Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Distributional Correspondence Indexing for Cross-Lingual and Cross-Domain Sentiment Classification (Extended Abstract)*

Alejandro Moreo Fernández, Andrea Esuli and **Fabrizio Sebastiani**
Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche
{alejandro.moreo,andrea.esuli,fabrizio.sebastiani}@isti.cnr.it

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 extended abstract we briefly describe a new DA method called *Distributional Correspondence Indexing* (DCI) for 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. The experiments we have conducted show that DCI obtains better performance than current state-of-the-art techniques for cross-lingual and cross-domain sentiment classification.

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. Deploying a model for a new *target domain* in the absence of high-quality annotated examples thus entails a substantial human labelling effort. *Transfer learning* (TL) [Pan and Yang, 2010] focuses on alleviating this problem by leveraging training examples from a different, although related, *source domain* for which the amount of available labelled examples is higher. TL thus operates in applicative scenarios in which the so-called "iid assumption" (the training and the test data are randomly drawn from the same distribution) no longer holds.

One such scenario of the utmost importance is *sentiment classification* [Liu, 2012], 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). In many contexts the amount of available, prelabelled opinions is scarce, or even null, in the case of new entities (e.g., products, policies, or political candidates). In such cases, promptly generating a sentiment classifier might become difficult, due to the considerable cost and time involved in producing a representative set 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].

Central to text classification is the representation of words and documents as numerical vectors that reflect, through their relative distances, the semantic relations among them, i.e., the smaller the distance between a pair of word vectors, the more semantically similar the words are assumed to be. While distributional semantic models like Latent Semantic Analysis (LSA) [Deerwester *et al.*, 1990] and Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003] have proven useful to represent similar words (e.g., *beautiful* and *nice*) close to each other (e.g., in terms of cosine similarity) in the vector space, one main difficulty in domain adaptation is to discover semantic correspondences that may exist *across domains* (e.g., *writer* and *director* when adapting from a source domain of books to a target domain of films).

Structural Correspondence Learning (SCL) was first adapted for the cross-domain case [Blitzer *et al.*, 2007] and then extended to the cross-lingual case in [Prettenhofer and Stein, 2011]. SCL discovers correspondences among terms from different domains that show similar distributions with respect to a set of *pivot* terms (highly predictive terms expected to behave in a similar way in both domains). Each pivot defines in SCL an auxiliary classification problem that brings to bear "structural" information of the adaptation problem. The intuition according to which the semantics of words is somehow determined by its distribution in text with respect to other terms is generally referred to as the *distributional hypothesis* [Harris, 1954]. The Distributional Correspondence Indexing (DCI) [Moreo Fernández *et al.*, 2016] method we propose also builds upon the distributional hypothesis, but it formalizes the notion of correspondence through the Distributional Correspondence Functions (DCF), which are much lighter, in terms of computational cost, than the auxiliary problems of SCL.

We experimentally show that DCI compares favourably to the state of the art in two popular sentiment datasets covering both cross-domain and cross-lingual adaptation, and at a

*This paper is an extended abstract of an article appeared as [Moreo Fernández *et al.*, 2016].

substantially smaller computational cost.

2 Distributional Correspondence Functions

DCFs are a family of real-valued functions that quantify the degree of correspondence between two features (terms, in our case) f^i and f^j by comparing their *context distribution vectors* $\mathbf{f}^i$ and $\mathbf{f}^j$ from any (unlabelled) collection U. A context distribution vector is a unit-length n-dimensional vector that models how a feature relates to a set of contexts (documents, in our case); $\mathbf{f}^i_c$ 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 (e.g., $tfidf$), and might lead to different interpretations of the DCF, e.g., as a probability function in an event space (Section 2.1), or as a kernel in a vector space (Section 2.2).

A DCF is thus 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. We force DCFs to be centered at 0 for *null correspondence*, i.e., $\eta(\mathbf{f}^i, \mathbf{f}^j) = 0$ means the correspondence between f^i and f^j is not different than the correspondence between any pair of context vectors which randomly distribute the same number of events as f^i and f^j.

2.1 Probability-Based DCF

Probability-based DCFs derive from information theory, and build upon the joint probability distributions of two features in the binomial event space.

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 called AMI, where in order to distinguish between positive and negative correspondence we swap the sign of MI by means of a function ρ if $tpr + tnr < 1$.

2.2 Kernel-Based DCFs

Kernel-based DCFs are rooted in the kernel functions typically used, e.g., within Support Vector Machines (SVM). In this case the values in the context vector can be numeric, thus indicating the relative importance of a term in a given context (we use $tfidf$ as the weighting function). We consider normalized context vectors, i.e., after weighting the document-by-term matrix we normalize the term vectors to unit length.

However (and differently from the probability-based DCFs we have considered above), not all kernels satisfy the *null correspondence* property. In general, a valid DCF η_K can be defined from a kernel $K(\cdot, \cdot)$ as:

$$\eta_K(\mathbf{u}, \mathbf{v}) = K(\mathbf{u}, \mathbf{v}) - \mathbb{E}_{\substack{\mathbf{u}' \sim P(U) \\ \mathbf{v}' \sim P(V)}}[K(\mathbf{u}', \mathbf{v}')] \quad (1)$$

where $\mathbf{u}'$ and $\mathbf{v}'$ are any two random context vectors with the same prevalences as $\mathbf{u}$ and $\mathbf{v}$ (random variables represented by U and V), respectively.

The second part of Table 1 shows the Kernel-based DCFs we have considered, based on the cosine, the polynomial and the radial basis function kernels.

	DCF	Mathematical form
Prob.-based	Linear	$P(f^i \mid f^j) - P(f^i \mid \overline{f^j})$
	PMI	$\log_2 \frac{P(f^i, f^j)}{P(f^i)P(f^j)}$
	AMI	$\rho(f^i, f^j) \sum_{x \in \{f^i, \overline{f^i}\}} \sum_{y \in \{f^j, \overline{f^j}\}} P(x, y) \log_2 \frac{P(x,y)}{P(x)P(y)}$
Kernel-based	Cosine	$\frac{\langle \mathbf{f}^i, \mathbf{f}^j \rangle}{\|\mathbf{f}^i\| \|\mathbf{f}^j\|} - \sqrt{p_i p_j}$
	Polynomial	$(a + \langle \mathbf{f}^i, \mathbf{f}^j \rangle)^b - (a + \sqrt{p_i p_j})^b$
	RBF	$\exp\{-\gamma \|\mathbf{f}^i - \mathbf{f}^j\|^2\} - \exp\left\{-4\gamma \left(1 - \sqrt{p_i p_j}\right)^2\right\}$

Table 1: Mathematical forms of DCFs discussed in this work. We use p_k to denote the prevalence of vector $\mathbf{f}^k$.

3 Distributional Correspondence Indexing

The working hypothesis of DCI is that words that play similar roles in their respective domains might present approximately invariant correspondences to the pivots. For example, consider the source domain of book reviews and the target domain of movie reviews. Consider also the list of pivots $[intriguing, annoying, captivating, \ldots]$ and a DCF η. We might thus expect to have:

$$\eta(book, intriguing) \approx \eta(movie, intriguing)$$
$$\eta(book, annoying) \approx \eta(movie, annoying)$$
$$\eta(book, captivating) \approx \eta(movie, captivating)$$

etc., since $book$ and $movie$ terms play analogous roles in the source and target domains. Therefore, by embedding each term with respect to its DCF values to the pivots, similar terms across domains might end up being represented by similar vectors in a common vector space. The cross-lingual adaptation is achieved by presenting each pivot by a translation equivalent in the target language.

3.1 Pivot Selection

[Blitzer *et al.*, 2006; 2007] defined pivots as highly predictive terms which occur frequently in the source and target domains and behave similarly in both domains. A good pivot should be highly *task-dependent*, and also present a similar degree of *domain-dependence* in the two domains. We thus look for pivots by selecting the top n terms (where n is a parameter) ranked by their pivot strength $\Psi(f^i)$ defined as:

$$\Psi(f^i) = MI_s(f^i) \frac{min\{p^s_i, p^t_i\}}{max\{p^s_i, p^t_i\}} \quad (2)$$

where $MI_s(f^i)$ is the mutual information (as previously done in [Prettenhofer and Stein, 2011]) of the term f^i to the label (to be estimated on the training set Tr_s), and the rightmost factor is the *cross-consistency*, measured as the drift in prevalence (p_i) of the term f^i across the two domains.

3.2 Term Profiles

Given n pivots and a DCF η, each source and target term (including the pivot terms) is embedded as an n-dimensional *profile* vector

$$\vec{f} = (\eta(\mathbf{f}, \mathbf{p}^1), \eta(\mathbf{f}, \mathbf{p}^2), \ldots, \eta(\mathbf{f}, \mathbf{p}^m)) \quad (3)$$

where $\mathbf{f}$ and $\mathbf{p}^i$ are the context distribution vectors (typically estimated in unlabelled collections) of the term f being profiled and the i^{th} pivot, respectively.

In order to avoid pivots with high prevalence to generate high DCF values which could lead to dominant dimensions, we center each profile dimension on its expected value and then rescale by the standard deviation, so that the values for all profile dimensions are approximately normally distributed in $\mathcal{N}(0, 1)$, and then rescale them to unit length.

As we assume pivot terms behave similarly in the two domains, we *unify* their term profiles by averaging the source and the target profiles in order to correct the possible misalignment between the source and target views of the pivot.

3.3 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

$$\vec{d_j} = \sum_{f_i \in d_j} w_{ij} \cdot \vec{f_i'} \quad (4)$$

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 term profile vector for f_i.

4 Experiments

In this section we experimentally compare DCI, equipped with different DCFs, to other state-of-the-art methods proposed in the literature.

We test these methods on two popular, publicly available sentiment datasets: Multi-Domain Sentiment Dataset[1] (MDS) [Blitzer *et al.*, 2007] and Webis-CLS-10[2] [Prettenhofer and Stein, 2011]. MDS is frequently used for evaluating cross-domain adaptation approaches and consists of English product reviews taken from Amazon.com for the four domains Books, DVDs, Electronics, and Kitchen appliances. 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. 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. In order to facilitate reproducibility and to allow for a fair comparison with the results reported in previous literature, we use the same pre-processed version of the dataset and the same experimental protocol used in [Blitzer *et al.*, 2007].

[1] http://www.cs.jhu.edu/ mdredze/datasets/sentiment/

[2] https://goo.gl/eppCro

The Webis-CLS-10 dataset, frequently used for evaluating cross-lingual methods, consists of Amazon product reviews written in four languages (English, German, French, and Japanese), covering three product domains (Books, DVDs, and Music). 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. We adhere to the exact experimental protocol described in [Prettenhofer and Stein, 2011] and used by many researchers in the field.

As the evaluation measure we use standard accuracy (the proportion of correctly classified documents over the total number of outcomes), following the practice common in the related literature.

For this extended abstract, we have updated the list of baseline methods by adding new approaches (TCT [Huang *et al.*, 2017], TrAdaB [Huang *et al.*, 2017], DANN [Ganin *et al.*, 2016], CL-TS [Zhou *et al.*, 2015], Bi-PV [Xu and Wan, 2017], BiDRL [Zhou *et al.*, 2016b], WSDNNs, [Zhou *et al.*, 2016a], CLDFA [Xu and Yang, 2017]) which have been published in the cross-domain and cross-lingual arena after our original work [Moreo Fernández *et al.*, 2016], and kept those which performed best in our original evaluation (SCL-MI [Blitzer *et al.*, 2007], SFA [Pan *et al.*, 2010], SDA [Glorot *et al.*, 2011], and SSMC [Xiao and Guo, 2014]). We also consider an upper bound that trains the classifier on the training set of the target domain ("Upper"), and a lower bound that trains the 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 ("No-Trans"). For cross-lingual adaptation we also report the machine translation baseline ("MT"), which first translates all target documents into the source language (English, in our experiments) before giving them as input to the classifier; we use the pre-translated documents provided by [Prettenhofer and Stein, 2011].

We implemented our method as part of the JaTeCs [Esuli *et al.*, 2017] framework using SVMs as the learning device. In all experiments we set the number of pivots to 100. To emulate the word oracle that translates a source pivot word to the target language we used the bilingual dictionaries created by [Prettenhofer and Stein, 2011].

Table 2 and 3 report the results for cross-domain adaptation and cross-lingual adaptation, respectively (bold indicates the best score for each dataset, while greyed-out cells indicate the DCI variants which outperform all other competitors). We refer the interested reader to our original paper [Moreo Fernández *et al.*, 2016], where other scenarios (such as simultaneously tackling cross-domain and cross-lingual) are explored.

Most variants of DCI perform comparably or better to all compared methods, even to the new approaches that have been published after our original paper. In this regard, it is interesting to note that many configurations of DCI outperform on average all the competitors. Particularly, the Cosine and Polynomial DCFs have shown to stand the test of time, delivering the best averaged results both in cross-domain and cross-lingual experiments.

Source	Target	NoTrans	Upper	SCL-MI	SFA	TCT	SDA	TrAdaB	DANN	Linear	PMI	AMI	Cos	Poly	RBF
Books	DVD	.772	.847	.758	.814	.818	.804	.796	**.829**	.808	.811	.806	.817	**.829**	.815
	Electronics	.708	.869	.759	.725	.757	.806	.749	.804	.810	.822	.793	.822	**.826**	.821
	Kitchen	.745	.902	.789	.788	.789	**.844**	.778	.843	.834	.839	.822	.835	**.844**	.835
DVD	Books	.728	.844	.797	.775	.792	.724	.747	.825	.825	.827	.811	.824	**.830**	.825
	Electronics	.730	.869	.741	.767	.778	**.872**	.759	.809	.822	.832	.812	.824	.833	.826
	Kitchen	.740	.902	.814	.808	.812	.803	.757	.849	.858	.856	.846	**.864**	.861	.863
Electronics	Books	.707	.844	.754	.757	.759	.768	.691	.774	.766	.763	.753	.764	**.776**	.765
	DVD	.706	.847	.762	.772	.773	**.902**	.718	.781	.768	.779	.765	.774	.799	.771
	Kitchen	.840	.902	.859	.868	.863	.777	.837	**.881**	.864	.864	.851	.868	.874	.867
Kitchen	Books	.709	.844	.686	.748	.748	**.807**	.706	.718	.783	.783	.769	.790	.791	.784
	DVD	.727	.847	.769	.766	.785	**.835**	.744	.789	.788	.789	.781	.799	.807	.798
	Electronics	.827	.869	.868	.851	.856	.802	.831	.856	.855	.851	.843	.858	**.863**	.857
	Average	.745	.866	.780	.786	.794	.812	.759	.813	.815	.818	.804	.820	**.828**	.819

Table 2: Cross-domain adaptation on the MDS dataset.

Target Language	Domain	NoTrans	Upper	CL-MT	SCL-MI	SSMC	CL-TS	Bi-PV	BiDRL	WSDNNs	CLDFA	Linear	PMI	AMI	Cos	Poly	RBF
German	Books	.541	.867	.808	.833	.819	.799	.796	**.841**	.813	.840	.798	.714	.797	.827	.837	.829
	DVD	.556	.835	.800	.809	.823	.819	.786	**.841**	.824	.831	.826	.819	.800	.822	.833	.788
	Music	.539	.859	.790	.829	.813	.796	.825	.847	.807	.790	.844	.850	.837	**.856**	.844	.801
French	Books	.589	.861	.820	.813	.831	.826	.843	**.844**	.835	.834	.746	.761	.768	.842	.819	**.844**
	DVD	.519	.871	.794	.804	.827	.827	.796	.836	.836	.826	.823	.823	.801	.827	.806	**.846**
	Music	.551	.889	.764	.781	.805	.802	.801	.826	.810	.833	.816	.827	.818	**.844**	.840	.803
Japanese	Books	.484	.812	.692	.770	.738	.735	.718	.732	.745	.774	**.779**	.731	.711	.758	.754	.782
	DVD	.471	.834	.722	.764	.776	.771	.754	.768	.783	.805	**.822**	.768	.797	.801	.795	.761
	Music	.535	.842	.714	.773	.775	.768	.755	.788	.775	.765	.826	.816	.807	**.839**	.832	.826
	Average	.532	.852	.767	.797	.801	.794	.786	.814	.803	.811	.809	.790	.793	**.824**	.818	.809

Table 3: Cross-lingual adaptation on the Webis-CLS-10 dataset.

4.1 Embeddings

The intuitions behind DCI bear strong resemblance to those behind "word embeddings" [Mikolov *et al.*, 2013], as from deep learning. Although neural models typically require large quantities of text data and expensive resources in terms of computation, DCI delivers meaningful representations in a fraction of the time they require (see [Moreo Fernández *et al.*, 2016] for a more detailed discussion on efficiency).

Table 4 illustrates the semantic properties captured by our term profiles; it lists the most similar (via cosine similarity) target terms to a given source term.

beautifully	classical	delightful
schöne (beautiful) .635	adagio .767	魅力(attractive) .610
liebevoll (loving) .596	Martenot .746	描き出さ(portrayed) .546
sehnsucht (longing) .533	Charles-Marie .736	風景(scenes) .545
ungewöhnlich (unusual) .510	violoncelle (cello) .727	繊細(delicate) .542
phantastisch (fantastic) .507	soliste (soloist) .720	味わえる(taste) .538

Table 4: Five most similar terms in German, French, Japanese given three terms (beautifully, classical, delightful) in English for the Music domain.

5 Conclusions and Future Work

Distributional Correspondence Indexing is 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 the distributional hypothesis [Harris, 1954] and the notion of a "pivot term" [Blitzer *et al.*, 2006]; the method indexes documents from different domains into a common vector space based on their semantic correspondence. Unlike other distributional semantic methods, DCI gives a lighter interpretation to the hypothesis through the distributional correspondence functions, resulting in a computationally cheap approach to domain adaptation.

Empirical evaluation we have carried out 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. DCI has remained unbeaten since its appearance in 2016.

In future research, we 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.

References

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Blitzer *et al.*, 2006] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 4th Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 120–128, Sydney, AU, 2006.

[Blitzer *et al.*, 2007] John Blitzer, Mark Dredze, and Fernando Pereira. 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)*, pages 440–447, Prague, CZ, 2007.

[Deerwester *et al.*, 1990] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[Esuli *et al.*, 2017] Andrea Esuli, Tiziano Fagni, and Alejandro Moreo Fernández. Jatecs an open-source java text categorization system. *arXiv preprint arXiv:1706.06802*, 2017.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[Glorot *et al.*, 2011] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, pages 513–520, Bellevue, US, 2011.

[Harris, 1954] Zellig S. Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

[Huang *et al.*, 2017] Xingchang Huang, Yanghui Rao, Haoran Xie, Tak-Lam Wong, and Fu Lee Wang. Cross-domain sentiment classification via topic-related tradaboost. In *AAAI*, pages 4939–4940, 2017.

[Liu, 2012] Bing Liu. *Sentiment Analysis and Opinion Mining*. Morgan and Claypool Publishers, San Rafael, US, 2012.

[Mikolov *et al.*, 2013] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[Moreo Fernández *et al.*, 2016] Alejandro Moreo Fernández, Andrea Esuli, and Fabrizio Sebastiani. Distributional correspondence indexing for cross-lingual and cross-domain sentiment classification. *Journal of artificial intelligence research*, 55(1):131–163, 2016.

[Pan and Yang, 2010] Sinno J. Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[Pan *et al.*, 2010] Sinno J. Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on the World Wide Web (WWW 2010)*, pages 751–760, Raleigh, US, 2010.

[Pan *et al.*, 2012] Weike Pan, Erheng Zhong, and Qiang Yang. Transfer learning for text mining. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 223–258. Springer, Heidelberg, DE, 2012.

[Prettenhofer and Stein, 2011] Peter Prettenhofer and Benno Stein. Cross-lingual adaptation using structural correspondence learning. *ACM Transactions on Intelligent Systems and Technology*, 3(1):Article 13, 2011.

[Xiao and Guo, 2014] Min Xiao and Yuhong Guo. Semi-supervised matrix completion for cross-lingual text classification. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*, pages 1607–1614, Québec City, CA, 2014.

[Xu and Wan, 2017] Kui Xu and Xiaojun Wan. Towards a universal sentiment classifier in multiple languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 511–520, 2017.

[Xu and Yang, 2017] Ruochen Xu and Yiming Yang. Cross-lingual distillation for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1415–1425, 2017.

[Zhou *et al.*, 2015] Guangyou Zhou, Tingting He, Jun Zhao, and Wensheng Wu. A subspace learning framework for cross-lingual sentiment classification with partial parallel data. In *IJCAI*, pages 1426–1433, 2015.

[Zhou *et al.*, 2016a] Guangyou Zhou, Zhao Zeng, Jimmy Xiangji Huang, and Tingting He. Transfer learning for cross-lingual sentiment classification with weakly shared deep neural networks. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 245–254. ACM, 2016.

[Zhou *et al.*, 2016b] Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1403–1412, 2016.