HIATUS

HUMAN INTERPRETABLE ATTRIBUTION OF T
UNDERLYING STRUCTURE

HIATUS Virtual Proposers’ Day
Wednesday, 19 January 2022

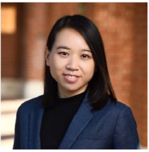
Presenter: Mahsa Yarmohammadi

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

UNIVERSITY OF
MARYLAND

NYU

Our team

Participant	**Capabilities**
Mark Dredze, Associate Professor Johns Hopkins University	Large language models, explainable AI, information extraction, multilingual models
Ben Van Durme, Associate Professor Johns Hopkins University	Controllable text generation, Human evaluation of text generation, Information Extraction, Natural Language Understanding
Anqi Liu, Assistant Professor Johns Hopkins University 	Domain generalization/adaptation, uncertainty estimation, active learning, fair machine learning
Jordan Boyd-Graber, Associate Professor University of Maryland	Evaluation of topic model interpretability, unsupervised machine learning, interactive modeling, interfaces for machine learning feedback
Mahsa Yarmohammadi, Asst. Research Scientist Johns Hopkins University	Topic classification, data annotation, Information Extraction and Retrieval, system integration
João Sedoc, Assistant Professor New York University	Evaluation of natural language generation systems, evaluation methodology, personality, empathy and emotion prediction

Our Team's Accomplishments

Examples of salient prior work

- Text generation
- Uncertainty Calibration and Bias Mitigation
- Human-in-the-loop generation of adversarial examples

Examples of evaluation

- Chatbot evaluation
- Efficient Annotation of Scalar Labels (EASL)
- Human-centric measure for evaluating topic models

Examples of Explainable AI

- Proxy models for faithful and plausible explanations
- Empirical study of explanations and feedback in interactive ML

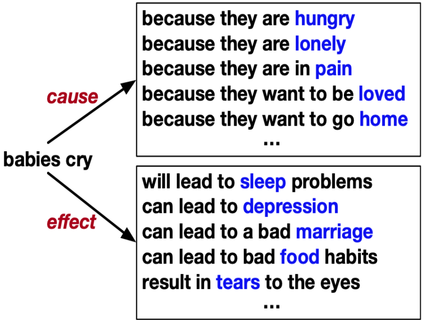
Text Generation

- Paraphrasing: differing textual realizations of the same meaning
  - ParaPhrase DataBase (**PPDB**). Pre-neural dominant artifact for enabling automatic paraphrasing.
    - over 100 million automatically constructed paraphrases
  - **ParaBank**. More recent line of work on building resources and systems for automatic paraphrastic text rewriting.
    - large-scale English paraphrase dataset containing 79.5 million references

Paraphrase.org

https://nlp.jhu.edu/parabank/

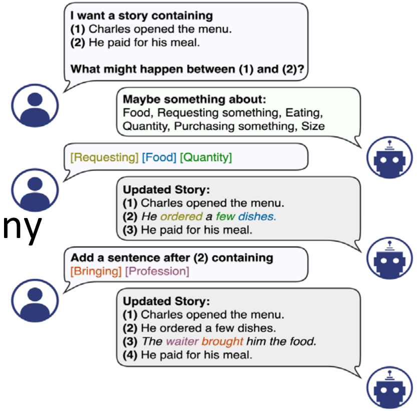
- Beyond paraphrasing
  - **CausalBank**. 314 million pairs of cause-effect statements scraped from the Common Crawl corpus using causal lexical patterns.



because they are hungry
because they are lonely
because they are in pain
because they want to be loved
because they want to go home
...

will lead to sleep problems
can lead to depression
can lead to a bad marriage
can lead to bad food habits
result in tears to the eyes
...

- **InFillmore.** Take an existing text and add sentences at any position of the input, guided by human constraints.



I want a story containing
(1) Charles opened the menu.
(2) He paid for his meal.

What might happen between (1) and (2)?

Maybe something about:
Food, Requesting something, Eating, Quantity, Purchasing something, Size

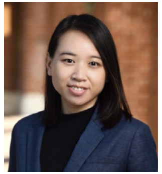
[Requesting] [Food] [Quantity]

Updated Story:
(1) Charles opened the menu.
(2) *He ordered a few dishes.*
(3) He paid for his meal.

Add a sentence after (2) containing
[Bringing] [Profession]

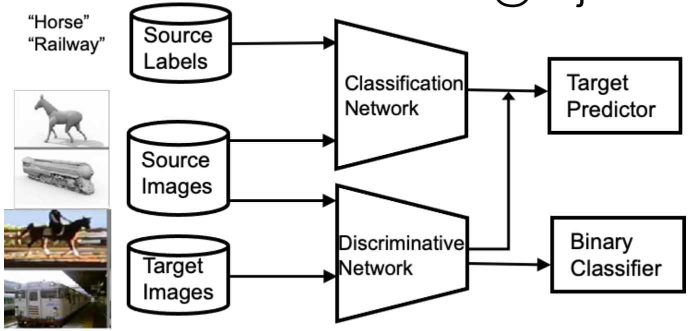
Updated Story:
(1) Charles opened the menu.
(2) He ordered a few dishes.
(3) *The waiter brought him the food.*
(4) He paid for his meal.

Uncertainty Calibration and Bias Mitigation under Distribution Shift

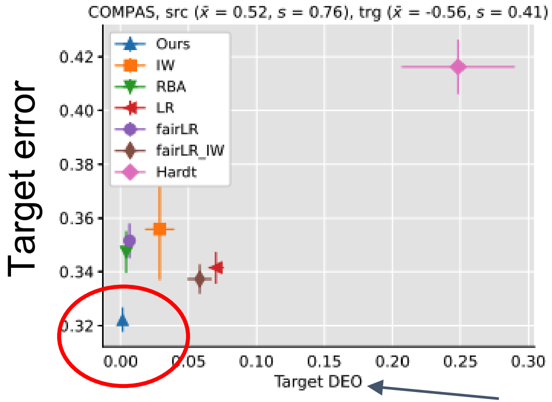


aliu@cs.jhu.edu

- **A distributionally robust learning framework** for dealing with domain shift, which generates more calibrated uncertainty estimates and interpretable density ratios.



- **Fair classification** with regard to protected attributes (race, gender, age, etc.) when there is a distribution shift, achieving both lower error and group fairness metrics.

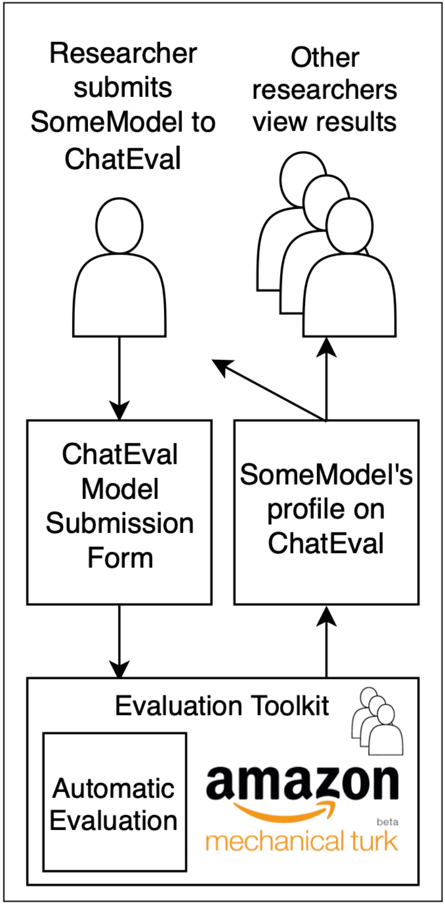


Difference of equalized opportunity

Wang, Haoxuan, et al. "Deep Distributionally Robust Learning for Calibrated Uncertainties under Domain Shift." *arXiv preprint arXiv:2010.05784* (2020).
Rezaei, Ashkan, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. "Robust Fairness Under Covariate Shift." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, pp. 9419-9427. 2021.

ChatEval

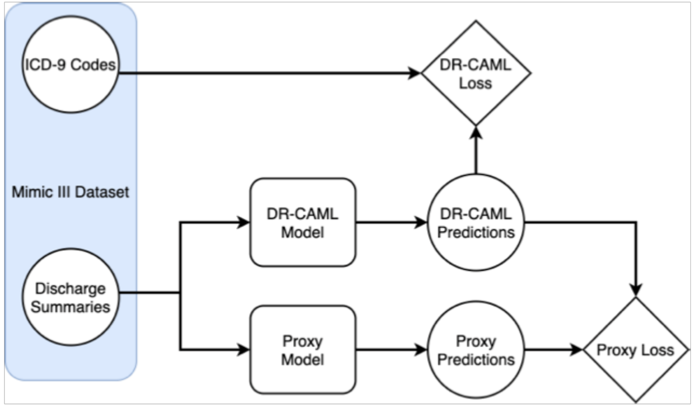
- We created a framework for the systematic evaluation of conversational agents.
- We have released a dozen datasets and run two shared tasks:
    - Dialogue Breakdown Detection Challenge
    - Dialog System Technology Challenge Track 5 on dialogue system evaluation.
- ChatEval has grown beyond just evaluating systems pairwise and we can now also assess pointwise scales between systems.


Researcher submits SomeModel to ChatEval
Other researchers view results
ChatEval Model Submission Form
SomeModel's profile on ChatEval
Evaluation Toolkit
Automatic Evaluation
amazon mechanical turk
beta

João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. ChatEval: A Tool for Chatbot Evaluation. In *Proceedings of NAACL (Demonstrations)*.

Proxy Models for Faithful and Plausible Explanations

- Developed a proxy mode to mimic the behavior of AI system
- Model provides fine-grained control between:
  - Faithful explanation: true to model’s decision making
  - Plausible: makes sense to domain experts
- Explanations are both locally meaningful and globally consistent
- Compared to popular interpretable AI methods on RNN task:
  very fast, achieves global coherence



Wood-Doughty, Z., Cachola, I. and Dredze, M., 2021. Faithful and Plausible Explanations of Medical Code Predictions. *arXiv preprint arXiv:2104.07894*.

Zach Wood-Doughty, Isabel Cachola, Mark Dredze. Proxy Model Explanations for Time Series RNNs. *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2021.

List of publications (1/3)

- Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. *Neural Information Processing Systems*, 2009.
- Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain Natural Language Inference. In *Proceedings of Association of Computational Linguistics (ACL)*.
- Ryan Culkin, J. Edward Hu, Elias Stengel-Eskin, Guanghui Qin, and Benjamin Van Durme. 2021. Iterative Paraphrastic Augmentation with Discriminative Span Alignment. In *Transactions of the Association of Computational Linguistics 9: 494-509*.
- Julian Martin Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. Fool Me Twice: Entailment from Wikipedia Gamification. *North American Association of Computational Linguistics (NAACL)*, 2021.
- Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence. *Advances in Neural Information Processing Systems, 34*.
- J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. ParaBank: Monolingual Bitext Generation and Sentential Paraphrasing via Lexically-constrained Neural Machine Translation. In *Proceedings of AAAI*.
- Huda Khayrallah and João Sedoc. 2021. Measuring the 'I don't know' Problem through the Lens of Gricean Quantity. In *Proceedings of NAACL-HLT*.
- Zachary C. Lipton. 2016. The Mythos of Model Interpretability. *arXiv preprint arXiv: 1606.03490 (2016).*
- Shi Feng and Jordan Boyd-Graber. 2019. What AI can do for me: Evaluating Machine Learning Interpretations in Cooperative Play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*.

List of publications (2/3)

- Shi Feng, Eric Wallace, Alvin Grissom II, Pedro Rodriguez, Mohit Iyyer, and Jordan Boyd-Graber. 2018. Pathologies of Neural Models Make Interpretation Difficult. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.
- Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of ACL*.
- Jeifu Ou, Nathaniel Weir, Anton Belyy, Felix Yu, and Benjamin Van Durme. 2021. InFillmore: Frame-Guided Language Generation with Bidirectional Context. In *Proceedings of the 10th Conference on Lexical and Computational Semantics*.
- Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of ACL*.
- Ashakn Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. 2021. Robust Fairness Under Covariate Shift. In *AAAI 2021*.
- Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation Examples Are Not Equally Informative: How Should That Change NLP Leaderboards?. In *Proceedings of ACL*.
- Pedro Rodriguez and Jordan Boyd-Graber. 2021. Evaluation Paradigms in Question Answering. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*.
- Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient Elicitation of Annotations for Human Evaluation of Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.
- Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient Online Scalar Annotation with Bounded Support. In *Proceedings of ACL*.

List of publications (3/3)

- João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. ChatEval: A Tool for Chatbot Evaluation. In *Proceedings of NAACL (Demonstrations)*.
- João Sedoc and Lyle Ungar. 2020. Item Response Theory for Efficient Human Evaluation of Chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*.
- Alison Smith, Jordan Boyd-Graber, Ron Fan, Melissa Birchfield, Tongshuang Wu, Dan Weld, and Leah Findlater. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. *Computer-Human Interaction*, 2020.
- Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick Me If You Can: Human-in-the-loop Generation of Adversarial Question Answering Examples . In *Transactions of the Association of Computational Linguistics*.
- Haoxuan Wang, Anqi Liu, Zhiding Yu, Junchi Yan, Yisong Yue, and Anima Anandkumar. 2021. Deep Distributionally Robust Learning for Calibrated Uncertainties under Domain Shift. *arXiv:2010.05784*.
- Zach Wood-Doughty, Isabel Cachola, and Mark Dredze. 2021a. Faithful and Plausible Explanations of Medical Code Predictions. *ArXiv, abs/2104.07894*.
- Zach Wood-Doughty, Isabel Cachola, and Mark Dredze. 2021b. Proxy Model Explanations for Time Series RNNs. *IEEE International Conference on Machine Learning and Applications (ICMLA)*.
- Chen Zhang, João Sadoc, Luis Fernando D'Haro, Rafael Banchs, and Alexander Rudnicky. 2021. Automatic Evaluation and Moderation of Open-domain Dialogue Systems. *arXiv:2111.02110*.