Neural Unsupervised Domain Adaptation in NLP—A Survey

Alan Ramponi,[1,2] **Barbara Plank**[3]
[1]Department of Inf. Eng. and Computer Science, University of Trento, Italy
[2]Fondazione The Microsoft Research – University of Trento
Centre for Computational and Systems Biology (COSBI), Italy
[3]Department of Computer Science, ITU Copenhagen, Denmark
ramponi@cosbi.eu, bplank@itu.dk

Abstract

Deep neural networks excel at learning from labeled data and achieve state-of-the-art results on a wide array of Natural Language Processing tasks. In contrast, learning from unlabeled data, especially under domain shift, remains a challenge. Motivated by the latest advances, in this survey we review neural unsupervised domain adaptation techniques which do not require labeled target domain data. This is a more challenging yet a more widely applicable setup. We outline methods, from early traditional non-neural methods to pre-trained model transfer. We also revisit the notion of *domain*, and we uncover a bias in the type of Natural Language Processing tasks which received most attention. Lastly, we outline future directions, particularly the broader need for *out-of-distribution generalization* of future NLP.[1]

1 Introduction

Deep learning has undoubtedly pushed the frontier in Natural Language Processing (NLP). Particularly large pre-trained language models have improved results for a wide range of NLP applications. However, the lack of portability of NLP models to new conditions remains a central issue in NLP. For many target applications, labeled data is lacking (Y scarcity), and even for pre-training general models data might be scarce (X scarcity). This makes it even more pressing to revisit a particular type of transfer learning, namely domain adaptation (DA). A default assumption in many machine learning algorithms is that the training and test sets follow the same underlying distribution. When these distributions do not match, we face a *dataset shift* (Gretton et al., 2007) – in NLP typically referred to as a *domain shift*. In this setup, the *target* domain and the *source* training data differ, they are not sampled from the same underlying distribution. Consequently, performance drops on the target, which undermines the ability of models to truly generalize *into the wild*. Domain adaptation is closely tied to a fundamental bigger open issue in machine learning: generalization beyond the training distribution. Ultimately, intelligent systems should be able to adapt and robustly handle any test distribution, without having seen any data from it. This is the broader need for *out-of-distribution generalization* (Bengio, 2019), and a more challenging setup targeted at handling *unknown* domains (Volpi et al., 2018; Krueger et al., 2020).

Work on domain adaptation focused largely on *supervised* domain adaptation (Daumé III, 2007; Plank, 2011). In such a classic supervised DA setup, a small amount of labeled target domain data is available, along with some larger amount of labeled source domain data. The task is to adapt from the source to the specific target domain in light of limited target domain data. However, annotation is a substantial time-requiring and costly manual effort. While annotation directly mitigates the lack of labeled data, it does not easily scale to new application targets. In contrast, DA methods aim to shift the ability of models from the traditional interpolation of similar examples to models that extrapolate to examples outside the original training distribution (Ruder, 2019). *Unsupervised domain adaptation* (UDA) mitigates the domain shift issue by learning only from *unlabeled* target data, which is typically available for both source and target domain(s). UDA fits the classical real-world scenario better, in which labeled data in the target domain is absent, but unlabeled data might be abundant. UDA thus provides an elegant and scalable solution. We believe these advances in UDA will help for out-of-distribution generalization.

[1]Accompanying repository: https://github.com/bplank/awesome-neural-adaptation-in-NLP

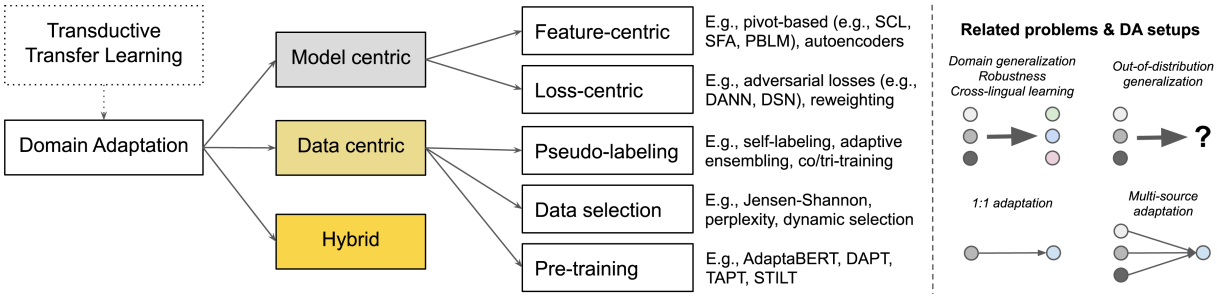

Figure 1: Taxonomy of DA as special case of transductive transfer learning (left). Related problems (e.g., domain and out-of-distribution generalization) and DA setups (1:1 and multi-source adaptation) (right).

A categorization of domain adaptation in NLP We categorize research into model-centric, data-centric and hybrid approaches, as shown in Figure 1. *Model-centric* methods target approaches to augment the feature space, alter the loss function, the architecture or model parameters (Blitzer et al., 2006; Pan et al., 2010; Ganin et al., 2016). *Data-centric* methods focus on the data aspect and either involve pseudo-labeling (or bootstrapping) to bridge the domain gap (Abney, 2007; Zhu and Goldberg, 2009; Ruder and Plank, 2018; Cui and Bollegala, 2019), data selection (Axelrod et al., 2011; Plank and van Noord, 2011; Ruder and Plank, 2017) and pre-training methods (Han and Eisenstein, 2019; Guo et al., 2020). As some approaches take elements of both, we include a *hybrid* category.[2] A comprehensive overview of UDA methods and the tasks each method is applied to is provided in Table 1.

Other surveys Comprehensive reviews on DA exist, each with a different focus: visual applications (Csurka, 2017; Patel et al., 2015; Wilson and Cook, 2020), machine translation (MT) (Chu and Wang, 2018), pre-neural DA methods in NLP (Jiang, 2008; Margolis, 2011). Seminal surveys in machine learning on transfer learning include Pan and Yang (2009), Weiss et al. (2016), and Yang et al. (2020).

Contributions In this survey, we (i) comprehensively review neural approaches to unsupervised domain adaptation in NLP,[3] (ii) we analyze and compare the strengths and weaknesses of the described approaches, and (iii) we outline potential challenges and future directions in this field.

2 Background

First we introduce the classic learning paradigm with its core assumption, then we outline DA setups. Given $\{x_1, ..., x_n\} = X$ the training instances and $\{y_1, ..., y_n\} = Y$ the corresponding class labels, the goal of machine learning is to learn a function f that generalizes well to unseen instances. In *supervised learning*, training data consists of tuples $\{(x_i, y_i)\}_{i=1}^{n}$, where n is the number of instances, while in *unsupervised learning* we only have $\{(x_i)\}_{i=1}^{n}$. A general assumption in supervised machine learning is that the test data follows the same distribution as the training data. Formally, training and test data are assumed to be independently and identically (i.i.d.) sampled from the same underlying distribution. In practice, this assumption does not hold, which translates into a drop in performance when the model f trained on a source domain S is tested on a different but related target domain T.

2.1 Domain adaptation and transfer learning: notation

Formally, a domain is defined as $\mathcal{D} = \{\mathcal{X}, P(X)\}$ where $\mathcal{X}$ is the feature space (e.g., the text representations), and $P(X)$ is the marginal probability distribution over that feature space. A task (e.g., sentiment classification) is defined as $\mathcal{T} = \{\mathcal{Y}, P(Y|X)\}$, where $\mathcal{Y}$ is the label space. Estimates for the prior distribution $P(Y)$ and the likelihood $P(Y|X)$ are learned from the training data $\{(x_i, y_i)\}_{i=1}^{n}$.

Domain adaptation aims to learn a function f from a source domain $\mathcal{D}_S$ that generalizes well to a target domain $\mathcal{D}_T$, where $P_S(X) \neq P_T(X)$. DA is a particular case of transfer learning, namely *transductive transfer learning* (Pan and Yang, 2009; Ruder, 2019). In inductive learning, the source and

[2] We take inspiration of the data-centric and model-centric terms from Chu and Wang (2018) in MT, and add hybrid.

[3] We disregard methods which are task-specific (like leveraging a sentiment thesaurus).

target tasks differ (Pan and Yang, 2009). In transductive DA, the source and target tasks $\mathcal{T}_S$ and $\mathcal{T}_T$ remain the same, but the source and target domains $\mathcal{D}_S$ and $\mathcal{D}_T$ differ in their underlying probability distributions. Given two distributions $P_S(X, Y)$ and $P_T(X, Y)$, DA typically addresses the shift in marginal distribution $P_S(X) \neq P_T(X)$, also known as covariate shift. A related problem is the problem of label shift, $P_S(Y) \neq P_T(Y)$. Since we do not assume any labeled target data, we focus on the former.[4]

3 What is a *domain*? From the notion of domain to variety space and related problems

Despite the formal definition of *domain* above, the term is quiet loosely used in NLP and there is no common ground on what constitutes a domain (Plank, 2016). Typically in NLP, domain is meant to refer to some coherent type of corpus, i.e., predetermined by the given dataset (Plank, 2011). This may relate to topic, style, genre, or linguistic register. The notion of domain and what plays into it has though significantly changed over the last years, leading to relevant research lines.

First, the Penn Treebank WSJ corpus (Marcus et al., 1993) and the Brown corpus (Francis and Kucera, 1979) are prototypical examples, with the WSJ being considered widely as the canonical newswire domain. In the recent decade, there has been considerable work on what is considered *non-canonical* data. The dichotomy between canonical (typically considered well-edited English newswire) and non-canonical data arose with the increasing interest of working with *social media* with all its challenges related to the 'noisiness' of the domain (Eisenstein, 2013; Baldwin et al., 2013). Models trained on canonical data failed in light of the challenges on, e.g., Twitter (Gimpel et al., 2011; Foster et al., 2011).

The general quest to understand the implications of variations of language on model performance led to lines of work on how human factors impact data in a covert or overt way, e.g., on how latent socio-demographic factors impact NLP performance (Hovy, 2015; Nguyen et al., 2016), or how direct data collection strategies like crowdsourcing impact corpus composition (Geva et al., 2019) or frequency effects impact NLP performance (Zhang et al., 2019). However, *what is a domain?* Is, say, Twitter, its own domain? Or is it a set of subdomains? Similarly, do language samples of social groups (e.g., sociolects) form a domain or a set of subdomains?

Variety space We believe it is time to reconsider the notion of *domain*, the use of the term itself, and raise even more awareness of the underlying variation in the data samples NLP works with. NLP is pervasively facing heterogeneity in data along many underlying (often unknown) dimensions. A theoretical notion put forward by Plank (2016) is the *variety space*. In the variety space a *corpus* is seen as a subspace (subregion), a sample of the variety space. A corpus is a set of instances drawn from the underlying unknown high-dimensional variety space, whose dimensions (or latent factors) are fuzzy language and annotation aspects. These latent factors can be related to the notions discussed above, such as genre (e.g., scientific, newswire, informal), sub-domain (e.g., finance, immunology, politics, environmental law, molecular biology) and socio-demographic aspects (e.g., gender), among other unknown factors, as well as stylistic or data sampling impacts (e.g., sentence length, annotator bias).

In spirit of the variety space (Plank, 2016), we suggest to use the more general term *variety*, rather than domain, which pinpoints better to the underlying linguistic differences and their implications rather than the technical assumptions. Each corpus is inevitably biased towards a specialized language and some latent aspects. Understanding bias sources and effects, besides effects only (Shah et al., 2020), and documenting the known are the first important steps (Bender and Friedman, 2018), as is building broader, more varied corpora (Ide and Suderman, 2004). What we need more work on is to link the known to the unknown, and studying its impact. Doing so will ultimately help to not only overcome overfitting to overrepresented domains (e.g., the newswire bias (Plank, 2016)), but also work on robustness and ultimately out-of-distribution generalization, as described later on.

Treating data as 'just another input' to machine learning is very problematic. For example, it is less known that the well-known Penn Treebank consists of multiple genres (Webber, 2009; Plank and van

[4] As outlined in Plank (2011), there exists a three-way distinction for domain adaptation: supervised DA, unsupervised DA but also *semi-supervised* DA. The latter was coined around 2010 to distinguish purely unsupervised DA from cases where a small amount of labeled data is available in addition to the unlabeled target data. As this setup still assumes some labeled data and it has not received much attention, it is not discussed in the current survey.

Noord, 2011), including reviews and some prose. It has almost universally been treated as prototypical news domain. Similarly, social media is typically considered only non-canonical data, but an analysis revealed the data to lie on a "continuum of similarity" (Baldwin et al., 2013). This has implications on NLP performance. As we have seen, there are a multitude of dimensions to consider in corpus composition and annotations, which are tied to the theoretical notion of a variety space. They challenge the true generalization capabilities of current models. *What remains is to study what variety comprises, how covert and overt factors impact results, and take them into consideration in modeling and evaluation.*

Related problems Following the idea of the *variety space*, we discuss three related notions: *cross-lingual learning*, *domain generalization/robustness*, and *out-of-distribution generalization*.

In *cross-lingual learning* the feature space drastically changes, as alphabets, vocabularies and word order can be different. It can be seen as extreme adaptation scenario, for which parallel data may exist and can be used to build multilingual representations (Ruder et al., 2019; Artetxe et al., 2020). Second, instead of adapting to a particular target, there is some work on *domain generalization* aimed at building a single system which is robust on several known target domains. One example is the SANCL shared task (Petrov and McDonald, 2012), where participants were asked to build a single system that can robustly parse reviews, weblogs, answers, emails, newsgroups. In this setup, the DA problem boils down to finding a more robust system for given targets. It can be seen as optimizing for both in-domain and out-of-domain(s) accuracy.

If domains are unknown a priori, robustness can be taken a step further towards *out-of-domain generalization*, to unknown targets, the most challenging setup. A recent solution is *distributionally robust optimization* (Oren et al., 2019), i.e., optimizing for worst-case performance without the knowledge of the test distribution. To do so, it assumes a *subpopulation shift*, where the test population is a subpopulation mix of the training distribution. A model is then trained to do well over a wide range of potential test distributions. Some early work in dialogue (Bod, 1999) and parsing (Plank and Sima'an, 2008) adopted a similar idea of *subdomains*, however, with manually identified subpopulations. This bears some similarity to early work on leveraging general background knowledge (embeddings trained on general data) for domain adaptation (Plank and Moschitti, 2013; Nguyen and Grishman, 2015; Li et al., 2018a), and also relates to recent work on pre-training (Section 5.3). An alternative and complementary interesting line of research is to *predict test set performance* for new data varieties (Ravi et al., 2008; Van Asch and Daelemans, 2010; Elsahar and Gallé, 2019; Xia et al., 2020).

4 Model-centric approaches

Model-centric approaches redesign parts of the model: the feature space, the loss function or regularization and the structure of the model. We categorize them into feature-centric and loss-centric methods.

4.1 Feature-centric methods

Two lines of work can be found within feature-centric methods: *feature augmentation* and *feature generalization* methods. The former use *pivots* (common shared features) to construct an aligned feature space. The latter use *autoencoders* to find latent representations that transfer better across domains.

Pivots-based DA Seminal *pivot-based* methods include: *structural correspondence learning* (SCL) (Blitzer et al., 2006) and *spectral feature alignment* (SFA) (Pan et al., 2010). They both aim at finding features which are common across domains by using unlabeled data from both domains. The two approaches differ in the specifics of the method to construct the shared space. SCL uses auxiliary functions inspired by Ando and Zhang (2005), while SFA uses a graph-based spectral learning method. Creating domain-specific and domain-general features is the key idea of EasyAdapt (Daumé III, 2007), a seminal supervised DA method. A recent line of work (Ziser and Reichart, 2017; Ziser and Reichart, 2018a; Ziser and Reichart, 2018b; Ziser and Reichart, 2019) brings SCL back to neural networks.

In particular, Ziser and Reichart (2017) propose to combine the strengths of pivot-based methods with autoencoder neural networks in an *autoencoder structural correspondence learning* (AE-SCL) model. Autoencoders are used to learn latent representations to map non-pivots to pivots, and these encodings

Work	Method	classif./inference				struct. prediction			
		SA	LI	TC	NLI	POS	DEP	NER	RE
Model-centric:									
(Ziser and Reichart 2017; 2018a; 2018b; 2019)	Neural SCL	✓							
(Miller, 2019)	Neural SCL (Joint AE-SCL)	✓							
(Glorot et al., 2011)	SDA	✓							
(Chen et al., 2012)	MSDA	✓							
(Yang and Eisenstein, 2014)	MSDA					✓			
(Clinchant et al., 2016)	MSDA	✓		✓					
(Ganin et al., 2016)	DANN	✓							
(Li et al., 2017)	DANN+SCL MemNet	✓							
(Kim et al., 2017)	DANN/DSN			✓				✓	
(Sato et al., 2017)	DANN						✓		
(Wu et al., 2017)	DANN								✓
(Yasunaga et al., 2018)	DANN					✓			
(Shen et al., 2018)	DANN	✓							
(Li et al., 2018a)	DANN	✓	✓						
(Alam et al., 2018a)	DANN			✓					
(Wang et al., 2019)	DANN			✓					
(Shah et al., 2018)	DANN+Wasserstein			✓					
(Fu et al., 2017)	DANN								✓
(Rios et al., 2018)	DANN								✓
(Xu et al., 2019)	DANN			✓					
(Shi et al., 2018)	DSN (GSN)								✓
(Rocha and Lopes Cardoso, 2019)*	DANN, Shared encoders	✓			✓				
(Ghosal et al., 2020)	DANN (concept embeddings)	✓							
(Naik and Rose, 2020)	DANN (context embeddings)							✓	
Data-centric:									
(Ruder and Plank, 2018)	SSL, Multitask tri-training	✓				✓			
(Lim et al., 2020)	SSL					✓			
(Rotman and Reichart, 2019)	Deep self-training						✓		
(Han and Eisenstein, 2019)	AdaptaBERT◇					✓		✓	
(Li et al., 2019)	Adaptive pre-training						✓		
(Gururangan et al., 2020)	Adaptive pre-training (incl. multi-phase)	✓		✓	✓				✓
Hybrid:									
(Saito et al., 2017)	Asymmetric tri-training	✓							
(Desai et al., 2019)	Adaptive (temporal) ensembling			✓					
(Jia et al., 2019)	Cross-domain LM							✓	
(Cui and Bollegala, 2019)	SelfAdapt (pivots+co-training)	✓							
(Peng and Dredze, 2017)	Multi-task-DA◇					✓		✓	
(Guo et al., 2020)	DistanceNet-Bandit	✓							
(Ben-David et al., 2020)	PERL (pivots+context embeddings)	✓							

Table 1: Overview of neural UDA in NLP: method and task(s). Methods: SCL = structural correspondence learning; AE = autoencoder; SDA = stacked denoising autoencoder; MSDA = marginalized SDA; DANN = domain-adversarial neural network; DSN = domain separation network; GSN = genre separation network; SSL = semi-supervised learning; LM = language modeling. Tasks: SA = sentiment analysis; LI = language identification; TC = binary text classification (incl. machine reading, duplicate question detection, stance detection, intent classification, political data identification); NLI = natural language inference; POS = part-of-speech (incl. Chinese word segmentation); DEP = dependency parsing; NER = named entity recognition (incl. slot tagging, event trigger identification, named entity segmentation); RE = relation extraction. *with *cross-lingual* adaptation. ◇ applicable to UDA but main focus is supervised DA.

are then used to augment the training data. The main drawback of this approach is that the output vector representations of the text are unique and not context-dependent. To solve this problem, a *pivot-based language modeling* (PBLM) method has been proposed (Ziser and Reichart, 2018a; Ziser and Reichart, 2018b). PBLM effectively combines SCL with a neural language model based on long short-term memory (LSTM) networks which predicts the presence of pivots and non-pivots, thus making representations structure-aware. A weakness of the PBLM approach relies in the large number of pivots needed. To remedy this issue, Ziser and Reichart (2019) adopted a *task refinement learning* approach using PBLM (called TRL-PBLM), showing gains in both accuracy and stability over different hyperparameters selection choices. The approach is an iterative training process where the network is trained using an

increasingly larger amount of pivots. Recent hybrid UDA work extends pivots with contextual embeddings (Ben-David et al., 2020), as we discuss in Section 6.

A common issue with the aforementioned methods is that they involve two independent steps: one for representation learning and one for task learning. To tackle this issue, recent studies propose training the two tasks jointly (i.e., pivot prediction and sentiment) (Miller, 2019) and learn pivots *automatically* via attention (Li et al., 2017), similar to work on automatic non-pivot identification (Li et al., 2018b).

To the best of our knowledge, neural pivot-based UDA approaches have been solely applied to sentiment classification, cf. Table 1. Notably, Ziser and Reichart (2018a) went a step further, and applied neural SCL cross-lingually; the NLP task is still sentiment classification. The effectiveness of pivot-based methods in neural models remains to be tested. Early non-neural work applied SCL to structure prediction problems with mixed results, i.e., POS (Blitzer et al., 2006) and parsing (Plank, 2011).

Autoencoder-based DA Early neural approaches for UDA have been based on autoencoders. Autoencoders are neural networks that are employed to learn latent representations from raw data in an unsupervised fashion by learning with an input reconstruction loss. Motivated by the *denoising autoencoders* (Vincent et al., 2008), the first work in this line is by Glorot et al. (2011), who introduced the *stacked denoising autoencoder* (SDA) for domain adaptation. Basically, a SDA automatically learns a robust and unified feature representation for all domains by stacking multiple layers, and artificially corrupts the inputs with a Gaussian noise that the decoder needs to reconstruct. However, SDAs showed issues in speed and scalability to high-dimensional data. To mitigate these limitations, a more efficient *marginalized stacked denoising autoencoder* (MSDA) that marginalizes the noise was proposed (Chen et al., 2012). MSDAs have been further extended by Yang and Eisenstein (2014) with marginalized structured dropout, and by Clinchant et al. (2016), which improved the regularization of MSDAs following the insights from the domain adversarial training of neural networks (Ganin and Lempitsky, 2015; Ganin et al., 2016) (described in Section 4.2). The main drawback of autoencoder approaches is that the induced representations do not make use of any linguistic information.

4.2 Loss-centric methods

Loss-centric approaches can be divided into methods which employ domain adversaries, and instance-level reweighting methods. We outline these two strands of work in the following.

Domain adversaries The most widespread methods for neural UDA are based on the use of *domain adversaries* (Ganin and Lempitsky, 2015; Ganin et al., 2016). Inspired by the way generative adversarial networks (GANs) (Goodfellow et al., 2014) minimize the discrepancies between training and synthetic data distributions, domain adversarial training aims at learning latent feature representations that serve at reducing the discrepancy between the source and target distributions. The intuition behind these methods puts its ground on the theory on domain adaptation (Ben-David et al., 2010), which argues that cross-domain generalization can be achieved by means of feature representations for which the origin (domain) of the input example cannot be identified.

The seminal approach in this category are DANNs: *domain-adversarial neural networks* (Ganin and Lempitsky, 2015; Ganin et al., 2016). The aim is to estimate an accurate predictor for the task while maximizing the confusion of an auxiliary domain classifier in distinguishing features from the source or the target domain. To learn domain-invariant feature representations, DANNs employ a loss function via a *gradient reversal layer* which ensures that feature distributions in the source and target domains are made similar. The strength of this approach is in its scalability and generality; however, DANNs only model feature representations that are shared across both domains, and suffer from a vanishing gradient problem when the domain classifier accurately discriminates source and target representations (Shen et al., 2018). *Wasserstein* methods (Martin Arjovsky and Bottou, 2017) are more stable training methods than gradient reversal layers. Instead of learning a classifier to distinguish domains, they attempt to reduce the approximated Wasserstein distance (also known as Earth Mover's Distance). A recent study on question pair classification shows that the two adversarial methods reach similar performance, but Wasserstein enables more stable training (Shah et al., 2018).

DANNs have been applied in many NLP tasks in the last few years, mainly to sentiment classification (e.g., Ganin et al. (2016), Li et al. (2018a), Shen et al. (2018), Rocha and Lopes Cardoso (2019), Ghoshal et al. (2020), to name a few), but recently to many other tasks as well: language identification (Li et al., 2018a), natural language inference (Rocha and Lopes Cardoso, 2019), POS tagging (Yasunaga et al., 2018), parsing (Sato et al., 2017), trigger identification (Naik and Rose, 2020), relation extraction (Wu et al., 2017; Fu et al., 2017; Rios et al., 2018), and other (binary) text classification tasks like relevancy identification (Alam et al., 2018a), machine reading comprehension (Wang et al., 2019), stance detection (Xu et al., 2019), and duplicate question detection (Shah et al., 2018). This makes DANNs the most widely used UDA approach in NLP, as illustrated in Table 1.

To model features that also belong to either the source or target domain, *domain separation networks* (DSNs) (Bousmalis et al., 2016) have been proposed. DSNs separate latent representations in i) separate private encoders (i.e., one for each domain) and ii) a shared encoder (in charge to reconstruct the input instance using these representations). This bears similarities to a traditional supervised method (Daumé III, 2007). The main drawback of DSNs is that domain-specific representations are solely used in the decoder, leaving the classifier to be trained on the domain-invariant representations only.

DSNs have seen a notable success in Computer Vision (CV) (Bousmalis et al., 2016). In NLP, Shi et al. (2018) propose the *genre separation networks* (GSNs) as a variant of the DSNs, introducing a novel reconstruction component that leverages both shared and private feature representations in the learning process. As noted also by Han and Eisenstein (2019), a downside of adversarial methods is that they require careful balancing between objectives (Kim et al., 2017; Alam et al., 2018a) to avoid instability during learning (Arjovsky et al., 2017).

Reweighting This family of methods is an instance-level adaptation method. The core idea of *instance weighting* (also known as importance weighting) is to assign a weight to each training instance proportional to its similarity to the target domain (Jiang and Zhai, 2007). We can see instance weighting as an alternative to domain adversaries. While domain adversaries distinguish the domains to learn domain invariant representations in a joint model, instance weighting decouples domain detection for a-priori weight estimation of an instance.

Methods that explicitly reweight the loss based on domain discrepancy information include *maximum mean discrepancy* (MMD) (Gretton et al., 2007) and its more efficient version called *kernel mean matching* (KMM) (Gretton et al., 2009). KMM reweights the training instances such that the means of the training and test points in reproducing a kernel Hilbert space are close to each other. Jiang and Zhai (2007) introduced instance weighting in NLP and proposed to learn weights by first training domain classifiers. The effectiveness of the method in neural setups remains to be seen. An early study reports non-significant improvements for POS tagging (Plank et al., 2014b).

5 Data-centric methods

Recently, data-centric approaches are on a rise, due to rapid growth of data and the gain in popularity of pre-training methods. We summarize data-centric strands next, which differ whether they use pseudo-labeling, select relevant data or use large unlabeled data or auxiliary tasks for model pre-training.

5.1 Pseudo-labeling

The main idea of pseudo-labeling is to apply a trained classifier to predict labels on unlabeled instances, which are then treated as 'pseudo' gold labels. Pseudo-labeling applies semi-supervised methods (Abney, 2007; Zhu and Goldberg, 2009) such as bootstrapping methods like self-training, co-training and tri-training or methods such as temporal ensembling (Charniak, 1997; McClosky et al., 2006; Blum and Mitchell, 1998; Steedman et al., 2003; Zhou and Li, 2005; Søgaard and Rishøj, 2010; Saito et al., 2017; Laine and Aila, 2016) by using either the same model, a teacher model, or multiple bootstrap models which may include slower but more accurate hand-crafted models (Petrov et al., 2010) to guide pseudo-labeling. Most pseudo-labeling works date back to traditional non-neural learning methods. Bootstrapping methods for domain adaptation are well-studied in parsing (McClosky et al., 2006;

Reichart and Rappoport, 2007; Yu et al., 2015). They include models trained on other grammar formalisms to improve dependency parsing on Twitter (Foster et al., 2011). Recently, this line of classics has been revisited (Ruder and Plank, 2018; Rotman and Reichart, 2019; Lim et al., 2020). For example, classic methods such as tri-training constitute a strong baseline for domain shift in neural times (Ruder and Plank, 2018). Pseudo-labeling has recently been studied for parsing with contextualized word representations (Rotman and Reichart, 2019; Lim et al., 2020) and a recent work proposes *adaptive ensembling* (Desai et al., 2019) as extension of temporal ensembling (see *hybrid* methods in Section 6).

5.2 Data selection

A relatively unexplored area is data selection for adaptation, which is gaining traction again in light of large pre-trained models (which data should they be trained on?) and the related problem of cross-lingual learning (what is/are the best source language(s) to transfer from?). Data selection aims to select the best matching data for a new domain, typically by using perplexity (Moore and Lewis, 2010) or using domain similarity measures such as Jensen-Shannon divergence over term or topic distributions (Plank and van Noord, 2011). This has mostly been studied for MT (Moore and Lewis, 2010; Axelrod et al., 2011; van der Wees et al., 2017; Aharoni and Goldberg, 2020), but also for parsing (Plank and van Noord, 2011; Ruder and Plank, 2017) and sentiment analysis (Remus, 2012) though for supervised domain adaptation setups only. For parsing and sentiment analysis, the simple Jensen-Shannon divergence on term distribution constitutes a strong baseline (Plank, 2011; Ruder and Plank, 2017). Within MT, van der Wees et al. (2017) propose a dynamic data selection approach which changes the subset of data in each epoch for MT. Data selection is gaining attention, in light of the abundance of data. Recent work investigates data representation and cosine similarity for MT data selection (Aharoni and Goldberg, 2020). Similarly, distance metrics have been been recently used for multi-source domain adaptation of sentiment classification models using a bandit-based approach (Guo et al., 2020). For morphosyntactic cross-lingual work, simple overlap metrics are indicative (Üstün et al., 2019; Lin et al., 2019). Another line explores whether tailoring large pre-trained models to the domain of a target task is still beneficial, and use of data selection to overcome costly expert selection. They propose two multi-phase pre-training methods (Gururangan et al., 2020) (as discussed further below) with promising results on text classification tasks.

5.3 Pre-training—And:—Is bigger better? Are domains (or: *varieties*) still relevant?

Large pre-trained models have become ubiquitous in NLP (Howard and Ruder, 2018; Peters et al., 2018; Devlin et al., 2019). *Fine-tuning* a transformer-based model with a small amount of labeled data often reaches high performance across NLP tasks and has become a de-facto standard. It means starting from the pre-trained model weights and training a new task-specific layer on supervised data. A natural question which arises is how universal such large models are. Is bigger better? And are domains (or varieties) still relevant? We return to these questions after depicting pre-training strategies. We delineate:

1. **Pre-training**: pre-training alone (e.g., multilingual BERT; language-specfic BERTs from scratch);
2. **Adaptive pre-training**: This encompasses pre-training, followed by secondary stages of pre-training on unlabeled data or on labeled data from intermediate higher-resource auxiliary tasks:
   (a) **Multi-phase pre-training**: two or more phases of secondary pre-training, from broad-coverage to domain-/task-adaptive pre-training (i.e., BioBERT, AdaptaBERT, DAPT, TAPT). They differ by the source of unlabeled data: broad-domain $\succ$ domain-specific $\succ$ task-specific;
   (b) **Auxiliary-task pre-training**: pre-training, followed by (possibly multiple stages of) *auxiliary-task* pre-training (e.g., supplementary training on intermediate labeled-data tasks, STILTs).

Pre-training (option 1) can be seen as straightforward adaptation, analogous to zero-shot in cross-lingual learning. The key idea is to train encoders with self-supervised objectives like (masked) language model and related unsupervised objectives (Peters et al., 2018; Devlin et al., 2019; Beltagy et al., 2019).

In light of a domain shift, *adaptive pre-training* is beneficial, in which in one instantiation contextualized embeddings are adapted to text from the target domain by masked language modeling, as introduced

by Han and Eisenstein (2019). More broadly, we distinguish two variants of *adaptive pre-training*. They differ whether unlabeled data or some form of auxiliary labeled data (or intermediate tasks data) is used. These variants can also be combined, and fine-tuning applies to all setups, if data is available. The key idea of *multi-phase pre-training* (option 2a) is to use secondary-stage unsupervised pre-training, such as broad-coverage domain-specific BERT variants (e.g., BioBERT). Gururangan et al. (2020) propose *domain-adaptive pre-training* (DAPT) from a broader corpus, compared to (Han and Eisenstein, 2019), and *task-specific pre-training* (TAPT) which uses unlabeled data closer-and-closer to the task distribution. As these studies show, domain-relevant data is important for pre-training (Han and Eisenstein, 2019; Gururangan et al., 2020) in both high and low resource setups. Similar adaptive pre-training work has been shown to be effective for dependency parsing (Li et al., 2019). This suggests that there exists a spectrum of domains of varying granularity, confirming ideas around domain similarity (Plank, 2011; Baldwin et al., 2013). Domains (*varieties*) do still matter in today's models.

An alternative line of work (option 2b) is *auxiliary-task pre-training* and use labeled auxiliary tasks either via multi-task learning (MTL) (Peng and Dredze, 2017) or intermediate-task transfer (Phang et al., 2018; Phang et al., 2020). The latter proposed *supplementary training on intermediate labeled-data tasks for transfer* (STILT) (Phang et al., 2018), and recently adopted this idea to cross-lingual learning, where English is used as intermediate-task for zero-shot transfer (Phang et al., 2020).

The choice of data used for pre-training (or the auxiliary tasks) do matter. Current transformer models are trained on either large general data like BookCorpus and Wikipedia in BERT (Devlin et al., 2019) or target-specific samples, like papers from Semantic Scholar in SciBERT (Beltagy et al., 2019), and PubMed abstracts and PMC full-text articles in BioBERT (Lee et al., 2020). What denotes *relevant* data is an open question. Today, it is either general background knowledge, domain-specific target data, or a combination thereof, possibly via auxiliary tasks or intermediate training stages. Most of these have been carefully selected manually, raising interesting connections to data selection (Section 5.2) and finding better curricula (Tsvetkov et al., 2016) to learn under domain shift (Ruder and Plank, 2017).

While large pre-trained models have shown to work well, many questions and challenges remain. Recent work has shown that these models degrade on out-of-domain data, maximum likelihood training makes them too over-confident (Oren et al., 2019) and particularly calibration is important for out-of-domain generalization (Hendrycks et al., 2020). An acknowledged issue with fine-tuning is the brittleness of the process (Phang et al., 2018; Dodge et al., 2020). Even with the same hyperparameters, distinct runs can lead to drastically different results and training data order and seed choice have a considerably impact (Dodge et al., 2020). Deeper investigations into what such models capture, how they can be robustly trained in light of known test distributions or out-of-domain conditions are interesting issues.

6 Hybrid approaches

Work on the intersection of data-centric and model-centric methods can be plentiful. It currently includes combining semi-supervised objectives with an adversarial loss (Lim et al., 2020; Alam et al., 2018b), combining pivot-based approaches with pseudo-labeling (Cui and Bollegala, 2019) and very recently with contextualized word embeddings (Ben-David et al., 2020), and combining multi-task approaches with domain shift (Jia et al., 2019), multi-task learning with pseudo-labeling (multi-task tri-training) (Ruder and Plank, 2018), and *adaptive ensembling* (Desai et al., 2019), which uses a student-teacher network with a consistency-based self-ensembling loss and a temporal curriculum. They apply adaptive ensembling to study temporal and topic drift in political data classification (Desai et al., 2019).

7 Challenges and future directions

While recent work has made important progress in neural UDA, our survey reveals i) an over-representation and bias of work on sentiment analysis (cf. column bias in Table 1) and ii) a general lack of testing across tasks (row sparsity in Table 1) and multiple adaptation methods.

Comprehensive UDA benchmarks Concretely, we recommend a) to create new benchmarks for UDA with multiple tasks and of increasing complexity, setups beyond 1:1 adaptation, and datasets which

document known *variety facets* of the data (Bender and Friedman, 2018). This will help to learn about the known and unknown (Section 3) as 'variety' (domain) matters; b) to release unlabeled data from the broader distribution from which annotated data was sampled, in line with Gururangan et al. (2020); this allows studying diachronic effects, as labeled evaluation data lacks diversity in terms of topics and time (Desai et al., 2019; Derczynski et al., 2016); and c) to release unaggregated, multiple annotations to study divergences in annotations (Plank et al., 2014a).

Back to the roots and how knowledge transfers Revisiting classics in neural times is beneficial, as shown for example in recent work which brings back SCL and pseudo-labeling methods (see Table 1), but much is left to see how these methods generalize. This can be linked to the question on what representations capture (Belinkov and Glass, 2019) and how knowledge transfers (Rethmeier et al., 2020).

X scarcity Even unlabeled data can be scarce (X scarcity), particularly in highly-specialized language varieties (e.g., clinical data) (Rethmeier and Plank, 2019). This is often due to data sharing restrictions. In some cases, only a trained source model could be available instead of raw or labeled texts (Laparra et al., 2020). Together with the quest for more efficient learning methods, the general question of how to adapt in light of X scarcity or absence becomes important.

8 Conclusion

In this survey, we review strands of unsupervised domain adaptation, summarized into *model-centric*, *data-centric*, and *hybrid* methods, including trends in pre-training. We also revisit the notion of *domain* and suggest to use the term *variety* instead, to better capture the multitude of dimensions of variation. Our survey identifies a limited focus on sentiment benchmarks and single-task evaluation for UDA. Lastly, we outline future directions, linking to the broader challenges related to learning beyond 1:1 scenarios and out-of-distribution generalization. This also calls for new directions on benchmarks and learning under scarce data.

Acknowledgements

We thank Nils Rethmeier, Raffaella Bernardi, Rob van der Goot and Sebastian Ruder for precious feedback on earlier drafts of this survey. This research is supported by a visit grant to Alan supported by COSBI and a research leader sapere aude grant to Barbara by the Independent Research Fund Denmark (Danmarks Frie Forskningsfond, grant number 9063-00077B MultiVaLUe).

References

Steven Abney. 2007. *Semisupervised learning for computational linguistics*. CRC press, 1st edition.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online, July. Association for Computational Linguistics.

Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018a. Domain Adaptation with Adversarial Training and Graph Embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 1077–1087.

Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018b. Domain adaptation with adversarial training and graph embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1077–1087, Melbourne, Australia, July. Association for Computational Linguistics.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug. PMLR.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online, July. Association for Computational Linguistics.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November. Association for Computational Linguistics.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175.

Eyal Ben-David, Carmel Rabinovitz, and Roi Reichart. 2020. Perl: Pivot-based domain adaptation for pre-trained deep contextualized embedding models. *Transactions of the Association for Computational Linguistics (To Appear)*.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Yoshua Bengio. 2019. Deep learning for ai. Turing Award Lecture, Heidelberg Laureate Forum. `http://www.iro.umontreal.ca/~bengioy/HLF-Turing-23sept2019.pdf`. Last accessed on 2020-05-26.

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

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

Rens Bod. 1999. Context-sensitive spoken dialogue processing with the dop model. *Natural Language Engineering*, 5(4):309–323.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain Separation Networks. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 343–351.

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. *AAAI/IAAI*, 2005(598-603):18.

Minmin Chen, Kilian Q Weinberger, Fei Sha, and Los Angeles. 2012. Marginalized Denoising Autoencoders for Domain Adaptation. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 767–774.

Chenhui Chu and Rui Wang. 2018. A Survey of Domain Adaptation for Neural Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319.

Stephane Clinchant, Gabriela Csurka, and Boris Chidlovskii. 2016. A Domain Adaptation Regularization for Denoising Autoencoders. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 26–31.

Gabriela Csurka. 2017. A comprehensive survey on domain adaptation for visual applications. In *Domain Adaptation in Computer Vision Applications*, pages 1–35. Springer International Publishing, Cham.

Xia Cui and Danushka Bollegala. 2019. Self-adaptation for unsupervised domain adaptation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 213–222, Varna, Bulgaria, September. INCOMA Ltd.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Shrey Desai, Barea Sinno, Alex Rosenfeld, and Junyi Jessy Li. 2019. Adaptive ensembling: Unsupervised domain adaptation for political document analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4718–4730, Hong Kong, China, November.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 359–369.

Hady Elsahar and Matthias Gallé. 2019. To annotate or not? predicting performance drop under domain shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173, Hong Kong, China, November. Association for Computational Linguistics.

Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef Van Genabith. 2011. # hardtoparse: Pos tagging and parsing the twitterverse. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

W. N. Francis and H. Kucera. 1979. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Lisheng Fu, Thien Huu Nguyen, Min Bonan, and Ralph Grishman. 2017. Domain Adaptation for Relation Extraction with Domain Adversarial Neural Network. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, pages 425–429.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1180–1189.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17:1–35.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China, November. Association for Computational Linguistics.

Deepanway Ghosal, Devamanyu Hazarika, Abhinaba Roy, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2020. KinGDOM: Knowledge-Guided DOMain Adaptation for Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3198–3210, Online, July. Association for Computational Linguistics.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 513–520.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pages 2672–2680.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. 2007. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. 2009. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2020. Multi-source domain adaptation for text classification via distancenet-bandits. In *AAAI*, pages 7830–7838.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July. Association for Computational Linguistics.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China, November. Association for Computational Linguistics.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China, July. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July. Association for Computational Linguistics.

Nancy Ide and Keith Suderman. 2004. The American national corpus first release. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, Florence, Italy, July. Association for Computational Linguistics.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271. Association for Computational Linguistics.

Jing Jiang. 2008. A literature survey on domain adaptation of statistical classifiers. `http://www.mysmu.edu/faculty/jingjiang/papers/da_survey.pdf`, Last accessed on 2020-05-26.

Young-Bum Kim, Karl Stratos, and Dongchan Kim. 2017. Adversarial Adaptation of Synthetic or Stale Data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, July. Association for Computational Linguistics.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. 2020. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*.

Samuli Laine and Timo Aila. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Egoitz Laparra, Steven Bethard, and Timothy A Miller. 2020. Rethinking domain adaptation for machine learning over clinical language. *JAMIA Open*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-End Adversarial Memory Network for Cross-domain Sentiment Classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2237–2243, Melbourne, Australia, August. International Joint Conferences on Artificial Intelligence Organization.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018a. What's in a Domain? Learning Domain-Robust Text Representations using Adversarial Training. In *Proceedings of NAACL-HLT 2018*, pages 474–479.

Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018b. Hierarchical attention transfer network for cross-domain sentiment classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhenghua Li, Xue Peng, Min Zhang, Rui Wang, and Luo Si. 2019. Semi-supervised Domain Adaptation for Dependency Parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July. Association for Computational Linguistics.

KyungTae Lim, Jay Yoon Lee, Jaime Carbonell, and Thierry Poibeau. 2020. Semi-Supervised Learning on Meta Structure: Multi-Task Tagging and Parsing in Low-Resource Scenarios. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy, July. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Anna Margolis. 2011. A literature review of domain adaptation with unlabeled data. Tech. Rep., University of Washington (USA).

SC Martin Arjovsky and Leon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34 th International Conference on Machine Learning*, Sydney, Australia.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Reranking and Self-Training for Parser Adaptation. *International Conference on Computational Linguistics (COLING) and Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 337–344, July.

Timothy Miller. 2019. Simplified Neural Unsupervised Domain Adaptation. In *Proceedings of NAACL-HLT 2019*, pages 414–419.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden, July. Association for Computational Linguistics.

Aakanksha Naik and Carolyn Rose. 2020. Towards open domain event trigger identification using adversarial domain adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7618–7624, Online, July. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation Extraction: Perspective from Convolutional Neural Networks. In *Proceedings of NAACL-HLT 2015*, pages 39–48, Denver, Colorado.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593.

Yonatan Oren, Shiori Sagawa, Tatsunori Hashimoto, and Percy Liang. 2019. Distributionally robust language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4227–4237, Hong Kong, China, November. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Sinno Jialin Pan, Xiaochuan Ni, Jian-tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-Domain Sentiment Classification via Spectral Feature Alignment. In *Proceedings of the 19th international conference on World wide web. ACM*, pages 751–760.

Vishal M. Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. 2015. Visual Domain Adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69.

Nanyun Peng and Mark Dredze. 2017. Multi-task domain adaptation for sequence tagging. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 91–100, Vancouver, Canada, August. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on Syntactic Analysis of NonCanonical Language (SANCL)*.

Slav Petrov, Pi-Chuan Chang, Michael Ringgaard, and Hiyan Alshawi. 2010. Uptraining for accurate deterministic question parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 705–713, Cambridge, MA, October. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Jason Phang, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, Iacer Calixto, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. *arXiv preprint arXiv:2005.13013*.

Barbara Plank and Alessandro Moschitti. 2013. Embedding Semantic Similarity in Tree Kernels for Domain Adaptation of Relation Extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1498–1507.

Barbara Plank and Khalil Sima'an. 2008. Subdomain sensitive statistical parsing using raw corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576, Portland, Oregon, USA, June. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014a. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland, June. Association for Computational Linguistics.

Barbara Plank, Anders Johannsen, and Anders Søgaard. 2014b. Importance weighting and unsupervised domain adaptation of POS taggers: a negative result. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–973, Doha, Qatar, October. Association for Computational Linguistics.

Barbara Plank. 2011. *Domain Adaptation for Parsing*. Ph.D. thesis, University of Groningen.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*.

Sujith Ravi, Kevin Knight, and Radu Soricut. 2008. Automatic prediction of parser accuracy. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 887–896, Honolulu, Hawaii, October. Association for Computational Linguistics.

Roi Reichart and Ari Rappoport. 2007. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 616–623.

Robert Remus. 2012. Domain adaptation using domain similarity- and domain complexity-based instance selection for cross-domain sentiment analysis. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops*, ICDMW '12, page 717–723, USA. IEEE Computer Society.

Nils Rethmeier and Barbara Plank. 2019. MoRTy: Unsupervised learning of task-specialized word embeddings by autoencoding. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 49–54, Florence, Italy, August. Association for Computational Linguistics.

Nils Rethmeier, Vageesh Kumar Saxena, and Isabelle Augenstein. 2020. TX-Ray: Quantifying and explaining model-knowledge transfer in (un-)supervised nlp. In *UAI*.

Anthony Rios, Ramakanth Kavuluru, and Zhiyong Lu. 2018. Generalizing biomedical relation classification with neural adversarial domain adaptation. *Bioinformatics*, 34(17):2973–2981.

Gil Rocha and Henrique Lopes Cardoso. 2019. A comparative analysis of unsupervised language adaptation methods. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 11–21, Hong Kong, China, November. Association for Computational Linguistics.

Guy Rotman and Roi Reichart. 2019. Deep contextualized self-training for low resource dependency parsing. *Transactions of the Association for Computational Linguistics*, 7(0):695–713.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382, Copenhagen, Denmark, September. Association for Computational Linguistics.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054, Melbourne, Australia, July. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Sebastian Ruder. 2019. *Neural Transfer Learning for Natural Language Processing*. Ph.D. thesis, National University of Ireland, Galway.

Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 2988–2997.

Motoki Sato, Hitoshi Manabe, Hiroshi Noji, and Yuji Matsumoto. 2017. Adversarial training for cross-domain universal dependency parsing. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 71–79, Vancouver, Canada, August. Association for Computational Linguistics.

Darsh J Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. 2018. Adversarial Domain Adaptation for Duplicate Question Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1056–1063.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online, July. Association for Computational Linguistics.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein Distance Guided Representation Learning for Domain Adaptation. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 4058–4065.

Ge Shi, Chong Feng, Lifu Huang, Boliang Zhang, Heng Ji, Lejian Liao, and Heyan Huang. 2018. Genre Separation Network with Adversarial Training for Cross-genre Relation Extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1018–1023.

Anders Søgaard and Christian Rishøj. 2010. Semi-supervised dependency parsing using generalized tri-training. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1065–1073.

Mark Steedman, Rebecca Hwa, Stephen Clark, Miles Osborne, Anoop Sarkar, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003. Example selection for bootstrapping statistical parsers. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Brian MacWhinney, and Chris Dyer. 2016. Learning the curriculum with Bayesian optimization for task-specific word representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 130–139, Berlin, Germany, August. Association for Computational Linguistics.

Ahmet Üstün, Rob van der Goot, Gosse Bouma, and Gertjan van Noord. 2019. Multi-team: A multi-attention, multi-decoder approach to morphological analysis. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 35–49, Florence, Italy, August. Association for Computational Linguistics.

Vincent Van Asch and Walter Daelemans. 2010. Using domain similarity for performance estimation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 31–36, Uppsala, Sweden, July. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark, September. Association for Computational Linguistics.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and Composing Robust Features with Denoising. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. 2018. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, pages 5334–5344.

Huazheng Wang, Zhe Gan, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, and Hongning Wang. 2019. Adversarial Domain Adaptation for Machine Reading Comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2510–2520, Hong Kong, China, November. Association for Computational Linguistics.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682, Suntec, Singapore, August. Association for Computational Linguistics.

Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big Data*, 3(1):9.

Garrett Wilson and Diane J. Cook. 2020. A Survey of Unsupervised Deep Domain Adaptation. *arXiv preprint, arXiv:1812.02849*, February.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783, Copenhagen, Denmark, September. Association for Computational Linguistics.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. 2020. Predicting performance for natural language processing tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8625–8646. Association for Computational Linguistics, July.

Brian Xu, Mitra Mohtarami, and James Glass. 2019. Adversarial Domain Adaptation for Stance Detection. *arXiv preprint arXiv:1902.02401*, February.

Yi Yang and Jacob Eisenstein. 2014. Fast Easy Unsupervised Domain Adaptation with Marginalized Structured Dropout. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 538–544, Baltimore, Maryland, June. Association for Computational Linguistics.

Qiang Yang, Yu Zhang, Wenyuan Dai, and Sinno Jialin Pan. 2020. *Transfer Learning*. Cambridge University Press.

Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. 2018. Robust multilingual part-of-speech tagging via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 976–986, New Orleans, Louisiana, June. Association for Computational Linguistics.

Juntao Yu, Mohab Elkaref, and Bernd Bohnet. 2015. Domain adaptation for dependency parsing via self-training. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 1–10, Bilbao, Spain, July. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zhi-Hua Zhou and Ming Li. 2005. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541.

Xiaojin Zhu and Andrew B Goldberg. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.

Yftah Ziser and Roi Reichart. 2017. Neural Structural Correspondence Learning for Domain Adaptation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 400–410.

Yftah Ziser and Roi Reichart. 2018a. Deep Pivot-Based Modeling for Cross-language Cross-domain Transfer with Minimal Guidance. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 238–249.

Yftah Ziser and Roi Reichart. 2018b. Pivot Based Language Modeling for Improved Neural Domain Adaptation. In *Proceedings of NAACL-HLT 2018*, pages 1241–1251.

Yftah Ziser and Roi Reichart. 2019. Task Refinement Learning for Improved Accuracy and Stability of Unsupervised Domain Adaptation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.