Gendered Ambiguous Pronouns (GAP) Shared Task at the Gender Bias in NLP Workshop 2019

Kellie Webster
Google Research
111 8th Avenue
New York, NY, USA
webster@google.com

Marta R. Costa-jussà
Universitat Politecnica de Catalunya
Barcelona, Spain
marta.ruiz@upc.edu

Christian Hardmeier
Uppsala Universitet
Sweden
christian.hardmeier@lingfil.uu.se

Will Radford
Canva
110 Kippax Street
Surry Hills, Australia
will.r@canva.com

Abstract

The 1st ACL workshop on Gender Bias in Natural Language Processing included a shared task on gendered ambiguous pronoun (GAP) resolution. This task was based on the coreference challenge defined in Webster et al. (2018), designed to benchmark the ability of systems to resolve pronouns in real-world contexts in a gender-fair way. 263 teams competed via a Kaggle competition, with the winning system achieving logloss of 0.13667 and near gender parity. We review the approaches of eleven systems with accepted description papers, noting their effective use of BERT (Devlin et al., 2019), both via fine-tuning and for feature extraction, as well as ensembling.

1 Introduction

Gender bias is one of the typologies of social bias (e.g. race, politics) that is alarming the Natural Language Processing (NLP) community. An illustration of the problematic behaviour are the recurrently appearing occupational stereotypes that *homemaker* is to *woman* as *programmer* is to *man* (Bolukbasi et al., 2016). Recent studies have aimed to detect, analyse and mitigate gender bias in different NLP tools and applications including word embeddings (Bolukbasi et al., 2016; Gonen and Goldberg, 2019), coreference resolution (Rudinger et al., 2018; Zhao et al., 2018), sentiment analysis (Park et al., 2018; Bhaskaran and Bhallamudi, 2019) and machine translation (Vanmassenhove et al., 2018; Font and Costa-jussà, 2019). One of the main sources of gender bias is believed to be societal artefacts in the data from which our algorithms learn. To address this, many have created gender-labelled and gender-balanced datasets (Rudinger et al., 2018; Zhao et al., 2018; Vanmassenhove et al., 2018).

We present the results of a shared task evaluation conducted at the 1st Workshop on Gender Bias in Natural Language Processing at the ACL 2019 conference. The shared task is based on the gender-balanced GAP coreference dataset (Webster et al., 2018) and allows us to test the hypothesis that *fair datasets would be enough to solve the gender bias challenge in* NLP.

The strong results of submitted systems tend to support this hypothesis and gives the community a great starting point for mitigating bias in models. Indeed, the enthusiastic participation we saw for this shared task has yielded systems which achieve near-human accuracy while achieving near gender-parity at 0.99, measured by the ratio between F1 scores on feminine and masculine examples. We are excited for future work extending this success to more languages, domains, and tasks. However, we especially note future work in algorithms which achieve fair outcomes given biased data, given the wealth of information from existing unbalanced datasets.

2 Task

The goal of our shared task was to encourage research in gender-fair models for NLP by providing a well-defined task that is known to be sensitive to gender bias and an evaluation procedure addressing this issue. We chose the GAP resolution task (Webster et al., 2018), which measures the ability of systems to resolve gendered pronoun reference from real-world contexts in a gender-fair way. Specifically, GAP asks systems to resolve a target personal pronoun to one of two names, or neither name. For instance, a perfect resolver would resolve that *she* refers to *Fujisawa* and not to *Mari Motohashi* in the Wikipedia excerpt:

(1) In May, *Fujisawa* joined *Mari Motohashi*'s rink as the team's skip, moving back from Karuizawa to Kitami where **she** had spent her junior days.

The original GAP challenge encourages fairness by balancing its datasets by the gender of the pronoun, as well as using disaggregated evaluation

Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing, pages 1–7
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

with separate scores for masculine and feminine examples. To simplify evaluation, we did not disaggregate evaluation for this shared task, but instead encouraged fairness by not releasing the balance of masculine to feminine examples in the final evaluation data.[1]

The competition was run on Kaggle[2], a well-known platform for competitive data science and machine learning projects with an active community of participants and support.

2.1 Setting

The original GAP challenge defines four evaluation settings, depending on whether the candidate systems have to identify potential antecedents or are given a fixed choice of antecedent candidates, and whether or not they have access to the entire Wikipedia page from which the example was extracted. Our task was run in *gold-two-mention* with *page-context*. This means that, for our task, systems had access to the two names being evaluated at inference time, so that the systems were not required to do mention detection and full coreference resolution. For each example, the systems had to consider whether the target pronoun was coreferent with the *first*, the *second* or *neither* of the two given antecedent candidates. A valid submission consisted of a probability estimate for each of these three cases. The systems were also given the source URL for the text snippet (a Wikipedia page), enabling unlimited access to context. This minimized the chance that systems could cheat, intentionally or inadvertently, by accessing information outside the task setting.

2.2 Data

To ensure blind evaluation, we sourced 760 new annotated examples for official evaluation[3] using the same techniques from the original GAP work (Webster et al., 2018), with three changes. To ensure the highest quality of annotations for this task, we (i) only accepted examples on which the three raters provided unanimous judgement, (ii) added heuristics to remove cases with errors in entity span labeling, and (iii) did an additional, manual round to remove assorted errors. The final set of 760 clean examples was dispersed in a larger set of 11,599 unlabeled examples to produce a set of 12,359 examples that competing systems had to rate. This augmentation was to discourage submissions based on manual labeling.

	logloss	F1	Bias
Attree (2019)	0.13667	96.2	0.99
Wang (2019)	0.17289	95.7	0.99
Abzaliev (2019)	0.18397	95.4	0.99

Table 1: Performance of prize-winning submissions on the blind Kaggle evaluation set. logloss was the official task metric, and correlates well with F1 score, which was used in the original GAP work.

We note many competing systems used the original GAP evaluation data[4] as training data for this task, given that the two have the same format, base domain (Wikipedia), and task definition.

2.3 Evaluation

The original GAP work defined two official evaluation metrics, F1 score and Bias, the ratio between the F1 scores on feminine and masculine examples. Bias takes a value of 1 at gender parity; a value below 1 indicates that masculine entities are resolved more accurately than feminine ones.

In contrast, the official evaluation metric of the competition was the logloss of the submitted probability estimates:

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log p_{ij}, \tag{1}$$

where N is the number of samples in the test set, $M = 3$ is the number of classes to be predicted, y_{ij} is 1 if observation i belongs to class j according to the gold-standard annotations and 0 otherwise, and p_{ij} is the probability estimated by the system that observation i belongs to class j.

Table 1 tabulates results based on the original and shared task metrics. Logloss and GAP F1 both place the winners in the same order.

2.4 Prizes

A total prize pool of USD 25,000 was provided by Google. The pool was broken down into prizes of USD 12,000, 8,000, and 5,000 for the top three systems, respectively. This attracted submissions

[1] We used 1:1 masculine to feminine examples.

[2] https://www.kaggle.com/c/gendered-pronoun-resolution

[3] Official evaluation ran in Stage 2, following an initial, development stage evaluated on the original GAP data, available at https://github.com/google-research-datasets/gap-coreference

[4] https://github.com/google-research-datasets/gap-coreference

from 263 teams, covering a wide diversity of geographic locations and affiliations, see Section 3.1. Table 1 lists results for the three prize-winning systems: Attree (2019), Wang (2019), and Abzaliev (2019).

3 Submissions

In this section, we describe the diverse set of teams who competed in the shared task, and the systems they designed for the GAP challenge. We note effective use of BERT (Devlin et al., 2019), both via fine-tuning and for feature extraction, and ensembling. Despite very little modeling targeted at debiasing for gender, the submitted systems narrowed the gender gap to near parity at 0.99, while achieving remarkably strong performance.

3.1 Teams

We accepted ten system description papers, from 11 of the 263 teams who competed (Ionita et al. (2019) is a combined submission from the teams placing 5 and 22). Table 2 characterises the teams by their number of members, whether their affiliation is to industry or an academic institution, and the geographic location of their affiliation. Details about participant gender were not collected.

Our first observation is that 7 of the top 10 teams submitted system descriptions, which allows us good insight into what approaches work well for the GAP task (see next, Section 3.2). Also, All these teams publicly release their code, promoting transparency and further development.

We note the geographic diversity of teams: there is at least one team from each of Africa, Asia, Europe, and USA, and one team collaborating across regions (Europe and USA). Five teams had industry affiliations and four academic; the geographically diverse team was diverse here also, comprising both academic and industry researchers.

There is a correlation between team size and affiliation: industry submissions were all from individual contributors, while academic researchers worked in groups. This correlation is somewhat indicative of performance: individual contributors from industry won all three monetary prizes, and only one academic group featured in the top ten submissions. A possible factor in this was the concurrent timing of the competition with other conference deadlines.

3.2 Systems

All system descriptions were from teams who used BERT (Devlin et al., 2019), a method to create context-sensitive word embeddings by pretraining a deep self-attention neural network on a training objective optimizing for cloze word prediction and recognition of adjacent sentences. This is perhaps not surprising, given the recent success of BERT for modeling a wide range of NLP tasks (Tenney et al., 2019; Kwiatkowski et al., 2019) and the small amount of training data available for GAP resolution (which makes LM pretraining particularly attractive). The different models built from BERT are summarized in Table 3.

Eight of the eleven system descriptions used BERT via fine-tuning, the technique recommended in Devlin et al. (2019). To do this, the original GAP data release was used as a tuning set to learn a classifier on top of BERT to predict whether the target pronoun referred to Name A, Name B, or Neither. Abzaliev (2019) also made use of the available datasets for coreference resolution: OntoNotes 5.0 (Pradhan and Xue, 2009), WinoBias (Zhao et al., 2018), WinoGender (Rudinger et al., 2018), and the Definite Pronoun Resolution Dataset (Rahman and Ng, 2012). Given the multiple BERT models available, it was possible to learn multiple such classifiers; teams marked *ensemble* fine-tuned multiple base BERT models and ensembled their predictions, while teams marked *single* produced just one, from a BERT-Large variant.

An alternative way to use BERT in NLP modeling is as a feature extractor. Teams using BERT in this capacity represented mention spans as input vectors to a neural structure (typically a linear structure, e.g. feed-forward network) that learned some sort of mention compatibility, via interaction or feature crossing. To derive mention-span representations from BERT subtoken encodings, Wang (2019) found that pooling using an attention-mediated process was more effective than simple mean-pooling; most teams pooled using AllenAI's SelfAttentionSpanExtractor[5]. An interesting finding was that certain BERT layers were more suitable for feature extraction than others (see Abzaliev (2019); Yang et al. (2019) for an exploration).

The winning solution (Attree, 2019) used a

[5] https://github.com/allenai/allennlp/blob/master/allennlp/modules/span_extractors/self_attentive_span_extractor.py

	Place	logloss	Members	Affiliation	Region
Attree (2019)	1	0.13667	1	Industry	USA
Wang (2019)	2	0.17289	1	Industry	Asia
Abzaliev (2019)	3	0.18397	1	Industry	Europe
Yang et al. (2019)	4	0.18498	4	Academic	Asia
Ionita et al. (2019)*	5	0.19189	1	Other	Africa
Liu (2019)	7	0.19473	1	Industry	USA
Chada (2019)	9	0.20238	1	Industry	USA
Bao and Qiao (2019)	14	0.20758	2	Academic	Europe
Ionita et al. (2019)*	22	0.22562	4	Mixed	Mixed
Lois et al. (2019)	46	0.30151	3	Academic	Europe
Xu and Yang (2019)	67	0.39479	2	Academic	USA

Table 2: Teams with accepted system description papers. *Note the two teams placing 5 and 22 submitted a combined system description paper.

	Rank	logloss	Fine-tuning	Feature Crossing	Resources
Attree (2019)	1	0.13667	single	–	syntax, coref, URL
Wang (2019)	2	0.17289	single	linear	–
Abzaliev (2019)	3	0.18397	ensemble	linear	synax, URL
Yang et al. (2019)	4	0.18498	ensemble	siamese	–
Ionita et al. (2019)*	5	0.19189	ensemble	linear	syntax, NER, coref
Liu (2019)	7	0.19473	–	linear	–
Chada (2019)	9	0.20238	ensemble	–	–
Bao and Qiao (2019)	14	0.20758	single	SVM & BIDAF	–
Ionita et al. (2019)*	22	0.22562	ensemble	linear	synax, NER, coref
Lois et al. (2019)	46	0.30151	–	–	–
Xu and Yang (2019)	67	0.39479	–	R-GCN	syntax

Table 3: Highlights of systems with accepted description papers. *Note the two teams placing 5 and 22 submitted a combined system description paper.

novel *evidence pooling* technique, which used the output of off-the-shelf coreference resolvers in a way that combines aspects of ensembling and feature crossing. This perhaps explains the system's impressive performance despite its relative simplicity. Two other systems stood out as novel in their approach to the task: Chada (2019) reformulated GAP reference resolution as a question answering task, and Lois et al. (2019) used BERT in a third way, directly applying the masked language modeling task to predicting resolutions.

Despite the scarcity of data for this challenge, there was little use of extra resources. Only two teams made use of the URL given in the example, with Attree (2019) using it only indirectly as part of a coreference heuristic fed into evidence pooling. Two teams augmented the GAP data by using name substitutions (Liu, 2019; Lois et al., 2019) and two automatically created extra examples of the minority label Neither (Attree, 2019; Bao and Qiao, 2019).

4 Discussion

Running the GAP shared task has taught us many valuable things about reference, gender, and BERT models. Based on these, we make recommendations for future work expanding from this shared task into different languages and domains.

GAP Given the incredibly strong performance of the submitted systems, it is tempting to ask whether GAP resolution is solved. We suggest the answer is no. Firstly, the shared task only tested one of the four original GAP settings. A more challenging setting would be *snippet-context*, in which use of Wikipedia is not allowed, which we would extend to LM pre-training. Also, GAP only targets particular types of pronoun usage, and the time is ripe for exploring others. We are particularly excited for future work in languages with different pronoun systems (esp. prodrop languages including Portuguese, Chinese, Japanese), and gender neutral personal pronouns, e.g. English *they*, Spanish *su* or Turkish *o*.

Gender It is encouraging to see submitted systems improve the gender gap so close to parity at 0.99, particularly as no special modeling strategies were required. Indeed, Abzaliev (2019) reported that a handcrafted pronoun gender feature had no impact. Moreover, Bao and Qiao (2019) report that BERT encodings show no significant gender bias on either WEAT (Caliskan et al., 2017) or SEAT (May et al., 2019). We look forward to studies considering potential biases in BERT across more tasks and dimensions of diversity.

BERT The teams competing in the shared task made effective use of BERT in at least three distinct methods: fine-tuning, feature extraction, and masked language modeling. Many system papers noted the incredible power of the model (see, e.g. Attree (2019) for a good analysis), particularly when compared to hand-crafted features (Abzaliev, 2019). We also believe the widespread use of BERT is related to the low rate of external data usage, as it is easier for most teams to reuse an existing model than to clean and integrate new data. As well as the phenomenal modeling power of BERT, one possible reason for this observation is that the public releases of BERT are trained on the same domain as the GAP examples, Wikipedia. Future work could benchmark non-Wikipedia BERT models on the shared task examples, or collect more GAP examples from different domains.

5 Conclusion

This paper describes the insights of shared task on GAP coreference resolution held as part of the 1st ACL workshop on Gender Bias in Natural Language Processing. The task drew a generous prize pool from Google and saw enthusiastic participation across a diverse set of researchers. Winning systems made effective use of BERT and ensembling, achieving near human accuracy and gender parity despite little efforts targeted at mitigating gender bias. We learned where the next research challenges in gender-fair pronoun resolution lie, as well as promising directions for testing the robustness of powerful language model pre-training methods, especially BERT.

Acknowledgements

We would like to extend very many thanks to the Kaggle team (especially Julia Elliot and Will Cukierski) and the Google Data Compute team (especially Daphne Luong and Ashwin Kakarla) who made this shared task possible.

This work is supported in part by the Spanish Ministerio de Economía y Competitividad and the European Regional Development Fund and the Agencia Estatal de Investigación, through the post-doctoral senior grant Ramón y Cajal and by the Swedish Research Council through grant 2017-930.

References

Artem Abzaliev. 2019. On GAP coreference resolution shared task: insights from the 3rd place solution. In *Proceedings of the First Workshop on Gender Bias for Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

Sandeep Attree. 2019. Gendered Pronoun Resolution Shared Task: Boosting Model Confidence by Evidence Pooling. In *Proceedings of the First Workshop on Gender Bias for Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

Xingce Bao and Qianqian Qiao. 2019. Transfer Learning from Pre-trained BERT for Pronoun Resolution. In *Proceedings of the First Workshop on Gender Bias for Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good Secretaries, Bad Truck Drivers? Occupational Gender Stereotypes in Sentiment Analysis. In *Proceedings of the First Workshop on Gender Bias for Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4356–4364, USA. Curran Associates Inc.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Rakesh Chada. 2019. Gendered Pronoun Resolution using BERT and an extractive question answering formulation. In *Proceedings of the First Workshop on Gender Bias for Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques. In *Proceedings of the First Workshop on Gender Bias for Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Matei Ionita, Yury Kashnitsky, Ken Krige, Vladimir Larin, and Atanas Atanasov. 2019. Gender-unbiased BERT-based Pronoun Resolution. In *Proceedings of the First Workshop on Gender Bias for Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics*.

Bo Liu. 2019. Anonymized BERT: An Augmentation Approach to the Gendered Pronoun Resolution Challenge. In *Proceedings of the First Workshop on Gender Bias for Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

Felipe Alfaro Lois, José A. R. Fonollosa, and Marta R. Costa-jussà. 2019. BERT Masked Language Modeling for Coreference Resolution. In *Proceedings of the First Workshop on Gender Bias for Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Sameer S. Pradhan and Nianwen Xue. 2009. OntoNotes: The 90% solution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 11–12, Boulder, Colorado. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Zili Wang. 2019. MSnet: A BERT-based Network for Gendered Pronoun Resolution. In *Proceedings of the First Workshop on Gender Bias for Natural Language Processing*, 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.

Yinchuan Xu and Junlin Yang. 2019. Look Again at the Syntax: Relational Graph Convolutional Network for Gendered Ambiguous Pronoun Resolution. In *Proceedings of the First Workshop on Gender Bias for Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

Kai-Chou Yang, Timothy Niven, and Hung-Yu Kao. 2019. Fill the GAP: Exploiting BERT for Pronoun Resolution. In *Proceedings of the First Workshop on Gender Bias for Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.