Domain Divergences: A Survey and Empirical Analysis

Abhinav Ramesh Kashyap[†], **Devamanyu Hazarika**[†], **Min-Yen Kan**[†], **Roger Zimmermann** [†]
[†]National University of Singapore, Singapore
{abhinav, hazarika, kanmy, rogerz}@comp.nus.edu.sg

Abstract

Domain divergence plays a significant role in estimating the performance of a model in new domains. While there is a significant literature on divergence measures, researchers find it hard to choose an appropriate divergence for a given NLP application. We address this shortcoming by both surveying the literature and through an empirical study. We develop a taxonomy of divergence measures consisting of three classes — Information-theoretic, Geometric, and Higher-order measures and identify the relationships between them. Further, to understand the common use-cases of these measures, we recognise three novel applications – 1) Data Selection, 2) Learning Representation, and 3) Decisions in the Wild – and use it to organise our literature. From this, we identify that Information-theoretic measures are prevalent for 1) and 3), and Higher-order measures are more common for 2). To further help researchers choose appropriate measures to predict drop in performance – an important aspect of Decisions in the Wild, we perform correlation analysis spanning 130 domain adaptation scenarios, 3 varied NLP tasks and 12 divergence measures identified from our survey. To calculate these divergences, we consider the current contextual word representations (CWR) and contrast with the older distributed representations. We find that traditional measures over word distributions still serve as strong baselines, while higher-order measures with CWR are effective.

1 Introduction

Standard machine learning models do not perform well when tested on data from a different target domain. The performance in a target domain largely depends on the domain divergence (Ben-David et al., 2010) – a notion of distance between the two domains. Thus, efficiently measuring and reducing divergence is crucial for adapting models to the new domain — the topic of *domain adaptation*. Divergence also has practical applications in predicting the performance drop of a model when adapted to new domains (Van Asch and Daelemans, 2010), and in choosing among alternate models (Xia et al., 2020).

Given its importance, researchers have invested much effort to define and measure domain divergence. Linguists use register variation to capture varieties in text – the difference between distributions of the prevalent features in two registers (Biber and Conrad, 2009). Other measures include probabilistic measures like $\mathcal{H}$-divergence (Ben-David et al., 2010), information theoretic measures like Jenssen-Shannon and Kullback-Leibler divergence (Plank and van Noord, 2011; Van Asch and Daelemans, 2010) and measures using higher-order moments of random variables like Maximum Mean Discrepancy (MMD) and Central Moment Discrepancy (CMD) (Gretton et al., 2007; Zellinger et al., 2017). The proliferation of divergence measures challenges researchers in choosing an appropriate measure for a given application.

To help guide best practices, we first comprehensively review the NLP literature on domain divergences. Unlike previous surveys, which focus on domain adaptation for specific tasks such as machine translation (Chu and Wang, 2018) and statistical (non-neural network) models (Jiang, 2007; Margolis, 2011), our work takes a different perspective. We study domain adaptation through the vehicle of *domain divergence measures*. First, we develop a taxonomy of divergence measures consisting of three groups: Information-Theoretic, Geometric, and Higher-Order measures. Further, to find the most common group used in NLP, we recognise three novel application areas of these divergences — Data Selection, Learning Representations, and Decisions in the Wild and organise the literature under them. We find that Information-Theoretic measures over word distributions are popular for Data Selection and Decisions in the wild, while Higher-order measures over continuous features are frequent for Learning representations.

Domain divergence is a major predictor of performance in the target domain. A better domain divergence metric ideally predicts the corresponding performance drop of a model when applied to a target domain – a practical and important component of *Decisions in the Wild*. We further help researchers identify appropriate measures for predicting performance drops, through a correlation analysis over 130 domain adaptation scenarios and three standard NLP tasks: Part of Speech Tagging (POS), Named Entity Recognition (NER), and Sentiment Analysis and 12 divergence metrics from our literature review. While information-theoretic measures over traditional word distributions are popular in the literature, are higher-order measures calculated over modern contextual word representations better indicators of performance drop? We indeed find that higher-order measures are superior, but traditional measures are still reliable indicators of performance drop. The closest to our work is (Elsahar and Gallé, 2019) who perform a correlation analysis. However, they do not compare against different divergence measures from the literature. Comparatively, we consider more tasks and divergence measures.

In summary, our contributions are:

- We review the literature from the perspective of domain divergences and their use-cases in NLP.
- We aid researchers to select appropriate divergence measure that indicate performance-drops, an important application of divergence measures.

2 A Taxonomy of Divergence Measures

We devise a taxonomy for domain divergence measures, shown in Figure 1. It contains three main classes. Individual measures belong to a single class, where relationships can exist between measures from different classes. We provide detailed description of individual measures in Appendix A.

Geometric measures calculate the distance between two vectors in a metric space. As a divergence measure, they calculate the distance between features ($tf.idf$, continuous representations, etc.) extracted from instances of different domains. The P-norm is a generic form of the distance between two vectors, where Manhattan *(p=1)* and Euclidean distance *(p=2)* are common. Cosine (Cos) uses the cosine of the angle between two vectors to measure similarity and 1-Cos measures distance. Geometric measures are easy to calculate, but are ineffective in a high dimensional space as all distances appear the same (Aggarwal et al., 2001).

Information-theoretic measures captures the distance between probability distributions. For example, cross entropy over n-gram word distributions are extensively used in domain adaptation for machine translation. f-divergence (Csiszár, 1972) is a general family of divergences where f is a convex function. Different formulations of the f function lead to KL and JS divergence. Chen and Cardie (2018) show that reducing f-divergence measure is equivalent to reducing the PAD measures (see next section). Another special case of f-divergence is the family of α divergences, where KL-Div is a special case of α divergence. Renyi Divergence is a member of the α-divergences and tends towards KL-Div as $\alpha \rightarrow 1$ (Edge Ⓐ); Often applied to optimal transport problems, Wasserstein distance measures the amount of work needed to convert one probability distribution to the other as distance and is used extensively for domain adaptation. KL-Div is also related to Cross Entropy (CE). In this paper, CE refers to measures based on entropy.

Higher-Order measures consider matching higher order moments of random variables or divergence in a projected space. Their properties are amenable to end-to-end learning based domain adaptation and recently have been extensively adopted. Maximum Mean Discrepancy (MMD) is one such measure which considers matching first order moments of variables in a Reproducible Kernel Hilbert Space. On the other hand, CORAL (Sun et al., 2017) considers second order moments and CMD (Zellinger et al., 2017) considers higher order moments. CORAL and CMD are desirable because they avoid computationally expensive kernel matrix computations. KL-Div can also be considered as matching the first-order moment (Zellinger et al., 2017); Edge Ⓑ. Proxy-A-Distance (PAD) measures the distance between source and target distributions via the error of a classifier in target domain samples as source domain samples (Ben-David et al., 2007).

A few other measures do not have ample support in the literature. These include information-theoretic measures such as Bhattacharya coefficient, higher-order measures like PAD* (Elsahar and Gallé, 2019), Word Vector Variance (WVV), and Term Vocabulary Overlap (TVO) (Dai et al., 2019). Our taxonomy synthesises the diversity and

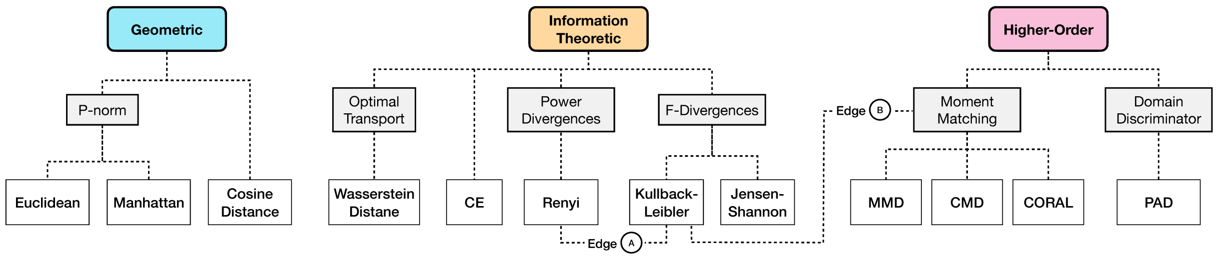


Figure 1: Taxonomy for divergence measures. i) **Geometric** measures the distance between vectors in a metric space ii) **Information- theoretic** measures the distance between probability distributions and iii) **Higher-order** measures the distance between distributions considering higher moments or the distance between representations or their projections in a nonlinear space. Edge Ⓐ indicates that Renyi divergence tends towards KL divergence as $\alpha \to 1$ and Edge Ⓑ indicates KL-Div can be considered as matching first-order moment.

the prevalence of the divergence measures in NLP.

3 Applications of Divergence Measures

Our key observation of the literature is that there are three primary families of applications of divergences (cf. Table 1 in the appendix): (i) **Data Selection**: selects a subset of text from a source domain that shares similar characteristics as target domain. The selected subset is then used to learn a target domain model. (ii) **Learning Representations**: aligns source and target domain distributions and learn domain-invariant representations. (iii) **Decisions in the Wild**: helps practitioners predict the performance or drops in performance of a model in a new target domain.

We limit the scope our survey to works that focus on divergence measures. We only consider unsupervised domain adaptation (UDA) – where there is no annotated data available in the target domain. It is more practical yet more challenging. For a complete treatment of neural networks and UDA in NLP, refer to (Ramponi and Plank, 2020). Also, we do not treat multilingual work. While cross-lingual transfer can be regarded as an extreme form of domain adaptation, measuring the distance between languages requires different divergence measures, outside our purview.

3.1 Data Selection

Divergence measures are used to select a subset of text from the source domain that shares similar characteristics to the target domain. Since the source domain has labelled data, the selected data serves as supervised data to train models in the target domain. We note that the literature pays closer attention to data selection for machine translation compared to other tasks. This can be attributed to its popularity in real-world applications and the difficulty of obtaining parallel sentences for every pair of language.

Simple word-level and surface-level text features like word and n-gram frequency distributions and $tf.idf$ weighted distributions have sufficient power to distinguish between text varieties and help in data selection. Geometric measures like cosine, used with word frequency distributions, are effective for selecting data in parsing and POS tagging (Plank and van Noord, 2011). Instead of considering distributions as (sparse) vectors, one can get a better sense of the distance between distributions using information-theoretic measures. Remus (2012) find JS-Div effective for sentiment analysis. While word-level features are useful to select supervised data for an end-task, they also can be used to select data to pre-train language-models subsequently used for NER. Dai et al. (2019) use Term Vocabulary Overlap for selecting data for pretraining language models. Geometric and Information-theoretic measures with word level distributions are inexpensive to calculate. However, the distributions are sparse and continuous word distributions help in learning denser representations.

Continuous or distributed representations of words, such as CBOW, Skip-gram (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), address shortcomings of representing text as sparse, frequency-based probability distributions by transforming them into dense vectors learned from free-form text. A geometric measure (e.g., Word Vector Variance used with static word embeddings) is useful to select pre-training data for NER (Dai et al., 2019). Such selected data is found to be similar in *tenor* (the participants in a discourse, the relationships between them, etc.) to the source data. But static embeddings do not change according to the context of use. In contrast, contextual word representations (CWR) — mostly derived from neural networks (Devlin et al., 2019; Peters et al., 2018)

Paper	Task(s)	Information-Theoretic					Geometric		Higher-Order			Others
		KL	JS	Renyi	CE	Wass.	Cos	P-Norm	PAD	CMD	MMD	-
DATA SELECTION												
(Plank and van Noord, 2011)	Par, POS	✔	✔	✔			✔					
(Dai et al., 2019)	NER											✔
(Ruder and Plank, 2017)	SA, NER, Par		✔	✔			✔	✔	✔			✔
(Ruder et al., 2017)	SA		✔				✔		✔			✔
(Remus, 2012)	SA		✔									
(Lü et al., 2007)	SMT						✔					
(Zhao et al., 2004)	SMT						✔					
(Yasuda et al., 2008)	SMT				✔							
(Moore and Lewis, 2010)	SMT				✔							
(Axelrod et al., 2011)	SMT				✔							
(Duh et al., 2013)	SMT				✔							
(Liu et al., 2014)	SMT				✔							
(van der Wees et al., 2017)	NMT				✔							
(Silva et al., 2018)	NMT											✔
(Aharoni and Goldberg, 2020)	NMT						✔					
(Wang et al., 2017)	NMT							✔				
(Carpuat et al., 2017)	NMT											✔
(Vyas et al., 2018)	NMT											✔
(Chen and Huang, 2016)	SMT											✔
(Chen et al., 2017)	NMT											✔
LEARNING REPRESENTATIONS												
(Ganin et al., 2015)	SA								✔			
(Kim et al., 2017)	Intent-clf								✔			
(Liu et al., 2017)	SA								✔			
(Li et al., 2018)	Lang-ID								✔			
(Chen and Cardie, 2018)	SA									✔		
(Zellinger et al., 2017)	SA									✔		
(Peng et al., 2018)	SA									✔		
(Wu and Guo, 2020)	SA								✔			
(Ding et al., 2019)	Intent-Clf								✔			
(Shah et al., 2018)	Question sim					✔			✔			
(Zhu et al., 2019)	Emo-Regress					✔						
(Gui et al., 2017)	POS								✔			
(Zhou et al., 2019)	NER								✔			
(Cao et al., 2018)	NER								✔			
(Wang et al., 2018)	NER										✔	
(Gu et al., 2019)	NMT								✔			
(Britz et al., 2017)	NMT								✔			
(Zeng et al., 2018)	NMT								✔			
(Wang et al., 2019)	NMT								✔			
DECISIONS IN THE WILD												
(Ravi et al., 2008)	Parsing				✔							
(Elsahar and Gallé, 2019)	SA, POS								✔			✔
(Ponomareva and Thelwall, 2012)	SA	✔	✔				✔					✔
(Van Asch and Daelemans, 2010)	POS	✔		✔				✔				✔

Table 1: Prior works using divergence measures for *Data Selection*, *Learning Representations* and *Decisions in the Wild*. Tasks can be *Par*: dependency parsing, *POS*: Parts of Speech tagging, *NER*: Named Entity Recognition, *SA*: Sentiment Analysis, *SMT*: Statistical and *NMT*: Neural Machine Translation, *Intent-Clf*: Intent classification, *Lang-ID*: Language identification, *Emo-Regress*: Emotional regression. *Wass.* denotes Wasserstein.

— capture contextual similarities between words in two domains. That is, the same word used in two domains in different contexts will have different embeddings. CWRs can be obtained from hidden representations of pretrained neural machine translation (NMT) models. (McCann et al., 2017) have found such representations along with P-norm effective for data selection in MT (Wang et al., 2017). Compared to representations from shallow NMT models, hidden representations of deep neural network language models (LM) like BERT have further improved data selection for NMT (Aharoni and Goldberg, 2020).

Divergences can be measured by comparing the probabilities of a language model, in contrast to directly using its hidden representations. If a LM trained on the target domain assigns high probability to a sentence from the source domain, then the sentence should have similar characteristics to the target domain. Cross Entropy (CE) between probability distributions from LMs capture this notion of similarity between two domains. They have been extensively used for data selection in statistical machine translation (SMT) (Yasuda et al., 2008; Moore and Lewis, 2010; Axelrod et al., 2011; Duh et al., 2013; Liu et al., 2014). However, CE based methods for data selection are less effective for neural machine translation (van der Wees et al., 2017; Silva et al., 2018). Instead, van der Wees et al. (2017) come up with a dynamic subset selection where new subset is chosen every epoch during training. We note again the common refrain that sufficient amount of data should be available; here, to train good language models in the target domain.

Similar to language models, probabilistic scores from classifiers — which distinguish between samples from two domains — can aid data selection. The probabilities assigned by such classifiers in construing source domain text as target domain has been used as a divergence measures in machine translation (Chen and Huang, 2016). However, the classifiers require supervised target domain data which is not always available. As an alternative, Chen et al. (2017) train a classifier and selector in an alternating optimisation manner.

From this literature review, we find that distinct measures are effective for different NLP tasks. Ruder and Plank (2017) argue that owing to their varying task characteristics, different measures should apply. They show that learning a linear combination of measures is useful for NER, parsing and sentiment analysis. However, this is not always possible, especially in unsupervised domain adaptation where there is no supervised data in target domain. We observe that information theoretic measures and geometric measures based on frequency distributions and continuous representations are common for text and structured prediction tasks (cf. Table 1 in the appendix). The effectiveness of higher order measures for these tasks are yet to be ascertained.

Further, we find that for SMT data selection, variants of Cross Entropy (CE) measures are used extensively. However, the conclusions of van der Wees et al. (2017) are more measured regarding the benefits of CE and related measures for NMT. Contextual word representations with cosine similarity has found some initial exploration for neural machine translation (NMT), with higher order measures yet to be explored for data selection in NMT.

3.2 Learning Representations

One way to achieve domain adaptation is to learn representations that are domain-invariant which are sufficiently powerful to perform well on an end task (Ganin et al., 2015; Ganin and Lempitsky, 2015). The theory of domain divergence (Ben-David et al., 2010) shows that the target domain error is bounded by the source domain error and domain divergence ($\mathcal{H}$-divergence) and reducing the domain divergence results in domain-invariant representation. The theory also proposes a practical alternative to measure $\mathcal{H}$-divergence called PAD. The idea is to learn a representations that confuses a domain discriminator sufficiently to make samples from two domains indistinguishable.

Ganin et al. (2015) operationalise PAD in a neural network named Domain Adversarial Neural Networks (DANN). The network employs a min–max game — between the representation learner and the domain discriminator — inspired by Generative Adversarial Networks (Goodfellow et al., 2014). The representation learner is not only trained to minimise a task loss on source domain, but also maximise a discriminator's loss, by reversing the gradients calculated for the discriminator. Note that this does not require any supervised data for target domain. In later work, Bousmalis et al. (2016) argue that domain-specific peculiarities are lost in a DANN, and propose *Domain Separation Networks* (DSN) to address this shortcoming. In DSN, both domain-specific and -invariant representations are captured in a *shared–private* network. DSN is flexible in its choice of divergence measures and they find PAD performs better than MMD. Here, we limit our review to works utilising divergence measures. We exclude feature-based UDA methods such as Structural Corresponding Learning (SCL) (Blitzer et al., 2006), Autoencoder-SCL and pivot based language models (Ziser and Reichart, 2017, 2018, 2019; Ben-David et al., 2020).

Obtaining domain invariant representations is desirable for many different NLP tasks, especially for sequence labelling where annotating large amounts of data is hard. They are typically used when there is a single source domain and a single target domain — for sentiment analysis (Ganin et al., 2016), NER (Zhou et al., 2019), stance detection (Xu et al., 2019), machine translation (Britz et al., 2017; Zeng et al., 2018). The application of DANN and DSN to a variety of tasks are testament of their generality.

DANN and DSN are applied in other innovative situations. Text from two different periods of time can be considered as two different domains for intent classification (Kim et al., 2017). Gui et al. (2017) consider clean formal newswire data as source domain and noisy, colloquial, unlabeled Twitter data as the target domain and use adversarial learning to learn robust representations for POS. Commonsense knowledge graphs can help in learning domain-invariant representations as well. Ghosal et al. (2020) condition DANN with an external commonsense knowledge graph using graph convolutional neural networks for sentiment analysis. In contrast, Wang et al. (2018) use MMD outside the adversarial learning framework. They use MMD to learn to reduce the discrepancy between neural network representations belonging to two domains. Such concepts have been explored in computer vision (Tzeng et al., 2014).

While single source and target domains are common, complementary information available in multiple domains can help to improve performance in a target domain. This is especially helpful when there is no large-scale labelled data in any one domain, but where smaller amounts are available in several domains. DANN and DSN have been extended to such multi-source domain adaptation: for intent classification (Ding et al., 2019), sentiment analysis (Chen and Cardie, 2018; Li et al., 2018; Guo et al., 2018; Wright and Augenstein, 2020) and machine translation (Gu et al., 2019; Wang et al., 2019).

DANN and DSN can also help in multitask learning which considers two complementary tasks (Caruana, 1997). A key to multitask learning is to learn a shared representation that captures the common features of two tasks. However, such representations might still contain task-specific information. The shared-private model of DSN helps in disentangling such representations and has been used for sentiment analysis (Liu et al., 2017), Chinese NER and word segmentation (Cao et al., 2018). Also, although beyond the scope of our discussion here, DANN and DSN have been used to learn language-agnostic representations for text classification and structured prediction in multilingual learning (Chen et al., 2018; Zou et al., 2018; Yasunaga et al., 2018).

Most works that adopt DANN and DSN framework reduce either the PAD or MMD divergence. However, reducing the divergences, combined with other auxiliary task specific loss functions, can result in training instabilities and vanishing gradients when the domain discriminator becomes increasingly accurate (Shen et al., 2018). Using other higher order measures can result in more stable learning. In this vein, CMD has been used for sentiment analysis (Zellinger et al., 2017; Peng et al., 2018), and Wasserstein distance has been used for duplicate question detection (Shah et al., 2018) and to learn domain-invariant attention distributions for emotional regression (Zhu et al., 2019).

The review shows that most works extend the DSN framework to learn domain invariant representations in different scenarios (cf. Table 1, in the appendix). The original work from (Bousmalis et al., 2016) includes MMD divergence besides PAD, which is not adopted in subsequent works, possibly due to the reported poor performance. Most works require careful balancing between multiple objective functions (Han and Eisenstein, 2019), which can affect the stability of training. The stability of training can be improved by selecting appropriate divergence measures like CMD (Zellinger et al., 2017) and Wasserstein Distance (Arjovsky et al., 2017). We believe additional future works will adopt such measures.

3.3 Decisions in the Wild

Models can perform poorly when they are deployed in the real world. The performance degrades due to the difference in distribution between training and test data. Such performance degradation can be alleviated by large-scale annotation in the new domain. However, annotation is expensive, and — given thousands of domains — quickly becomes infeasible. Predicting the performance in a new domain, where there is no labelled data, is thus important. Much recent work provides theory (Rosenfeld et al., 2020; Chuang et al., 2020; Steinhardt and Liang, 2016). As models are put into production in the real world, this application becomes practically important as well. Empirically, NLP considers the divergence between the source and the target domain to predict performance drops.

Simple measures based on word level features have been used to predict the performance of a machine learning model in new domains. Information theoretic measures like Renyi-Div and KL-Div has been used for predicting performance drops in POS (Van Asch and Daelemans, 2010) and Cross-Entropy based measure has been used for dependency parsing (Ravi et al., 2008). Prediction of performance can also be useful for machine translation where obtaining parallel data is hard. Based on distance between languages, (Xia et al., 2020) predict performance of the model on new languages for MT, among other tasks. Such performance prediction models have also been done in the past for SMT (Birch et al., 2008; Specia et al., 2013). However, Ponomareva and Thelwall (2012) argue that predicting *drops in performance* is more appropriate compared to raw performance. They find that JS-Div effective for predicting performance drop of Sentiment Analysis systems.

Only recently, predicting model failures in practical deployments from an empirical viewpoint has regained attention. Elsahar and Gallé (2019) find the efficacy of higher-order measures to predict the drop in performance for POS and SA and do not rely on hand crafted measures as in previous works. However, analysing performance drops using CWR is still lacking. We tackle this in the next section.

4 Experiments

A practical use case of domain divergences is to predict the performance drop of a model applied to a new domain. We ask how relevant are traditional measures over word distributions compared to higher-order measures like CMD and MMD over contextual word representations like BERT, Elmo, DistilBERT (Devlin et al., 2019; Peters et al., 2018; Sanh et al., 2019)? We perform an empirical study to assess their suitability to predict performance drops for three important NLP tasks: POS, NER, and SA leaving machine translation to future work.

Performance difference between the source and the target domain depends on the divergence between their feature distributions (Ben-David et al., 2010). We assume a co-variate shift, as in (Ganin et al., 2016), where the marginal distribution over features change, but the conditional label distributions does not — i.e., $P_{\mathcal{D}_s}(y|x) = P_{\mathcal{D}_T}(y|x)$ $\mathcal{P}_{\mathcal{D}_s}(x) \neq P_{\mathcal{D}_T}(x)$. Although difference in conditional label distribution can increase the $\mathcal{H}$-Divergence measure (Wisniewski and Yvon, 2019), it requires labels in the target domain for assessment. In this work, we assume no labelled data in the target domain, to best mimic realistic settings.

4.1 Experimental Setup

Datasets: For POS, we select 5 different corpora from the English Word Tree Bank of Universal Dependency corpus (Nivre et al., 2016)[1] and also include the GUM, Lines, and ParTUT datasets. We follow Elsahar and Gallé (2019) and consider these as 8 domains. For NER, we consider CONLL 2003 (Tjong Kim Sang and De Meulder, 2003), Emerging and Rare Entity Recognition Twitter (Derczynski et al., 2017) and all 6 categories in OntoNotes v5 (Hovy et al., 2006)[2], resulting in 8 domains. For SA, we follow Guo et al. (2020), selecting the same 5 categories[3] for experiments (Liu et al., 2017).

Divergence Measures: We consider 12 divergences. For Cos, we follow the instance based calculation (Ruder et al., 2017). For MMD, Wasserstein and CORAL, we randomly sample 1000 sentences and average the results over 3 runs. For MMD, we experiment with different kernels (*cf.* Appendix A) and use default values of σ from the GeomLoss package (Feydy et al., 2019). For TVO, KL-div, JS-div, Renyi-div, based on word frequency distribution we remove stop-words and consider the top 10k frequent words across domains to build our vocabulary (Ruder et al., 2017; Gururangan et al., 2020). We use α=*0.99* for Renyi as found effective by Plank and van Noord (2011). We do not choose CE as it is mainly used in MT and ineffective for classification and structured prediction (Ruder et al., 2017).

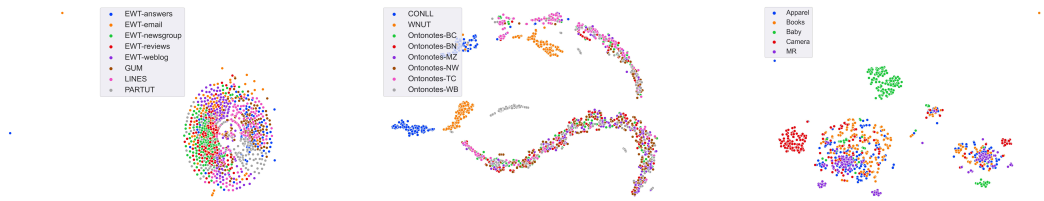
Model Architecture: For all our experiments, unless otherwise mentioned, we use the pre-trained DistilBERT (Sanh et al., 2019) model. It has competitive performance to BERT, but has faster inference times and lower resource requirements. For every text segment, we obtain the activations from the final layer and average-pool the representations. We train the models on the source domain training split and test the best model — picked from validation set grid search — on the test dataset of the same and other domains (cf. Appendix C).

For POS and NER, we follow the original BERT model where a linear layer is added and a prediction is made for every token. If the token is split into

[1] Yahoo! Answers, Email, NewsGroups, Reviews and Weblogs.

[2] Broadcast News (BN), Broadcast Conversation (BC), Magazine (MZ), Telephone Conversation (TC) and Web (WB).

[3] Apparel, Baby, Books, Camera and MR.



(a) (POS)-MMD-Gaussian (b) (NER)-MMD-RQ (c) (SA)-JS-Div

Figure 2: t-SNE plots for select measures. The complete set of diagrams are available in Appendix D.

multiple tokens due to Byte Pair Encoding, the label for the first token is predicted. For SA and domain discriminators, we pool the representation from the last layer of DistilBERT and add a linear layer for prediction (Appendix B).

4.2 Are traditional measures still relevant?

For POS, the PAD measure has the best correlation with performance drop (cf. Table 2). Information-theoretic measures over word frequency distributions, such as JS-div, KL-div, and TVO, which have been prevalent for data selection and performance drop use cases (cf. Table 1) are comparable to PAD. Plank et al. (2014) claim that the errors in POS are dictated by out of vocabulary words. Our findings validate their claim, as we find strong correlation between POS performance drop and word probability distribution measures For NER, MMD-RQ provides the best correlation of 0.495. CORAL — a higher-order measure — and JS-div are comparable. For SA, Renyi-div and other information-theoretic measures provide considerably better correlation compared to higher-order measures. Cos is a widely-used measure across applications, however it *did not* provide significant correlation for either task. TVO is used for selecting pretraining data for NER (Dai et al., 2019) and as a measure to gauge the benefits of fine-tuning pre-trained LMs on domain-specific data (Gururangan et al., 2020). Although TVO does not capture the nuances of domain divergences, it has strong, reliable correlations for performance drops. PAD has been suggested for data selection in SA by Ruder and Plank (2017) and for predicting drop in performance by Elsahar and Gallé (2019). Our analysis confirms that PAD provides good correlations across POS, NER, and SA.

We find no single measure to be superior across all tasks. However, information theoretic measures consistently provide good correlations. Currently, when contextual word representations dictate results in NLP, simple measures based on frequency distributions are strong baselines for predicting performance drop. Although higher-order measures do not always provide the best correlation, they are differentiable, thus suited for end-to-end training of domain-invariant representations.

4.3 Discussion

Why are some divergence measures better at predicting drops in performance? The *one-dataset-one-domain* is a key assumption in such works. However, many works have questioned this assumption (Plank and van Noord, 2011). Multiple domains may exist within the same domain (Webber, 2009) and two different datasets may not necessarily be considered different domains (Irvine et al., 2013). Recently Aharoni and Goldberg (2020) show that BERT representations reveal their underlying domains. They qualitatively show that a few text segments from a dataset actually belong to another domain. However the degree to which the samples belong to different domains is unclear.

We first test the assumption that different datasets are different domains using Silhouette scores (Rousseeuw, 1987) which quantify the separability of clusters. We initially assume that a dataset is in its own domain. A positive score shows that datasets can be considered as well-separated domains; a negative score shows that most of the points within a dataset can be assigned to a nearby domain; and 0 signifies overlapping domains. We calculate Silhouette scores and t-SNE plots (Maaten and Hinton, 2008) for different divergence measures. Refer to the plots (Figures 3a to 3c) and calculation details in Appendix D.

Almost all the measures across different tasks have negative values close to 0 (Table 2, (r)).

- For POS, CORAL, Wasserstein and Cos strongly indicate that text within a dataset belongs to other

Measure	Correlations			Silhouette Coefficients		
-	**POS**	**NER**	**SA**	**POS**	**NER**	**SA**
Cos	0.018	0.223	-0.012	-1.78×10^{-1}	-2.49×10^{-1}	-2.01×10^{-1}
KL-Div	0.394	0.384	0.715	-	-	-
JS-Div	0.407	0.484	0.709	-8.50×10^{-2}	-6.40×10^{-2}	$+2.04 \times 10^{-2}$
Renyi-Div	0.392	0.382	**0.716**	-	-	-
PAD	**0.477**	0.426	0.538	-	-	-
Wasserstein	0.378	0.463	0.448	-2.11×10^{-1}	-2.36×10^{-1}	-1.70×10^{-1}
MMD-RQ	0.248	**0.495**	0.614	-4.11×10^{-2}	-3.04×10^{-2}	-1.70×10^{-2}
MMD-Gaussian	0.402	0.221	0.543	$+4.25 \times 10^{-5}$	$+2.37 \times 10^{-3}$	-8.42×10^{-5}
MMD-Energy	0.244	0.447	0.521	-9.84×10^{-2}	-1.14×10^{-1}	-8.48×10^{-2}
MMD-Laplacian	0.389	0.273	0.623	-1.67×10^{-3}	$+4.26 \times 10^{-4}$	-1.08×10^{-3}
CORAL	0.349	0.484	0.267	-2.34×10^{-1}	-2.78×10^{-1}	-1.41×10^{-1}
TVO	-0.437	-0.457	-0.568	-	-	-

Table 2: (l): Correlation of performance drops with divergence measures. Measures with higher correlations are better indicators of performance drops. (r): Silhouette coefficients considering different divergence measures. We randomly sample 200 points for calculation and average the results over 5 runs. Only certain divergences which are metrics are allowed. The colours are from the taxonomy of divergence measures in Figure 1.

domains. However, for MMD-Gaussian the domains overlap (Figure 2a).

- For NER, MMD-Gaussian and MMD-Laplacian indicate that the clusters overlap while all other metrics have negative values.
- For SA, JS-Div has positive values compared to other measures, and as seen in Figure 2c, we can see a better notion of distinct clusters.

The Silhouette scores along with the t-SNE plots show that datasets are, in fact, not distinct domains. Considering data-driven methods for defining domains is needed (Aharoni and Goldberg, 2020).

If there are indeed separate domains, does it explain why some measures are better than the others? We see better notions of clusters for NER and sentiment analysis (*cf.* Figures 2b and 2c). We can expect the drop in performance to be indicative of these domain separations. Comparing the best correlations from Table 2, correlations for NER and sentiment analysis are higher compared with POS. For POS, there are no indicative domain clusters and the correlation between domain divergence and performance may be less; whereas for SA, both the t-SNE plot and the Silhouette scores for JS-Div (*cf.* Figure 2c) corroborate comparatively better separation. If datasets are indeed different domains, these divergence measures are reliable indicators of performance drops. If they are not, there might be other confounding factors (such as differences in label distribution) and one has to be cautious in using them.

Domain overlap also has consequences for data selection strategies. For example, Moore and Lewis (2010) select *pseudo in-domain data* from source corpora (cf Section 3.1). As the Silhouette coefficients are negative and close to 0, many data points in a dataset belong to nearby domains. Data selection strategies thus may be effective. If the Silhouette coefficients are more negative and if more points in the source aptly belong to the target domain, we should expect increased sampling from such source domains to yield additional performance benefits in the target domain.

5 Conclusion

We survey domain adaptation works, focusing on divergence measures and their usage for *data selection*, *learning domain-invariant representations*, and *making decisions in the wild*. We synthesised the divergence measures into a taxonomy of *information theoretic*, *geometric* and *higher-order* measures. While traditional measures are common for data selection and making decisions in the wild, higher-order measures are prevalent in learning representations. Based on our correlation experiments, silhouette scores, and t-SNE plots, we make the following recommendations:

- PAD is a reliable indicator of performance drop. It is best used when there are sufficient examples to train a domain discriminator.
- JS-Div is symmetric and a formal metric. It is related to PAD, easy to compute, and serves as a strong baseline.
- While Cosine is popular, it is an unreliable indicator of performance drop.
- One-dataset-is-not-one-domain. Instead, cluster representations and define appropriate domains.

Acknowldgements

We would also like to acknowledge the support of the NExT research grant funds,supported by the National Research Foundation,Prime Ministers Office, Singapore under its IRC@ SG Funding Initiative, and to gratefully acknowledge the support of NVIDIA Corporation with the donation of the GeForce GTX Titan XGPU used in this research.

References

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional spaces. In *Proceedings of the 8th International Conference on Database Theory*, ICDT '01, page 420–434, Berlin, Heidelberg. Springer-Verlag.

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. Association for Computational Linguistics.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *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. PMLR.

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. Association for Computational Linguistics.

Eyal Ben-David, Carmel Rabinovitz, and Roi Reichart. 2020. Perl: Pivot-based domain adaptation for pre-trained deep contextualized embedding models.

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.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 137–144. MIT Press.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii. Association for Computational Linguistics.

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*, pages 120–128, Sydney, Australia. Association for Computational Linguistics.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 343–351. Curran Associates, Inc.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126, Copenhagen, Denmark. Association for Computational Linguistics.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2018. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 182–192, Brussels, Belgium. Association for Computational Linguistics.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, Vancouver. Association for Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Mach. Learn.*, 28(1):41–75.

Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost weighting for neural machine translation domain adaptation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 40–46, Vancouver. Association for Computational Linguistics.

Boxing Chen and Fei Huang. 2016. Semi-supervised convolutional networks for translation adaptation with tiny amount of in-domain data. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 314–323, Berlin, Germany. Association for Computational Linguistics.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. 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 1226–1240, New Orleans, Louisiana. Association for Computational Linguistics.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ching-Yao Chuang, Antonio Torralba, and Stefanie Jegelka. 2020. Estimating generalization under distribution shifts via domain-invariant representations. *CoRR*, abs/2007.03511.

I. Csiszár. 1972. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2(1):191–213.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2019. Using similarity measures to select pretraining data for NER. 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 1460–1470, Minneapolis, Minnesota. Association for Computational Linguistics.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

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. Association for Computational Linguistics.

X. Ding, Q. Shi, B. Cai, T. Liu, Y. Zhao, and Q. Ye. 2019. Learning multi-domain adversarial neural networks for text classification. *IEEE Access*, 7:40323–40332.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 678–683, Sofia, Bulgaria. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. 2019. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1180–1189. JMLR.org.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2015. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:59:1–59:35.

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, ACL 2020, Online, July 5-10, 2020*, pages 3198–3210. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J. Smola. 2007. A kernel method for the two-sample-problem. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press.

Shuhao Gu, Yang Feng, and Qun Liu. 2019. Improving domain adaptation translation with domain invariant and specific information. 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 3081–3091, Minneapolis, Minnesota. Association for Computational Linguistics.

Tao Gui, Qi Zhang, Haoran Huang, Minlong Peng, and Xuanjing Huang. 2017. Part-of-speech tagging for twitter with adversarial neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2411–2420, Copenhagen, Denmark. Association for Computational Linguistics.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2020. Multi-source domain adaptation for text classification via distancenet-bandits. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7830–7838. AAAI Press.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703, Brussels, Belgium. Association for Computational Linguistics.

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. 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. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Munteanu. 2013. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1:429–440.

James J. Jiang. 2007. A literature survey on domain adaptation of statistical classifiers.

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)*, pages 1297–1307, Vancouver, Canada. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. What's in a domain? learning domain-robust text representations using adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 474–479, New Orleans, Louisiana. Association for Computational Linguistics.

Le Liu, Yu Hong, Hao Liu, Xing Wang, and Jianmin Yao. 2014. Effective selection of translation model training data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 569–573, Baltimore, Maryland. Association for Computational Linguistics.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada. Association for Computational Linguistics.

Yajuan Lü, Jin Huang, and Qun Liu. 2007. Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 343–350, Prague, Czech Republic. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Anna Margolis. 2011. A literature review of domain adaptation with unlabeled data.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6294–6305.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*

- *Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

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. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Minlong Peng, Qi Zhang, Yu-gang Jiang, and Xuanjing Huang. 2018. Cross-domain sentiment classification with target domain specific information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2505–2513, Melbourne, Australia. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. 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. Association for Computational Linguistics.

Barbara Plank, Anders Johannsen, and Anders Søgaard. 2014. 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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Natalia Ponomareva and Mike Thelwall. 2012. Biographies or blenders: Which resource is best for cross-domain sentiment analysis? In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 488–499. Springer.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP - A survey. *CoRR*, abs/2006.00632.

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. Association for Computational Linguistics.

Robert Remus. 2012. Domain adaptation using domain similarity- and domain complexity-based instance selection for cross-domain sentiment analysis. In *12th IEEE International Conference on Data Mining Workshops, ICDM Workshops, Brussels, Belgium, December 10, 2012*, pages 717–723. IEEE Computer Society.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. 2020. A constructive prediction of the generalization error across scales. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Peter Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2017. Data selection strategies for multi-domain sentiment analysis. *ArXiv*, abs/1702.02426.

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. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Darsh 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, Brussels, Belgium. Association for Computational Linguistics.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4058–4065. AAAI Press.

Catarina Cruz Silva, Chao-Hong Liu, Alberto Poncelas, and Andy Way. 2018. Extracting in-domain training corpora for neural machine translation using data selection methods. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 224–231, Brussels, Belgium. Association for Computational Linguistics.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. Association for Computational Linguistics.

Jacob Steinhardt and Percy S Liang. 2016. Unsupervised risk estimation using only conditional independence structure. In *Advances in Neural Information Processing Systems*, volume 29, pages 3657–3665. Curran Associates, Inc.

Baochen Sun, Jiashi Feng, and Kate Saenko. 2017. Correlation alignment for unsupervised domain adaptation. In Gabriela Csurka, editor, *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, pages 153–171. Springer.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474.

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. Association for Computational Linguistics.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. 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 1503–1515, New Orleans, Louisiana. Association for Computational Linguistics.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada. Association for Computational Linguistics.

Yong Wang, Longyue Wang, Shuming Shi, Victor O.K. Li, and Zhaopeng Tu. 2019. Go from the general to the particular: Multi-domain translation with domain transformation networks. *ArXiv*, abs/1911.09912.

Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018. Label-aware double transfer learning for cross-specialty medical named entity recognition. 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 1–15, New Orleans, Louisiana. 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. 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. Association for Computational Linguistics.

Guillaume Wisniewski and François Yvon. 2019. How Bad are PoS Tagger in Cross-Corpora Settings? Evaluating Annotation Divergence in the UD Project. 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 218–227, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Dustin Wright and Isabelle Augenstein. 2020. Transformer based multi-source domain adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7963–7974, Online. Association for Computational Linguistics.

Yuan Wu and Yuhong Guo. 2020. Dual adversarial co-learning for multi-domain text classification. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 6438–6445. AAAI Press.

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, Online. Association for Computational Linguistics.

Brian Xu, Mitra Mohtarami, and James R. Glass. 2019. Adversarial domain adaptation for stance detection. *ArXiv*, abs/1902.02401.

Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto, and Eiichiro Sumita. 2008. Method of selecting training data to build a compact and efficient translation model. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

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. Association for Computational Linguistics.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (CMD) for domain-invariant representation learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao. 2018. Multi-domain neural machine translation with word-level domain context discrimination. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 447–457, Brussels, Belgium. Association for Computational Linguistics.

Bing Zhao, Matthias Eck, and Stephan Vogel. 2004. Language model adaptation for statistical machine translation via structured query models. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 411–417, Geneva, Switzerland. COLING.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. 2019. Dual adversarial neural transfer for low-resource named entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3461–3471, Florence, Italy. Association for Computational Linguistics.

Suyang Zhu, Shoushan Li, and Guodong Zhou. 2019. Adversarial attention modeling for multi-dimensional emotion regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 471–480, Florence, Italy. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Yftah Ziser and Roi Reichart. 2018. Pivot based language modeling for improved neural domain adaptation. 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 1241–1251, New Orleans, Louisiana. Association for Computational Linguistics.

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*, pages 5895–5906, Florence, Italy. Association for Computational Linguistics.

Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. 2018. Adversarial feature adaptation for cross-lingual relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 437–448, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A Domain Divergence Measures

This section provides the necessary background on different kinds of divergence measures used in the literature. They can be either information-theoretic – which measure the distance between two probability distributions, geometric - which measure the distance between two vectors in a space, or higher-order which capture similarity in a projected space and consider higher order moments of random variables.

A.1 Information-Theoretic Measures

Let P and Q be two probability distributions. These information-theoretic measures are used to capture differences between P and Q.

Kullback-Leibler Divergence (KL-Div) Q is called the reference probability distribution[4]. More precisely, KL is defined if only for all $Q(x)$ st $Q(x) = 0$, $P(x)$ is also 0; and undefined if ∃ x, $Q(x) = 0$ and $P(x) > 0$.

$$D_{KL}(P||Q) = \sum_x P(x) log\left(\frac{P(x)}{Q(x)}\right) \quad (1)$$

Renyi Divergence (Renyi-Div) Renyi Divergence is a generalisation of the KL Divergence and is also called α-power divergence:

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} log\left(\sum_x \frac{P(x)^\alpha}{Q(x)^{\alpha-1}}\right) \quad (2)$$

Here $\alpha \geq 0$ and $\alpha \neq 1$. Renyi divergence is equivalent to KL divergence in the limit where $\alpha \to 1$.

Jensen Shannon Divergence (JS-Div) Jensen Shannon divergence (JS-divergence) is a symmetric version of KL-Divergence. It has many advantages. The square root of the Jensen Shannon Divergence is a metric and it can be used for non-continuous probabilities:

$$D_{JS}(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M)$$
$$M = \frac{1}{2}(P + Q) \quad (3)$$

[4]KL divergence is asymmetric and cannot be considered a metric

Entropy-Related - (CE) Let H_T, H_S assign entropy to a sentence using a language model trained on the target and source domain, respectively. If s is a text segment from the source domain, then the difference in entropy, as shown below, gives the similarity of a source domain segment to the target domain. Some works just use H_T, ignoring H_S. MT related work (Moore and Lewis, 2010), consider only the source language. Axelrod et al. (2011) extend to consider both the source and the target language of machine translation, which performs better for data selection. We present these variations in the formulae below and attribute the same name **CE** to both these variations in the literature review.

$$D_{CE} = H_T(s) - H_S(s) \quad (4)$$

$$D_{CE} = [H_T^{src-lang}(s) - H_S^{src-lang}(s)] + [H_T^{trg-lang}(s) - H_S^{trg-lang}(s)] \quad (5)$$

A.2 Geometric Measures

Let $\vec{p}$ and $\vec{q}$ be two vectors in $\mathbb{R}^n$. Domain adaptation works use geometric metrics for continuous representations like word vectors.

Cosine Similarity (Cos): It calculates the cosine of the angle between vectors. To measure the cosine distance between two points, we use $1 - Cos$:

$$cos(\vec{p}, \vec{q}) = \frac{\vec{p}.\vec{q}}{||p|| \cdot ||q||} \quad (6)$$

l_p-norm (Norm): **Euclidean distance or l_2 distance** measures the straight line distance between vectors and **Manhattan or l_1** measures the sum of the difference between their projections.

$$d_2(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (7)$$

$$d_1(p, q) = \sum_{i=1}^{n} |p_i - q_i| \quad (8)$$

A.3 Higher-Order Measures

$\mathcal{H}$-divergence and Proxy-A-Distance (PAD): Ben-David et al. (2010) state that the error of a machine learning classifier in a target domain is bound by its performance on the source domain and the $\mathcal{H}$-divergence between the source and the target distributions. $\mathcal{H}$-divergence is expensive to calculate. An approximation of $\mathcal{H}$ is called *Proxy-A-Distance*. This definition has been adopted from (Elsahar and Gallé, 2019). Here $G : \mathcal{X} \rightarrow [0, 1]$ is a supervised machine learning model that classifies examples to the source and target domains, D_s, D_t. $|D|$ is the size of the training data and $\mathbb{1}$ is an indicator function:

$$PAD = 1 - 2\epsilon(G_d) \tag{9}$$

$$\epsilon(G_d) = 1 - \frac{1}{|D|} \sum_{x_i \in D_s, D_t} |G(x_i) - \mathbb{1}(x_i \in D_s)| \tag{10}$$

Wasserstein Distance: Wasserstein Distance (also called Earth Mover's distance) is another metric for two probability distributions. Intuitively, it measures the least amount of work done to transport probability mass from one probability distribution to another to make them equal. The work done in this case is measured as the mass transported multiplied by the distance of travel. It is known to be better than Kullback-Leibler Divergence and Jensen-Shannon Divergence when the random variables are high dimensional or otherwise. The Wasserstein metric is defined as:

$$D_{Wasserstein} = \underset{\gamma \in \pi}{inf} \sum_{x,y} \|x - y\| \, \gamma(x, y)$$

Here $\gamma \in \pi(P, Q)$ where $\pi(P, Q)$ is the set of all distributions where the marginals are P and Q.

Maximum Mean Discrepancy (MMD): MMD is a non-parametric method to estimate the distance between distributions based on Reproducing Kernel Hilbert Spaces (RKHS). Given two random variables $X = \{x_1, x_2, ..., x_m\}$ and $Y = \{y_1, y_2, ..., y_n\}$ that are drawn from distributions P and Q, the empirical estimate of the distance between distribution P and Q is given by:

$$MMD(X, Y) = \left\| \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) - \frac{1}{n} \sum_{i=1}^{n} \phi(y_i) \right\|_{\mathcal{H}} \tag{11}$$

Here $\phi : \mathcal{X} \rightarrow \mathcal{H}$ are nonlinear mappings of the samples to a feature representation in a RKHS. In this work, we map the contextual word representations of the text to RKHS. The different kinds of kernels we use in this work are given below. We use the default values of $\sigma = 0.05$ of the GeomLoss package (Feydy et al., 2019).

Rational Quadratic Kernel

$$\phi(x, y) = \left(1 + \frac{1}{2\alpha}(x - y)^T \Theta^{-2}(x - y)\right)^{-\alpha}$$

Energy

$$\phi(x, y) = - \|x - y\|_2$$

Gaussian

$$\phi(x, y) = exp(-\frac{\|x - y\|_2^2}{2\sigma^2})$$

Laplacian

$$\phi(x, y) = exp(-\frac{\|x - y\|_2}{\sigma})$$

Correlation Alignment (CORAL): Correlation alignment is the distance between the second-order moment of the source and target samples. If d is the representation dimension, $\|\|_F$ represents Frobenius norm and Cov_S, Cov_T is the covariance matrix of the source and target samples, then CORAL is defined as:

$$D_{CORAL} = \frac{1}{4d^2} \|Cov_S - Cov_T\|_F^2 \tag{12}$$

Central Moment Discrepancy (CMD): Central Moment Discrepancy is another metric that measures the distance between source and target distributions. It not only considers the first moment and second moment, but also other higher-order moments. While MMD operates in a projected space, CMD operates in the representation space. If P and Q are two probability distributions and $X = \{X_1, X_2, ..., X_N\}$ and $Y = \{Y_1, Y_2, ..., Y_N\}$ are random vectors that are independent and identically distributed from P and Q and every component of the vector is bounded by $[a, b]$, CMD is then defined by:

$$\begin{aligned} CMD(P, Q) = \frac{1}{|b - a|} \|E(X) - E(Y)\|_2 \\ + \sum_{k=2}^{\infty} \frac{1}{|b - a|^k} \|c_k(X) - c_k(Y)\|_2 \end{aligned} \tag{13}$$

where $E(X)$ is the expectation of X and c_k is the $k - th$ order central moment, defined as:

$$c_k(X) = E\left(\prod_{i=1}^{N} (X_i - E(X_i))^{r_i}\right) \tag{14}$$

and $r_1 + r_2 + r_N = k$ and $r_1....r_N \geq 0$

A.4 Other Measures

Bhattacharya Coefficient: If P and Q are probability distributions, then the Bhattacharya coefficient and Bhattacharya distance are defined as:

$$Bhattacharya(P,Q) = \sum_{x} \sqrt{P(x)Q(x)} \quad (15)$$

$$D_{Bhattacharya} = -log(Bhattacharya(P,Q)) \quad (16)$$

Term Vocabulary Overlap (TVO): This measures the proportion of target vocabulary that is also present in the source vocabulary. If V_S is the source domain vocabulary and V_T is the target domain vocabulary, then the Term Vocabulary Overlap between the source domain (D_S) and the target domain (D_T) is given by:

$$TVO(D_S, D_T) = \frac{|V_S \bigcap V_T|}{|V_T|} \quad (17)$$

Word Vector Variance: Different contexts in which a word is used in two different datasets can be used as an indication of the divergence between two datasets. Let $\vec{w}^{i}_{src}$ denote the word embedding of word i in source domain and $\vec{w}^{i}_{trg}$ is the word embedding of the same word in the target domain. Let d be the dimension of the word embedding. The word vector variance between the source domain (D_S) and the target domain (D_T) is given by:

$$WVV(D_S, D_T) = \frac{1}{|V_S| * d} \sum_{i}^{|V_s|} \left\| w^{i}_{src} - w^{i}_{trg} \right\|_2^2 \quad (18)$$

B Model Hyperparameters

For POS, NER and Sentiment Analysis models, we do a grid search of learning rate in {1e-01, 1e-05, 5e-05} and dropout in $\{0.2, 0.3, 0.4, 0.5\}$ and number of epochs in $\{25, 50\}$. PAD requires a domain discriminator. We sample as many samples in the target domain as the source domain (Ruder et al., 2017) and train a DistilBERT based classifier. For every domain discriminator we do a grid search of learning rate in {*1e-05, 5e-05*}, dropout in $\{0.4, 0.5\}$ and number of epochs in $\{10, 25\}$. For POS and NER, we monitor the macro F-Score; for domain discrimination, we monitor the accuracy scores. We chose the best model after the grid search for all subsequent calculations. For training the models we use the Adam Optimiser (Kingma and Ba, 2015) with the $\beta_1 = 0.9$ and $\beta_2 = 0.99$ and ϵ as 1e-8. We use HuggingFace Transformers (Wolf et al., 2019) for all our experiments.

C Cross-Domain Performances

C.1 Parts of speech tagging

Table 3 shows the hyper parameters for the best model for POS and Table 4 shows the cross domain performances.

C.2 Named Entity Recognition

Table 5 shows the hyper parameters for the best model for NER and Table 6 shows the cross domain performances.

C.3 Sentiment Analysis

Table 7 shows the hyper parameters for the best model for Sentiment analysis and Table 8 shows the cross domain performances.

D Silhouette Scores and t-SNE Plots

For calculating Silhouette scores we use a subset of domain divergence measures that are metrics (a requirement of Silhouette scores) and can be calculated between single instances of text. We sample 200 points for each dataset as the time complexity increases exponentially with number of points. We average the results over 5 runs.

We plot the t-SNE plots for POS (Figure 3a), NER (Figure 3b) and Sentiment Analysis (SA) (Figure 3c). We sample 200 points from each of the datasets for the plot. Wherever relevant, we use DistilBERT (Sanh et al., 2019) representations for calculations.

Dataset	Epochs	Learning Rate	Dropout	Fscore
EWT-answers	50	5×10^{-5}	0.4	95.38
EWT-email	50	1×10^{-5}	0.3	96.62
EWT-newsgroup	50	5×10^{-5}	0.5	95.92
EWT-reviews	50	5×10^{-5}	0.4	96.97
EWT-weblog	50	5×10^{-5}	0.3	97.03
GUM	50	1×10^{-5}	0.3	95.73
LINES	50	5×10^{-5}	0.3	97.38
PARTUT	50	1×10^{-5}	0.4	97.06

Table 3: Model performance and hyper-parameters producing the best model for Parts of Speech Tagging trained using DistilBERT as the base model. The datasets are from the Universal Dependencies Corpus (UD) (Nivre et al., 2016). 5 corpora are from the English Word Tree (EWT) portion which are EWT-answers, EWT-email, EWT-newgroup, EWT-reviews, EWT-weblog.

Source/Target	EWT-answers	EWT-email	EWT-newsgroup	EWT-reviews	EWT-weblog	GUM	LINES	PARTUT
EWT-answers	95.38	93.96	94.02	95.83	95.64	93.58	93.86	92.06
EWT-email	94.11	96.62	94.40	95.42	95.37	93.08	93.98	93.47
EWT-newsgroup	94.71	95.07	95.92	95.31	96.80	93.82	93.83	92.74
EWT-reviews	94.99	94.51	94.56	96.97	95.55	93.07	94.27	92.62
EWT-weblog	95.38	93.96	94.02	95.83	95.64	93.58	93.87	92.06
GUM	91.63	92.59	91.75	93.55	93.56	95.73	93.54	93.12
LINES	89.79	89.77	88.76	92.39	90.77	91.75	97.38	92.68
PARTUT	89.27	89.54	89.56	91.28	92.27	90.65	92.97	96.65

Table 4: Cross-domain performance for POS tagging. The best model for each source domain is tested on the test dataset of the same domain and all other domains.

Dataset	Epochs	Learning Rate	Dropout	Fscore
CONLL-2003	50	5×10^{-5}	0.5	0.90
WNUT	25	5×10^{-5}	0.5	0.50
Onto-BC	50	5×10^{-5}	0.5	0.82
Onto-BN	50	1×10^{-5}	0.3	0.89
Onto-MZ	50	1×10^{-5}	0.3	0.86
Onto-NW	25	5×10^{-5}	0.4	0.89
Onto-TC	50	1×10^{-5}	0.5	0.75
Onto-WB	50	5×10^{-5}	0.4	0.63

Table 5: Model performance and hyper-parameters for Named Entity Recognition trained using DistilBERT as the base model. The datasets are CONLL-2003, Emerging and Rare Entity Recognition twitter dataset (WNUT), and six different sources of text in Ontonotes v5 (Hovy et al., 2006)

Source/Target	CONLL	WNUT	ONTO-BC	ONTO-BN	ONTO-MZ	ONTO-NW	ONTO-TC	WB
CONLL 2003	0.90	0.37	0.54	0.65	0.59	0.54	0.51	0.41
WNUT	0.66	0.50	0.40	0.44	0.49	0.42	0.49	0.33
ONTO-BC	0.48	0.31	0.82	0.81	0.77	0.74	0.72	0.45
ONTO-BN	0.53	0.37	0.77	0.89	0.76	0.79	0.76	0.47
ONTO-MZ	0.49	0.29	0.72	0.78	0.86	0.75	0.69	0.45
ONTO-NW	0.52	0.32	0.73	0.86	0.73	0.89	0.76	0.46
ONTO-TC	0.51	0.37	0.61	0.64	0.57	0.55	0.75	0.41
ONTO-WB	0.43	0.12	0.52	0.63	0.54	0.57	0.52	0.63

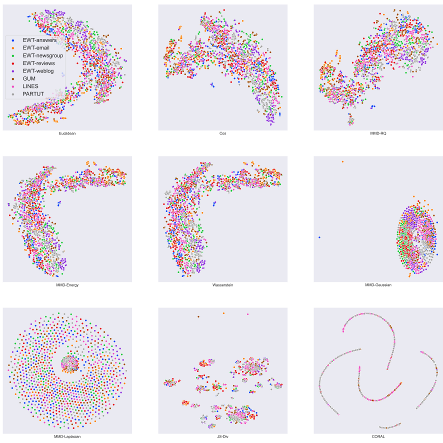
Table 6: Cross-domain performance for NER. The best model for each source domain is tested on the test dataset of the same domain and all other domains.

Dataset	Epochs	Learning Rate	Dropout	Fscore
Apparel	25	1×10^{-5}	0.4	91.25
Baby	50	5×10^{-5}	0.4	93.75
Books	50	1×10^{-5}	0.4	92
Camera/Photo	25	1×10^{-5}	0.4	92
MR	50	5×10^{-5}	0.3	82.5

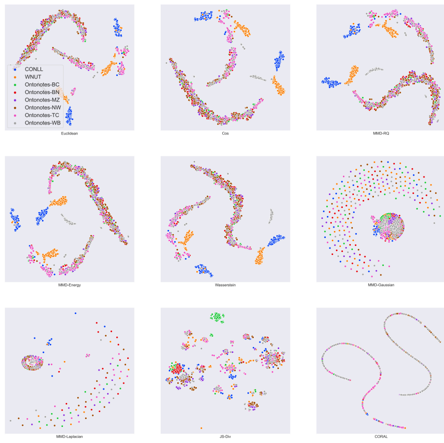
Table 7: Model performance and hyper-parameters for Sentiment Analysis with DistilBERT as the base model. We chose 5 out of 16 datasets from (Liu et al., 2017) which are Apparel, Baby, Books, Camera/Photo, and MR.

Source/Target	Apparel	Baby	Books	Camera/Photo	MR
Apparel	0.91	0.9100	0.85	0.87	0.77
Baby	0.89	0.9375	0.86	0.89	0.75
Books	0.88	0.8875	0.92	0.87	0.79
Camera/Photo	0.89	0.89	0.86	0.92	0.75
MR	0.76	0.76	0.8375	0.74	0.83

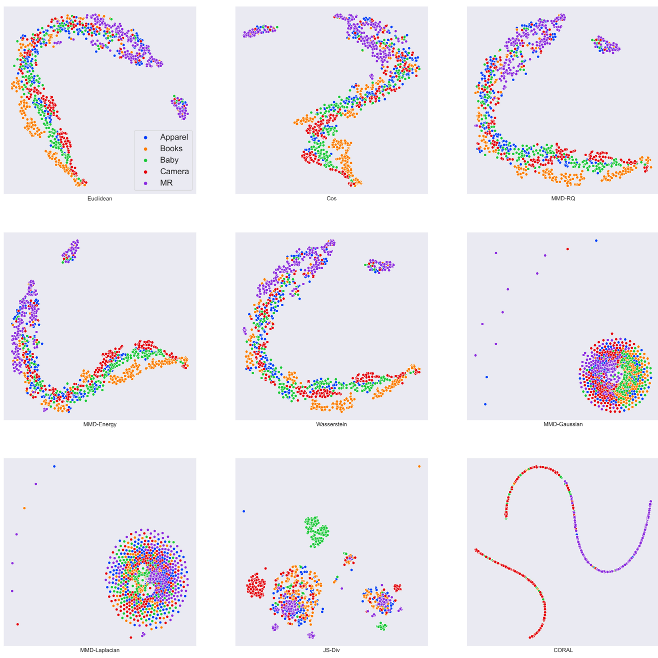
Table 8: Cross-domain performance for Sentiment Analysis. The best model for each source domain is tested on the test dataset of the same domain and all other domains.



(a) POS



(b) NER



(c) SA

Figure 3: t-SNE plots for different tasks.