Generalizable, Reliable, and Interpretable Language Technologies

Anjalie Field, anjalief@cs.cmu.edu *Carnegie Mellon University*

Even though natural language processing (NLP) is an applied field, many NLP technologies do not translate to real-world settings. Machine learning models have become prevalent in classification tasks, including child-welfare screening and recidivism prediction, but these models rarely incorporate text features. While data like notes written by hot-line operators or parole officers could improve performance, they are difficult to process and can introduce bias. Biases are additionally problematic in machine translation, coreference resolution, and other language technologies, because they cause models to perform disproportionately poorly for underrepresented minorities [20, 17]. In other scenarios, bias in text could be leveraged to study deeper social issues: gender bias is prevalent in social media comments and state-owned newspapers contain government propaganda. However, NLP traditionally uses supervised machine learning models and carefully curated data sets that do not generalize to the diverse types of issues prevalent in society.

As a PhD student at Carnegie Mellon University, I aim to build machine learning models capable of analyzing social issues and incorporating text into high-stakes AI systems. Over the past 2 years, I have worked with my advisor, Prof. Yulia Tsvetkov, on several projects in this area by investigating topics like propaganda, media bias, and gender bias [8, 7, 9]. In the next few years, I intend to focus on 3 primary strategies that are essential for shifting NLP research towards societal applicability: (1) Developing distant supervision and domain adaptation methods that *generalize* models to diverse types of tasks, (2) introducing methods that control for confounds to improve model *reliability*, and (3) ensuring model decisions and outputs are *interpretable* to downstream users. While I focus on these aims in NLP models, they are also directly applicable to other machine learning systems.

Generalizability

Current reliance on supervised models restricts the usefulness of NLP research, because obtaining annotated data can be difficult or impossible in many scenarios. We cannot answer a question like "Does this newspaper article contain bias?" using a supervised classification model. First, concepts like bias are difficult to define and require expertise to identify, which makes annotating data prohibitively time-consuming and expensive. Second, it is challenging even for expert annotators to prevent their own biases from influencing the annotations [16]. Third, annotations are often specific to tasks: even if we collected annotations over one set of articles, they cannot be reused in other scenarios [4].

In preliminary work, I have addressed these challenges by building NLP models that adapt simple generalizable annotations to task-specific questions. I investigated how to measure difficult-to-define complex phenomena in two primary domains: how *people* are described and how *events* are framed. In the first, I examined how people are portrayed differently across outlets in media coverage of the #MeToo movement (ICWSM 2019 [9]), as well as in newspaper articles more generally (ACL 2019 [7]). How people are portrayed is a broad research question that cannot be approached as a supervised prediction task. Instead, I drew from behavioral science theories that have identified power, agency, and sentiment as the 3 most important axes of affective meaning [13].

I then combined word-level annotations along these dimensions with pre-trained contextualized word embeddings, ultimately identifying bias; for example, even though the #MeToo movement has been viewed as empowering, women are often portrayed as less powerful than men. Figure 1 shows a more specific example: an anonymous woman using the pseudonym Grace was often portrayed has having lower agency than the man she accused of sexual harassment, Aziz Ansari. Both Grace and Aziz Ansari have lower agency than 3 journalists who reported on the event (Caitlin Flannagan, Katie Way, Ashleigh Banfield). In ongoing extensions to this project, I am working on analyzing how members of the LGBTQ community are portrayed in different cultures, primarily using data from Wikipedia in several languages. Multilingual models that can project annotations across languages and cultures allows these annotations to generalize even further.

In the second domain, I similarly combined theories from various disciplines with distant supervision in order to analyze a complex phenomena: how autocratic governments manipulate public opinion (EMNLP 2018 [8]). I analyzed a corpus of Russian newspaper articles by drawing two propaganda theories, namely framing and agenda-setting, from political science literature. In order to measure these abstract concepts, I compared news coverage to economic indicators (agenda-setting) and developed a distant supervision approach that projects English framing annotations into Russian (framing). I am currently extending this project to examine Indian newspaper articles, particularly to what extent news coverage in India follows the same patterns as coverage in countries with state-controlled media like Russia and China. This ongoing work focuses on how to find annotation labels that were not included in the original annotation task.

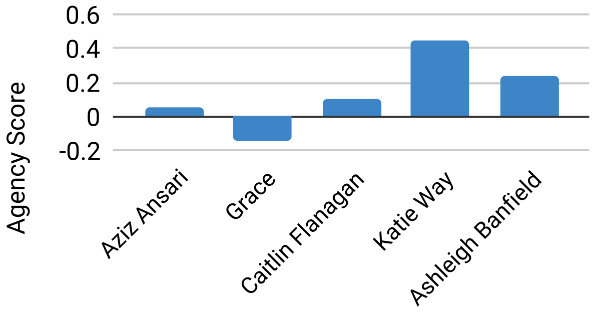

Figure 1: Agency scores for entities in media coverage of the #MeToo movement, obtained using our methodology [9]. Anonymous accuser Grace is portrayed with lower agency than accused Aziz Ansari. Both have lower agency than journalists Caitlin Flannagan, Katie Way, Ashleigh Banfield.

Both of these lines of work involve developing generalizable methodologies to measure complex phenomena. Rather than creating a task-specific annotation scheme for each research question, I focused on developing distant supervision and domain adaptation methods that allow general annotations to be useful for task-specific questions. A crucial part of this process involves drawing from existing social science theory to understand what general annotations are most informative and how to adapt them to a particular task. This framework for NLP allows us to build models that are applicable to a broad range of topics.

Reliability

Reducing the dependency on direct supervision makes models broadly applicable, but to deploy these models in high-stakes settings, it is essential to ensure that they measure target values without becoming biased by other confounds. NLP research overly emphasizes performance metrics, e.g. accuracy over a standardized test set, without considering what features led to the model's performance. More specifically, models that improve accuracy by exploiting spurious correlations in the data are often considered better than models that limit focus to target features. This viewpoint is harmful to NLP research, as it results in models that overfit to singular tasks, and it is detrimental when models are deployed, since they can introduce biases. Researchers from Google Research, Google Brain, Jigsaw, as well as academic institutions have shown that models for abusive language detection, co-reference resolution, and machine translation exhibit bias, including models deployed by Alphabet [16, 14, 20, 17]. While only recently acknowledged in NLP, this problem is prevalent in machine learning more generally: models for predicting risk of recidivism have been heavily criticized for implicitly learning racism [5].

I am currently working on preventing models from absorbing spurious correlations by drawing from substantial work on controlling for confounds in statistics and fairness and incorporating these concepts into machine learning models. In an ongoing project, I am using this approach to identify gender bias in second-person text. My primary data set consists of comments addressed towards individuals, where each comment is annotated with the gender of the addressee [19]. I reveal systemic differences in comments addressed towards men and women by training a model to predict the gender of the addressee and examining which features the model uses to make predictions. The main challenge in this project is preventing the model from focusing on overt predictive features, like names and pronouns, in order to identify subtle features that are indicative of bias. The difficulty in controlling for these types of confounds is one of the reasons that most work on gender bias in NLP has focused on corpus-level analyses [22, 3], rather than phrase-level detection, which is more practically applicable. My method for demoting overt features uses text-based propensity matching inspired by causality literature [18] and also draws from recent work in NLP on demoting latent confounds [11]. I expect to publish preliminary results from this work within the next year.

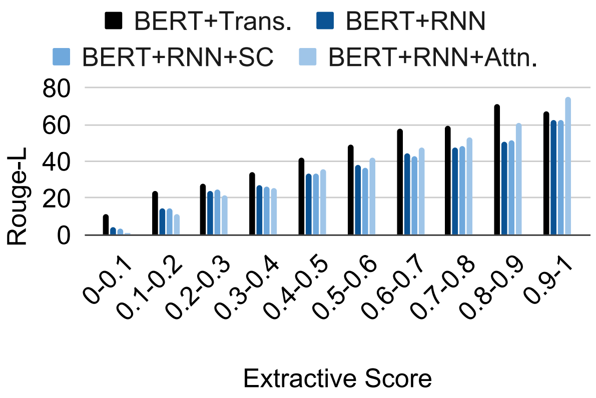

Figure 2: Rouge-L scores for our models over the Gigaword test data of length 5-10 tokens, segmented according to extractive score. Model performance is strongly correlated with how extractive the data sample is (how many tokens can be directly copied from the input into the output).

Furthermore, as a Research Intern with the Structured Data team at Google this summer, I worked on understanding how models achieve high performance in text generation tasks, focusing on architectures initialized with pre-trained language models. This work exemplifies how models can achieve high scores on performance metrics by exploiting patterns in the data. More specifically, I showed that decoders with self-attention mechanisms, such as a transformer, outperform decoders without self-attention, but only over highly extractive data samples, where there is high token overlap between the input text and output text. Figure 2 displays this trend using results from four of our models. While the concept of extractiveness has been explored previously in text summarization research, I used state-of-the art transformer models and additionally evaluated models' internal states for their ability to capture semantic information. I found that self-attention mechanisms achieve high performance scores when they can copy input tokens into the output, but they can discourage models from learning deeper semantics. My work was primarily supervised by Abe Ittycheriah and Cong Yu, and we are in the process of submitting results from this project for publication.

Interpretability

Much of the resistance around adopting machine learning models for societal issues traces to concerns about “black-boxes” that output values with no context or explanations, suggesting that NLP systems cannot become widely applicable without transparency and interpretabiliy. Understanding why models make decisions can lead to higher accuracy on tasks like coreference resolution and machine translation, and it is essential in settings like child-welfare screening where incorrect decisions have severe consequences [6]. This need has prompted much recent interest in the NLP community [1]. Furthermore, methodologies that improve interpretability can be leveraged to identify bias in models and data sets. My work on comment-level gender bias described in the previous section relies on developing and leveraging this type of explainability. By training a model to predict gender and examining what features influence the model’s decision, we can reveal the systemic differences in comments addressed towards men and towards women. My methodology for identifying these indicators of bias builds on prior work about interpretability, including saliency maps, attention scores, and generated rationales [12, 21].

All of these approaches generate explanations using input test data, but in many cases, models must leverage training data in order to provide meaningful explanations. For instance, in a model that predicts the risk of recidivism, we can infer that the model makes biased decisions if it frequently equates input profiles with training profiles of people of the same race. In future work on this topic, I will use data sets for analyzing bias in coreference resolution systems [15], and extend existing work on influence functions [10] to understand which training points are most influential in a model’s decision and *why*. While recent work has sought to expose and mitigate bias in coference resolutions systems –this task was a focus of the Gender Bias in NLP workshop at ACL 2019, which included organizers and sponsors from Google Research [2]– little work has examined the role of interpretability in revealing and mitigating this bias. For example, given an input sentence “The doctor asked the nurse to help her”, can we show that the model incorrectly predicts “her” refers to “nurse”, because in most training samples, nurses are women? Developing this level of interpretability serves as a starting point for improving the performance of NLP systems by preventing them from making biased errors. Additionally, it would allow us to deploy models in high-stakes settings, since we could monitor them for signs of bias.

My work addresses some of the major limitations in NLP models that make them unusable for real-world tasks. While I highlight several limitations in existing systems, these concepts are intended to be a starting point, and a component of the proposed work will involve identifying additional important limitations. One related concept that I intend to address throughout these projects is ethics. While ethics is a relatively new focus area in NLP, it is a prominent component of other fields, such as philosophy, psychology, and fairness in machine learning. As a teaching assistant for the course created by my advisor, titled “Computational Ethics for NLP”, I have established partnerships with experts in these fields, as well as fostered discussions about ethics among NLP researchers. Overall, I am excited to continue researching these topics, and I hope that my work will result in a lasting positive impact on society.

References

[1] *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W18-5400`.

[2] *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, Florence, Italy, Aug. 2019. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W19-3800`.

[3] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*, pages 4349–4357, 2016.

[4] A. E. Boydstun, J. H. Gross, P. Resnik, and N. A. Smith. Identifying media frames and frame dynamics within and across policy issues. In *TADA Workshop, London*, 2013.

[5] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[6] A. Chouldechova, D. Benavides-Prado, O. Fialko, and R. Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *FAT**, 2018.

[7] A. Field and Y. Tsvetkov. Entity-centric contextual affective analysis. In *ACL*, 2019.

[8] A. Field, D. Kliger, S. Wintner, J. Pan, D. Jurafsky, and Y. Tsvetkov. Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies. In *EMNLP*, 2018.

[9] A. Field, G. Bhat, and Y. Tsvetkov. Contextual affective analysis: A case study of people portrayals in online #metoo stories. In *ICWSM*, 2019.

[10] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *ICML*, 2017.

[11] S. Kumar, S. Wintner, N. A. Smith, and Y. Tsvetkov. Topics to avoid: Demoting latent confounds in text classification. In *EMNLP*, 2019.

[12] T. Lei, R. Barzilay, and T. Jaakkola. Rationalizing neural predictions. In *EMNLP*, 2016.

[13] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. *The measurement of meaning*. Number 47. University of Illinois press, 1957.

[14] V. Prabhakaran, B. Hutchinson, and M. Mitchell. Perturbation sensitivity analysis to detect unintended model biases. In *EMNLP*, 2019.

[15] R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme. Gender bias in coreference resolution. In *NAACL*, 2018.

[16] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. The risk of racial bias in hate speech detection. In *ACL*, 2019.

[17] G. Stanovsky, N. A. Smith, and L. Zettlemoyer. Evaluating gender bias in machine translation. In *ACL*, June 2019.

[18] E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[19] R. Voigt, D. Jurgens, V. Prabhakaran, D. Jurafsky, and Y. Tsvetkov. Rtgender: A corpus for studying differential responses to gender. In *LREC*, 2018.

[20] K. Webster, M. Recasens, V. Axelrod, and J. Baldridge. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *TACL*.

[21] S. Wiegreffe and Y. Pinter. Attention is not not explanation. In *EMNLP*, 2019.

[22] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.