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

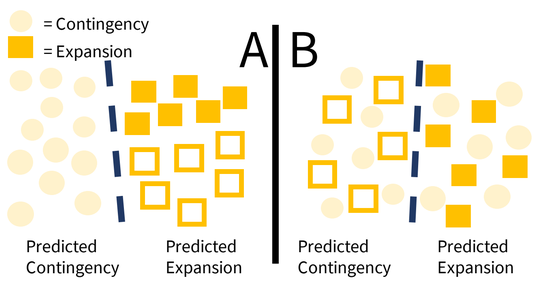
Katherine Atwell[1], Anthony Sicilia[2], Seong Jae Hwang[3], Malihe Alikhani[1]
[1]Computer Science and [2]Intelligent Systems Program, University of Pittsburgh; [3]Department of Artificial Intelligence, Yonsei University

Discourse Parsing

- Discourse parsing allows us to make high-level inferences across sentence boundaries
- The goal of discourse parsing is to label **relations** between multiple text units
- Two examples, marked with their relation, are below:
  - While **the earnings picture confuses**, observers say *the major forces expected to shape the industry in the coming year are clearer*. Contrast
  - *Just as the 1980s bull market transformed the U.S. securities business*, **so too will the more difficult environment of the 1990s**," says Christopher T. Mahoney, a Moody's vice president. Similarity
- **Argument 1** is in italics, *argument 2* is in bold, and the connective is underlined

Domain Shift in Parsing

- Differences in train and test set can severely impact parser performance
- To make use of existing discourse datasets, we need to be able to quantify the impact of data differences
- Proper quantification can aid in model and dataset selection



Why not just capture distribution shift?

- Feature distribution shift does not always correlate with changes in classifier error
- E.g., in A, when shift is present, the classifier does well on both domains (solid/hollow shapes). In B there is no feature shift and parser transfers poorly

Results

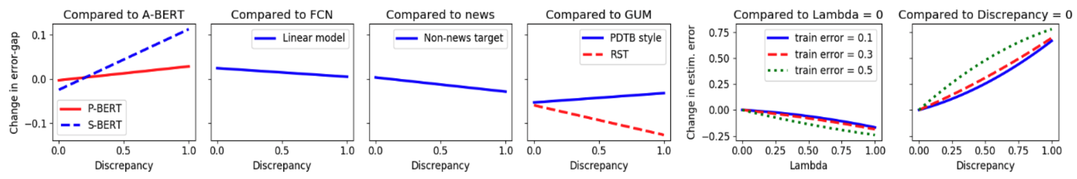
We demonstrate the benefit of using tools from theoretical domain adaptation to select models and datasets for discourse parsing.

Spearman (Rank) Correlation

Split	FRS	Energy	MMD	BBSD	*h*-disc
All	.5394	.6059	.5051	.4054	**.8299**
PDTB	.5451	.6359	.5472	.4746	**.8265**
RST-DT	.2166	.3059	-.0011	.2087	**.7625**
News	.5262	.6356	.5507	.5759	**.8517**
Other	.3760	.4517	.2767	.1737	**.8386**
WD	.0884	.5735	-.0324	.2368	**.7890**
OOD	.4597	.5249	.3917	.2813	**.7666**

Pearson Correlation

Split	FRS	Energy	MMD	BBSD	*h*-disc
All	.4986	.4396	.3413	.4404	**.7628**
PDTB	.5295	.4704	.3709	.4274	**.7642**
RST-DT	.2853	.1660	-.1605	.1677	**.7599**
News	.7079	.6302	.5558	.5386	**.8890**
Other	.3420	.2791	.1760	.2051	**.7072**
WD	.1075	.5831	-.0515	.4853	**.9519**
OOD	.4342	.3909	.2761	.3745	**.6976**



Correlation Analysis

- In the top two tables, we present correlations for various statistics.
- Our proposed, theoretically motivated statistic (h-disc) has higher correlation with error gap than other statistics.
- Our statistic makes use of classifier information in a theoretically motivated way (see right panel).

Regression Analysis

- In the bottom figure, expected change in error gap is plotted, controlling for all features of the experiment not explicitly listed in the figure
- More complex models are more difficult to transfer as difference in domains increases.
- Some datasets are harder to transfer to (e.g., news test sets and more variable test sets).
- More results are given in the paper.

Quantifying Meaningful Domain Shift

Problems with common two-sample test statistics

- Statistics such as **FRS**, **Energy**, and **MMD** test shift in feature distribution, which is problematic (see bottom left panel).

Our goal

- Design a statistic capable of estimating changes in error rate across datasets by using more information than just feature shift.

Mathematical Setup

- Consider a random source sample *S*, a target distribution **T**, and random sample from that distribution *T*
- We use **R** to denote the risk (i.e., error rate) of a classifier on a distribution or sample.
- We use Δ to denote the change in error rates across sample(s) and distribution(s).

Proposed Statistic

- We propose a statistic *D*, which we call the *h*-discrepancy.
- It generalizes previous proposals in domain adaptation such as the source-guided discrepancy (Kuroki et al., 2018).
- We tailor the statistic to make use of important classifier information to estimate impact of domain-shift.
- To theoretically study, we prove bounds on the bias of this statistic as an estimator for change in parser error.

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 $h \in \mathcal{H}$,*

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

where $\lambda = \min_{h' \in \mathcal{H}} \mathbf{R}_S(h') + \mathbf{R}_T(h')$.

- We interpret the bounds to determine reasonable cases when our statistic should correlate with changes in error.

The QR code on the left can be used to access our manuscript, as can this link: https://arxiv.org/pdf/2203.11317.pdf

Our GitHub can be accessed at the link below: https://github.com/anthonysicilia/change-that-matters-ACL2022

References

Our references can be found in the manuscript (see left).

University of Pittsburgh