The Change that Matters in Discourse Parsing: Estimating the Impact of Domain Shift on Parser Error

Anonymous ACL submission

Abstract

Discourse analysis allows us to attain high-level inferences of a text document beyond the sentence-level. However, currently the performance of discourse models is very low on texts outside of the training distribution's coverage. There is need for a measure that can inform us to what extent our model generalizes from the training to the test sample when these samples may be drawn from distinct distributions. While this can be estimated via distribution shift, we argue that this does not directly correlate with change in the observed error of a classifier (i.e. error-gap). Thus, we propose to use a statistic from the theoretical domain adaptation literature which *can* be directly tied to error-gap. We study the bias of this statistic as an estimator of error-gap both theoretically and through a large-scale empirical study of over 2400 experiments on 6 discourse datasets from domains including, but not limited to: news, biomedical texts, TED talks, Reddit posts, and fiction. Our results not only motivate our proposal and help us to understand its limitations, but also provide insight on the properties of discourse models and datasets which improve performance in domain adaptation. For instance, we find that non-news datasets are slightly easier to transfer to than news datasets when the training and test sets are very different. We plan to release our code as a Python package to allow practitioners to make more informed model and dataset choices.

1 Introduction

Computational approaches to discourse aim to learn inferences in text and a representation of the structure of discourse. Discourse parsing models are trained on a dataset annotated based on a *discourse framework*, where two textual units are identified that exhibit some discourse relation, and the type of discourse relation is labeled. Discourse parsing has been shown to be helpful in several downstream NLP tasks (Marcu, 1999, 2000; Bhatia et al., 2015; Narasimhan and Barzilay, 2015; Cohan et al., 2018). However, in other cases, discourse relations have also been found not to improve, or even to hurt, the performance of many of these downstream tasks (Zhong et al., 2020). There are several hypothesized reasons for this. For one, due to the difficulty of the annotation task, datasets labeled with these discourse relations are typically small, and the most widely used datasets consist only of text from news articles. In general, the performance of discourse models trained on these datasets is very low and even slight domain shift has been shown to worsen the performance (Atwell et al., 2021). When deciding on a model to use, it is important to select one which can generalize well between the training and test samples. The same holds true for the dataset; it is important to train on a dataset that allows the model to transfer well to the test set.

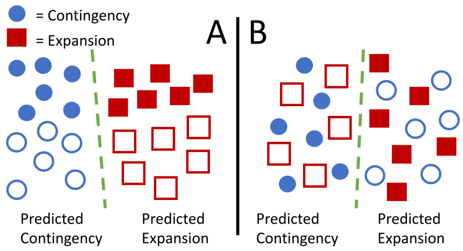

Figure 1: Solid/hollow shapes indicate training/test set, while circles/squares indicate the correct labels. (A) Vertical shift is easily identified, but the classifier (dotted line) does well on both domains. (B) In the feature space, shift is imperceptible, but the classifier assigns the incorrect relation label to each point in the test set. In both, identifiable shift does *not* correlate with the classifier's ability to correctly predict the discourse relation

To estimate the extent of a model's generalizability on a particular train/test pair, common proposals suggest using two-sample statistics which capture distributional shift in the feature space (Rabanser et al., 2019). However, the working hypothesis of this paper is that *changes in feature-distribution* *do not necessarily equate to changes in a classifier's error*; i.e., from train to test sample. Figure 1 captures this idea by illustrating some examples in simple 2D-space where domain shift may occur without high error, and vice versa.

Motivated by this hypothesis, we look to existing theoretical domain adaptation literature. We propose to use a statistic which has not only been designed to incorporate information about the classifier we would like to transfer, but has also been shown (theoretically) to directly relate to model performance on the test set. Namely, we consider a slight generalization of the source-guided discrepancy (Kuroki et al., 2019) which we call the h-discrepancy defined for any classifier h (we introduce and define this metric in Section 4). We provide (novel) theoretical analysis of the errors of this statistic in estimating adaptation performance and, based on this, hypothesize this statistic will correlate more substantially with the classifiers' generalization ability than the two-sample statistics previously mentioned. We support this hypothesis by illustrating these correlations across several different widely-used discourse datasets (described in Section 3). We also provide a detailed empirical analysis of the estimation error of this statistic in predicting adaptation performance using a regression model. In doing so, we provide interesting insights on the effect of various properties of different discourse models/datasets on performance in domain adaptation, which we enumerate in Section 6. We expand on these contributions next.

First, we contribute a new theoretical analysis to characterize the bias of the h-discrepancy as an estimator of performance in domain adaptation. Although this discrepancy is typically biased, we provide upper and lower bounds on this bias and interpret them to provide insight on the use of this statistic in practice. In particular, we show that a small h-discrepancy often means the practitioner can be confident in transferring the model from the train- to the test-set. Our theoretical analysis motivates our hypothesis that the h-discrepancy should outperform common two-sample statistics.

Next, we empirically study the aforementioned hypothesis. We compare correlation of the h-discrepancy with performance in domain adaptation against correlation of various two-sample statistics across multiple discourse datasets. As we are aware, this large-scale comparison has never been done for discourse relation classification.

We also perform a regression analysis of the estimation errors of the h-discrepancy as an estimator for domain adaptation performance. This analysis not only allows us to understand the properties and pitfalls of our estimator, but also to gain more insight about how different types of datasets, genres, feature representations, and models influence domain adaptation performance.

Lastly, we plan to release our code for calculating h-divergence as a Python package, so that other practitioners can easily make use of it for their own model setups and datasets.

2 Related Work

2.1 Discourse and Domain Shift

Computational analysis of discourse has been the focus of several shared tasks (Xue et al., 2015, 2016), and there have been several discourse-annotated corpora for multiple languages (Zeyrek and Webber, 2008; Meyer et al., 2011; Danlos et al., 2012; Zhou and Xue, 2015; Zeyrek et al., 2020). However, discourse models have been shown not to perform well under even gradual domain shift (Atwell et al., 2021), which may be the result of the limited timeframe and distribution of the articles contained in the most commonly used discourse datasets, the Penn Discourse Treebank (Miltsakaki et al., 2004; Prasad et al., 2008; Webber et al., 2019) and the RST (Carlson et al., 2003). These datasets are both made up of Wall Street Journal articles spanning a three-year period, and thus do not contain much variation with respect to linguistic distribution.

Several works have quantified domain shift in the context of natural language processing, mostly in the task of sentiment analysis. For instance, Plank and Van Noord (2011) use word frequencies and topic models to measure domain similarity, while Wu and Huang (2016) use sentiment graphs.

Blitzer et al. (2007) and Elsahar and Gallé (2019) use $\mathcal{H}$-divergence to analyze a sentiment classification task on the Amazon Reviews dataset, while Ruder et al. (2017) use $\mathcal{H}$-divergence to select the source datasets for transfer. However, none of these works have studied the h-discrepancy we study here, nor have they studied this with respect to discourse parsing.

To the best of our knowledge, no works have yet studied the correlation of statistics from the theoretical domain adaptation literature with the adaptation performance of discourse parsers. This is especially true given the wide array of different datasets and distributional shifts we consider as well as the theoretical and empirical tools we propose to conduct our study. Both our novel theoretical result (Theorem 1) and our large-scale regression analysis (Section 5), provide new, practical insights on domain-shift in discourse parsing.

2.2 Domain Adaptation Theory

Statistics that relate to domain adaptation performance have long been studied in the theoretical literature. Kifer et al. (2004); Ben-David et al. (2007, 2010a) initiate this investigation with a modification of the total variation distance (the $\mathcal{H}$-divergence) that depends on the set of classifiers $\mathcal{H}$; this statistic can be directly related to adaptation performance through a finite-sample bound. Mansour et al. (2009) extend this discussion from classification error to general loss functions. Certain two-sample statistics can also be related to adaptation performance through finite sample bounds, but only under stringent assumptions on the space of classifiers and the computation of the two-sample statistic (Fukumizu et al., 2009; Gretton et al., 2012; Long et al., 2015; Redko et al., 2020). Assumptions, in general, play a large role in successful domain adaptation. In fact, common adaptation algorithms can actually *worsen* performance if important assumptions are not met (Zhao et al., 2019). Different assumptions have led to diverse theories disjoint from the $\mathcal{H}$-divergence, including the proposals of Lipton et al. (2018), Johansson et al. (2019), and Tachet des Combes et al. (2020). Under certain strict and untestable assumptions, it is even possible to derive unbiased estimators of adaptation performance (Sugiyama et al., 2007; You et al., 2019). We later discuss our own assumptions on the *ideal-joint error* λ which are typical when using the $\mathcal{H}$-divergence and its descendants. We find this assumption to be more mild than those in other mentioned works. In any case, this assumption has been shown to be *necessary* for the adaptation algorithms we study (Ben-David et al., 2010b).

3 Methods

Data In Table 1, we describe the discourse datasets we use. For information on the distinctions between these frameworks, see Appendix A.

Data splitting For each individual dataset, we randomly split the dataset in half based on 3 different seeds. For example, PDTB 2.0 (10K examples)

Dataset	Genre	Label schema	Intra-sent
RST (Carlson et al., 2003)	News	RST	Yes
PDTB 2.0 (Prasad et al., 2008)	News	PDTB	No
PDTB 3.0 (Webber et al., 2019)	News	PDTB	Yes
BioDRB (Ramesh and Yu, 2010)	Bio	PDTB	No
TED-MDB (Zeyrek et al., 2020)	TED talks	PDTB	Yes
GUM (Zeldes, 2017)	Multiple	RST	Yes

Table 1: Characteristics of each discourse dataset used in our study. The "multiple" domains in the GUM corpus are as follows: Academic, Biography, Fiction, Interview, News, Reddit, Travel, and How-to guides. For the datasets with the PDTB label schema, we use only the top-level sense labels (Expansion, Contingency, Comparison, and Temporal). We use the top-level RST classes for the datasets with the RST label schema, and map the GUM corpus classes to the RST classes using Braud et al. (2017).

is randomly split into to disjoint sets of about 5K examples. This allows us to test variability in the transfer from source to target datasets as well as test transfer in case of within-distribution shifts.

Features To encode argument pairs, we concatenate and tokenize them using the BERT (Devlin et al., 2018) tokenizer. We then feed these tokens through the pretrained base BERT model and experiment with two different ways of capturing the model output: using the pooled output, e.g. the output of the [CLS] token, and averaging the hidden states. We will refer to these encodings as P-BERT and A-BERT respectively. We also experiment with encoding our argument pairs using Sentence-BERT (Reimers and Gurevych, 2019) which we will refer to as S-BERT.

Experiments Each data point in all of our results (e.g., when computing correlation or doing regression analysis) corresponds to a particular *experiment* done on a source (train) dataset S and target (test) dataset T using a classifier h. The classifier h is trained on the source S and evaluated on target T. This is meant to mimic a common domain adaptation scenario in which the NLP practitioner would like to transfer a pre-trained discourse classification model to a new unlabeled dataset (i.e., this is discussed again in Section 4). For each experiment, h is trained using a standard optimization procedure to have low error on S. We discuss this procedure and its competitiveness with respect to the state-of-the-art in Section 5. To account for variability introduced by this procedure, we test 3 random seeds (i.e., producing 3 different models).

The pair S and T are taken from the set of data splits outlined previously using each of the different BERT representations. We restrict the pair to have a common set of discourse labels. For example, we only transfer from S using the PDTB label schema to T using the same schema. We discuss specific (S, T) pairs in detail in Appendix B. Accounting for each pair and each random seed for model training, the number of (S, T, h) triples we study totals more than 2400.

4 Quantifying Meaningful Domain Shift

Identifying and quantifying domain shift is a classical problem. Perhaps, the most widely used mechanism for this task is the two-sample test; i.e., a test designed to indicate difference of distribution between two samples. We begin this section by discussing a few of the statistics used in these tests. We observe a common problem in using these statistics to predict adaptation performance, and following this, discuss the aforementioned h-discrepancy.

4.1 Common Two-Sample Test Statistics

We now informally discuss some common statistics used in two-sample tests. These statistics can be easily adapted to infer adaptation performance under the assumption that changes in distribution perfectly correlate with changes in error. As mentioned earlier, we do *not* agree with this hypothesis. Still, these types of statistics serve as a good point of comparison. In our experiments, we compute each of these statistics using the PyTorch library `torch_two_sample` (Cruceru et al., 2020).

- **FRS**: (Friedman and Rafsky, 1979) counts edges from S to T in a graph representation.
- **Energy**: (Székely and Rizzo, 2013) compares dissimilarity of points within/across S and T.
- **MMD**: (Gretton et al., 2012) compares *similarity* of points within/across S and T.
- **BBSD**: (Lipton et al., 2018) applied MMD to softmax output (i.e., scores) of classifier h.

For more computational details, see Appendix D. The important thing is to understand the shortcoming of these approaches which we discuss next.

A Common Problem The majority of these statistics share the common trait that they were originally designed to test differences in feature distribution – *not* differences in hypothesis error. As such, while we do expect them to be sensitive to changes in error – in so far as changes in feature distribution relate to changes in error – we have no theoretical reason to expect this should be the case. As we saw in Figure 1, these two changes can be very different: large changes to the distribution of features may not hurt performance in every case and imperceptible changes to the distribution of features can have large impact when the labeling function changes. In fact, most of these statistics do not even incorporate information about the classifier we use for inference. While BBSD does, we are not aware of any theoretical arguments linking it to adaptation performance in the same way as the h-discrepancy (discussed next).

4.2 Identifying the Change that Matters

Contrary to those statistics described above, the statistic we give in this section is directly related to adaptation performance by theoretical means. Before beginning our description of this metric, we need to formalize our mathematical setup and a particular notion of adaptation performance.

Mathematical Setup We measure adaptation performance through the error-gap which is defined:

$$\Delta_h(S, \mathbb{T}) = |\mathbf{R}_S(h) - \mathbf{R}_{\mathbb{T}}(h)| \tag{1}$$

where S is a sample and $\mathbb{T}$ is a distribution – both over a space $\mathcal{X} \times \mathcal{Y}$. In this paper, $\mathcal{X}$ is usually the space of real-valued vectors (i.e., BERT representations for argument pairs) and $\mathcal{Y}$ corresponds to a set of possible discourse labels. h is a classifier $h : \mathcal{X} \to \mathcal{Y}$ and the risk $\mathbf{R}_{\mathbb{D}}(h)$ is defined for distribution $\mathbb{T}$ as $\mathbf{R}_{\mathbb{T}}(h) = \mathbf{Pr}(h(\tilde{X}) \neq \tilde{Y}),\ (\tilde{X}, \tilde{Y}) \sim \mathbb{T}$. For sample $S = (X_i, Y_i)_{i=1}^n$, we instead write $\mathbf{R}_S(h) = n^{-1} \sum_i 1[h(X_i) \neq Y_i]$ where $1[\cdot]$ is the indicator function. To compute each statistic which we would like to use to infer the error-gap, we assume access to the mentioned sample S drawn i.i.d from some distribution $\mathbb{S}$. We also assume access to a new unlabeled sample $T_X = (\tilde{X}_i)_{i=1}^m$ drawn i.i.d from the $\mathcal{X}$-marginal $\mathbb{T}_X$ of the distribution $\mathbb{T}$. In general, we do not know whether $\mathbb{T} \neq \mathbb{S}$ or $\mathbb{T} = \mathbb{S}$, but may have reason to suspect $\mathbb{T} \neq \mathbb{S}$.

Roadmap In the next part, we give the statistic we would like to use to predict adaptation performance. We then quantify its bias as an estimator for the error-gap with a theoretical result. We also propose a technique to study the relationship between this statistic and the error-gap empirically through a regression analysis. Finally, we show how this technique can be used to study the impact certain attributes of a model or dataset have on error-gap.

Source-Guided Discrepancy The source-guided discrepancy was proposed by Kuroki et al. (2019) with a similar conceptualization given independently by Zhang et al. (2019). These statistics improve upon a long history of domain adaptation statistics (Kifer et al., 2004; Blitzer et al., 2007; Ben-David et al., 2007, 2010a), specifically, by incorporating information on the source-labels. We consider a generalization of the source-guided discrepancy which we call the h-discrepancy, defined for any classifier h. For samples S and T_X, a binary label space $\mathcal{Y}$, a space of classifiers $\mathcal{H}$ over $\mathcal{X} \times \mathcal{Y}$, and any[1] fixed classifier $h \in \mathcal{H}$, it is defined as:

$$\begin{aligned} &D = \max_{g \in \mathcal{H}} |\mathbf{R}_U(g) - \mathbf{R}_V(g)| \quad \text{where} \\ &U = ((X_i, h(X_i))_{i=1}^n, \quad V = ((\tilde{X}_i, h(\tilde{X}_i))_{i=1}^m, \end{aligned} \tag{2}$$

and recall, $S_X = (X_i)_i$ and $T_X = (\tilde{X}_i)_i$. In the binary case, Kuroki et al. (2019) show that this may be approximated by learning a classifier (i.e., g) which *agrees* with h on the source sample S_X and *disagrees* with h on the target sample T_X. Their procedure extends naturally to the multi-class case as well, but we must disambiguate between the possible ways in which g can disagree with h. In our experiments, we do so by training g to pick the next most likely label according to the scores of h.

Theoretical Motivation Here, we provide our primary motivation for the h-discrepancy as an estimator of error-gap. Our result makes use of the work of Crammer et al. (2007), Ben-David et al. (2010a), and Kuroki et al. (2019). It distinguishes itself from these finite-sample bounds in that it explicitly concerns itself with the bias of D as an estimator of error-gap. Proof is given in Appendix E.

Theorem 1. *Let $\mathcal{Y}$ be a binary space and let $\mathcal{H}$ be a subset of classifiers in $\mathcal{Y}^{\mathcal{X}}$. Then, for any realization of S, for all $h \in \mathcal{H}$,*

$$-\mathbf{E}_T[\lambda] \leq \mathbf{E}_T[D] - \Delta_h(S, \mathbb{T}) \leq \mathbf{E}_T[D] \tag{3}$$

where $\lambda = \min_{h' \in \mathcal{H}} \mathbf{R}_S(h') + \mathbf{R}_T(h')$ is called the ideal-joint error.

Notice, when $\mathbf{E}[\lambda]$ is small and $\mathbf{E}[D]$ is also small we know the bias must be small because it is "sandwiched" between these two. In this situation, the practitioner can very confidently transfer h from S to T. In practice we cannot compute λ since it requires labels from T, but we often do expect $\mathbf{E}[\lambda]$ to have small magnitude. As first observed by Ben-David et al. (2010a) (i.e., concerning a similar term), λ will be small whenever there is *any* classifier in $\mathcal{H}$ which does well on S and T simultaneously. This is not an overly strong requirement as neural-networks, for example, have been shown to perfectly fit even random labeling (Zhang et al., 2016). Thus, we are primarily concerned with the positive bias of D. When $\mathbf{E}[D]$ is larger, the positive bias of D can also be larger. Intuitively, D might have more "false positives" where it reports a high value but the error-gap is actually comparatively small. In this sense, it is a conservative statistic. It plays things on the "safe side." So, while D will possibly have some bias, it is at least described by the above bounds. As we are aware, the two-sample statistics discussed previously do not have such a description.

Regression Analysis of Errors of D From Theorem 1, we do not expect the random estimation error $D - \Delta_h(S, \mathbb{T})$ to be zero. So, in our experimentation, we propose to study this quantity through a regression analysis. Namely, suppose $\mathbf{X} \in \mathbb{R}^{N \times p}$ is some fixed, non-singular design matrix whose rows each represent one of N experiments and whose columns represent one of p features for each experiment. An experiment corresponds to an (S, T, h) triple as disucssed in Section 3. The features are dependent on properties of the datasets and models used in each experiment as well as realizations of h-discrepancy, ideal-joint error, and training error. Then, we assume

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4}$$

where the randomness in the outcome $\mathbf{Y}$ comes from the error-terms $\boldsymbol{\epsilon}_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, $\sigma > 0$. The response $\mathbf{Y} = (D_i - \Delta_h(S, \mathbb{T})_i)_{i=1}^N$ are the realizations of the error across the N experiments.[2] We give details of the design matrix $\mathbf{X}$ in Appendix F; it is selected manually using domain knowledge and to best meet the model assumptions model. Model diagnostics are also provided in Appendix F.

Regression analysis is particularly useful because standard techniques allow us to understand and isolate the impact of individual columns (i.e., features) in $\mathbf{X}$ on the estimation errors of D. In

[1] The source-guided discrepancy originally proposed by Kuroki et al. (2019) considers only one particular h.

[2] We do not have access to $\mathbb{T}$, so we use sample T instead.

particular, we can use this model to determine the expected change in estimation error as a function of a particular feature, while controlling (i.e., holding constant) all other features in $\mathbf{X}$:

$$\mathbf{E}[\mathbf{Y}_i \mid \mathbf{X}_i = \mathbf{x}] - \mathbf{E}[\mathbf{Y}_i \mid \mathbf{X}_i = \mathbf{x}'] \tag{5}$$

where $\mathbf{x}$ is any setting of the features and $\mathbf{x}'$ is identical to $\mathbf{x}$ except every component involving the feature of interest is modified (e.g., increased) systematically. For a specific example using Eq. (5), consider inspecting the change in estimation error as a function of increase in h-discrepancy (controlling for all other features). In this case, Eq. (5) evaluates to a polynomial[3] in the coefficients $\boldsymbol{\beta}$ and components of $\mathbf{x}'$, so we can estimate this result in an unbiased manner using the OLS estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$. To empirically validate our theoretical analysis, we might check if this polynomial is an increasing, positive function; i.e., because our theory predicts increases in the expected h-discrepancy allow for increases in bias.

Regression Analysis of Error-Gap Given $\mathbf{X}$ and $\boldsymbol{\beta}$, rearranging Eq. (4) lets us also write

$$\Delta_h(S, \mathbb{T})_i = D_i - \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{6}$$

where $\mathbf{X}_i$ is the i^{th} row of $\mathbf{X}$; i.e., the features of the i^{th} experiment. Similar to before, this type of analysis lets us draw interesting insights. In particular, we can isolate the impact of features in $\mathbf{X}$ on the error-gap. Since our design matrix $\mathbf{X}$ controls for training error, the error-gap can be interpreted to act as a measure of performance in domain adaptation (DA). Those features which are positively associated with error-gap can be said to be worse for DA. Likewise, those with negative association are "better" for DA. As before, we isolate the impact of a feature by checking the change in error-gap as a function of change in this feature (i.e., similar to Eq. 5). In Appendix G Example 2, we use this technique to compare the impact of different BERT representations on error-gap, while also controlling for the other features in $\mathbf{X}$.

5 Results

5.1 Analysis of Transfer Error

Comparison to Other Work *Our experimental setup produces results comparable to current discourse models.* In Appendix C, Figure 3 shows the distribution of the error rates when transferring on within- and out-of-distribution datasets. To validate whether our setup is comparable to other discourse parsing models, we compare error rates to current implicit sense classifiers; e.g., Kishimoto et al. (2020) who achieve an error rate of ≈ 0.38 under a comparable setup. Our PDTB within-distribution results often improve upon this.

[3] For details, please see Appendix G, Example 1.

Error Analysis Across Genres *Fiction and How-To Guides are the most difficult to transfer to, while Academic Journals and Biographies are the easiest.* Figure 4 in Appendix C shows the error rates for multi-source adaptation on the GUM corpus across S-BERT, P-BERT, and A-BERT. Although the error rates differ across these three representations, the relative order of the GUM corpus domains with respect to transfer error is fairly consistent across all of them. For all three, the highest mean error rate occurred in the How-to Guide and Fiction domains, and the lowest mean error rate occurred in the Academic and Biography domains.

5.2 Analysis of Correlations

In Table 2, we show linear and rank correlation of each statistic with the error-gap. This tests the ability of each statistic to discern scenarios where domain adaptation performance may be either good or bad. In practice, a statistic with good rank correlation can be used in model-selection or (source) dataset selection. A statistic with good linear correlation may also be used and will be more easy to interpret since we expect changes in the statistic to be proportional to changes in the error-gap.

Comparison of Statistics *h-discrepancy is consistently, most strongly correlated with error-gap.* The overarching trend is that the h-discrepancy is far better than every other statistic with regards to both types of correlation. In fact, the linear correlations are not much worse than the rank correlations (in some cases they are even better). This validates our opening hypothesis that domain-shift does not always correlate with domain adaptation performance (i.e., error-gap). It is important to also consider the classifier we use. Still, BBSD – another statistic that relies on the classifier – is also somewhat ineffective compared to the h-discrepancy. Importantly, despite depending on the classifier, BBSD was still designed with identification of feature-distribution shift in mind. In some sense, this observation validates our theoretical motivations for the h-discrepancy (i.e., Theorem 1) which directly relates it to error-gap. Our results indicate

Split	Spearman (Rank) Correlation FRS	Energy	MMD	BBSD	h-disc	Pearson (Linear) Correlation FRS	Energy	MMD	BBSD	h-disc
All	0.5394	0.6059	0.5051	0.4054	**0.8299**	0.4986	0.4396	0.3413	0.4004	**0.7628**
PDTB	0.5451	0.6359	0.5472	0.4746	**0.8265**	0.5295	0.4704	0.3709	0.4274	**0.7642**
RST	0.2166	0.3059	-0.0011	0.2087	**0.7625**	0.2853	0.1660	-0.1605	0.1677	**0.7599**
News	0.5262	0.6356	0.5507	0.5759	**0.8517**	0.7079	0.6302	0.5558	0.5386	**0.8890**
Other	0.3760	0.4517	0.2767	0.1737	**0.8386**	0.3420	0.2791	0.1760	0.2051	**0.7072**
WD	0.0884	0.5735	-0.0324	0.2368	**0.7890**	0.1075	0.5831	-0.0515	0.4853	**0.9519**
OOD	0.4597	0.5249	0.3917	0.2813	**0.7666**	0.4342	0.3909	0.2761	0.3745	**0.6976**

Table 2: Correlations with error-gap for each statistic. Data splits indicate the subset of data used. h-discrepancy consistently yields the largest correlation with error-gap; i.e., difference in Pearson correlations are all significant at level $\alpha = 0.001$ using test of Steiger (1980) implemented by Diedenhofen and Musch (2015).

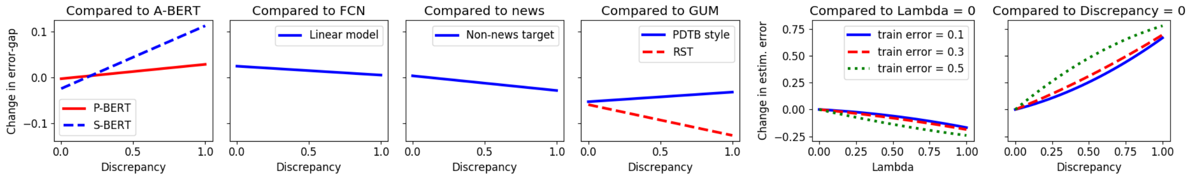

Figure 2: (Left, 1-4) Expected change in error-gap when changing properties of the dataset or model. Shown as a function of discrepancy and controls for all other features of the experiment. Reference category is indicated in title. (Right, 5-6) Expected change in estimation error of h-discrepancy shown as a function of λ (5th) and discrepancy (6th). Left assumes use of A-BERT and FCN on a GUM non-news target, but trends are consistent in other cases.

that, at least for the task of discourse parsing, h-discrepancy is the most effective statistic to use with regards to predicting error-gap.

Additional Trends *Experiments using RST label schemas and non-news targets show very low correlation between distributional shift and error-gap.* If we look at particular experiment subsets, we also see some interesting trends. First, most statistics are better correlated with error-gap datasets that use the PDTB label schema than those that use the RST label schema. The difference is less pronounced for the h-discrepancy than for the other statistics, suggesting that it is especially important to use statistics tied directly to the error-gap when working with datasets that use the RST schema. The same is true when the test dataset is comprised of news articles instead of other types of text.

The h-discrepancy has highest linear correlation on similar distributions. We observe much stronger linear correlation between the h-discrepancy and error-gap on within-distribution adaptation scenarios (**WD**) as compared to out-of-distribution adaptation scenarios (**OOD**). We believe this is because the h-discrepancy is typically small when S and T follow a similar distribution. As Theorem 1 notes, the bias of the h-discrepancy as an estimator for error-gap can be near zero if both $\mathbf{E}[D]$ and $\mathbf{E}[\lambda]$ are small; i.e., we expect the linear correlation of a nearly unbiased estimator to be fairly high.

5.3 Regression Analysis of Estimation Error

Figure 2 shows expected change in estimation error of h-discrepancy (used as an estimator for error-gap). Trend lines indicate expected change as a function of the ideal joint error λ and the discrepancy D compared to the case where each is 0.[4] Trend lines are computed using a similar technique for regression analysis as described in Appendix G Example 1. The key takeaway is that these empirical results are consistent with our theoretical discussion surrounding Theorem 1. As λ increases, the estimation error decreases. Similarly, Theorem 1 shows us that we incur the possibility of negative bias when λ is large. As D increases, the estimation error does the same. Theorem 1 agrees here too, predicting the possibility of positive bias proportional to the discrepancy D.

5.4 Regression Analysis of Error-Gap

Figure 2 also shows expected change in error-gap when modifying categorical features of the experiment; e.g., use of S-BERT vs. A-BERT. Trend lines indicate expected change as a function of

[4] Note, if both are 0 in expectation, D is unbiased.

h-discrepancy and are computed using a similar technique for regression analysis as described in Appendix G Example 2. Since we control for training set error, positive changes in error-gap indicate a setting is better for domain adaptation, while negative indicates the opposite.

BERT Features *S-BERT is better for similar train and test sets, while A-BERT is better for more divergent sets.* As a function of discrepancy, S-BERT is better for DA when the discrepancy is small. As the difference between the train and test set increases, the reference category (i.e., A-BERT) is better for DA. Comparing P-BERT to A-BERT we do not see large differences; marginally, A-BERT is better as the domain divergence increases. These results are consistent with typical rules of thumb on model complexity. A more complex feature representation (i.e., from S-BERT or P-BERT) is beneficial when the training and test distributions align, but allows for the possibilty of overfitting when discrepancy increases.

Classifier *Linear classifiers perform marginally worse than neural-networks.* In general, the FCN appears to be slightly better for domain adaptation. Possibly, this is due to increased modelling capacity. This benefit wanes as the discrepancy between the training/test sample increases. As before, this may be explained by overfitting, as overfitting and class imbalance are known problems in discourse parsing (Atwell et al., 2021).

News Test Set *It is slightly harder to transfer to news datasets.* We consider a "news" corpus to be any of PDTB, RST, or the news domain of GUM. When the target (test) dataset consists of news texts, we see adaptation performance consistent with non-news targets for small discrepancy. As the discrepancy between training and test set grows, the non-news targets are actually better suited for domain adaptation. That is, it is slightly easier to transfer to a non-news target. Possibly, this is related to the length and complexity of news texts.

Dataset *A more variable label schema results in a more difficult task, even when adding variability during training.* In general, we see that the GUM dataset presents a more challenging adaptation task than the other datasets. This is sensible due to the larger selection of domains present in GUM. Assuming this larger selection induces increased variability in the annotations for each domain, we should expect larger error-gap when doing adaptation. Based on our results, increased variability at train-time does not seem to counteract this issue, because adaptation experiments in the GUM corpus are all multi-source (i.e., see Appendix B for more details). For PDTB, as the discrepancy increases, performance is more similar to GUM. On the other hand, RST presents the easiest adaptation task. This is expected as all test sets in the RST splits should follow the same distribution as the training set because both sets are drawn from the same news corpus (see Appendix B). The continued improvement as discrepancy decreases does not make as much sense. Likely, this is due to a lack of experiments with large discrepancy in RST, since all RST experiments are within-distribution.

6 Conclusion

In the experiments above, we find several interesting results relevant to the field of discourse analysis. For one, analysis of the correlations indicates that, for datasets with the RST labeling schema, the statistics not tied directly to error-gap are very weakly correlated with error-gap. This also holds for non-news targets, and indicates that the h-discrepancy is especially useful in these cases.

Additionally, our regression analysis provides the following insights, all of which may be useful to future discourse researchers: (1) more variability in the labeling schema appears to make domain adaptation more difficult, even if the training set contains a similar level of variability; (2) S-BERT is better than A-BERT when domains are similar, but A-BERT outperforms S-BERT when the domains further diverge; (3) non-news texts are easier to adapt to than news texts.

This is the first computational and empirical study that looks at distribution shifts across different discourse datasets and evaluates the performance of various models under these shifts. This is also the first work that examines the efficacy of different two-sample tests for predicting the error-gap when compared to a metric that is theoretically tied to error gap. Future work can extend these results by using the h-discrepancy metric to predict the error-gap for other NLP tasks or for other components needed for discourse parsing, such as constructing the RST dataset. We hope practitioners will find both our insights and our code useful for model/dataset selection.

References

Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. Where are we in discourse relation recognition? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–325.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010a. 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 *Advances in Neural Information Processing Systems*, pages 137–144.

Shai Ben-David, Tyler Lu, Teresa Luu, and David Pal. 2010b. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 129–136, Chia Laguna Resort, Sardinia, Italy. PMLR.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal. Association for Computational Linguistics.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. 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 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Koby Crammer, Michael Kearns, and Jennifer Wortman. 2007. Learning from multiple sources. In *Advances in Neural Information Processing Systems*, pages 321–328.

Calin Cruceru, Brady Neal, and Github User `josipd`. 2020. Torch two sample package (commit 23aa002).

Laurence Danlos, Diégo Antolinos-Basso, Chloé Braud, and Charlotte Roze. 2012. Vers le fdtb: French discourse tree bank. In *TALN 2012: 19ème conférence sur le Traitement Automatique des Langues Naturelles*, volume 2, pages 471–478. ATALA/AFCP.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Birk Diedenhofen and Jochen Musch. 2015. cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one*, 10(4):e0121945.

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.

Jerome H Friedman and Lawrence C Rafsky. 1979. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, pages 697–717.

Kenji Fukumizu, Arthur Gretton, Gert Lanckriet, Bernhard Schölkopf, and Bharath K. Sriperumbudur. 2009. Kernel choice and classifiability for rkhs embeddings of probability distributions. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. 2019. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536. PMLR.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. 2004. Detecting change in data streams. In *VLDB*, volume 4, pages 180–191.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France. European Language Resources Association.

Seiichi Kuroki, Nontawat Charoenphakdee, Han Bao, Junya Honda, Issei Sato, and Masashi Sugiyama. 2019. Unsupervised domain adaptation based on source-guided discrepancy. In *Proceedings of the*

AAAI Conference on Artificial Intelligence, volume 33, pages 4122–4129.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. 2018. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2009. Domain adaptation with multiple sources. In *Advances in neural information processing systems*, pages 1041–1048.

Daniel Marcu. 1999. Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, 293:123–136.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT press.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual annotation and disambiguation of discourse connectives for machine translation. In *Association for Computational Linguistics-Proceedings of 12th SIGdial Meeting on Discourse and Dialogue*, CONF.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Karthik Narasimhan and Regina Barzilay. 2015. Machine comprehension with discourse relations. 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 1253–1262.

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.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. 2019. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32:1396–1408.

Balaji Polepalli Ramesh and Hong Yu. 2010. Identifying discourse connectives in biomedical text. In *AMIA Annual Symposium Proceedings*, volume 2010, page 657. American Medical Informatics Association.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. 2020. A survey on domain adaptation theory. *ArXiv*, abs/2004.11829.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sebastian Ruder, Parsa Ghaffari, and John G Breslin. 2017. Data selection strategies for multi-domain sentiment analysis. *arXiv preprint arXiv:1702.02426*.

James H Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. 2007. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5).

Gábor J Székely and Maria L Rizzo. 2013. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. 2020. Domain adaptation with conditional distribution matching and generalized label shift. In *Advances in Neural Information Processing Systems*, volume 33, pages 19276–19289. Curran Associates, Inc.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 annotation manual.

Fangzhao Wu and Yongfeng Huang. 2016. Sentiment domain adaptation with multiple sources. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 301–310.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task*, pages 1–16.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. ConLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19.

Kaichao You, Ximei Wang, Mingsheng Long, and Michael Jordan. 2019. Towards accurate model selection in deep unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 7124–7133. PMLR.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. Ted multilingual discourse bank (tedmdb): a parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, 54(2):587–613.

Deniz Zeyrek and Bonnie Webber. 2008. A discourse resource for Turkish: Annotating discourse connectives in the METU corpus. In *Proceedings of the 6th workshop on Asian language resources*.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530v2*.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. 2019. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. 2019. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532. PMLR.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9709–9716.

Yuping Zhou and Nianwen Xue. 2015. The Chinese discourse treebank: a Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397–431.

A Dataset Descriptions

A.1 Frameworks

The Penn Discourse Treebank (Miltsakaki et al., 2004; Prasad et al., 2008; Webber et al., 2019) is a dataset that consists of Wall Street Journal articles labeled with shallow discourse relations, or relations that occur only between two argument spans with no additional context needed. It consists of several different types of coherence relations: explicit, implicit, AltLex, EntRel, and NoRel. Explicit discourse relations are ones in which a *connective* between the arguments provides some indication of the correct discourse sense label. Implicit discourse relations, the main ones that we will be focusing on in this paper due to the difficulty of classifying them, are ones in which a connective can be *inserted* that indicates the correct sense.

The RST Discourse Treebank (Carlson et al., 2003) is a corpus containing Wall Street Journal articles annotated in the style of Rhetorical Structure Theory, where a document is split into *elementary discourse units* (EDUs) and relations made up of these EDUs form a tree structure. The RST Discourse Treebank does not differentiate between explicit and non-explicit discourse relations, nor does it label discourse connectives.

B Adaptation Scenarios

For experiments involving PDTB label schema, we consider single-source domain adaptation. Specifically, single-source adaptation involves simply pairing one data split S with another T; for instance, the first half of the TED-MDB and the second half of the BioDRB. Or, the first half of BioDRB and the second half of BioDRB. The former allows us to investigate scenarios with significant domain-shift while the latter allows us to investigate scenarios where there is likely less domain-shift. The former also allows us to investigate variability arising from a particular sample (i.e., data split).

For experiments involving RST label schema, we used both single-source and multi-source domain adaptation setups. We use the multi-source setup for domains in the GUM corpus. Here, T is derived from a single domain and S from all of the other domains contained in the corpus (i.e., S would contain 7 of the GUM domains and T would contain the remaining one). Although we continue to split the domains in half, we only use one of the halves for our experiments in order to prevent samples from the target distribution from appearing in the source. We use the single-source setup for RST itself. Here, S is one split of RST while T is another. So, the multi-source setup within the RST style allow us to test cases with larger domain-shift, while the single-source setup within the RST style allow us to test much smaller domain-shift.

C Model Training and Transfer Results

Optimization Parameters We use SGD on an NLL loss with momentum set to 0.9 to train all of our models. We use a batch size of 250. We start training with a learning of 1×10^{-2} for 100 epochs and then train for another 50 epochs using a learning rate of 1×10^{-3}. If a model achieves a training error lower than 5×10^{-4}, we stop training.

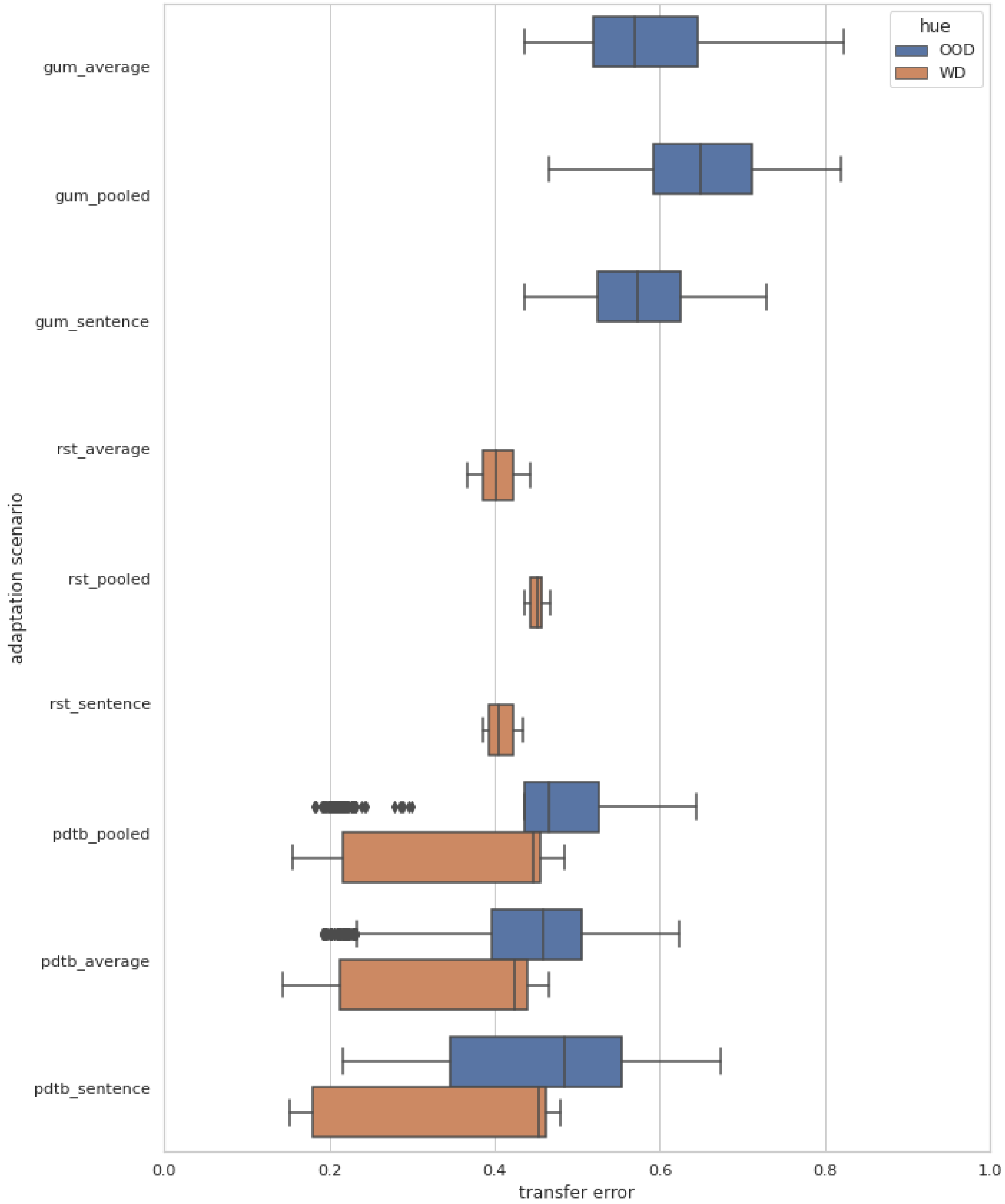

Figure 3: Transfer error within and out of distribution for each dataset

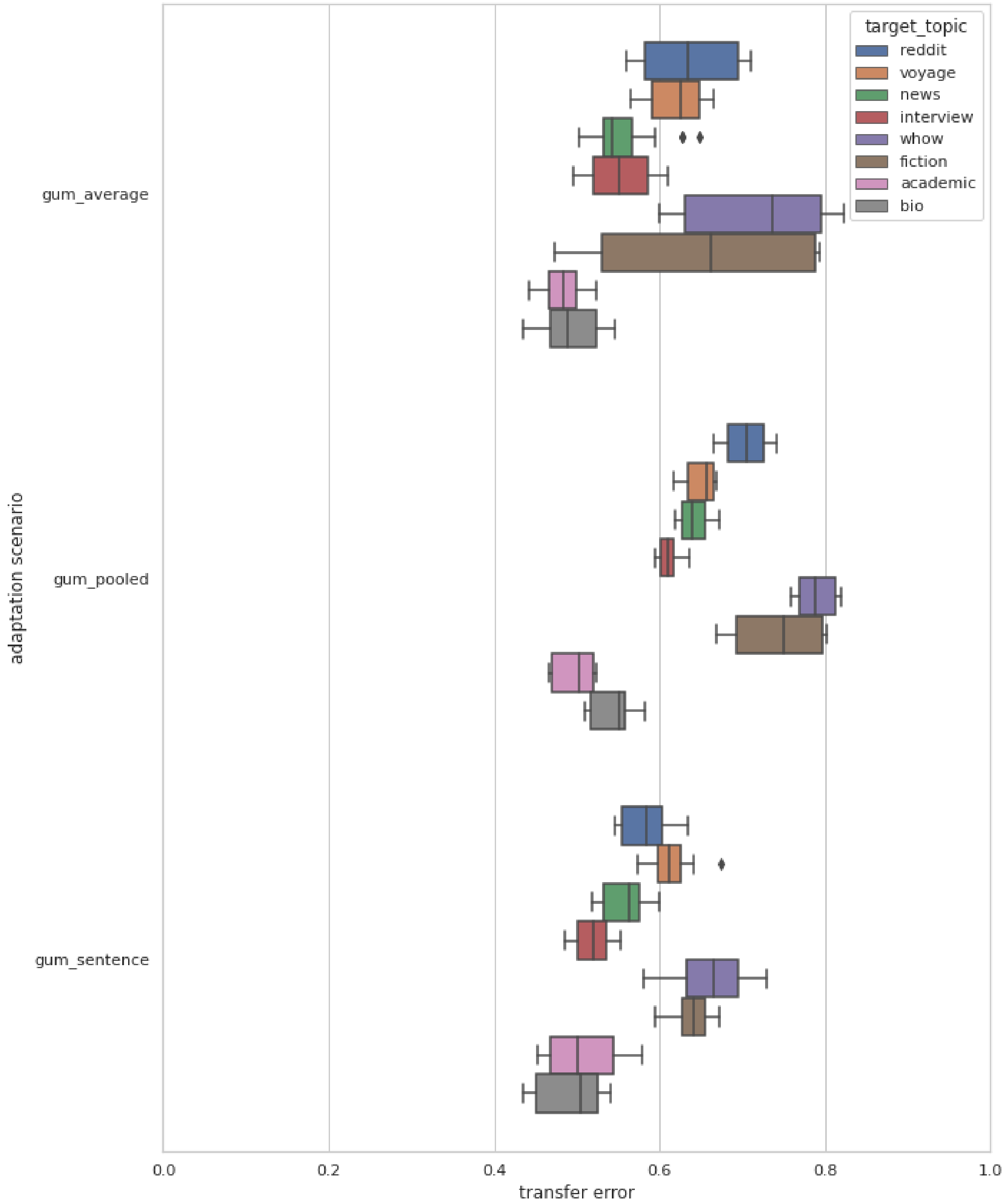

Figure 4: Transfer error for each topic within the GUM corpus

D Two-Sample Statistics

Here, we describe in detail the common two-sample statistics listed in Section 4 and studied in Section 5

Friedman-Rafsky Test Statistic The Friedman-Rafsky Test Statistic R (Friedman and Rafsky, 1979) is computed by forming a minimum-spanning tree (MST) using the pooled sample $P = (X_i \mid (X_i, Y_i) \in S) +_{\mathrm{c}} (\tilde{X}_i \mid \tilde{X}_i \in T_X)$ of marginal features. Here, $+_{\mathrm{c}}$ is the concatenation operation. To form the tree, we form a weighted graph G_P by treating each point $Z_i \in P$ as vertex and assigning an edge between each pair of vertices whose weight is the distance between the data-points. When $\mathcal{X} = \mathbb{R}^d$ for some d, this is usually the Euclidean distance or L2 norm. The MST is then precisely the MST of G_P. The statistic R is computed as the number of edges whose endpoints originally belonged to the same sample. For example, R increases by 1 for each edge whose endpoints both originally belong to T_X. Likewise, R increases by 1 for each edge whose endpoints are both the features of points in S. When endpoints originally belonged to distinct samples, R remains unmodified. We report modified statistic below which is normalized to account for sample size $R_{\text{normed}} = R/(n + m - 2)$. Since the size of the MST is $n + m - 1$ and there is always at least one edge between S and T_X, this statistic has a maximum value of 1.

Energy Statistic Given samples S and T_X as before, the energy statistic may be computed as below

$$\begin{aligned} E = &\frac{2}{nm} \sum\nolimits_{i,j} ||X_i - \tilde{X}_j|| - \frac{1}{n^2} \sum\nolimits_{i,j} ||X_i - X_j|| \\ &- \frac{1}{m^2} \sum\nolimits_{i,j} ||\tilde{X}_i - \tilde{X}_j|| \end{aligned} \tag{7}$$

where $|| \cdot ||$ gives the Euclidean norm (distance). Originally proposed by Székely and Rizzo (2013), the statistic is motivated by Newton's potential energy between heavenly bodies. Intuitively, it is fairly easy to understand as a comparison of dissimilarity within samples and across samples. If the dissimilarity across samples (i.e., the first term) is much higher than the dissimilarity within samples, then the two samples are likely drawn from different distributions.

Maximum Mean Discrepancy (MMD) Given samples S and T_X as before, the MMD statistic (Gretton et al., 2012) may be computed as below

$$\begin{aligned} M = &\frac{\sum_{i \neq j} K(X_i, X_j)}{n(n-1)} + \frac{\sum_{i \neq j} K(\tilde{X}_i, \tilde{X}_j)}{m(m-1)} \\ &- \frac{2}{nm} \sum\nolimits_{i,j} K(X_i, \tilde{X}_j) \end{aligned} \tag{8}$$

where $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is the kernel for some RKHS. In our experiments, we use an Gaussian RBF kernel and select σ to be an approximate[5] median distance of the pooled sample as done by Rabanser et al. (2019). Intuitively, K behaves as a similarity metric between points in $\mathcal{X}$ and, in this sense, the MMD statistic compares samples in much the same way that the energy statistic does. Rather than dissimilarity, the MMD statistic looks at similarity of points within and across samples, modifying the order of the summands appropriately to retain direct proportionality with the difference in samples.

[5] Specifically, we use a smaller random sample of 100 data points to compute this median.

E Proof of Theorem 1

Proof. We use the triangle inequality of classification error (Crammer et al., 2007; Ben-David et al., 2007). For any realization of the sample S and any distribution $\mathbb{T}$ over $\mathcal{X} \times \mathcal{Y}$, for any classifiers $h, h' \in \mathcal{H}$, we have[6]

$$\begin{aligned}\mathbf{R}_{\mathbb{T}}(h) - \mathbf{R}_S(h) \leq \mathbf{R}_S(h') + \mathbf{R}_{\mathbb{T}}(h') \\ + |\mathbf{R}_S(h, h') - \mathbf{R}_{\mathbb{T}}(h, h')|\end{aligned} \tag{9}$$

where for $\mathbb{T}$ over $\mathcal{X} \times \mathcal{Y}$ we have

$$\mathbf{R}_{\mathbb{T}}(h, h') = \Pr_{\tilde{X} \sim \mathbb{T}_X}(h(\tilde{X}) \neq h'(\tilde{X})) \tag{10}$$

and for $S = (X_i, Y_i)_{i=1}^n$ we have

$$\mathbf{R}_S(h, h') = n^{-1} \sum_{i=1}^{n} 1[h(X_i) \neq h'(X_i)]. \tag{11}$$

Interchanging roles of $\mathbb{T}$ and S in Eq. (9) and using the definition of the absolute value, we see

$$\begin{aligned}\Delta_h(S, \mathbb{T}) \leq \mathbf{R}_S(h') + \mathbf{R}_{\mathbb{T}}(h') \\ + |\mathbf{R}_S(h, h') - \mathbf{R}_{\mathbb{T}}(h, h')|.\end{aligned} \tag{12}$$

For brevity, for any distribution $\mathbb{D}$, set

$$\xi(\mathbb{D}) = |\mathbf{R}_S(h, h') - \mathbf{R}_{\mathbb{D}}(h, h')|. \tag{13}$$

Then, using the common "addition of zero" trick, we arrive at

$$\begin{aligned}\Delta_h(S, \mathbb{T}) \leq\ & \mathbf{R}_S(h') + \mathbf{R}_{\mathbb{T}}(h') \\ & - \mathbf{R}_T(h') + \mathbf{R}_T(h') + \xi(\mathbb{T}) \\ & - \xi(T) + \xi(T)\end{aligned} \tag{14}$$

Then, by monotonicity and linearity of the expectation we have

$$\begin{aligned}\Delta_h(S, \mathbb{T}) \leq\ & \mathbf{E}_T\Big[\mathbf{R}_S(h') + \mathbf{R}_T(h')\Big] \\ & + \mathbf{E}_T\big[\xi(T)\big] \\ & + \mathbf{R}_{\mathbb{T}}(h') - \mathbf{E}_T\Big[\mathbf{R}_T(h')\Big] \\ & + \xi(\mathbb{T}) - \mathbf{E}_T\big[\xi(T)\big].\end{aligned} \tag{15}$$

[6] A full derivation of Eq. (9) may be found in the proof of Theorem 2 of Ben-David et al. (2010a). In general, the relationship holds for any pair of distributions (i.e., not just a sample and a distribution). Our re-use of this result – for a sample and a distribution – relies on the fact that the sample S, itself, defines a unique *empirical* distribution over $\mathcal{X} \times \mathcal{Y}$ by assigning equal probability to each point in S. The expectation for this empirical distribution is precisely the sample average, so the definition of $\mathbf{R}_S$ coincides for both the sample and the empirical distribution. So, Eq. (9) holds in our case as well.

Let us consider some of these terms individually. Using linearity of expectation and the correspondence between probability and the expectation of an indicator function, we have

$$\begin{aligned}\mathbf{E}_T\Big[\mathbf{R}_T(h')\Big] &= \mathbf{E}\Big[m^{-1}\sum_{i=1}^{m} 1[h(\tilde{X}_i) \neq \tilde{Y}_i]\Big] \\ &= m^{-1}\sum_{i=1}^{m} \mathbf{E}\big[1[h(\tilde{X}_i) \neq \tilde{Y}_i]\big] \\ &= m^{-1}\sum_{i=1}^{m} \Pr_{(\tilde{X}_i, \tilde{Y}_i) \sim \mathbb{T}}\big(h(\tilde{X}_i) \neq \tilde{Y}_i\big) \\ &= m^{-1}\sum_{i=1}^{m} \mathbf{R}_{\mathbb{T}}(h) \\ &= \mathbf{R}_{\mathbb{T}}(h).\end{aligned} \tag{16}$$

Additionally, we have

$$\begin{aligned}\mathbf{E}_T\big[\xi(T)\big] &= \mathbf{E}_T\Big[|\mathbf{R}_S(h, h') - \mathbf{R}_T(h, h')|\Big] \\ &\geq |\mathbf{R}_S(h, h') - \mathbf{E}\big[\mathbf{R}_T(h, h')\big]| \\ &= \xi(\mathbb{T}).\end{aligned} \tag{17}$$

Here, the second line follows by Jensen's Inequality and linearity of the expectation. The last line follows using a similar derivation as in Eq. (16). Then,

$$\xi(\mathbb{T}) - \mathbf{E}_T[\xi(T)] \leq 0 \tag{18}$$

and

$$\mathbf{R}_{\mathbb{T}}(h') - \mathbf{E}_T\Big[\mathbf{R}_T(h')\Big] = 0. \tag{19}$$

Using these two facts in conjunction with Eq. (15) yields

$$\begin{aligned}\Delta_h(S, \mathbb{T}) \leq\ & \mathbf{E}_T\Big[\mathbf{R}_S(h') + \mathbf{R}_T(h')\Big] \\ & + \mathbf{E}_T\big[\xi(T)\big].\end{aligned} \tag{20}$$

Using h as in Eq. (2) to define the statistic D, for any $h' \in \mathcal{H}$, we know $\xi(T) \leq D$ (i.e., by definition of $\max$). So, monotonicity and linearity of expectation implies $\mathbf{E}_T[\xi(T)] \leq \mathbf{E}_T[D]$. For an appropriate choice of h', we then have

$$\Delta_h(S, \mathbb{T}) \leq \mathbf{E}_T\big[\lambda\big] + \mathbf{E}_T\big[D\big]. \tag{21}$$

Rearranging terms gives the lowerbound and the upperbound follows immediately from the fact that $\Delta_h(S, \mathbb{T})$ is non-negative.

$\square$

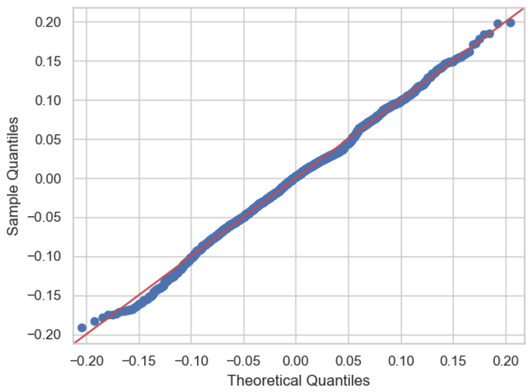

Figure 5: Quantile-Quantile plot. Red line shows ideal: sample quantiles should be the same as the theoretical quantiles of a normal distribution with same variance.

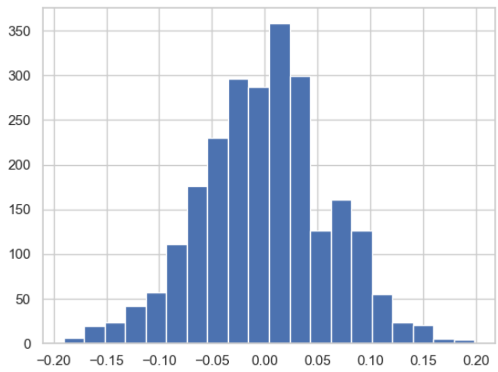

Figure 6: Histogram of realized error terms. Horizontal axis shows value of error term, while vertical axis shows count.

F Regression Diagnostics

Normal Errors Assumption Here, we give diagnostics for the regression model used to analyze data in the main text. Primarily, we would like to check the assumptions that our error terms (i.e., $\boldsymbol{\epsilon}$) are all identically and independently normally distributed. The Jarque-Bera (JB) test uses a statistic based on the skew and kurtosis of the observed errors to study this hypothesis. Assuming the residuals are i.i.d. normal, the probability of observing a JB statistic as extreme as observed is ≈ 0.25. So, we fail to reject the hypothesis that the residuals are i.i.d normal at significance level $\alpha = 0.05$. The assumption that error terms are normal distributed may also be visually checked using the qq-plot, histogram of errors, and the residual plots contained in Figures 5, 6, and 7, respectively. We do not see particularly strong evidence that the residuals are not i.i.d. normal. Albeit, some patterning in the residual plots and skew in the histogram of residuals may be of concern.

Other Possible Assumptions In any case, even if the normality assumption does not hold, our analysis can still be interpreted using more loose assumptions. The most important assumption is that the error terms all have mean 0. Empirically, we find this to be the case with the average residual being $\approx 2.4 \times 10^{-15}$. In fact, Figure 7 shows the line-of-best fit through the residuals (which is typically close to the zero line). As long as the assumption that the error terms have common mean 0 is true, the OLS estimates we use for the coefficients will be unbiased. The only possible short-coming of the OLS estimate is that it could have larger variance than some other estimate. In our analysis, we are most concerned with the unbiased property of our coefficient estimates, but a larger variance in our estimator decreases our confidence that this particular experiment produces estimates close to the truth. Either way, under our relaxed assumption of only a common mean 0 in the errors, we can expect our analysis in the main text to reveal the truth across repeated experiments.

G Regression Analysis Examples

In this section, we give detailed examples (i.e., Exampled 1 and 2) to clarify how we compute estimates in Figure 2. As noted, we use the unbiased OLS estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ in place of $\boldsymbol{\beta}$ as is standard.

Example 1. *Let column j of $\mathbf{X}$ contain the realizations of the h-discrepancy for each experiment and let column k contain the train error. Suppose column ℓ is the (element-wise) product of columns k and j, column q is the square of column j, and column r is the product of columns q and k. Then, controlling for all other features in $\mathbf{X}$, the expected change in estimation error per $\delta > 0$ increase in the h-discrepancy is*

$$\begin{aligned}&\mathbf{E}[\mathbf{Y}_i \mid \mathbf{X}_i = \mathbf{x}] - \mathbf{E}[\mathbf{Y}_i \mid \mathbf{X}_i = \mathbf{x}'] = \beta_j\delta + \beta_\ell\delta \\ &+ \beta_q(\delta^2 + 2\delta\mathbf{x}_j) + \beta_r(\delta^2\mathbf{x}_k + 2\delta\mathbf{x}_j\mathbf{x}_k)\end{aligned} \tag{22}$$

where $\mathbf{x}'$ is a fixed row-vector of features and $\mathbf{x}$ is

Dep. Variable:	est. error	**R-squared:**	0.944
Model:	OLS	**Adj. R-squared:**	0.944
Method:	Least Squares	**F-statistic:**	1949.
		Prob (F-statistic):	0.00
		Log-Likelihood:	3347.1
No. Observations:	2428	**AIC:**	-6650.
Df Residuals:	2406	**BIC:**	-6523.
Df Model:	21		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	-0.0206	0.034	-0.606	0.545	-0.087	0.046
hspace[T.lin]	-0.0239	0.006	-3.817	0.000	-0.036	-0.012
group[T.pdtb]	0.0536	0.016	3.340	0.001	0.022	0.085
group[T.rst]	0.0600	0.018	3.256	0.001	0.024	0.096
bert[T.pooled]	0.0034	0.006	0.601	0.548	-0.008	0.015
bert[T.sentence]	0.0250	0.009	2.872	0.004	0.008	0.042
news[T.notnews]	-0.0029	0.010	-0.289	0.773	-0.022	0.017
train_error	0.3262	0.080	4.054	0.000	0.168	0.484
lamb	-0.0150	0.048	-0.312	0.755	-0.109	0.079
hdisc	0.1545	0.081	1.906	0.057	-0.004	0.313
bert[T.pooled]:hdisc	-0.0313	0.009	-3.622	0.000	-0.048	-0.014
bert[T.sentence]:hdisc	-0.1370	0.013	-10.600	0.000	-0.162	-0.112
hspace[T.lin]:hdisc	0.0194	0.009	2.159	0.031	0.002	0.037
group[T.pdtb]:hdisc	-0.0210	0.021	-1.002	0.316	-0.062	0.020
group[T.rst]:hdisc	0.0671	0.028	2.410	0.016	0.013	0.122
news[T.notnews]:hdisc	0.0320	0.013	2.529	0.012	0.007	0.057
hdisc:train_error	1.9665	0.196	10.052	0.000	1.583	2.350
np.power(hdisc, 2)	0.4831	0.052	9.323	0.000	0.381	0.585
train_error:np.power(hdisc, 2)	-1.6867	0.152	-11.074	0.000	-1.985	-1.388
lamb:train_error	-0.5861	0.122	-4.803	0.000	-0.825	-0.347
np.power(lamb, 2)	-0.1346	0.071	-1.892	0.059	-0.274	0.005
train_error:np.power(lamb, 2)	0.4043	0.100	4.029	0.000	0.208	0.601

Omnibus:	2.707	**Durbin-Watson:**	1.548
Prob(Omnibus):	0.258	**Jarque-Bera (JB):**	2.718
Skew:	-0.046	**Prob(JB):**	0.257
Kurtosis:	3.136	**Cond. No.**	463.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Table 3: Full description of the regression model including all features, estimated coefficients, and relevant tests for diagnosis and inference. Tests involving standard errors (std err) are only valid if the model errors follow the assumed distribution. We believe most variables are self-explanatory, but we do provide some assistance to reader: **lamb** corresponds to λ, **hdisc** corresponds to the h-discrepancy, **train_error** corresponds to the error on the source sample, **np.power(**$\diamond$**, 2)** corresponds to the square of the feature $\diamond$, presence of **:** indicates a multiplication of features (i.e., an interaction-term), and **hspace** corresponds to the type of classifier used (i.e., linear model or fully-connected network).

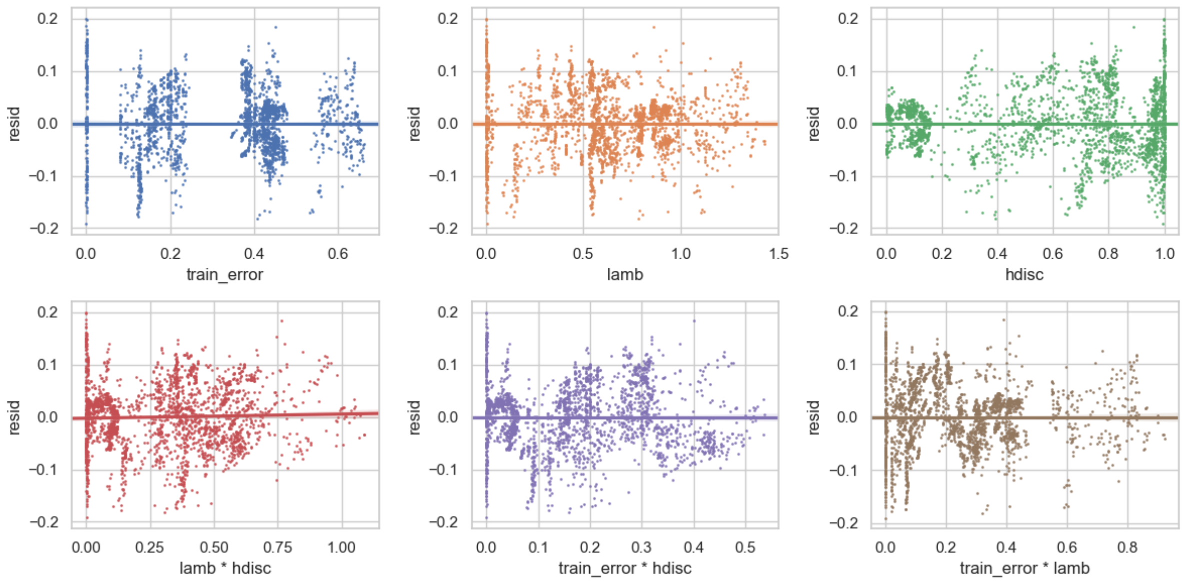

Figure 7: Residual plots. Vertical axes show realized error terms, while horizontal axes show value of some feature that may or may not be in our design matrix. Significant patterns may indicate a missing term in our model. While some patterning may exist, we choose not to include additional terms for reason of interpretability and to meet other (quantifiable) model assumptions.

defined by

$$\mathbf{x}_p = \begin{cases} \mathbf{x}'_p + \delta & \text{if } p = j, \\ \mathbf{x}'_k(\mathbf{x}'_j + \delta) & \text{if } p = \ell, \\ (\mathbf{x}'_j + \delta)^2 & \text{if } p = q, \\ \mathbf{x}'_k(\mathbf{x}'_j + \delta)^2 & \text{if } p = r, \\ \mathbf{x}'_p & \text{else} \end{cases} . \tag{23}$$

If this function of δ is positive, we know increasing the h-discrepancy increases the bias as suggested by our theory.

Example 2. *Let column j of $\mathbf{X}$ be 1 if we use S-BERT representations and 0 otherwise. Let column k of $\mathbf{X}$ indicate use of P-BERT in the same way and suppose the reference category*[7] *for the BERT representations is A-BERT. Let column ℓ of $\mathbf{X}$ contain discrepancy D_i for each experiment and let column q be the element-wise product of columns j and ℓ; i.e., interaction terms. Then, controlling for all other features in $\mathbf{X}$, the expected increase in error-gap using S-BERT instead of A-BERT is*

$$\begin{aligned} &\mathbf{E}[D_i - \mathbf{Y}_i \mid \mathbf{X}_i = \mathbf{x}] - \mathbf{E}[D_i - \mathbf{Y}_i \mid \mathbf{X}_i = \mathbf{x}'] \\ &= -(\beta_j + \beta_q D_i) \end{aligned} \tag{24}$$

where $\mathbf{x}'$ is a fixed row-vector of features such that $\mathbf{x}'_\ell = D_i$ and $\mathbf{x}'_j = \mathbf{x}'_k = 0$. The row-vector $\mathbf{x}$ is defined by $\mathbf{x}_r = \{1 \text{ if } r = j,\ \mathbf{x}'_\ell \text{ if } r = q,\ \mathbf{x}'_r \text{ else}\}$. When this function of D_i is positive, we know using S-BERT is expected to increase the error-gap.

[7] In regression, the reference is the single category from any group of categories which is not explicitly included in $\mathbf{X}$. It serves as a point of comparison for the other categories. For technical reasons, a point of comparison is typically needed to analyze impact of categorical features (i.e., so $\mathbf{X}$ is full rank).