The Change that Matters in Discourse:

Estimating the Impact of Domain Shift on Parser Error

Kate Atwell[*1], Anthony Sicilia[*2], Seong Jae Hwang[3], and Malihe Alikhani[1,2]
[1]Department of Computer Science and [2]Intelligent Systems Program, University of Pittsburgh,
[3]Department of Artificial Intelligence, Yonsei University
{kaa139, anthonysicilia}@pitt.edu, seongjae@yonsei.ac.kr, malihe@pitt.edu

*equal contribution

Discourse Analysis as a Ubiquitous Tool

Discourse analysis allows us to attain inferences **beyond** the sentence-level in a text.

The **output of discourse models** (i.e., trained to classify discourse relations) has been shown to **improve performance on downstream tasks** including natural language generation, machine comprehension, and question-answering tasks.

What is Discourse Analysis?

Discourse analysis studies higher-level inferences between units of text by capturing the **relation** between these text units.

Examples that capture the importance of discourse are shown 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**

Our Discourse Datasets

Dataset	Genre	Label schema
RST-DT (Carlson et al., 2001)	News	RST-DT
PDTB 2.0 (Prasad et al., 2008)	News	PDTB
PDTB 3.0 (Webber et al., 2019)	News	PDTB
BioDRB (Ramesh and Yu, 2010)	Bio	PDTB
TED-MDB (Zeyrek et al., 2020)	TED talks	PDTB
GUM (Zeldes, 2017)	Multiple	RST-DT

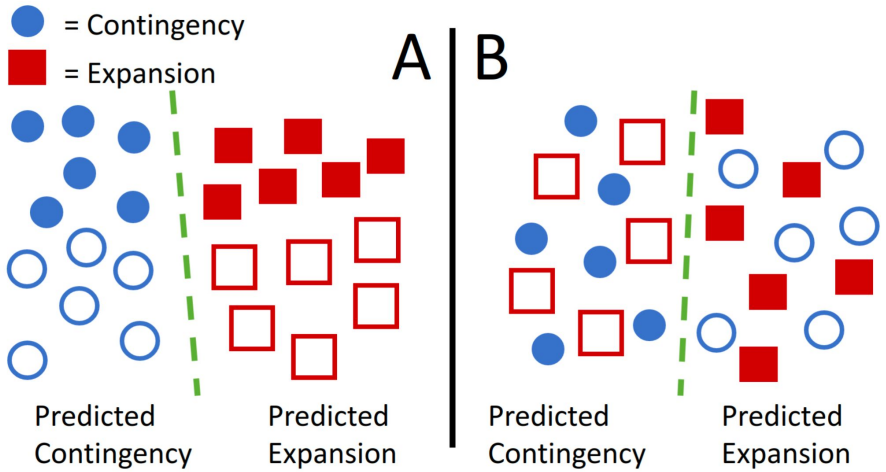
Problems with Applying Discourse Parsers

There are some cases in which discourse does not help, or only yields small improvements. This is likely because of the nature of the most widely-used discourse datasets; they contain **only Wall Street Journal articles over a three-year period**.

Thus, **most applications of discourse parsers require some domain shift**, which has been shown to be difficult for these parsers.

How Should We Estimate Parser Error?

While common practice is to measure **distribution shift in the feature space**, this shift **does not always correlate with parser error**



A Statistic Theoretically Tied to Parser Error

We **propose a statistic for estimating change in parser error** based on recent domain adaptation literature. We **derive bounds on the bias** of this statistic as an estimator of 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 $\lambda = \min_{h' \in \mathcal{H}} \mathbf{R}_S(h') + \mathbf{R}_T(h')$.

What's new about this result?

- Generalization of previous statistics, which utilizes vital information about the parser
- Focus on bias rather than sample-complexity bounds

Results from **2400**+ models on **6** Datasets

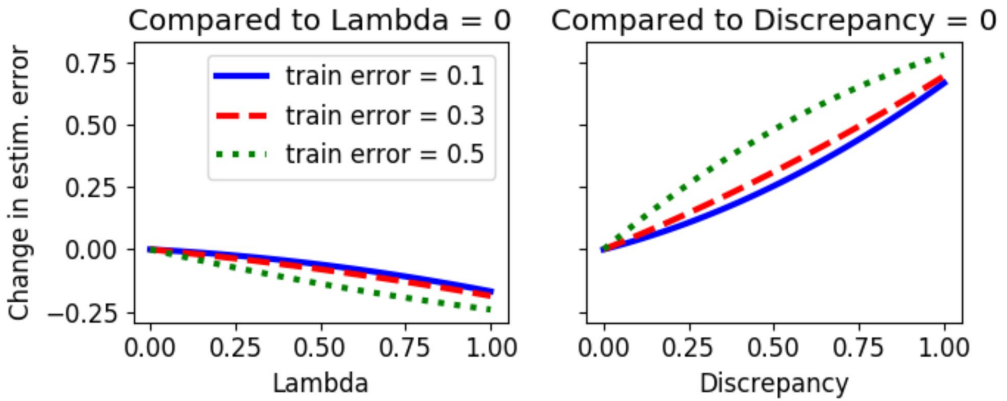
As hypothesized, **our proposed statistic correlates best** with parser error across a variety of data splits

Split	Spearman (Rank) Correlation					Pearson (Linear) Correlation				
	FRS	**Energy**	**MMD**	**BBSD**	***h*-disc**	**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-DT	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**

Ignore classifier information!

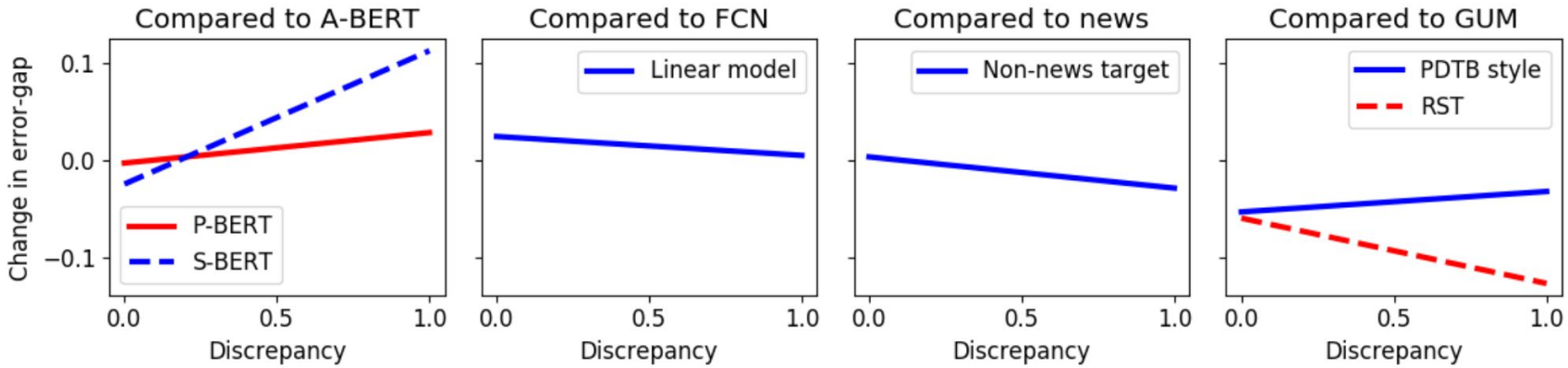
Results from **2400**+ models on **6** Datasets

Using a regression analysis of the estimation error, we are able to validate our theoretical result



Results from **2400**+ models on **6** Datasets

The regression analysis also allows us to study how advantageous different models and datasets are in presence of domain-shift



Key Takeaways and Conclusions

- Parser error does not always correlate with feature distribution shift
- Statistics theoretically related to parser error through domain adaptation bounds (e.g., as proposed) are better suited
- Large scale empirical analysis of such statistics can provide important practical insight
    - Model complexity is an import consideration when faced with domain shift
    - Some datasets are easier to transfer to/from
    - More in the paper!

Thanks!

Paper:

https://arxiv.org/pdf/2203.11317.pdf

Contact:

{kaa139, anthonysicilia}@pitt.edu, seongjae@yonsei.ac.kr, malihe@pitt.edu

Code:

https://github.com/anthonysicilia/change-that-matters-ACL2022