Detecting Independent Pronoun Bias with Partially-Synthetic Data Generation

Robert (Munro) Monarch
Machine Learning Consulting
San Francisco, CA
rmunro@alumni.stanford.edu

Alex (Carmen) Morrison
Amazon Web Services
Oakland, CA

Abstract

We report that state-of-the-art parsers consistently failed to identify "hers" and "theirs" as pronouns but identified the masculine equivalent "his". We find that the same biases exist in recent language models like BERT. While some of the bias comes from known sources, like training data with gender imbalances, we find that the bias is *amplified* in the language models and that linguistic differences between English pronouns that are not inherently biased can become biases in some machine learning models. We introduce a new technique for measuring bias in models, using Bayesian approximations to generate partially-synthetic data from the model itself.

1 Introduction

We share a negative result for the Natural Language Processing (NLP) community as a whole: for 20 years the major part-of-speech (POS) taggers and parsers missed that "hers" and "theirs" were pronouns, but it had gone unreported until this paper. This paper also shows that biases against "hers" and "theirs" are *amplified* in popular language models, predicted by the models with less frequency than expected given the training data.

Our solution for the parser problem is a new dataset with "hers" and "theirs" used in a syntactically diverse set of contexts, released in Universal Dependency format (Nivre et al., 2016).

For the language model problem, we introduce a novel use of Bayesian modeling for sentence generation, in our case using it to detect bias by alternating pronouns in different contexts. The contexts are suggested by the model, avoiding the problems in measuring bias that come from rare or pathological data (Feng et al., 2018). We conclude that this is a general technique that can be used for measuring other types of bias and for text generation more broadly.

While our contribution doesn't mitigate bias in language models, it improves the ability to detect and measure bias. We test on BERT (Devlin et al., 2019) because it is the most widely used pretrained model for which all the training data was also available. We find that other masked language models are also amplifying the bias in their data, but we cannot measure how much that bias is amplified without access to the training data. Where data for a pretrained model can't be shared, we encourage researchers to report on the biases in their models using the techniques in this paper.

1.1 The bias is hers...and singular theirs

Independent possessive pronouns (IDPs) are interesting problems for NLP because they are the only English pronouns to encode *two* long-distance relationships: the person possessing an attribute and the attribute being possessed (see Table 1).

The preference for "his" in language models will bias any text generation system against "hers" or "theirs", a problem that has led developers to remove gendered pronouns entirely from applications including *Gmail*'s predictive text (Dave, 2018).

Identifying IDPs as pronouns is also a necessary step for co-reference resolution, although recent shared-tasks for pronoun resolution did not include IDPs (Webster et al., 2018, 2019).

Unfortunately, major academic and commercial parsers including those from *AWS* and *Google* (Andor et al., 2016) wrongly labeled "hers" and "theirs" as adjectives or nouns.

Parsers from 20 years ago also missed these pronouns (Charniak, 2000; Charniak and Johnson, 2005), confirming that this is not a new bias that only surfaced with more recent dependency parsers. The syntactic information was also typically wrong, for example, parsers that labeled "hers" as an "adjective" wrongly classified the syntactic relationships as modifiers. This will perpetuate bias in any

	Subj	Obj	Dep	Ind
Feminine	she	**her**		hers
Masculine	he	him	**his**	
Neutral	they	them	their	theirs
1st Person	I	me	my	mine
2nd Person	**you**		your	yours

Table 1: Common English personal pronouns showing the irregular "her", "his" and "you". Key (Examples):
Subj: Subject (**they** saw a cat")
Obj: Object ("a cat saw **them**")
Dep: Dependent Possessive ("**their** cat")
Ind: Independent Possessive (IDP) ("a cat saw **theirs**")
In addition, a "-self/-selves" suffix on the Obj or Dep pronoun creates the Reflexive/Intensive pronoun.

downstream model using pronouns for co-reference or possession relationships.[1]

Two causes for biases are historical disempowerment resulting in less training data and linguistic differences in how "he/him/his" and "she/her/hers" pattern irregularly, as in Table 1.[2]

For the linguistic differences, the parsers correctly identify the *independent* "his" as a pronoun because they trained on the *dependent* "his" with the same form. Therefore, even without bias in the data, "hers" and "theirs" can be under-predicted because of *richer* grammatical distinctions.

Because "hers" and "theirs" have only one sense, a small amount of training data can fix the problem for the syntactic parsers. We solved this with the dataset introduced in Section 2. We recognize that more data would be needed to solve the problem for polysemous independent pronouns like "mine" and the singular/plural distinction for "theirs".

2 Process for detecting bias

In this section, we share our method for partially-synthetic data generation to measure bias in pretrained models (see the workflow in Figure 1).

The code is open-source and will reproduce the results in a single command:[3]

Step 1: Train a contextual model. We use the pretrained BERT (Devlin et al., 2019) English un-

[1]We tested more than a dozen systems that also failed to identify "hers" and "theirs" as pronouns, but limit our report to ones where we share responsibility for previously missing this because the authors have both worked at AWS and built NLP training data for Google.

[2]See this blog article from when we first announced this problem for more about why "hers" and "theirs" were missed by parsers until now: bit.ly/hers_theirs

[3]https://github.com/rmunro/masked_bias_detection

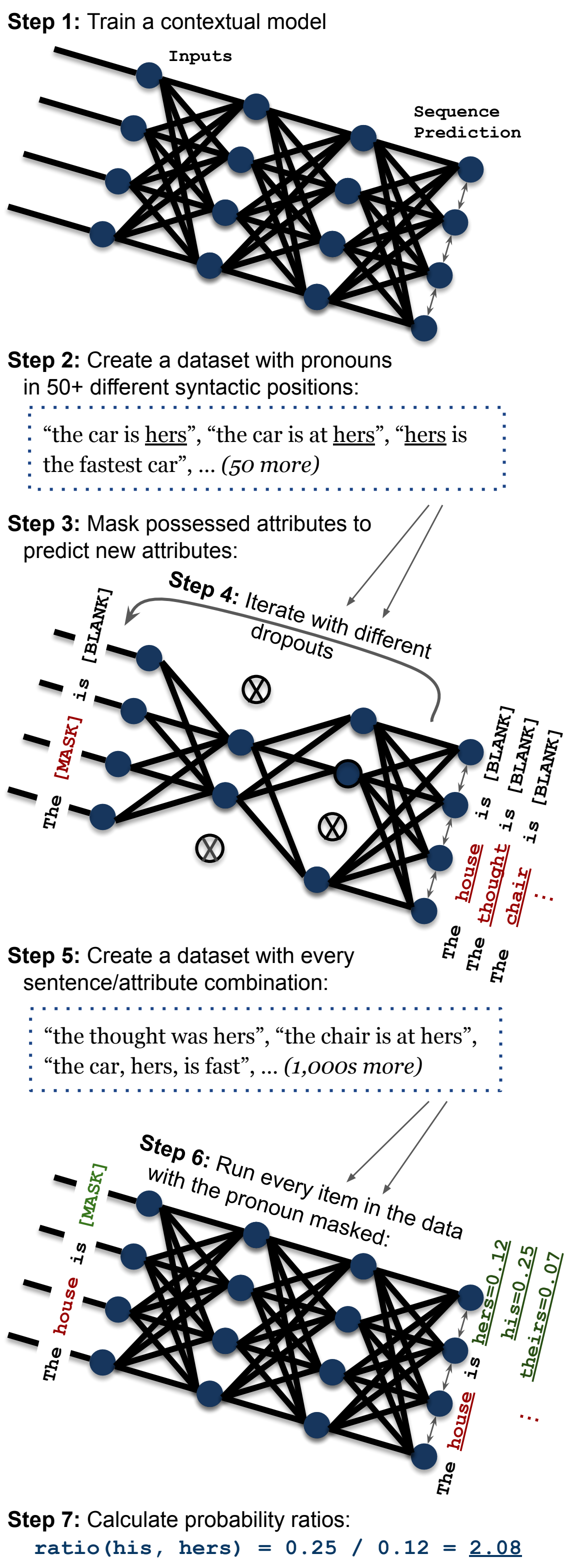

Figure 1: An overview of our workflow for bias-detection. We generate 1000s of unique sentences to test the bias, from an initial set of 50+ sentences created with maximally diverse contexts, utilizing masked models with dropouts to generate a list of candidate possessed attributes.

cased model with Whole Word Masking, 24-layers, 1024-hidden neurons, 16-heads, and 340M params.

Step 2: Create a starting dataset. We created a new dataset in Universal Dependency (UD) format that contained an independent possessive pronoun in different syntactic configurations (Pollard and Sag, 1994): Subject, Object, Extraction, Interjection, etc. and sentence types that are grammatically identical in English UD but different in other languages, like Transitive vs Intransitive sentences, prepositions ("in theirs", "at theirs", etc), formal/informal variations ("isn't theirs"/"aint theirs") and the IDP's special context ("[item] of theirs").

The dataset is sentence pairs, like "What color is Alex's car? Theirs is red" because natural-sounding single sentences were not always possible.[4]

Step 3: Mask attributes to predict new ones. Using the new dataset, mask attributes like "car(s)" so that BERT predicts the most likely tokens for where "car(s)" is masked in a sentence like "The [MASK] is hers".

Step 4: Iterate with Bayesian Deep Learning. Used random dropouts (Monte Carlo Sampling) to generate multiple attributes for each sentence. Dropouts at inference follow the same principles as for training, where an estimation function E that ignores neurons i for an input vector I and weight w at a dropout rate δ_i for least-squared loss is:

$$E = \frac{1}{2}\left(t - \sum_{i=1}^{n} \delta_i w_i I_i\right)^2 \qquad (1)$$

We use the same dropout profile in inference that the BERT model used for training and leave experiments with different dropout profiles as interesting potential future work.

To solve the problem of when to stop trying to generate new sentences, we use a modified Good-Turing estimate (Gale, 1995) where the core insight is that the number of items you have encountered just once is the main factor in predicting the likelihood of seeing new items. We calculate this as P_r when there are $C(att_1)$ attributes seen once, $C(att)$ attributes seen in total, and $sum_i att_i$ total attributes (including duplicates):

$$P_r = \frac{C(att_1) + C(att)}{\sum_i att_i} \qquad (2)$$

[4]https://github.com/UniversalDependencies/UD_English-Pronouns

We stop generating when P_r falls below a certain probability, 0.05 in the results presented here.

This process generated 115 attributes (see Figure 2) including concrete items like "camera" and "world" and abstract items like "night" and "instincts". Because these items are predicted to be the most likely item in a given context, we can be confident that they aren't low-frequency items that will make BERT produce erroneous results.

Step 5: Create a dataset with combinations. 11 of the 115 attributes produced grammatically incorrect sentences and were removed manually. The 104 remaining attributes were combined with the initial sentences, resulting in thousands of unique sentences.

Step 6: Predict the probability of different pronouns. With the thousands of sentences from Step 5, we generate sentences that use each attribute to predict the *pronoun*. For example, BERT is asked to guess what the masked (blank) word would be in 1000s of sentences like "the camera is ___", "the world is ___", "the night is ___", etc.

Step 7: Calculate the bias. Measure the ratio between the relative probabilities of "hers", "his" or "theirs" in the softmax output, following the bias-detection methods of Kurita et al. (2019).

3 Bias analysis

Figure 2 compares the model predictions to the training data which is Wikipedia and BookCorpus. BERT is trained on cross-entropy loss (Devlin et al., 2019), so if a token is four times more frequent than another in the data and occurs in the same contexts, then softmax should converge on a 4:1 ratio in the predictive model. While there can be a lot of variation in models that will change the ratio of a given prediction, especially for rare and pathological cases (Feng et al., 2018), we avoid rare tokens by having the model generate the contexts and by testing across thousands of sentences.

3.1 Possession is 99% of the flaw

For 103/104 attributes, "his" was preferred over "hers" or "theirs". For 92/104 attributes the model ratios were 3+ standard errors higher than the maximum data measurement, despite high variance. See Figure 2 for the full results.

The results in Figure 2 rule out that there is something special about possessives compared to other pronouns because ratios for objects in gendered sentences are consistently lower. The few exceptions

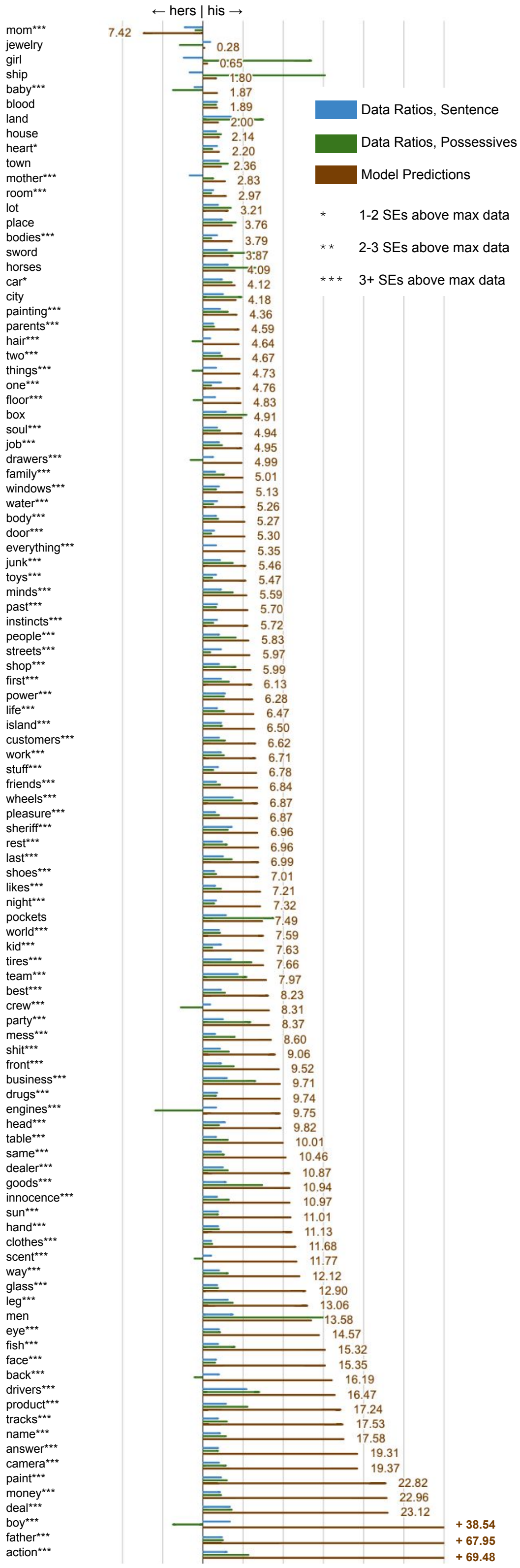

Figure 2: Comparison of model predictions and training data, showing the bias for "his" over "hers".
"Sentence": ratio of sentences with a gendered word ("man", "woman", etc) that contains the attribute.
"Possessives": ratio of explicit possessive structures like "his car".
"Model Predictions": the ratio between the pronouns in masked predictions.

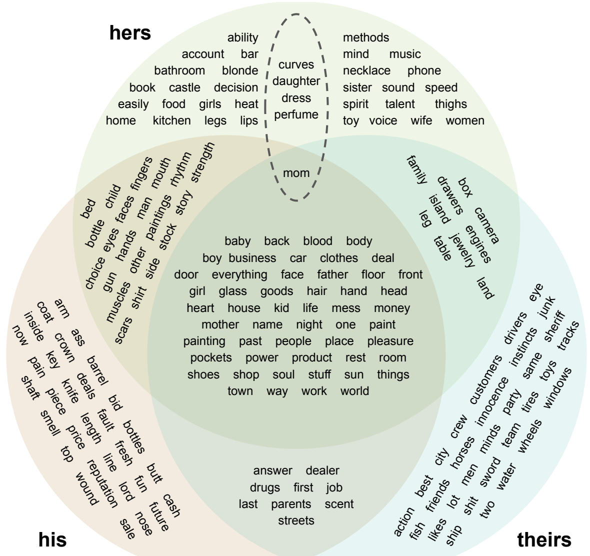

Figure 3: The words generated as possessions of "hers", "his", and/or "theirs" in our data. The dashed circle shows the five words predicted as possessed by "hers" when the pronoun is masked. The other words are predicted to be more likely possessed by "his", even for 70+ words that"his" did not generate as a possession.

in Figure 2 can be explained as being gendered directly ("girl", "men") or by convention ("she's a fast ship"). For the latter, the ownership shows that they are still male-dominated spaces ("*his* ship" and "*her* engines"), which is a novel observation for the relationship between the possessed and possessor.[5]

We also tried to bias the model in favor of "hers" and "theirs" by including them in Step 3 (see Figure 3), but even then, "theirs" was *never* the preference, and "hers" was rarely selected. For example "land" was generated in sentences like the "the [MASK] is hers/theirs", but never generated with sentences like "the [MASK] is his". However, when we try "the land is [MASK]", "his" is significantly predicted above "hers" and "theirs".

We conclude that the model has amplified the bias in the data that it is trained on.

4 Related Work

Kurita et al. (2019) introduced the method for measuring bias in contextualized word representations with two template sentences and a set of nouns. We extend Kurita et al. by using 1000s of sentences generated from 57 templates, instead of two, and by automatically expanding the context attributes with

[5]We thank an anonymous reviewer for pointing this out.

BERT itself to avoid pathological cases. Kurita et al. use the *difference* of log probabilities, but we use the (mathematically equivalent) *ratio* of actual probabilities because ratios allow more transparent comparisons with the corpus ratios.

Dropouts at inference generate a Gaussian distribution and is therefore known as *Bayesian Deep Learning* (Gal and Ghahramani, 2016). For Human-in-the-Loop methods, like this paper, it known as *Deep Active Bayesian Learning* (Shen et al., 2018; Siddhant and Lipton, 2018). Our paper is the first application of Bayesian Deep Learning to bias detection and, more broadly, to text generation.

Gonen at al. (2019) concluded that careful use of a language-specific morphological analyzer is needed to avoid bias in embeddings in gendered languages like Italian and German.

The recent Workshop on Gender Bias in Natural Language Processing (Costa-jussà et al., 2019) had a shared task for English gender-ambiguous pronouns (Webster et al., 2018, 2019), but the dataset and task did not include possessive pronouns.

Hahn (2020) shows limitations in how transformer models can learn finite-state and hierarchical structures. Therefore, some language models might be unable to fully distinguish the two forms of 'his' because they can not fully capture the syntax in the same way as humans, leading the dependent form of "his" in the training data to bias in favor of the independent form.

5 Accepting the responsibility themself

We found an additional problem when we first shared the "hers/theirs" problem one year before this paper: *some parsers didn't always recognize "themself" as a pronoun.*

The systems with "themself" errors did *not* fix the problem when they fixed "hers/theirs".

Our linguistic understanding of (neo)pronouns and inclusive NLP development are areas of ongoing research (Ackerman, 2019; Bradley et al., 2019; Cao and Daumé III, 2020; Conrod, 2020; Denton et al., 2020; Mitchell et al., 2020).

However, we argue that no system should have still missed "themself" after we alerted everyone to the "hers/theirs" error and recommended that every pronoun be investigated.

English is one of the simplest languages in terms of the paradigms like those in Table 1 (Bresnan, 2001), and the most well-studied in NLP (Bender and Friedman, 2018). If we can't identify unambiguous pronouns in our NLP, systems given a year's notice and clear instructions for how to find the errors, what biases are we missing elsewhere?

6 Discussion and Recommendations

People who use "hers", "theirs" and "themself" to align their current social gender(s) with their pronouns' grammatical gender are marginalized when applications fail to identify those pronouns. This is especially timely with singular "they" as Merriam-Webster's 2019 word of the year (Dwyer, 2019).

We find that pretrained models *amplify* biases in the data because linguistic differences that are not biases can become biases in the models. This has significant implications for bias in tasks like co-reference resolution and text generation.

From Gonen et al., (2019), it is not clear that debiasing the model itself would solve the problem. However, it might be possible to encode the data with the grammatical categories to mitigate some bias, for example, encoding the two "his" pronouns, like "his{DEP}" and "his{IND}". A pretrained model would therefore treat the two forms separately, without one amplifying the other even when a language model can't capture the syntactic differences.

We recommend that creators of widely used English syntactic parsers and part-of-speech taggers ensure that all unambiguous pronouns, including "hers", "theirs", and "themself", are correctly identified as pronouns. This will support applications the rely on technologies like conference resolution.

We recommend that creators of language models include identity words as full tokens. BERT's tokenizer includes all pronouns in this paper *except* "themself", thus exhibiting unintended feature-design bias by needing to be constructed as "them-s-elf" or "them-se-lf", presumably because all other forms with a "-self" suffix are already full tokens. That might include words other than pronouns, especially for multilingual models.

We recommend that creators of language models use the methods introduced in this paper for partially-synthetic data generation to diagnose potential bias in their models and that this text generation strategy is explored for other applications.

Acknowledgements

Thanks to Kellie Webster for detailed feedback on this paper and Emily Bender and Will Radford for detailed feedback on earlier versions. No faults for errors or omissions are theirs.

References

Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: A Journal of General Linguistics*, 4.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Evan D. Bradley, Julia Salkind, Ally Moore, and Sofi Teitsort. 2019. Singular 'they' and novel pronouns: gender-neutral, nonbinary, or both? In *Proceedings of the LSA*, volume 4.

Joan Bresnan. 2001. *Lexical-Functional Syntax*. Blackwell, Malden, MA.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference (NAACL'00)*, pages 132–139, Seattle, Washington. Association for Computational Linguistics.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Kirby Conrod. 2020. Predicative Pronouns. In *Linguistics Society of America Annual Meeting*, New Orleans, LA.

Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster. 2019. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. (Editors). Association for Computational Linguistics, Florence, Italy.

Paresh Dave. 2018. Fearful of bias, Google blocks gender-based pronouns from new AI tool. *Reuters*.

Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets. In *ICML Workshop on Participatory Approaches to Machine Learning*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.

Colin Dwyer. 2019. *Merriam-Webster Singles Out Nonbinary 'They' For Word Of The Year Honors*. National Public Radio (NPR).

Kawin Ethayarajh. 2020. Is your classifier actually biased? measuring fairness under uncertainty with bernstein bounds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2914–2919. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout As a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1050–1059, New York, NY, USA.

William A. Gale. 1995. Good-Turing smoothing without tears. *Journal of Quantitative Linguistics*, 2.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *The 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL-HLT*.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does Grammatical Gender Affect Noun Representations in Gender-Marking Languages? In *Proceedings of the 2019 Workshop on Widening NLP*, pages 64–67, Florence, Italy. Association for Computational Linguistics.

Michael Hahn. 2020. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*.

Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 117–123, New York, NY, USA. Association for Computing Machinery.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. 2018. Deep Active Learning for Named Entity Recognition. In *6th International Conference on Learning Representations, ICLR*.

Aditya Siddhant and Zachary C. Lipton. 2018. Deep Bayesian Active Learning for Natural Language Processing: Results of a Large-Scale Empirical Study. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kellie Webster, Marta R. Costa-jussà, Christian Hardmeier, and Will Radford. 2019. Gendered ambiguous pronoun (GAP) shared task at the gender bias in NLP workshop 2019. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Florence, Italy. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.