RESEARCH PROPOSAL: EVALUATING AND ENABLING HUMAN–AI COLLABORATION

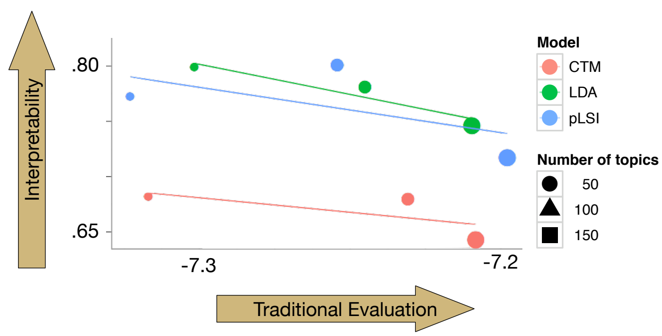
JORDAN BOYD-GRABER, UNIVERSITY OF MARYLAND

Artificial intelligence[1] (AI) is ubiquitous: detecting spam e-mails, flagging fraudulent purchases, and providing the next movie in a Netflix binge. But they do not exist in a vacuum: as Shneiderman [39] argues, AI must exist alongside humans. My goal is to create metrics to measure whether AI methods make sense to users, helping users craft examples to advance AI, and applying AI to applications that help illuminate complex social science applications.

1. Evaluating Interpretability

My journey with evaluating interpretability began over ten years ago with topic models. Topic models are sold as a tool for understanding large data collections: lawyers scouring Nordstream e-mails for a smoking gun, journalists making sense of Wikileaks, or humanists characterizing the oeuvre of Lope de Vega. But topic models' proponents never asked what those lawyers, journalists, or humanists needed. Instead, they optimized *held-out likelihood*.

When my colleagues and I developed the *interpretability* measure to assess whether topic models' users understood their outputs, interpretability and held-out likelihood were negatively correlated [6]! The topic modeling community (including me) had fetishized complexity at the expense of usability…and topic modeling is not alone.



Since this humbling discovery, I've built topic models that are a collaboration between humans and computers. The computer starts by proposing an organization of the data. The user separates confusing clusters or joins similar clusters together [20], an improvement over the "take it or leave it" philosophy of most machine learning algorithms.

Focusing on collaboration also requires algorithms that are low latency (not just high throughput). We extended the geometric interpretations of admixture models developed by Arora et al. [1] to multi-anchor topics [27] and multi-lingual topics [46]. These are much faster than traditional probabilistic topic models—they can handle millions of documents in seconds—but they are less well understood theoretically and less used in practice. Thus, we also developed better understanding of the projections of multilingual representations via graph theory [12] and the convergence of alternating projections [47].

After we proposed our "reading tea leaves" evaluation, it's heartening that Lau et al. [26] and their "machine reading tea leaves" (which correlate with our human measures) became a standard topic model evaluation: a survey of forty recent topic modeling papers, **all but four** use a form of their coherence evaluation. However, as we argue in Hoyle et al. [19], you cannot just use this evaluation forever and forget about humans. In that same survey, **none** of those papers do a human evaluation. As topic models evolve

Date: Updated December 2022.

[1] I take a broad interpretation of AI; some of my examples might be better characterized machine learning. But rather than distracting boundary policing, I will embrace the general term but will be specific in describing particular tools/models.

(e.g., incorporating neural components), you need to validate that these automatic metrics still correlate with whether it is useful for a human–computer collaboration.

2. Teaming as an Evaluation

Within the HCI community, we have argued for the foundations of what should go into human–computer collaborations: computers that incorporate users' suggestions [25]; explanations with accountability [41]; and stable explanations [42].

In addition to these human-centered understanding of users' needs and desires, we've developed machine learning approaches to measure how well users complete a task. For example, for a question answering task, we measured how much the accuracy of the human–computer *team* increases with different explanations and found that explanations help all users but that novices are easily overwhelmed [10]. In follow-on work, we learned how to explicitly optimize explanations for individual users [11].

3. Connecting with Social Science: Pedagogy, Framing, and Deception

The reverse of cooperation is human competition; it also has much to teach computers. I've increasingly looked at language-based games whose clear goals and intrinsic fun speed research progress. For example, in the board game *Diplomacy*, users chat with each other while marshaling armies for world conquest. Alliances are fluid: friends are betrayed and enemies embraced as the game develops. However, users' conversations let us predict when friendships break.

Thus, we argued that Diplomacy would be an exciting testbed for natural language processing, and our 2015 paper is—to the best of our knowledge—the first NLP research on Diplomacy. We discovered that betrayers write ostensibly friendly messages before a betrayal become more polite, stop talking about the future, and change how *much* they write [33]. In follow-on work, we developed a dataset that predict both when users lie to each other and when recipients of lies detect deception [34]. Diplomacy may be a nerdy game, but it is a fruitful testbed to teach computers to understand messy, emotional human interactions. We are continuing to look into these questions with researchers from across the nation in a new DARPA program: SHADE, which focuses on Diplomacy as a testbed for understanding deception.

Recently, the use of NLP methods in the game of Diplomacy has been the subject of highly-publicized papers by DeepMind in Nature Communications [24] and Meta in Science [3]. The Meta paper, like our 2020 paper, used a classifier to detect deceptive statements. The DeepMind paper built a game theoretic understanding of when betrayal should happen, building on our descriptive investigation of deception in human games.

A game with higher stakes is politics. However, just like Diplomacy, the words that people use reveal their underlying goals; computational methods can help expose the "moves" political players can use. With collaborators in political science, we've built models that: show when politicians in debates strategically change the topic to influence others [29, 31]; frame topics to reflect political leanings [30]; use subtle linguistic phrasing to express their political leaning [22]; or create political subgroups with larger political movements [32].

Because political discourse is built on a common set of commonly accepted facts, we have focused on developing fact checking: datasets for general knowledge fact checking [8] and climate change fact checking [7]. However, because fact checking is part of an information arms race, we need to build these examples as part of a human-in-the-loop adversarial process, which I'm exploring in an ongoing collaboration with journalism professor Naeemul Hassan that extends my question answering work, which I talk about next.

4. Human-in-the-Loop Adversarial Examples

One of the most fun aspects of my research has been building trivia-playing robots [5, 21, 23]; in addition to the research, it has faced off against former Jeopardy champions in front of hundreds high school students[2] and against researchers at NeurIPS 2015 (which won the best demonstration award). But after defeating some of the smartest trivia players, did I actually believe that our system was better at question answering? No!

[2] https://www.youtube.com/watch?v=LqsUaprYMOw

Adversarial examples first came out of the vision community: add a small epsilon to an example and suddenly a object detector calls a turtle a gun [2].[3] While others have attempted to create adversarial examples for language using paraphrasing, it's hard to know if the changes are perceptually negligible ("who wrote the invisible man" is fundamentally different from "who wrote the man you can't see") and it's hard to "add epsilon" to a discrete word.

Consistent with the theme of my research, my NSF CAREER grant added a *human in the loop* to generate novel adversarial language examples that can provide new training examples to make AI more robust and to expose what AI cannot (yet) do. With Eric Wallace, an undergraduate student, we built a system that could help an expert trivia question writer to stump a computer: as the author writes the question, it shows the author what the system is thinking [43]. And it worked, even generalizing across models [44] (an example written with an IR model still stumps a neural model). After we introduced human-in-the-loop adversarial example generation, Meta Facebook adopted this framework with gusto [4] in their Dynabench framework, the Dynamic Adversarial Data Collection workshop, and call for proposals (which I'm grateful is funding our continuing research in this area).

5. But wait, there's more!

Many of our best-cited papers are "traditional" papers that do better on some task:

- We developed the deep averaging network [23, DAN], an incredibly simple model that is still being used even in the age of transformers [45].
- In question answering we have proposed new evaluation mechanisms for knowing if an answer is correct [40] or to improve unsupervised retrieval of information to answer complicated questions [9, 37, 38].
- We also introduced reinforcement learning to *simultaneous machine interpretation* [14], a language-based task that requires significant human intuition, insight, and—for those who want to become interpreters—training.[4] We learned tricks from professional human interpreters—passivizing sentences and guessing the verb—to translate sentences sooner [17], letting speakers and algorithms cooperate together and enabling more natural cross-cultural communication. We also use reinforcement learning to learn machine translation feedback from noisy supervision such as star ratings on a webpage [28].

This work doesn't *yet* fit nicely into the human–computer collaboration narrative, but these more complex tasks are part of my broader vision for where my research will go: state-of-the-art models built to support human decisions, not replace them. And that requires the low-latency models built to react to input "like a human" described above.

6. Future Work

To advance AI, we need to ask better questions. Existing datasets are not diverse in the questions that they ask about: Google's Natural Questions, SQuAD, and others contain entities that are overwhelmingly male and either American on British [13]. More importantly Rogers et al. [36] outline an ontology of what skills a computer answering questions *should* possess, and adversarial QA generation systems do not probe these skills. Using item response theory and skilled authors, we will probe and explicate just how capable modern AI is and where it still needs more work.

Where computers have strengths that go beyond what a human can do, we will build interactive systems that help users come to a correct answer through a process they trust. This requires basic engineering—ensuring all of the components of a system are efficient and low latency—and user modeling, as we cannot assume that every user will have the same knowledge and capabilities. Then it will require careful vetting in diverse domains to validate that users' skills and knowledge are actually augmented by the help of the computer. We aim to focus on multiple areas: question answering in a single language [18], question answering in a language [15] or culture [35] you are unfamiliar with, and the strategic game of Diplomacy.

[3] Point of personal pride: I mentored Kevin on another previous research project [16], but I myself had nothing to do with this later adversarial work.

[4] This framework—using reinforcement learning to capture human strategies—was featured in Liang Huang's ACL keynote.

As these systems become more capable and usable, we can no longer assume that our model of the user should remain static: the user will learn and adapt to the system. This makes modeling more complicated, but it also allows for employing these models in educational settings through examples ordered in a curriculum: expanding the frontier of what the user knows, reinforcing weaker knowledge, and using strategies to both educate and explain information from the AI. This will require matching users' capabilities with what models can offer.

But interactions with individual people are not how AI will be a part of 21^{st} society: it will be interactions with *populations*. Thus, we need to have models and systems that capture population-level interactions. Helping detect misinformation online, working collaboratively with authors to craft effective countermeasures, and to propagate that within a social network. This builds on our fake news systems, our deception detection work, but will also require deeper collaboration with social scientists and journalists to develop the interfaces and the models to build human–computer AI that informs and helps society as a whole.

Full list of my publications at http://boydgraber.org/dyn-pubs/year.html

References

[1] Arora, S., Ge, R., Moitra, A.: Learning topic models—going beyond svd. In: Proceedings of Foundations of Computer Science (2012)

[2] Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: Proceedings of the International Conference of Machine Learning (2018)

[3] Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A.P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A.H., Mitts, S., Renduchintala, A., Roller, S., Rowe, D., Shi, W., Spisak, J., Wei, A., Wu, D., Zhang, H., Zijlstra, M.: Human-level play in the game of <i>diplomacy</i> by combining language models with strategic reasoning. Science 378(6624), 1067–1074 (2022), https://www.science.org/doi/abs/10.1126/science.ade9097

[4] Bartolo, M., Roberts, A., Welbl, J., Riedel, S., Stenetorp, P.: Beat the AI: Investigating adversarial human annotation for reading comprehension. Transactions of the Association for Computational Linguistics 8, 662–678 (2020), https://aclanthology.org/2020.tacl-1.43

[5] Boyd-Graber, J., Satinoff, B., He, H., III, H.D.: Besting the quiz master: Crowdsourcing incremental classification games. In: Proceedings of Empirical Methods in Natural Language Processing (2012)

[6] Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Proceedings of Advances in Neural Information Processing Systems (2009)

[7] Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., Leippold, M.: CLIMATE-FEVER: A dataset for verification of real-world climate claims. In: NeurIPS Workshop on Tackling Climate Change with Machine Learning (2020)

[8] Eisenschlos, J.M., Dhingra, B., Bulian, J., Börschinger, B., Boyd-Graber, J.: Fool me twice: Entailment from wikipedia gamification. In: North American Association of Computational Linguistics (2021), http://umiacs.umd.edu/~jbg//docs/2021_naacl_fm2.pdf

[9] Elgohary, A., Peskov, D., Boyd-Graber, J.: Can you unpack that? learning to rewrite questions-in-context. In: Empirical Methods in Natural Language Processing (2019)

[10] Feng, S., Boyd-Graber, J.: What AI can do for me: Evaluating machine learning interpretations in cooperative play. In: International Conference on Intelligent User Interfaces (2019)

[11] Feng, S., Boyd-Graber, J.: Learning to explain selectively: A case study on question answering. In: Empirical Methods in Natural Language Processing (2022), http://umiacs.umd.edu/~jbg//docs/2022_emnlp_augment.pdf

[12] Fujinuma, Y., Paul, M., Boyd-Graber, J.: A resource-free evaluation metric for cross-lingual word embeddings based on graph modularity. In: Association for Computational Linguistics (2019), http://umiacs.umd.edu/~jbg//docs/2019_acl_modularity.pdf

[13] Gor, M., Webster, K., Boyd-Graber, J.: Toward deconfounding the influence of subject's demographic characteristics in question answering. In: Empirical Methods in Natural Language Processing. p. 6 (2021), http://umiacs.umd.edu/~jbg//docs/2021_emnlp_qa_fairness.pdf

[14] Grissom II, A., He, H., Boyd-Graber, J., Morgan, J.: Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In: Proceedings of Empirical Methods in Natural Language Processing (2014)

[15] Han, H., Carpuat, M., Boyd-Graber, J.: Simqa: Detecting simultaneous mt errors through word-by-word question answering. In: Empirical Methods in Natural Language Processing (2022), http://umiacs.umd.edu/~jbg//docs/2022_emnlp_simqa.pdf

[16] He, H., Boyd-Graber, J., Kwok, K., Daumé III, H.: Opponent modeling in deep reinforcement learning. In: Proceedings of the International Conference of Machine Learning (2016)

[17] He, H., Grissom II, A., Boyd-Graber, J., Daumé III, H.: Syntax-based rewriting for simultaneous machine translation. In: Empirical Methods in Natural Language Processing (2015), http://umiacs.umd.edu/~jbg//docs/2015_emnlp_rewrite.pdf

[18] He, W., Mao, A., Boyd-Graber, J.: Cheater's bowl: Human vs. computer search strategies for open-domain qa. In: Findings of Empirical Methods in Natural Language Processing (2022), http://umiacs.umd.edu/~jbg//docs/2022_emnlp_cheaters.pdf

[19] Hoyle, A., Goel, P., Peskov, D., Hian-Cheong, A., Boyd-Graber, J., Resnik, P.: Is automated topic model evaluation broken?: The incoherence of coherence. In: Neural Information Processing Systems (2021), http://umiacs.umd.edu/~jbg//docs/2021_neurips_incoherence.pdf

[20] Hu, Y., Boyd-Graber, J., Satinoff, B., Smith, A.: Interactive topic modeling. Machine Learning 95(3), 423–469 (Jun 2014), http://dx.doi.org/10.1007/s10994-013-5413-0

[21] Iyyer, M., Boyd-Graber, J., Claudino, L., Socher, R., Daumé III, H.: A neural network for factoid question answering over paragraphs. In: Proceedings of Empirical Methods in Natural Language Processing (2014)

[22] Iyyer, M., Enns, P., Boyd-Graber, J., Resnik, P.: Political ideology detection using recursive neural networks. In: Proceedings of the Association for Computational Linguistics (2014)

[23] Iyyer, M., Manjunatha, V., Boyd-Graber, J., Daumé III, H.: Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of the Association for Computational Linguistics (2015), http://www.cs.colorado.edu/~jbg/docs/2015_acl_dan.pdf

[24] KramÃąr, J., Eccles, T., Gemp, I., Tacchetti, A., McKee, K.R., Malinowski, M., Graepel, T., Bachrach, Y.: Negotiation and honesty in artificial intelligence methods for the board game of diplomacy. Nature Communications 13(1), 10.1038/s41467-022-34473-5
[25] Kumar, V., Smith, A., Findlater, L., Seppi, K., Boyd-Graber, J.: Why didn't you listen to me? comparing user control of human-in-the-loop topic models. In: Proceedings of the Association for Computational Linguistics (2019)
[26] Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: Proceedings of the European Chapter of the Association for Computational Linguistics (2014)
[27] Lund, J., Cook, C., Seppi, K., Boyd-Graber, J.: Tandem anchoring: A multiword anchor approach for interactive topic modeling. In: Association for Computational Linguistics (2017)
[28] Nguyen, K., Boyd-Graber, J., Daumé III, H.: Reinforcement learning for bandit neural machine translation with simulated human feedback. In: Empirical Methods in Natural Language Processing (2017), http://umiacs.umd.edu/~jbg//docs/2017_emnlp_bandit_mt.pdf
[29] Nguyen, V.A., Boyd-Graber, J., Resnik, P.: SITS: A hierarchical nonparametric model using speaker identity for topic segmentation in multiparty conversations. In: Proceedings of the Association for Computational Linguistics (2012)
[30] Nguyen, V.A., Boyd-Graber, J., Resnik, P.: Lexical and hierarchical topic regression. In: Proceedings of Advances in Neural Information Processing Systems (2013)
[31] Nguyen, V.A., Boyd-Graber, J., Resnik, P., Cai, D., Midberry, J., Wang, Y.: Modeling topic control to detect influence in conversations using nonparametric topic models. Machine Learning 95, 381–421 (2014), http://umiacs.umd.edu/~jbg//docs/2014_mlj_influencer.pdf
[32] Nguyen, V.A., Boyd-Graber, J., Resnik, P., Miler, K.: Tea party in the house: A hierarchical ideal point topic model and its application to Republican legislators in the 112th Congress. In: Association for Computational Linguistics (2015)
[33] Niculae, V., Kumar, S., Boyd-Graber, J., Danescu-Niculescu-Mizil, C.: Linguistic harbingers of betrayal: A case study on an online strategy game. In: Association for Computational Linguistics (2015), http://umiacs.umd.edu/~jbg//docs/2015_acl_diplomacy.pdf
[34] Peskov, D., Cheng, B., Elgohary, A., Barrow, J., Danescu-Niculescu-Mizil, C., Boyd-Graber, J.: It takes two to lie: One to lie and one to listen. In: Association for Computational Linguistics (2020), http://umiacs.umd.edu/~jbg//docs/2020_acl_diplomacy.pdf
[35] Peskov, D., Hangya, V., Boyd-Graber, J., Fraser, A.: Adapting entities across languages and cultures. Findings of Empirical Methods in Natural Language Processing (2021), http://umiacs.umd.edu/~jbg//docs/2021_emnlp_adaptation.pdf
[36] Rogers, A., Gardner, M., Augenstein, I.: QA dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. In: ACM Computing Surveys (2022)
[37] Shi, T., Zhao, C., Boyd-Graber, J., Daumé III, H., Lee, L.: On the potential of lexico-logical alignments for semantic parsing to sql queries. In: Findings of EMNLP (2020), http://umiacs.umd.edu/~jbg//docs/2020_findings_qalign.pdf
[38] Shi, T., Zhao, C., Boyd-Graber, J., Daumé III, H., Lee, L.: On the potential of lexico-logical alignments for semantic parsing to sql queries. In: Findings of EMNLP (2020), http://umiacs.umd.edu/~jbg//docs/2020_findings_qalign.pdf
[39] Shneiderman, B.: Human-Centered AI: A New Synthesis. Springer-Verlag, Berlin, Heidelberg (2021), https://doi.org/10.1007/978-3-030-85623-6_1
[40] Si, C., Zhao, C., Boyd-Graber, J.: What's in a name? answer equivalence for open-domain question answering. In: Empirical Methods in Natural Language Processing (2021), http://umiacs.umd.edu/~jbg//docs/2021_emnlp_answer_equiv.pdf
[41] Smith, A., Boyd-Graber, J., Fan, R., Birchfield, M., Wu, T., Weld, D., Findlater, L.: No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In: Computer-Human Interaction (2020), http://umiacs.umd.edu/~jbg//docs/2020_chi_explanation.pdf
[42] Smith, A., Kumar, V., Boyd-Graber, J., Seppi, K., Findlater, L.: Digging into user control: Perceptions of adherence and instability in transparent models. In: Intelligent User Interfaces (2020), http://umiacs.umd.edu/~jbg//docs/2020_iui_control.pdf
[43] Wallace, E., Boyd-Graber, J.: Trick me if you can: Adversarial writing of trivia challenge questions. In: ACL Student Research Workshop (2018), http://aclweb.org/anthology/P18-3018
[44] Wallace, E., Rodriguez, P., Feng, S., Yamada, I., Boyd-Graber, J.: Trick me if you can: Human-in-the-loop generation of adversarial question answering examples. Transactions of the Association of Computational Linguistics 10 (2019)
[45] Ye, Q., Khabsa, M., Lewis, M., Wang, S., Ren, X., Jaech, A.: Sparse distillation: Speeding up text classification by using bigger student models. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2361–2375. Association for Computational Linguistics, Seattle, United States (Jul 2022), https://aclanthology.org/2022.naacl-main.169
[46] Yuan, M., Van Durme, B., Boyd-Graber, J.: Multilingual anchoring: Interactive topic modeling and alignment across languages. In: Neural Information Processing Systems (2018), http://umiacs.umd.edu/~jbg//docs/2018_neurips_mtanchor.pdf
[47] Zhang, M., Xu, K., Kawarabayashi, K.i., Jegelka, S., Boyd-Graber, J.: Are girls neko or shōjo? Cross-lingual alignment of non-isomorphic embeddings with iterative normalization. In: Association for Computational Linguistics (2019), http://umiacs.umd.edu/~jbg//docs/2019_acl_clwe.pdf