Computer Science
UNIVERSITY OF TORONTO

Quantifying the Task-Specific Information in Text-Based Classifications

Zining Zhu, Aparna Balagopalan, Marzyeh Ghassemi, Frank Rudzicz

Neural NLP models have high capacities

Playground

Chat | Save | View code | Share | ...

The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.

Human: Hello, who are you?
AI: I am an AI created by OpenAI. How can I help you today?
Human: Describe a city you live in.
AI: I live in a city of computers, so my city is made of circuits, routers, drives, servers, and a lot more.
AI: It is an ordinary city, nothing special.
AI: Is this satisfactory?

Assess model's abilities with classifications

15 DNN LM models

#16: Human baseline

Rank	Name	Model
1	AliceMind & DIRL	StructBERT + CLEVER
2	ERNIE Team - Baidu	ERNIE
3	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4
4	HFL iFLYTEK	MacALBERT + DKM
5	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS
6	liangzhu ge	Deberta + adv (ensemble)
7	T5 Team - Google	T5
8	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART
9	Huawei Noah's Ark Lab	NEZHA-Large
10	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)
11	ELECTRA Team	ELECTRA-Large + Standard Tricks
12	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)
13	Junjie Yang	HIRE-RoBERTa
14	Facebook AI	RoBERTa
15	Microsoft D365 AI & MSR AI	MT-DNN-ensemble
16	GLUE Human Baselines	GLUE Human Baselines

Classification datasets contain shortcuts

- Shortcuts allow NLP models to be "right for the wrong reasons" (McCoy et al.., 2019).
- Common shortcuts include punctuation marks, overlapped words.

Quora example

S1: What can make Physics easy to learn?

S2: How can you make physics easy to learn?

Label: True (similar question)

Correct reason: They have very similar meanings.

Shortcut: They both contain can, to, and "?".

MNLI example

S1: You have access to the facts.

S2: The facts are accessible to you.

Label: Entailment

Correct reason: S1 entails S2.

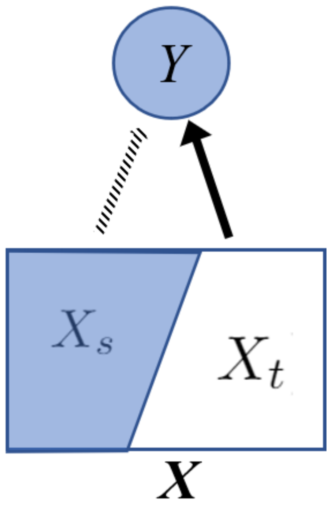
Shortcut: They both contain the, to and you.

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428–3448. https://doi.org/10.18653/v1/P19-1334

Computer Science
UNIVERSITY OF TORONTO

Shortcuts vs. the other part

- Input data as a random variable X
  - The identified shortcut: X_s
  - The remaining part: X_t
  - According to our definition of “shortcut”: $X_s \perp X_t$
- How much information does X_t contribute to the target Y?
  - Task-Specific Information (TSI)
  - We define TSI to be $I(Y; X_t)$



Quantifying the Task-Specific Information

- With the assumptions, we can arrive at the expression for TSI:

$$I(Y;X_t) = H(Y|X_s) - H(Y|X)$$

- Empirically: use cross entropy to approximate the entropy:

$$H(p) = \mathrm{E}_p \log\frac{1}{q} - \mathrm{E}_p \log\frac{p}{q} = NLL - KL(p||q)$$

  - Where NLL is the cross-entropy loss, and KL is the Kullback-Leibler divergence.
  - And $q(\cdot)$ is the distribution approximating the unknown true distribution $\mathrm{p}(\cdot)$

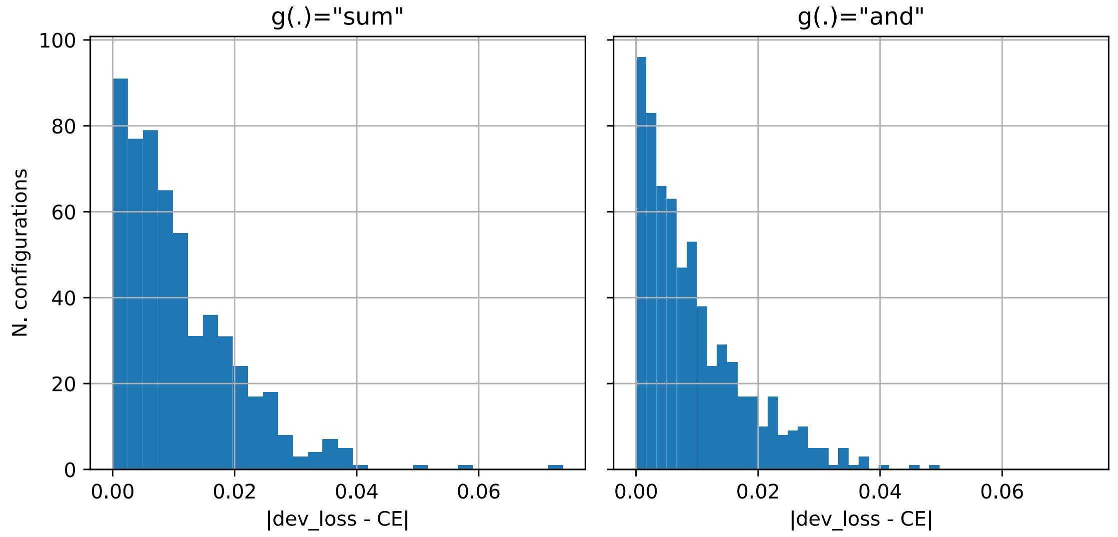
- This results in the proposed method:

$$TSI = NLL_{Y|X_s} - NLL_{Y|X}$$

How close is NLL to the conditional entropy?

- In 99.5% configurations, NLL is within 0.04 nats away from H(Y|X).

$$X_j \sim \text{Bernoulli}(p_x), \text{ where } j \in \{1, 2, .., m\}$$
$$X = [X_1, X_2, ..., X_m]$$
$$Y = g(X_1, ..., X_m) + \epsilon, \text{ where } \epsilon \sim \text{Bernoulli}(p_y)$$



Identified shortcuts

We identify the following shortcuts:

- Punctuation marks
- Occurrence of (non-negative) stopwords
- Count of overlapped words (for sentence pair tasks)

All shortcut features are normalized by sentence length.

Estimated TSI values

All TSI values are in nats.

Dataset	**Acc**$_{Y \mid X}$	**TSI**$^{\mathbf{P+S}}$	**TSI**$^{\mathbf{P+S+O}}$
MNLI	0.85	0.68	0.64
IMDB	0.92	0.43	–
Yelp	0.97	0.41	–
QQP	0.89	0.31	0.23

Ablation: using imperfect models

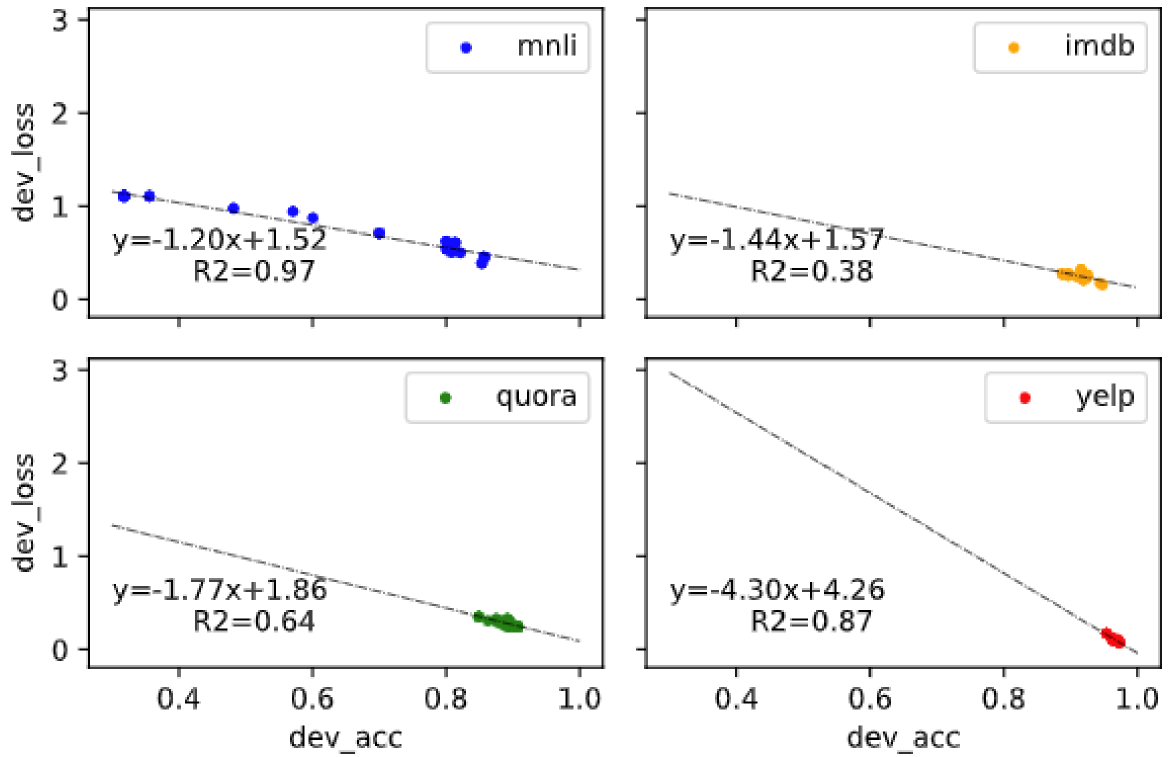


Figure 4: A scatter plot of the accuracy against dev loss of models trained on full datasets.

Ablation: stability to dataset sizes

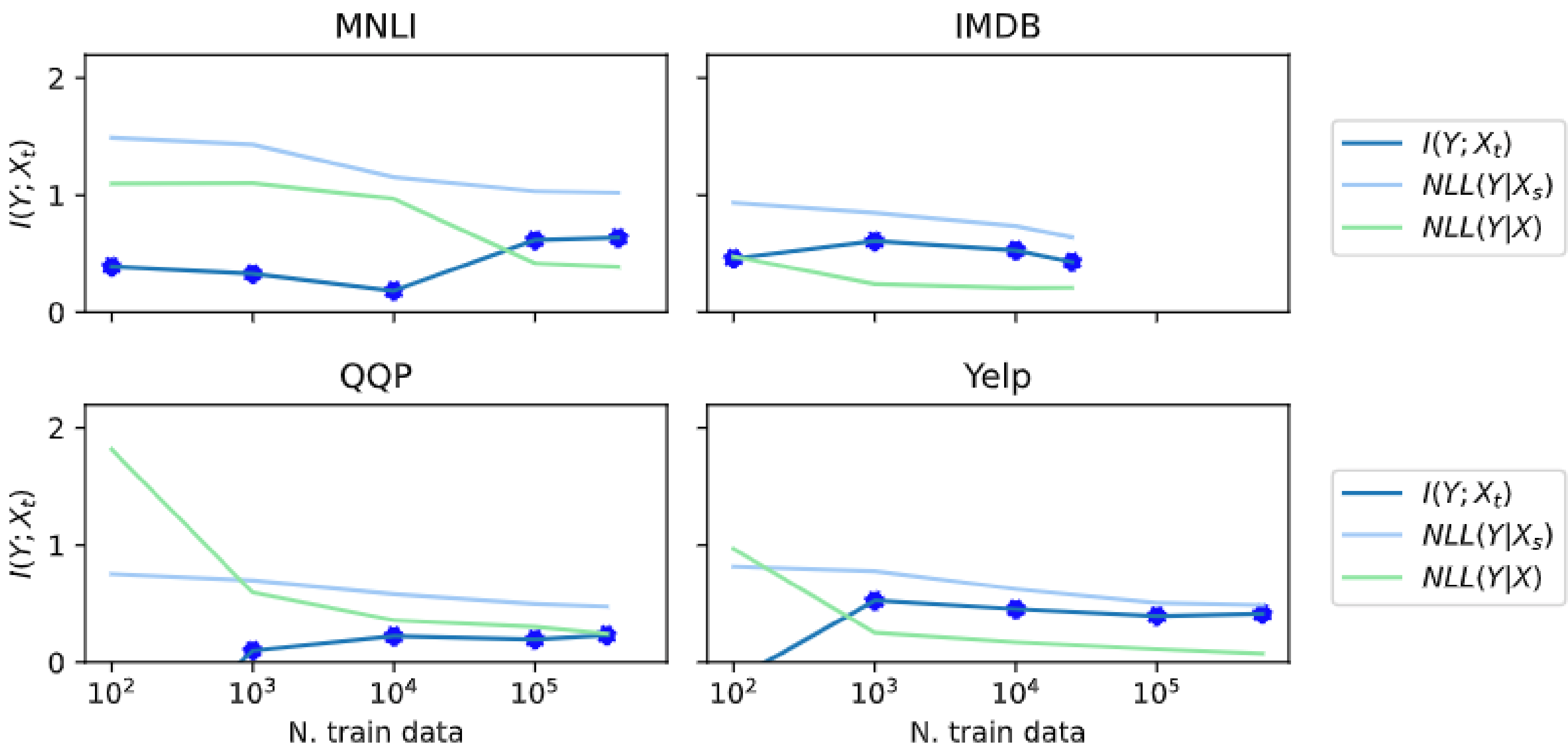


Figure 6: The $I(Y;X_t)$ estimation when we subsample different sizes of datasets.

Future work

The future work can be in these directions:

- Identifying the shortcut features.
- Leaderboard practices.
- Metrics for cross-task comparison.
- Use information-theoretic methods to understand text corpus.

Conclusion

- We identify the task-specific information (TSI) for text-based classification datasets.
- We propose a method to estimate TSI.

Thank you for listening! Any questions?