to a set of contexts. A context is any text unit in which a term could appear (e.g., a sentence, a fixed-size window, or an entire document); $\mathbf{f}^i_c$ denotes the value of the vector for term f^i in context c, with $\mathbf{f}^i_c = 0$ if f^i does not appear in context c. The cases in which $\mathbf{f}^i_c > 0$ are determined by the weighting function in use, and might lead to different interpretations of the DCF, e.g., as a probability function in an event space (Section 4.2), or as a kernel in a vector space (Section 4.3). We define p_i as the prevalence of f^i, i.e., the portion of contexts for which $\mathbf{f}^i_c > 0$, i.e.,

$$p_i = \frac{|\{c|\mathbf{f}^i_c > 0\}|}{n} \tag{1}$$

where n is the dimensionality of the vector space, i.e., the number of different contexts. In this work we take documents as contexts, so $\mathbf{f}^i_c = 0$ means that term f^i does not appear in document c, and p_i is the portion of documents of the unlabelled collection in which f^i appears.

A DCF is a function $\eta : \mathcal{R}^n \times \mathcal{R}^n \rightarrow \mathcal{R}$, where the sign of $\eta(\mathbf{f}^i, \mathbf{f}^j)$ indicates the polarity of the correspondence, i.e., positive values indicate positive correlation and negative values indicate negative correlation; $\eta(\mathbf{f}^i, \mathbf{f}^j) = 0$ indicates null correspondence. We will force DCFs to be such that $\eta(\mathbf{f}^i, \mathbf{f}^j) = 0$ for the expected correspondence measured between randomly chosen pairs of vectors once the prevalence p_i and p_j of terms f^i and f^j have been set. The rationale of this choice is that high-prevalence terms have a higher probability to have non-zero values in a number of common positions, and in this case some measures (see Section 4.3) will record a level of correspondence due to the non-perfect orthogonality of the vectors, which happens much more rarely for low-prevalence terms. We want to factor out this bias, centering the DFC on the expected correspondence value.

4.2 Probability-Based DCF

In this section we discuss some probability-based DCFs derived from information theory. The distribution $P(f^i)$ of a term f^i across the contexts is modelled using a binomial event space, thus considering only the presence or absence of the term in the context; $P(f^i)$ thus denotes the probability that f^i occurs in a random context, while $P(\overline{f}^i)$ denotes the probability that f^i does not occur in it.

The first part of Table 1 shows the probability-based DCFs we investigate. We consider *Pointwise Mutual Information* (PMI, the ratio between the joint distribution and the product of the marginal distributions), and a simple probabilistic function (here called *Linear*) that contrasts the probabilities of f^i conditioned on f^j and $\overline{f}^j$, respectively. We also consider Mutual Information (MI, the reduction in entropy of a distribution due to the observation of another distribution) in an asymmetric version. The rationale of this asymmetry is that MI *per se* is symmetric with respect to positive and negative correlation; that is, the two cases in which (a) f^i occurs in all and only the contexts in which f^j also occurs, and (b) f^i occurs in all and only the contexts in which f^j does *not* occur, obtain the same MI score. In our scenario these two kinds of correlation must be kept distinct, because a high positive correlation indicates semantic similarity, while a high negative correlation indicates a lack of semantic similarity. For this reason we invert the sign of the DCF in the negative correlation case, by defining a ρ function that detects the negative correlation by using the *true positive rate* ($tpr = P(f^i, f^j)/P(f^j)$) and the *true negative rate* ($tnr = P(\overline{f}^i, \overline{f}^j)/P(\overline{f}^j)$), i.e.,

$$\rho(f^i, f^j) = \begin{cases} +1 & if\ (tpr + tnr > 1) \\ -1 & otherwise \end{cases} \tag{2}$$

and by multiplying $\rho(f^i, f^j)$ by MI, to yield *Asymmetric Mutual Information* (AMI, see Table 1).

4.3 Kernel-Based DCFs

In this section we consider different kernel functions as DCFs. Kernel functions are similarity functions, typically used e.g., within support vector machines to operate in high- (and potentially infinite-) dimensional spaces. In this case the values in the context vector can be numeric, thus indicating the relative importance of a term in a given context, and are usually computed as a function of the frequency of the term in the context and in the corpus, e.g., $tfidf$. We consider normalized context vectors, i.e., after weighting the document-by-term matrix we normalize the term vectors to unit length.

The most popular vector similarity measure is probably cosine similarity, which measures the cosine of the angle between them, and is defined as

$$cos(\mathbf{f}^i, \mathbf{f}^j) = \frac{\langle \mathbf{f}^i, \mathbf{f}^j \rangle}{\|\mathbf{f}^i\| \|\mathbf{f}^j\|} \tag{3}$$

We also consider as DCFs other popular kernels: the polynomial kernel and the Radial Basis Function (RBF – a.k.a. Gaussian) kernel, i.e.,

$$\begin{aligned} polynomial_{a,b}(\mathbf{f}^i, \mathbf{f}^j) &= (a + \langle \mathbf{f}^i, \mathbf{f}^j \rangle)^b & (4) \\ RBF_\gamma(\mathbf{f}^i, \mathbf{f}^j) &= \exp\{-\gamma \|\mathbf{f}^i - \mathbf{f}^j\|^2\} & (5) \end{aligned}$$

where $\|\mathbf{f}^i - \mathbf{f}^j\| = \sqrt{\sum_{c=1}^{n} (\mathbf{f}^i_c - \mathbf{f}^j_c)^2}$ is the Euclidean distance between $\mathbf{f}^i$ and $\mathbf{f}^j$.

Since non-zero values in a frequency vector are always positive, it turns out that the expected value of both the dot product and the Euclidean distance between two random distributional vectors is greater than zero. In order to satisfy the necessary condition imposed to DCFs the kernels should be "centred" to zero by factoring out this bias. Let $\mathbf{r}^i$ and $\mathbf{r}^j$ be two unit-length vectors with prevalences p_i and p_j, whose non-zero values are independently distributed at random; the expected value of non-zero positions is $\sqrt{p_i n}^{-1}$ and $\sqrt{p_j n}^{-1}$, respectively. The expected value of the dot product and of the Euclidean distance between $\mathbf{r}^i$ and $\mathbf{r}^j$ are, respectively,

$$\begin{aligned} E[\langle \mathbf{r}^i, \mathbf{r}^j \rangle] &= p_i p_j n \frac{1}{\sqrt{p_i n}} \frac{1}{\sqrt{p_j n}} = \sqrt{p_i p_j} && (6) \\ E[\lVert \mathbf{r}^i - \mathbf{r}^j \rVert] &= n p_i p_j \left(\frac{1}{\sqrt{p_i n}} - \frac{1}{\sqrt{p_j n}} \right)^2 + n(p_i - p_i p_j) \frac{1}{p_i n} + n(p_j - p_i p_j) \frac{1}{p_j n} && (7) \\ &= 2(1 - \sqrt{p_i p_j}) \end{aligned}$$

The resulting DCFs, obtained by factoring out these expected values from the corresponding kernels, are reported in the second part of Table 1.

Table 1: Mathematical forms of DCFs discussed in this work.

Probability-based DCFs	**Mathematical form**
Linear	$P(f^i \mid f^j) - P(f^i \mid \overline{f^j})$
Pointwise Mutual Information	$\log_2 \frac{P(f^i, f^j)}{P(f^i)P(f^j)}$
Asymmetric Mutual Information	$\rho(f^i, f^j) \sum_{x \in \{f^i, \overline{f^i}\}} \sum_{y \in \{f^j, \overline{f^j}\}} P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)}$
Kernel-based DCFs	**Mathematical form**
Cosine	$\frac{\langle \mathbf{f}^i, \mathbf{f}^j \rangle}{\lVert \mathbf{f}^i \rVert \lVert \mathbf{f}^j \rVert} - \sqrt{p_i p_j}$
Polynomial	$(a + \langle \mathbf{f}^i, \mathbf{f}^j \rangle)^b - (a + \sqrt{p_i p_j})^b$
RBF	$\exp\{-\gamma \lVert \mathbf{f}^i - \mathbf{f}^j \rVert^2\} - \exp\left\{-4\gamma \left(1 - \sqrt{p_i p_j}\right)^2\right\}$

5. Distributional Correspondence Indexing

In this section we explain our Distributional Correspondence Indexing (DCI) method in detail, by paying special attention to each step of the workflow.

The DCI method works by first identifying a small set of pivot terms (or “pivots”, for short – Section 5.1), and then projects each term into a new vector space where each dimension measures the correspondence of the term to a pivot (Section 5.2). Each term will thus be assigned to a new vectorial representation that will be here referred to as *term profile*, or simply *profile*. The term space is then post-processed by first normalizing each dimension (Section 5.3) and then unifying the source and target term profiles for certain terms that we expect to behave similarly in both domains, such as pivots (Section 5.4). Documents are then projected by cumulating (i.e., summing) the profiles of the terms that occur in them (Section 5.5), after which a classifier can be trained as usual. In order to clarify the explanation we will use a running example throughout the different steps. To that aim we will be tracking the case Books → Electronics, i.e., the domain adaptation from the Books domain to the Electronics one (more details about the datasets will be given in Section 6.1).

5.1 Pivot Selection

Pivots are terms that are shared across the source and target domains, and are meant to link them, thus enabling the knowledge transfer process. Blitzer et al. (2006) defined pivots as terms which occur frequently in the source and target domains and behave similarly in both domains. Subsequent work (Blitzer et al., 2007; Prettenhofer & Stein, 2010; Pan et al., 2010) extended the definition of pivots to take into account also the co-occurrence relation between terms and class labels in the source domain, as it was shown that informative terms for the prediction task are better pivots. This idea was later adapted to the cross-lingual setting by Prettenhofer and Stein (2010), in which a fixed frequency threshold ϕ, called *support*, was used to filter infrequent pivot candidates, then ranking the remaining candidates by their mutual information with respect to labels in the source domain. Prettenhofer and Stein also introduced the notion of *word translation oracle* (WTO), i.e., a translator that, given a word in the source language, provides its translation in the target language. The method of Prettenhofer and Stein assumes the possibility of issuing a limited number of calls to such an oracle.

As pointed out by Pan et al. (2010), pivot selection using mutual information can help identify predictive terms for the source domain, but there is no guarantee that those terms act similarly in both domains. In this respect, we think that even though the support threshold might serve to filter out some problematic candidates, this strategy is suboptimal. For example, a candidate occurring 29 times and 31 times out of 50,000 in the source and in the target domain, respectively, will be discarded if the support is set to $\phi = 30$, while a pivot occurring 5000 times and 31 times could be chosen, even if it is clear that in this second case its role in the two domains is very different.

A good pivot should be highly *task-dependent*, and also present a similar degree of *domain-dependence* in the two domains. We have formalized this intuition via the function

$$\Psi(f^i) = I_s(f^i)\zeta_{s-t}(f^i) \tag{8}$$

where $\Psi(f^i)$ is the term's strength as a pivot, $I_s(f^i)$ quantifies the informativeness (to be estimated on the training set Tr_s) of term f^i for the classification task, and $\zeta_{s-t}(f^i)$ measures the *cross-consistency* of term f^i estimated on U_s and U_t, that quantifies how similarly the term behaves in the two domains.

Following previous research, we will instantiate $I_s(f^i)$ via mutual information. Ideally, $\zeta_{s-t} : F \rightarrow [0, 1]$ should be defined in such a way that $\zeta_{s-t}(f^i) \approx 1$ if the distribution of f^i is consistent across domains, and $\zeta_{s-t}(f^i) \approx 0$ if the importance of f^i varies a lot from one domain to another. Given the lack of labelling information in the target domain, we adopt a simple heuristic that relates the prevalence on both domains, as we might expect a comparable prevalence of use in text for terms that posses a similar degree of domain-dependence[1]. We thus define

$$\zeta_{s-t}(f^i) = \frac{min\{p_i^s, p_i^t\}}{max\{p_i^s, p_i^t\}} \tag{9}$$

where p_f^s (resp., p_i^t) is the prevalence of f^i as measured on set U_s (resp., U_t). The top-ranked m terms according to their Ψ value that have a frequency greater than ϕ in both U_s and U_t, are selected as pivots; m is a user defined parameter indicating the number of pivots to select. Table 2 exemplifies the pivot selection process in the Books → Electronics adaptation. For instance, the cross-consistency weight succeeds in penalizing adjective boring, which might be a good candidate for predicting the polarity of book reviews but is rather uninformative for electronic devices, which is thus pushed out of the top-100 list (only 10 elements are shown here due to space limitations).

Table 2: Top-10 terms ranked according to mutual information ($I_s(f^i)$) (left), and mutual information combined with cross-consistency ($\Psi(f^i) = I_s(f^i)\zeta_{s-t}(f^i)$) (right).

#	term	I_s score	term	Ψ score
1	waste	0.029	waste	0.028
2	boring	0.029	excellent	0.022
3	disappointing	0.029	bad	0.020
4	excellent	0.026	not	0.018
5	no	0.021	don't	0.017
6	waste_of	0.021	waste_of	0.017
7	bad	0.021	highly_recommend	0.014
8	don't	0.019	disappointing	0.013
9	not	0.018	great	0.012
10	your_money	0.018	your_money	0.012

5.2 Term Profiles

We implement a rather direct application of the distributional hypothesis, following the intuition that the semantics of a term can be captured by its distributional correspondence with pivots. We thus build a term profile $\vec{f}$ for each source and target term f (including pivot terms) as the m-dimensional vector

$$\vec{f} = (\eta(\mathbf{f}, \mathbf{p}^1), \eta(\mathbf{f}, \mathbf{p}^2), \ldots, \eta(\mathbf{f}, \mathbf{p}^m)) \tag{10}$$

where $\mathbf{f}$ and $\mathbf{p}^i$ are the context distribution vector from the unlabelled collection of the term f being profiled and the i^{th} pivot, respectively, and η is the selected DCF.

Table 3 displays the term profiles associated with four relevant terms from our running example, i.e., boring, excellent, waste, reliable. Note that excellent and waste are pivots with opposite polarity, while the other two are domain-dependent terms, i.e., boring is only

[1] Note that the $\zeta_{s-t}(\cdot)$ function is meant to capture the cross-domain drift, and thereby we will ignore it in the cross-lingual setting – by simply defining $\zeta_{s-t}(f^i) = 1$ in that case – as the domain of knowledge is the same for both collections.

informative for Books reviews while reliable is only informative for Electronics reviews. The DCF used in this example is the cosine kernel. Note that boring has no representation in the target side, as this term does not appear in the Electronics dataset. Note also that pivot terms are associated to vectorial representations that are somehow close to each other in both domains. The same does not happen for reliable, a domain-dependent term that plays different roles in the two domains. We can also note that correspondence tends to be positive between terms with similar polarity (e.g., between waste and bad), and negative otherwise (e.g., between excellent and bad). Finally, notice that $cos(\text{waste}, \text{waste}) \neq 1$ and $cos(\text{excellent}, \text{excellent}) \neq 1$, due to the correction factor introduced on the cosine formula (see Table 1).

Table 3: Term profiles generated for terms boring, excellent, waste, and reliable (rows) for the different dimensions (columns) in the source (left) and target (right). Only the first 5 dimensions, corresponding to pivots waste, excellent, bad, no, and don't, are shown due to space restrictions.

	Books						Electronics					
	waste	excellent	bad	not	don't	...	waste	excellent	bad	not	don't	...
boring	0.058	-0.042	0.029	-0.029	0.004	...	-	-	-	-	-	...
excellent	-0.030	0.938	-0.051	-0.082	-0.065	...	-0.042	0.927	-0.023	-0.021	-0.018	...
waste	0.957	-0.030	0.007	-0.013	0.161	...	0.959	-0.042	0.028	0.025	0.138	...
reliable	-0.012	-0.002	-0.007	0.004	0.016	...	0.001	0.014	0.005	-0.004	-0.008	...

5.3 Normalization of Term Profiles

Each dimension of the space reflects the distributional correspondence to a given pivot. Pivots with high prevalence are likely to generate high DCF values, which could lead to dominant dimensions in the profile vectors; this could be detrimental during the learning phase. To avoid this effect, we center each profile dimension on its expected value and then rescale by the standard deviation (Equation 11), so that the values for all profile dimensions are approximately normally distributed in $\mathcal{N}(0,1)$, i.e.,

$$\vec{f'} = \left(\frac{\vec{f_1} - \mu_1}{\sigma_1}, \frac{\vec{f_2} - \mu_2}{\sigma_2}, \ldots, \frac{\vec{f_m} - \mu_m}{\sigma_m} \right) \tag{11}$$

where $\vec{f_i}$ is the i^{th} dimension of the term profile $\vec{f}$ (see Equation 10), and μ_i and σ_i are the mean and the standard deviation for the same i^{th} dimension, respectively. After normalization, term profile vectors are rescaled to unit length.

Table 4 demonstrates the effect of term normalization for the example terms discussed in Table 3. Note that, after normalization, the source and target profiles for pivot terms seem to get closer in the vectorial space. Furthermore, the target representation of reliable turns out to be more consistent with our intuitions, as it reflects a negative correspondence to waste, and a stronger positive correspondence to excellent.

Table 4: Effect of term normalization for terms boring, excellent, waste, and reliable.

	Books						Electronics					
	waste	excellent	bad	not	don't	...	waste	excellent	bad	not	don't	...
boring	0.223	-0.163	0.105	-0.125	0.020	...	-	-	-	-	-	...
excellent	-0.021	0.861	-0.038	-0.080	-0.050	...	-0.035	0.909	-0.024	-0.038	-0.023	...
waste	0.555	-0.017	0.005	-0.008	0.094	...	0.572	-0.032	0.016	0.018	0.087	...
reliable	-0.091	0.020	-0.056	0.076	0.169	...	-0.011	0.119	0.005	-0.146	-0.138	...

5.4 Unification of Term Profiles

As we assume pivot terms behave similarly in the two languages, we *unify* their term profiles by simply averaging the source profile and the target profile and then normalizing the result to unit length. Unification is also applied to profiles of terms that appear in both the source and target domains with a frequency greater than the support ϕ. The rationale behind unification is to correct the possible misalignment between the source and target term profiles for terms that should receive the same vectorial representation in both domains, such as pivot terms or proper nouns. This is done in order to equalize across domains the contribution of the term to the document representation (see below).

Table 5 shows the term profiles of our running example after unification. Note that boring does not experience any change as its target counterpart does not even exist. Term reliable is also not affected by normalization, as its frequency in the Books domain does not exceed the support ϕ, set to 30 for this example. Finally, the term profiles of pivots excellent and waste are unified, i.e., are computed as the average of their respective source and target profile representations, and then normalized to unit norm.

Table 5: Term profile unification for terms boring, excellent, waste, and reliable.

	Books						Electronics					
	waste	excellent	bad	not	don't	...	waste	excellent	bad	not	don't	...
boring	0.223	-0.163	0.105	-0.125	0.020	...	-	-	-	-	-	...
excellent	-0.028	0.893	-0.031	-0.059	-0.037	...	-0.028	0.893	-0.031	-0.059	-0.037	...
waste	0.570	-0.025	0.010	0.005	0.091	...	0.570	-0.025	0.010	0.005	0.091	...
reliable	-0.091	0.020	-0.056	0.076	0.169	...	-0.011	0.119	0.005	-0.146	-0.138	...

5.5 Document Indexing

Finally, train and test documents are indexed in the profile space via a weighted sum of all profile vectors associated to their terms. That is, document d_j is represented as the m-dimensional vector

$$\vec{d_j} = \sum_{f_i \in d_j} w_{ij} \cdot \vec{f_i'} \tag{12}$$

where w_{ij} is the weight of term f_i in document d_j according to any weighting function (in our experiments we used the standard cosine-normalized $tfidf$), and $\vec{f_i'}$ is the normalized and unified term profile vector for f_i.

6. Experiments

In this section we experimentally compare our DCI method, implemented using different DCFs, to other state-of-the-art methods proposed in the literature.

6.1 Datasets

We test our method on two popular, publicly available sentiment datasets: Multi-Domain Sentiment Dataset (version 2.0) and Webis-CLS-10. The former is a dataset frequently used for evaluating cross-domain adaptation, while the latter has often been used for evaluating cross-lingual methods. We will also use Webis-CLS-10 to explore the cross-domain/cross-lingual setting.

6.1.1 Multi-domain Sentiment Dataset (version 2.0)

The Multi-Domain Sentiment (MDS) dataset, first proposed by Blitzer et al. (2007), contains English product reviews taken from Amazon.com for the four domains Books (B), DVDs (D), Electronics (E), and Kitchen (K) appliances. In order to facilitate reproducibility and to allow for a fair comparison with the results reported in previous literature, we used the same pre-processed version of the dataset as used in previous evaluations, made publicly available by Blitzer et al. (see MSD dataset, 2007) . In this pre-processed version, terms were extracted by taking unigrams and bigrams; reviews originally rated higher than "3 stars" were labelled as positive, and those rated lower than "3 stars" as negative; reviews with intermediate ratings were removed. The dataset comprises 1000 positive reviews and 1000 negative reviews for each of the four domains, and a set of unlabelled documents ranging from 3,586 to 5,945 documents for each domain. Table 6 shows the number of labelled and unlabelled documents, number of distinct terms and total number of terms for each dataset. According to the same evaluation procedure followed by the proposers of other methods we compare against, we randomly split each labelled dataset into a training set of 1600 instances and a test set of 400 instances.

Table 6: Main characteristics of the Multi-Domain Sentiment dataset (version 2.0).

Domain	Labelled	Unlabelled	Terms	Occurrences
Books	2,000	4,465	195,887	445,793
DVDs	2,000	3,586	188,778	370,844
Electronics	2,000	5,681	111,407	392,699
Kitchen	2,000	5,945	93,474	351,162

6.1.2 Webis-CLS-10

Webis-CLS-10, first proposed by Prettenhofer and Stein (2010), is a cross-lingual sentiment collection consisting of Amazon product reviews written in four languages (English (E), German (G), French (F), and Japanese (J)), covering three product domains (Books (B), DVDs (D), and Music (M)). For each language-domain pair there are 2,000 training documents, 2,000 test documents, and from 9,000 to 50,000 unlabelled documents depending on the language-domain combination (see Table 7 for further details). We used the preprocessed version of the dataset made publicly available by the authors (see Webis-CLS dataset, 2010), where terms correspond to uni-grams. Following the work by Prettenhofer and Stein, we consider English as the source language, since it is by far the most realistic scenario. Documents are either labelled as positive or negative, following the same procedure of Blitzer et al. (2007). Positive and negative examples are balanced in all sets (see Table 7 for details). Each labelled dataset was split into a perfectly balanced training set of 2,000 instances and a test set of 2,000 instances. This split was proposed by Prettenhofer and Stein; all baseline methods we here compare against use exactly the same corpus both for training and test.

Table 7: Main characteristics of the Webis-CLS-10 dataset.

Domain	Labelled	Unlabelled	Terms	Occurrences
EB	4,000	50,000	62,499	6,289,014
ED	4,000	30,000	50,124	4,001,678
EM	4,000	25,220	38,632	2,664,955
GB	4,000	50,000	105,360	6,618,037
GD	4,000	50,000	100,265	6,303,371
GM	4,000	50,000	95,952	5,688,874
FB	4,000	32,870	52,664	2,427,178
FD	4,000	9,358	26,117	714,105
FM	4,000	15,940	39,001	1,371,800
JB	4,000	50,000	51,179	7,637,325
JD	4,000	50,000	53,318	7,263,796
JM	4,000	50,000	53,078	6,284,653

6.2 Evaluation Metrics

Following the practice common in the related literature, we adopt standard accuracy as the evaluation measure. Accuracy measures the proportion of correctly classified documents over the total number of outcomes (Equation 13), i.e.,

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{13}$$

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

6.3 Baseline Methods

We experimentally compare DCI, using different DCFs, to different baseline methods proposed in the literature for cross-domain and cross-lingual domain adaptation. We limit the comparison to algorithms evaluated on the same corpora, and we report results taken from the original papers. When not explicitly mentioned, results for all baseline algorithms were obtained by using the same datasets, though obviously using different random splits. For the MDS dataset, following common practice, we run experiments on multiple random splits and average them. For the Webis-CLS-10 dataset we instead used the split proposed by Prettenhofer and Stein (2010) (more details below).

As an upper bound, we implemented a method (hereafter called "Upper") that trains the SVM classifier on the training set of the target domain. As a lower bound we instead implemented a method that trains the SVM classifier on the source domain and then applies the trained classifier directly in the target domain, i.e., without carrying out any sort of knowledge transfer ("NoTrans"). When considering various languages, we also report the machine translation baseline ("MT"), which first translates all target documents into the source language (i.e., English in our experiments) before giving them as input to the SVM classifier trained on the source domain; we used the pre-translated documents provided by Prettenhofer and Stein (2011).

For cross-domain adaptation the baselines we consider are Structural Correspondence Learning using Mutual Information to select pivots (SCL-MI – Blitzer et al., 2007), Spectral Feature Alignment (SFA – Pan et al., 2010), multiple sources Sentiment Sensitive Thesaurus (SST – Bollegala et al., 2011), and Stacked Denoising Autoencoder (Glorot et al., 2011) trained on domain pairs (SDA) and trained on 22 domains available in the former version of the MDS dataset (a method that the authors abbreviate as SDA$_{sh}$).

For cross-lingual adaptation the baselines we consider are cross-lingual Latent Semantic Indexing (LSI – Dumais et al., 1997), cross-lingual Kernel Canonical Correlation Analysis (KCCA – Vinokourov et al., 2002), Oriented Principal Component Analysis (OPCA – Platt et al., 2010), Two-Step Learning method (TSL – Xiao & Guo, 2013), Semi-Supervised Matrix Completion (SSMC – Xiao & Guo, 2014), and the cross-lingual version of Structural Correspondence Learning (SCL – Prettenhofer & Stein, 2011).

Since there are no published results to compare with for the cross-lingual/cross-domain adaptation, as a baseline we consider SCL-MI (Prettenhofer & Stein, 2011) by reusing the publicly available source code (Natural Language Understanding Toolkit, 2011) for running our own experiments.

6.4 Implementation Details and Parameter Setting

We implemented our method (see DCI-source, 2015) as part of the JaTeCS (2015) framework. We used the popular SVM^{light} (2008) implementation of Support Vector Machines as the learning device, with default parameters, for DCI and for baselines NoTrans, Upper, and MT.

In our experiments we set the support (see Section 5.1) to $\phi = 30$, following the indications of Prettenhofer and Stein (2010) for the Webis-CLS-10 dataset. Since the amount of unlabelled documents in the MDS Dataset is one order of magnitude smaller, in that case we set $\phi = 1$.

To emulate the word oracle – and for the sake of a fair comparison – we reused the bilingual dictionary[2] created for evaluating cross-lingual SCL by Prettenhofer and Stein (2010). This dictionary emulates a context-unaware word-translation oracle, i.e., each source word

[2] Note that we used a different pivot selection criterion, as detailed in Section 5.1, and therefore the oracle could be queried to translate some words that where never considered in the cited work, and thus might

is mapped into its "most likely" translation; potential problems arising from the ambiguity of single words are simply disregarded.

One important factor to take into consideration is the number of calls issued to the oracle; the oracle simulates a human translator, and this number is thus an indicator of the human effort required to perform domain adaptation. We limited the number of translations to the top $2m$ terms with the highest mutual information, where m is the number of pivots words. In order to perform comparisons with other methods, we fixed the number of pivots to $m = 100$, which corresponds to the minimal setup tested by Prettenhofer and Stein (2010). In Section 6.6 we then explore the impact in accuracy due to variations in the value of m.

Parameters for the polynomial and RBF kernel DCFs were optimized via a grid search on the Books → DVDs cross-domain adaptation for the MDS dataset, and on EnglishBooks → GermanBooks cross-lingual adaptation for Webis-CLS-10 dataset (as was done in previous research, Prettenhofer & Stein, 2010). The actual values we ended up in using were $b = 0.5$ and $\gamma = 0.82$ in MDS, and $b = 0.8$ and $\gamma = 0.88$ in Webis-CLS-10. We set $a = 0$ for the (homogeneous) polynomial DCF in both cases, as we did not perceive any consistent improvement that justifies a more complex grid search exploration on the two parameters.

We used the normalized $tfidf$ weighting criterion to represent the co-occurrence matrices in all experiments.

6.5 Experimental Results

In this section we present the results of our experiments using different DCFs. Experiments are presented by different domain adaptation setups, including cross-domain adaptation (Section 6.5.1), cross-lingual adaptation (Section 6.5.2), and cross-domain/cross-lingual adaptation (Section 6.5.3). Additional related experiments are then conducted in Section 6.6.

For the sake of brevity and consistently with the notation $L_sC_s \rightarrow L_tC_t$ introduced earlier, we will use a single upper-case character (as defined in Section 6.1) to denote the languages and the domains involved in the experimental setup. For example, EB → GD denotes an experiment in which EnglishBooks is used as the source and GermanDVDs is used as the target.

6.5.1 Cross-Domain Results

Table 8 reports results obtained on the MDS dataset, including (a) performance averages by product category, i.e., computed by averaging all the results obtained when the product category is considered as the target domain, and (b) global averages. All results reported correspond to the accuracy of the different methods[3] computed via 5-fold cross-validation, i.e., using 1,600 training documents and 400 test documents in each run. A direct comparison with many other methods (He et al., 2011; Xia & Zong, 2011; Denecke, 2009; Ponomareva & Thelwall, 2012) would also be feasible in principle. We omitted direct com-

not be present in the dictionary. Such cases happened rarely however, and we preferred to simply skip these candidates rather than completing the bilingual dictionary, in order to guarantee a fair comparison.

[3] Missing results are ones which were not reported in the original papers.

parisons with these methods since previous research has shown them to be comparable, but not superior, to SFA.

Table 8: Cross-domain adaptation on the MDS dataset.

Task	NoTrans	Upper	SCL-MI	SFA	SST	SDA	SDA$_{sh}$	Linear	PMI	AMI	Cos	Poly	RBF
ED → EB	0.728	0.844	0.797	0.775	-	0.724	0.768	0.825	0.827	0.811	0.824	**0.830**	0.825
EE → EB	0.707	0.844	0.754	0.757	-	0.768	**0.780**	0.766	0.763	0.753	0.764	0.776	0.765
EK → EB	0.709	0.844	0.686	0.748	-	0.807	**0.837**	0.783	0.783	0.769	0.790	0.791	0.784
EB → ED	0.772	0.847	0.758	0.814	-	0.804	**0.855**	0.808	0.811	0.806	0.817	0.829	0.815
EE → ED	0.706	0.847	0.762	0.772	-	0.902	**0.905**	0.768	0.779	0.765	0.774	0.799	0.771
EK → ED	0.727	0.847	0.769	0.766	-	0.835	**0.854**	0.788	0.789	0.781	0.799	0.807	0.798
EB → EE	0.708	0.869	0.759	0.725	-	0.806	0.824	0.810	0.822	0.793	0.822	**0.826**	0.821
ED → EE	0.730	0.869	0.741	0.767	-	0.872	**0.875**	0.822	0.832	0.812	0.824	0.833	0.826
EK → EE	0.827	0.869	**0.868**	0.851	-	0.802	0.820	0.855	0.851	0.843	0.858	0.863	0.857
EB → EK	0.745	0.902	0.789	0.788	-	0.844	**0.846**	0.834	0.839	0.822	0.835	0.844	0.835
ED → EK	0.740	0.902	0.814	0.808	-	0.803	0.821	0.858	0.856	0.846	**0.864**	0.861	0.863
EE → EK	0.840	0.902	0.859	0.868	-	0.777	0.811	0.864	0.864	0.851	0.868	**0.874**	0.867
Books	0.715	0.844	0.746	0.760	0.763	0.766	0.795	0.791	0.791	0.778	0.793	**0.799**	0.791
DVDs	0.735	0.847	0.763	0.784	0.788	0.847	**0.871**	0.788	0.793	0.784	0.797	0.811	0.795
Electronics	0.755	0.869	0.789	0.781	0.836	0.827	0.840	0.829	0.835	0.816	0.835	**0.841**	0.835
Kitchen	0.775	0.902	0.821	0.821	0.852	0.808	0.826	0.852	0.853	0.840	0.856	**0.860**	0.855
Average	0.745	0.866	0.780	0.786	0.810	0.812	**0.833**	0.815	0.818	0.804	0.820	0.828	0.819

Most configurations of DCI outperform all compared methods, with the exception of SDA$_{sh}$; only the MI-based DCF performed slightly worse than SST on average. It should be noted, however, that both SST and SDA$_{sh}$ use a different problem setting, since they train on multiple domains. More concretely, SST trained on three out of four domains from the MDS dataset to create the sentiment thesaurus, while SDA$_{sh}$ exploited 22 domains included in the previous version of the MDS dataset to train the auto-encoder. Furthermore, SDA and SDA$_{sh}$ rely on a deep learning approach, a paradigm that requires significant more computational power and many parameters to be tuned. Notwithstanding this, DCI with the polynomial kernel as the DCF (with only two parameters) obtained three best averaged results (i.e., Books, Electronics, and Kitchen) out of five, not leveraging any additional domain, and requiring low computational cost (as will be later discussed in Section 6.7).

Table 9 reports experiments on cross-domain adaptation using the Webis-CLS-10 dataset. These results are consistent with our previous observations, i.e., the polynomial and cosine-based DCFs are the best performers, followed by the PMI and RBF functions. In this case the best results obtained by DCI are very close to "Upper", and surprisingly surpass it in ED → EB and EM → EB. We conjecture that this improvement may be due to the larger size of the unlabelled sets, that are one order of magnitude greater with respect to the MDS dataset, thus allowing for more robust evaluations of the cross-domain consistency function $\zeta_{s-t}(\cdot)$ and of the DCF.

6.5.2 Cross-Lingual Results

Table 10 reports our results on the Webis-CLS-10 dataset for cross-lingual adaptation. As discussed earlier, the source language is always English, while the target languages include German, French, and Japanese.

DCI outperformed the MT baseline on average in all cases but the PMI DCF in the German case. All kernel-based DCFs outperformed all the compared methods in terms of

Table 9: Cross-domain performance on the Webis-CLS-10 dataset.

Task	NoTrans	Upper	SCL-MI	Linear	PMI	AMI	Cos	Poly	RBF
ED → EB	0.803	0.829	0.839	0.840	0.843	0.831	0.851	**0.855**	0.848
EM → EB	0.783	0.829	0.823	0.828	0.838	0.826	0.840	**0.841**	0.838
EB → ED	0.798	0.831	0.810	0.798	0.812	0.788	**0.818**	**0.818**	0.806
EM → ED	0.778	0.831	0.797	0.802	**0.821**	0.798	0.821	**0.822**	0.816
EB → EM	0.786	0.845	0.804	0.825	0.835	0.816	**0.838**	0.836	0.831
ED → EM	0.804	0.845	0.823	0.831	**0.833**	0.815	0.829	0.832	0.827
Books	0.793	0.829	0.831	0.834	0.841	0.829	0.846	**0.848**	0.843
DVDs	0.788	0.831	0.804	0.800	0.817	0.793	0.819	**0.820**	0.811
Music	0.795	0.845	0.814	0.828	**0.834**	0.816	**0.834**	**0.834**	0.829
Average	0.792	0.835	0.816	0.821	0.830	0.812	0.833	**0.834**	0.828

Table 10: Cross-lingual performance on the Webis-CLS-10 dataset.

Task	Upper	MT	SCL-MI	LSI	KCCA	OPCA	TSL	SSMC	Linear	PMI	AMI	Cos	Poly	RBF
EB → GB	0.868	0.808	0.833	0.776	0.791	0.747	0.792	0.819	0.798	0.714	0.797	0.827	**0.837**	0.829
ED → GD	0.835	0.800	0.809	0.796	0.776	0.766	0.819	0.823	0.826	0.819	0.800	0.822	**0.833**	0.788
EM → GM	0.859	0.791	0.829	0.727	0.695	0.714	0.726	0.813	0.844	0.850	0.837	**0.856**	0.844	0.801
EB → FB	0.862	0.821	0.813	0.792	0.767	0.746	0.813	0.831	0.746	0.761	0.768	0.842	0.819	**0.844**
ED → FD	0.872	0.795	0.804	0.778	0.782	0.705	0.820	0.827	0.823	0.823	0.801	0.827	0.806	**0.846**
EM → FM	0.890	0.765	0.781	0.726	0.748	0.718	0.766	0.805	0.816	0.827	0.818	**0.844**	0.840	0.803
EB → JB	0.812	0.692	0.770	0.738	0.792	0.745	0.794	0.738	0.779	0.731	0.711	0.758	0.754	**0.782**
ED → JD	0.834	0.722	0.764	0.754	0.782	0.737	0.793	0.776	**0.822**	0.768	0.797	0.801	0.795	0.761
EM → JM	0.842	0.714	0.773	0.734	0.735	0.750	0.762	0.775	0.826	0.816	0.807	**0.839**	0.832	0.826
German	0.854	0.800	0.824	0.766	0.754	0.742	0.779	0.818	0.823	0.794	0.811	0.835	**0.838**	0.806
French	0.875	0.794	0.799	0.765	0.766	0.723	0.800	0.821	0.795	0.804	0.796	**0.838**	0.822	0.831
Japanese	0.829	0.709	0.769	0.742	0.770	0.744	0.783	0.763	**0.809**	0.772	0.772	0.799	0.794	0.790
Average	0.852	0.767	0.797	0.758	0.763	0.736	0.787	0.801	0.809	0.790	0.793	**0.824**	0.818	0.809

average accuracy, and the best result was obtained by one of the kernel-based DCFs in 11 out of 12 cases. The best performing DCFs are again the cosine and polynomial DCFs.

6.5.3 Cross-Domain/Cross-Lingual Results

Table 11 reports our experiments in the cross-domain/cross-lingual setting.

This setting is arguably the most difficult one, since both the term space and the marginal probabilities of the domains differ, as reflected in the noticeable degradation of MT results. Notwithstanding this, consistent observations could be derived from these results. The cosine and polynomial DCFs confirm their superiority with respect to all other compared methods. The best result was obtained by DCI in 17 out of 18 cases; in half of the cases the cosine DCF obtained the best result.

6.5.4 Statistical Significance Tests

Statistical significance tests (paired t-test on the accuracy values from Table 8) indicate that all our DCI configurations, with the exception of MI, are significantly better with $p < 0.01$ than SCL, SCL-MI, and SFA in the MDS dataset. For the polynomial DCF, which obtained 10 best results out of 12, higher-confidence levels were obtained, i.e., $p < 0.001$. A t-test on all runs on Webis-CLS-10 reveals that all kernel-based DCFs and the Linear DCF are

Table 11: Cross-domain/cross-lingual accuracy on the Webis-CLS-10 dataset.

Task	Upper	MT	SCL-MI	Linear	PMI	AMI	Cos	Poly	RBF
ED → GB	0.868	0.789	0.823	0.823	0.764	0.811	**0.824**	0.818	**0.824**
EM → GB	0.868	0.751	**0.825**	0.791	0.821	0.705	0.812	0.791	0.800
EB → GD	0.835	0.774	0.784	0.790	0.796	0.788	**0.827**	0.825	0.783
EM → GD	0.835	0.773	0.792	0.778	0.829	0.772	**0.834**	0.814	0.808
EB → GM	0.859	0.768	0.811	0.786	0.812	0.793	**0.843**	0.833	0.807
ED → GM	0.859	0.768	0.824	**0.844**	**0.844**	0.828	0.816	0.835	0.832
ED → FB	0.862	0.788	0.790	0.744	0.798	0.747	0.848	0.846	**0.852**
EM → FB	0.862	0.765	0.784	0.810	0.833	0.785	**0.845**	0.843	0.789
EB → FD	0.872	0.783	0.780	0.810	0.816	0.788	0.823	0.793	**0.841**
EM → FD	0.872	0.780	0.745	0.798	0.822	0.761	**0.841**	0.829	0.775
EB → FM	0.889	0.771	0.762	0.822	0.753	0.794	**0.833**	0.824	0.829
ED → FM	0.889	0.745	0.757	0.836	0.826	0.827	0.847	0.849	**0.855**
ED → JB	0.812	0.700	0.725	0.738	0.675	0.715	**0.761**	0.741	0.741
EM → JB	0.812	0.642	0.708	0.711	0.621	0.636	0.721	0.689	0.**722**
EB → JD	0.834	0.708	0.742	**0.813**	0.663	0.710	0.805	0.789	0.782
EM → JD	0.834	0.693	0.756	0.792	**0.828**	0.721	0.790	0.763	0.711
EB → JM	0.842	0.673	0.742	0.826	0.699	0.811	**0.831**	0.826	0.827
ED → JM	0.842	0.710	0.776	**0.817**	0.804	0.762	0.816	**0.817**	0.804
German	0.854	0.771	0.810	0.802	0.811	0.783	**0.826**	0.819	0.809
French	0.874	0.772	0.770	0.803	0.808	0.784	**0.840**	0.831	0.824
Japanese	0.829	0.688	0.742	0.783	0.715	0.726	**0.787**	0.771	0.765
Books	0.847	0.739	0.776	0.770	0.752	0.733	**0.802**	0.788	0.788
DVDs	0.847	0.752	0.767	0.797	0.792	0.757	**0.820**	0.802	0.783
Music	0.863	0.739	0.779	0.822	0.790	0.803	**0.831**	**0.831**	0.826
Average	0.852	0.743	0.774	0.796	0.778	0.768	**0.818**	0.807	0.799

better than SCL-MI with statistical significance at confidence level $p < 0.01$; the Cosine DCF obtained $p = 0.854 \cdot 10^{-7}$ and the Polynomial DCF $p = 0.522 \cdot 10^{-5}$.

6.6 Further Experiments

This section presents further experiments aimed at testing the influence of different parameters and modules of DCI, and its performance in the standard text classification setting.

Regarding the effect of parameters, we show trend plots for some representative cases, considering the Linear function as a representative example of a probabilistic-based DCF and the Cosine function as an example of a kernel-based DCF. Each plot involves three settings, one for each scenario: cross-domain adaptation, cross-lingual adaptation, and cross-domain/cross-lingual adaptation. For the sake of brevity we selected some illustrative examples, omitting experiments showing similar behaviour. First, we investigated the sensitivity to the value of parameter m, which indicates the number of pivots to select. Figure 1 shows how performance varies at the variation of m, from 5 to 500 pivots.

The overall tendency displayed in the plots is that performance tends to stabilise as m increases. Adaptations involving the cross-lingual setting seem to be more strongly affected

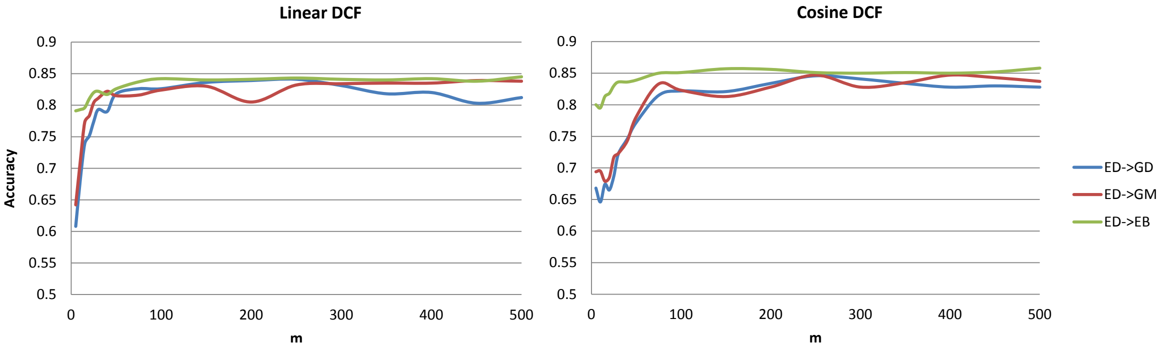

Figure 1: Variation of accuracy at the variation of the number of pivots for different setups.

by the number of pivots. We attribute this effect to the more limited capability of a pivot to reflect a term correspondence through imprecision introduced by the context-unaware single-word translations of the oracle. Such a negative effect seems however to reduce as the number of pivots increases. The method scales well on the number of pivots in terms of efficiency (see below), which might be an indication that simply increasing the pivot set size could be a feasible alternative rather than moving to more complicated definitions of the cross-lingual pivots in order to cover translation nuances. Larger fluctuations could be observed for $m < 75$, but also surprising peaks of performance for extremely small values of m. For example, DCI obtained 81.1% accuracy with the Linear DCF in the ED → GM adaptation with only 30 pivots, while other baselines obtained 76.8% (MT) or 82.4% (SCL-MI, which uses 450 pivots). A similar experiment was reported by Prettenhofer and Stein (2010) for CL-SCL, by varying parameter m in the range $[100, 800]$. A direct comparison shows that our method achieves better accuracy for smaller values of m. Given that the number of calls to the oracle is directly related to m (i.e., m calls in the cross-lingual case, and $2m$ in cross-domain/cross-lingual case due to the cross-consistency reweighting, see Section 5.1), it follows that DCI requires less human effort in creating the bilingual pivots.

The unlabelled collection plays a key role in the domain adaptation, as it is responsible for the term-distribution representation; we thus can expect better estimations of distributions for larger collections. We have investigated how the unlabelled set size affects the performance of the method. We plot the accuracy score obtained for different reduction ratios – preserving balance – in Figure 2.

As expected, the observed trend shows that accuracy is high for large unlabelled collections, and performance tends to stabilize with the addition of unlabelled examples. Better performance is observed in the cross-domain experiments, even with smaller distributional representations.

We also validated empirically each of the different elements that constitute the DCI method, including the cross-distortion factor in pivot selection, the dimensionality standardization, and the unification process. For reasons of conciseness we only report the global improvement for the linear DCF averaged in each dataset, since consistent variations are observable for other DCFs. We found a consistent improvement of 0.47% ± 1.02 in

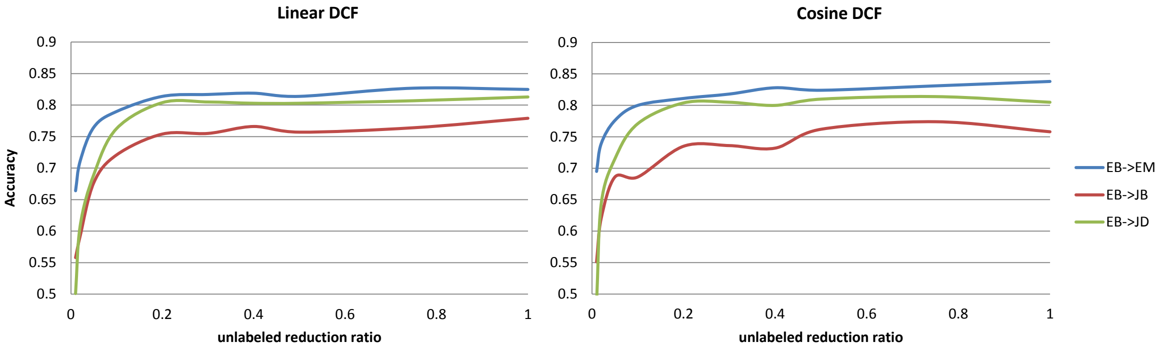

Figure 2: Variation of accuracy at the variation of the unlabelled corpus size for different setups.

accuracy due to cross-consistency in pivot selection, 1.785% ± 3.63 due to dimensionality standardization, and 1.261% ± 2.18 due to unification.

Our experiments reveal that classification performance seems to benefit when the adaptation involves semantically close domains, as is the case of Books→DVDs in the MDS dataset, or English→German in Webis-CLS-10. Analogously, the performance seems to degrade when the source and target domains are dissimilar, as for example Kitchen→Books in MDS or English→Japanese in Webis-CLS-10. It has been noticed in the literature that reducing the distance between the representations of source and target domains is crucial in order to allow better knowledge transfer. Given that the probability distributions are unknown, their distance is sometimes computed through an approximation (the *proxy* $\mathcal{A}$-*distance* Ben-David, Blitzer, Crammer, & Pereira, 2006) that considers the source and target documents as two samples drawn from each distribution. The proxy $\mathcal{A}$-distance is computed as $\hat{d}_{\mathcal{A}} = 2(1-2\epsilon)$, where ϵ is the error produced by an SVM trained to discriminate between the source and target domains.

Figure 3 graphically compares, for the MDS dataset, the proxy $\mathcal{A}$-distances between domains (i) for the raw representations, and (ii) for the DCI representations. The $\hat{d}_{\mathcal{A}}$ is clearly reduced in the cross-domain space generated by DCI, which contributes to explain the improvement in performance with respect to the baseline "NoTrans" on the raw representation. Such reduction is even more noticeable for semantically close domains such as Electronics→Kitchen and Books→DVDs. Hence, DCI projects both domains into a common vector space where the source and target distributions get effectively closer to each other, thus facilitating the transfer of knowledge between them.

Finally, Table 12 reports performance accuracy in the text classification setting, that is, assuming the test data follows the same marginal distribution and is represented in the same term space as the training data. In this case, we consider as baselines (a) the well-known BoW representation with $tfidf$ weighting, and (b) the SCL-MI.

Even though the amount of experiments for the text classification case is too small to allow any substantial claim, it is surprising that, in some runs, DCI with only 100 dimensions yielded better results than traditional BoW representation considering all terms. This is a topic we will investigate in future research.

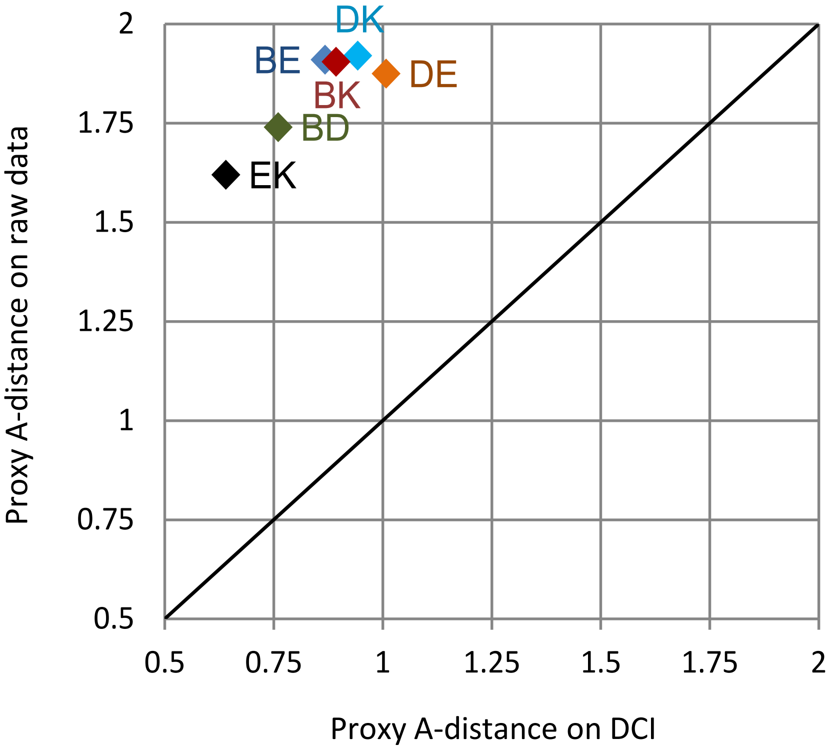

Figure 3: Proxy $\mathcal{A}$-distances between domains of the MDS dataset. The vertical axis displays $\hat{d}_{\mathcal{A}}$ in the raw data (NoTrans), while the horizontal axis displays $\hat{d}_{\mathcal{A}}$ in the vector space produced by DCI using cosine as the DCF. The abscissa coordinate for each point (e.g., **BK**) is the averaged $\hat{d}_{\mathcal{A}}$ produced by domain adaptation in both directions (e.g., **EB** → **EK** and **EK** → **EB**).

Table 12: Text Classification performance on the Webis-CLS-10 dataset.

Task	BoW	SCL-MI	Linear	PMI	AMI	Cos	Poly	RBF
EB → **EB**	0.829	0.828	0.848	0.855	0.836	0.854	**0.856**	0.853
ED → **ED**	0.831	0.815	0.819	**0.826**	0.798	0.818	0.821	0.809
EM → **EM**	0.845	0.832	0.838	**0.846**	0.825	0.841	0.844	0.835
Average	0.835	0.825	0.835	**0.842**	0.820	0.838	0.840	0.832

6.7 Efficiency

The computational cost of DCI is asymptotically bound by the cost of projecting f terms from the two domains into an m-dimensional space, that could be roughly estimated as $O(fmc)$, where the c component is due to the cost of comparing two term distribution models, that depends on the average prevalence c of terms in the unlabelled corpus –

typically much smaller than the effective number of unlabelled documents as a result of to sparsity.

Note that all probabilistic functions discussed in Section 4 can be implemented very efficiently by using sparse data structures. For example, calculating the joint probability $P(v, w)$ is achieved in $O(c)$ steps by intersecting two hash sets with c expected elements. The kernel-based DCFs discussed here depend on the dot product or the Euclidean distance, that can also be computed in $O(c)$ by iterating only over non-zero values. Thus, DCI has a computational cost of $O(fmc)$; note that m is a fixed parameter, so that the overall cost can also be considered to be $O(fc)$. Other relevant alternatives do typically involve singular value decomposition or matrix multiplication, thus resulting in $O(dfc)$ algorithms, where d is the number of documents or contexts in the collection.

We performed efficiency tests comparing DCI to SCL on the Webis-CLS-10 dataset. The test was run for all combinations of source and target classes and all target languages, which amounts to 36 runs. The same dedicated computer[4] run all experiments with the same number of 10 threads. Table 13 shows the averaged time scores obtained.

Table 13: Running time (in seconds) for DCI with two different DCFs (linear or cosine) and for SCL.

	Linear	Cosine	SCL-MI
Min (s)	6.406	7.501	449.988
Max (s)	22.119	27.032	859.719
Average (s)	11.553	17.774	678.834
Standard Deviation	3.976	5.516	98.324

SCL suffers from much higher computational costs than DCI. On average, DCI reduced by 98.3% and 97.4% the computational cost with respect to SCL for the linear and cosine DCFs, respectively. SCL required $m = 450$ binary optimization problems and translations, and required performing LSA on the predictive parameters. Our DCI-based method obtained better results in substantially less time. These efficiency tests suggest that DCI could scale well to larger datasets.

6.8 Embeddings

As a final note, the intuitions behind DCI have strong relationships with those behind "word embeddings", as from deep learning, a research area that has gained interest in the last renewed years. Neural language models are trained to obtain meaningful term representations that seem to capture interesting language regularities (Bengio, 2009). Although deep learning has been applied to cross-domain adaptation by Glorot et al. (2011) (a work we used as a baseline in Section 6.5.1), cross-lingual adaptation requires additional effort. That is, to consistently obtain bilingual word embeddings, large unlabelled datasets and aligned corpora (Zou, Socher, Cer, & Manning, 2013) or bilingual dictionaries (Mikolov, Le, & Sutskever, 2013) are typically required. Assuming a small set of words are translated (no

[4] Computer specifications: 64-bit Intel Core (TM) Genuine-Intel I7 with 12 processors at 3.47GH, 24GB RAM, running Ubuntu 14.04.2 LTS.

more than 200 words in our experiments), our method obtains term profiles that perform consistently in both languages for the classification task. Table 14 illustrates the semantic properties captured by our term profiles; it lists the most similar (via cosine similarity) target terms to a given source term.

Table 14: Five most similar terms in each of three target languages (German, French, Japanese) given three terms (beautifully, classical, delightful) in English for the Music domain.

beautifully		classical		delightful	
schöne (≈ beautiful)	0.635	adagio	0.767	魅力(≈ attractive)	0.610
liebevoll (≈ loving)	0.596	Martenot	0.746	描き出さ(≈ portrayed)	0.546
sehnsucht (≈ longing)	0.533	Charles-Marie	0.736	風景(≈ scenes)	0.545
ungewöhnlich (≈ unusual)	0.510	violoncelle (≈ cello)	0.727	繊細(≈ delicate)	0.542
phantastisch (≈ fantastic)	0.507	soliste (≈ soloist)	0.720	味わえる(≈ taste)	0.538

For word embeddings, even assuming that external resources are available, an additional optimization problem, posed as a geometrical transformation involving scaling and rotating the data matrices, is subsequently required in order to align the two embedding spaces. This was done by Mikolov et al. (2013) by forcing the embedding representations of some words of the bilingual dictionary to get closer to each other through the matrix transformation. Apart from the additional computational cost this may involve, we believe that such method might not be directly applicable to the scenario in which cross-domain and cross-lingual adaptations are tackled simultaneously. The main reason is that this final transformation aims at aligning the meaning of words taken from the bilingual dictionary in both domains, which could play different roles across domains, i.e., if they are not pivots. The embeddings generated by DCI do not require computationally expensive post-processing, and the correspondences in roles for different terms in both domains turn out to be directly captured by the DCF scores to the pivots.

To illustrate this, we used LSI to plot into a bidimensional space some important term profiles from the EB → GM adaptation obtained by DCI with cosine as the DCF. Figure 4 shows two zoomed-in areas of the bilingual space. Noticeably, the left-most part of the plot seems to represent the positive sentiment, while the right-most one seems to capture the negative sentiment. Some relevant semantic correspondences could be directly observed. Semantically related English words, such as expecting, expected, and hoping, are projected close together in the space. More interestingly, some related semantics seem to be preserved across languages, e.g., English words boring, irritating, bored and German words erschreckend (terrifying), acherlich (ridiculous), and schrecklich (terribly) are projected into regions of the space close to each other. Some interesting cases could be regarded as examples of cross-semantic correspondence. For example classic and love (book genres) from English reviews were projected close to folk and rock (music genres) from German reviews, an incidental semantic correspondence that emerged due to the juxtaposition of cross-domain and cross-lingual adaptation.

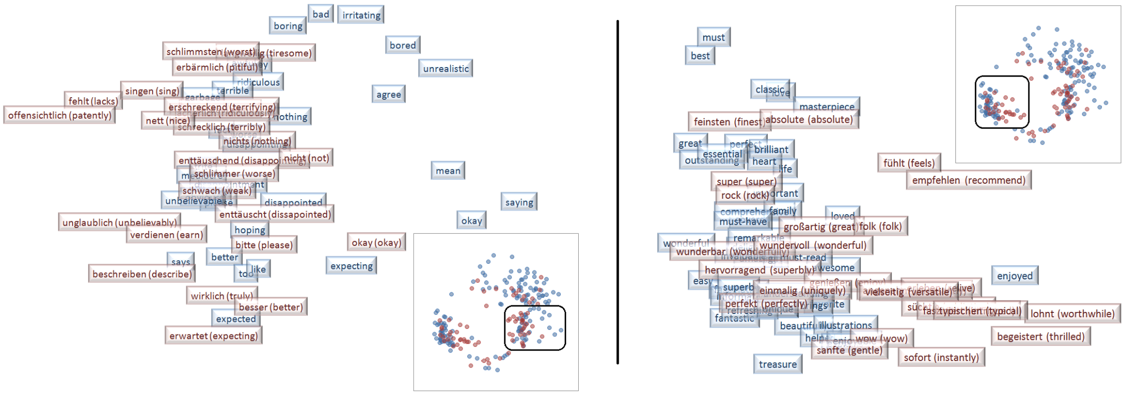

Figure 4: Vector profiles as word embeddings obtained for **EB** → **GM** adaptation. Zoom in the positive (left) and negative (right) sentiment area. The plot was obtained by applying LSI to terms deemed highly informative by mutual information.

These preliminary experiments suggest that DCI embeddings could be potentially useful for other tasks in natural language processing. This however will require a dedicated investigation that we defer to future work.

7. Conclusions and Future Work

We have proposed Distributional Correspondence Indexing, an efficient method for domain adaptation that represents terms in a vectorial space based on their distributional correspondence with respect to a small, fixed set of terms. This representation is motivated by Harris' distributional hypothesis and the notion of a "pivot term" of Blitzer et al. (2006); the method indexes documents from different domains into a common vector space based on their semantic correspondence.

Empirical evaluation on two popular sentiment analysis benchmarks shows that our method outperforms several state-of-the-art approaches in different domain-adaptation settings, including cross-domain and cross-lingual sentiment adaptation. We have also proposed an extended formulation of the domain adaptation problem, which tackles cross-domain adaptation and cross-language adaptation at the same time; on this we have present experiments where our system compares favourably to other related approaches. From the point of view of efficiency, we show our method to require modest computational resources, which is an indication that DCI can scale well to huge collections; in particular, in the cross-lingual case it required a smaller amount of human intervention than competing approaches in order to create the pivot set. We presented some high-performance DCFs that are parameter-free, which is a valuable characteristic in the domain adaptation setting, given that we are not expected to count on labelled data drawn from the target distribution on which parameters could be optimized.

The bilingual pivots created by a context-unaware word-translator oracle represent arguably an oversimplified naive approach of the translation problem. Notwithstanding this, DCI seems to compensate it with the aggregative contribution of the partial semantics scattered in several pivots. In this regard, we are interested in enhancing the concept of pivots for cross-lingual adaptation in a more general direction that better captures the context-aware multi-word translation, so as to attempt the polylingual case. Some possible directions might include enriching the term representation so as to incorporate part-of-speech tags and syntactic information, and also keeping track of the contexts in which a given term appeared. Moreover, and motivated by some empirical evidences in the cross-domain experiments that suggested comparable performance could be achieved even with extremely reduced sets of pivots, we will investigate more sophisticated pivot selection techniques by better characterizing the concept of pivot and the geometrical properties of the vector space they generate. We also plan to put to test DCI in other domains and settings, including multi-class multi- and single-label datasets, highly imbalanced classes, and transductive problems.

Acknowledgements

We are grateful to Xavier Glorot for sending us additional details on the experiments reported in their previous work (Glorot et al., 2011).
This paper is an extension of the short paper by Esuli and Moreo Fernández (2015).
Fabrizio Sebastiani is on leave from Consiglio Nazionale delle Ricerche, Italy.

References

Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, *6*, 1817–1853.

Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS 2006)*, pp. 137–144, Vancouver, CA.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, *2*(1), 1–127.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blitzer, J., Dredze, M., & Pereira, F. (2007). 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.

Blitzer, J., McDonald, R., & Pereira, F. (2006). 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.

Bollegala, D., Weir, D., & Carroll, J. (2011). Using multiple sources to construct a sentiment-sensitive thesaurus for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pp. 132–141, Portland, US.

Dai, W., Xue, G.-R., Yang, Q., & Yu, Y. (2007). Transferring naïve Bayes classifiers for text classification. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI 2007)*, pp. 540–545, Vancouver, CA.

DCI-source (2015) `http://hlt.isti.cnr.it/dciext/`.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Denecke, K. (2009). Are SentiWordNet scores suited for multi-domain sentiment classification?. In *Proceedings of the 4th International Conference on Digital Information Management (ICDIM 2009)*, pp. 33–38, Ann Arbor, US.

Dumais, S. T., Letsche, T. A., Littman, M. L., & Landauer, T. K. (1997). Automatic cross-language retrieval using latent semantic indexing. In *Working Notes of the AAAI Spring Symposium on Cross-language Text and Speech Retrieval*, pp. 18–24, Stanford, US.

Esuli, A., & Moreo Fernández, A. (2015). 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.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI 2007)*, pp. 1606–1611, San Francisco, US.

Gao, J., Fan, W., Jiang, J., & Han, J. (2008). Knowledge transfer via multiple model local structure mapping. In *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, pp. 283–291, Las Vegas, US.

Gliozzo, A., & Strapparava, C. (2005). Cross-language text categorization by acquiring multilingual domain models from comparable corpora. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pp. 9–16, Ann Arbor, US.

Gliozzo, A., & Strapparava, C. (2006). Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006)*, pp. 553–560, Sydney, AU.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, pp. 513–520, Bellevue, US.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(23), 146–162.

He, Y., Lin, C., & Alani, H. (2011). Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pp. 123–131, Portland, US.

JaTeCS (2015) `http://hlt.isti.cnr.it/jatecs/`.

Joachims, T. (1999). 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.

Kanerva, P., Kristofersson, J., & Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, p. 1036, Austin, US.

Koehn, P., & Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL 2002 Workshop on Unsupervised Lexical Acquisition*, pp. 9–16, Philadelphia, US.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Li, F., Pan, S. J., Jin, O., Yang, Q., & Zhu, X. (2012a). Cross-domain co-extraction of sentiment and topic lexicons. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pp. 410–419, Jeju Island, KR.

Li, L., Jin, X., & Long, M. (2012b). 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.

Ling, X., Dai, W., Xue, G.-R., Yang, Q., & Yu, Y. (2008). 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.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Morgan and Claypool Publishers, San Rafael, US.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. *ArXiv e-prints*, *arXiv:1309.4168 [cs.CL]*.

Moen, H., & Marsi, E. (2013). Cross-lingual random indexing for information retrieval. In *Proceedings of the 1st International Conference on Statistical Language and Speech Processing (SLSP 2013)*, pp. 164–175, Tarragona, ES.

MSD dataset (2007) `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/`.

Natural Language Understanding Toolkit (2011) `https://github.com/pprett/nut`.

Pan, S. J., Ni, X., Sun, J.-T., Yang, Q., & Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on the World Wide Web (WWW 2010)*, pp. 751–760, Raleigh, US.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

Pan, W., Zhong, E., & Yang, Q. (2012). Transfer learning for text mining. In Aggarwal, C. C., & Zhai, C. (Eds.), *Mining Text Data*, pp. 223–258. Springer, Heidelberg, DE.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1/2), 1–135.

Peirsman, Y., & Padó, S. (2010). Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Proceedings of the 8th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010)*, pp. 921–929, Los Angeles, US.

Platt, J. C., Toutanova, K., & Yih, W.-t. (2010). Translingual document representations from discriminative projections. In *Proceedings of the 8th Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pp. 251–261, Cambridge, US.

Ponomareva, N., & Thelwall, M. (2012). Do neighbours help? An exploration of graph-based algorithms for cross-domain sentiment classification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL 2012)*, pp. 655–665, Jeju Island, KR.

Prettenhofer, P., & Stein, B. (2010). Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pp. 1118–1127, Uppsala, SE.

Prettenhofer, P., & Stein, B. (2011). Cross-lingual adaptation using structural correspondence learning. *ACM Transactions on Intelligent Systems and Technology*, *3*(1), Article 13.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics (ACL 1995)*, pp. 320–322, Cambridge, US.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting on Association for Computational Linguistics (ACL 1999)*, pp. 519–526, College Park, US.

Rigutini, L., Maggini, M., & Liu, B. (2005). An EM-based training algorithm for cross-language text categorization. In *Proceedings of the 3rd IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005)*, pp. 529–535, Compiègne, FR.

Sahlgren, M. (2005). An introduction to random indexing. In *Proceedings of the Workshop on Methods and Applications of Semantic Indexing*, Copenhagen, DK.

Sorg, P., & Cimiano, P. (2008). Cross-language information retrieval with explicit semantic analysis. In *Working Notes of the 2008 Cross-Language Evaluation Forum (CLEF 2008)*, Aarhus, DE.

Sorg, P., & Cimiano, P. (2012). Exploiting Wikipedia for cross-lingual and multilingual information retrieval. *Data and Knowledge Engineering*, *74*, 26–45.

SVM^{light} (2008) `http://svmlight.joachims.org/`.

Vinokourov, A., Shawe-Taylor, J., & Cristianini, N. (2002). Inferring a semantic representation of text via cross-language correlation analysis. In *Proceedings of the 16th Annual Conference on Neural Information Processing Systems (NIPS 2002)*, pp. 1473–1480, Vancouver, CA.

Wan, C., Pan, R., & Li, J. (2011). Bi-weighting domain adaptation for cross-language text classification. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, pp. 1535–1540, Barcelona, ES.

Wan, X. (2009). Co-training for cross-lingual sentiment classification. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL/IJCNLP 2009)*, pp. 235–243, Singapore, SN.

Wang, P., Domeniconi, C., & Hu, J. (2008). Using Wikipedia for co-clustering-based cross-domain text classification. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008)*, pp. 1085–1090, Pisa, IT.

Webis-CLS dataset (2010) `http://www.uni-weimar.de/en/media/chairs/webis/research/corpora/corpus-webis-cls-10/`.

Xia, R., & Zong, C. (2011). A POS-based ensemble model for cross-domain sentiment classification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pp. 614–622, Chiang Mai, TH.

Xiang, E. W., Cao, B., Hu, D. H., & Yang, Q. (2010). Bridging domains using world-wide knowledge for transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(6), 770–783.

Xiao, M., & Guo, Y. (2013). A novel two-step method for cross-language representation learning. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS 2013)*, pp. 1259–1267, Lake Tahoe, US.

Xiao, M., & Guo, Y. (2014). Semi-supervised matrix completion for cross-lingual text classification. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*, pp. 1607–1614, Québec City, CA.

Xue, G.-R., Dai, W., Yang, Q., & Yu, Y. (2008). 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.

Zhuang, F., Luo, P., Xiong, H., He, Q., Xiong, Y., & Shi, Z. (2011). Exploiting associations between word clusters and document classes for cross-domain text categorization. *Statistical Analysis and Data Mining*, *4*(1), 100–114.

Zou, W. Y., Socher, R., Cer, D. M., & Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 11th Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pp. 1393–1398, Seattle, US.