The Balancing Act in Gendered Pronoun Resolution

Navya Sandadi
UC Berkeley School of Information
navya.sandadi@berkeley.edu

Abstract

Several downstream tasks in NLP, such as question answering and machine translation depend on effective coreference resolution. However, most state-of-the-art coreference models exhibit gender bias due to being trained on gender imbalanced datasets such as OntoNotes and Definite Pronoun Resolution (DPR), which have more male pronouns than female pronouns, thus, producing models that perform better on male pronouns. The purpose of this paper is to build a model that performs well irrespective of the pronoun gender.

The approach is to use a balanced corpus of gender ambiguous pronouns called GAP, to solve coreference resolution on a gold-two-mention task. Pre-trained BERT and its variants are used for the task. The best performing model, achieved using attention mechanism with pooling, is evaluated against gender imbalanced datasets to discover that a gender balanced training dataset eliminates gender bias on the task. The model reproduces state-of-the-art results of 92.9% accuracy score benchmark on DPR dataset produced by Ye et al. (2020), but using a much simpler architecture and less training.

1. Introduction

Coreference resolution is the task of identifying all mentions within a text that refer to the same real-world entity. It is a hard problem in Natural Language Processing and was proposed as an alternative to the Turing test (Levesque, 2013).

GAP task formulates the coreference resolution as a classification problem, where the model must resolve a given pronoun to either of the two given candidates or neither.

Models developed for coreference resolution can be broadly categorized as (1) Mention pair classifier model, (2) Entity centric model, (3) Ranking model. However, most state-of-the-art coreference models are trained on gender imbalanced datasets and perform poorly on GAP dataset.

GAP corpus is seen as an opportunity to reduce gender bias in coreference datasets and thus, promote equitable modeling of reference phenomena. The dataset contains sentences which have references to entities as proper nouns and ambiguous gendered pronouns (Webster et al., 2018).

Recent work shows that pre-trained language representation models work well for coreference resolution (Lee et al., 2018). The pre-training task allows the language model to capture semantic and syntactic relationnships between sentences, useful for Natural Language Inference. Since the GAP corpus is relatively small with 2454 records, transformer based models: BERT, RoBERTa and CorefRoBERTa; and their extensive pre-training is leveraged for this task. A mention score classifier model architecture is used

with pooling and attention mechanism (described in Section 3). Selecting embeddings from specific BERT layers further boosts model performance.

This paper also explores the importance of training on a balanced dataset in order to eliminate gender bias by testing on highly unbalanced datasets. The model is evaluated on GAP test dataset filtered to contain only male and only female pronoun records.

Definite Pronoun Resolution dataset (Rahman et al., 2012) is chosen as another dataset to be used for testing model performance. The dataset's training data has only 27% of the examples representing the female gender.

Dealing with gender bias is important from both an ethical and practical standpoint, which is why this paper has set out to address gender bias through GAP dataset.

2. Background

Gender bias has been studied in word embeddings, coreference resolution and recently, in datasets. NLP systems containing bias in training data can produce gender biased predictions and sometimes even amplify biases present in the training sets (Zhao et al., 2017).

Extensive work has been done in the space of pronoun identification before the release of GAP dataset. Different versions of Winograd schemas have been released. They contain a pair of sentences that differ in only a word and the pronoun in each of the sentences corresponds to a different entity. These datasets are, however, highly imbalanced. It is of value to understand these prior works which led to the need to create the GAP dataset to solve pronoun resolution in an unbiased manner.

Winograd Schemas require world knowledge to resolve the pronoun. There have been a number of Winograd schema challenges and datasets of which the largest dataset, WinoBias, contains 3,160 examples created by Zhao et al. (2018). Although the Winograd Schema datasets are an excellent source of ambiguous pronouns, they do not generalize well and were very carefully curated.

In terms of general coreference datasets, OntoNotes is a good collection of simpler, high frequency coreference examples, but does not have many examples of ambiguous pronouns. In order to solve this, Ghaddar and Langlais (2016) released WikiCoref, a corpus of 30 articles annotated with coreference. The examples in the GAP dataset are similar to Winograd schemas in that they contain two person named entities of the same gender with an ambiguous pronoun which could potentially refer to either. They are different in that they have no reference-flipping word, but still represent a comparable challenge and require similar inferential power.

As our dataset is similar in many ways to the Winograd Schemas datasets, it is of value to evaluate our models against these imbalanced datasets to understand the importance of gender balancing in training data.

A notable score on a Winograd dataset emerged when Rahman and Ng (2012) scored 73.05% on their dataset, Definite Pronoun Resolution (DPR), through the usage of narrative chains, web-based counts, and selectional preferences. This was improved upon by Peng et al. (2015) to a 76.41% by implementing triplets of (subject, verb, object) and (subject/object, verb, verb). This score was further improved by Kocijan et al., (2019) by introducing a large dataset, WikiCrem, and using it in combination with

a language-model based approach to achieve an accuracy score of 84.8%. In October 2020, Deming Ye et al. presented CorefBERT and CorefRoBERTa models and benchmarked a score of 92.9% on the DPR dataset using CorefRoBERTa Large.

CorefBERT and CorefRoBERTa models adopt the deep bi-directional transformer architecture and utilize two tasks for training. (1) Mention Reference Prediction (2) Masked Language Modeling (same as in BERT). The mention reference prediction uses mention reference masking to mask one of the repeated mentions and then predicts the masked tokens. Since we use a mention score classifier architecture, this pre-trained model is a great fit to our task.

3. Model

The task is to classify whether the pronoun refers to entity A, entity B or NEITHER. Hence, the model aims to learn the probability distribution from the input text over the candidate reference entity.

The candidate reference set of the pronouns is {A, B, Neither}. We make use of an entity span comprised of the following:

```
a_span = [a_offset,a_offset+len(a),'a']
b_span = [b_offset,b_offset+len(b),'b']
p_span = [p_offset,p_offset+len(p),'p']
spans = [a_span, b_span, p_span]
```

where a_span, b_span, p_span refer to span of entity A, entity B and pronoun respectively.

Heavily inspired by the mention score network model of Zili Wang (2019), a vector called similarity vector is used to represent the triple-wise semantic similarity among the pronoun and the entities of a specific layer in BERT (or its variant) and then a feed forward neural architecture is used to compute the mention scores, given the distance between the pronoun and its candidate entities and the concatenated similarity vector. Pre-trained BERT uncased, RoBERTa and Coref-RoBERTa are used to build the model.

3.1 Baseline

The baseline uses token embeddings from the last hidden layer to predict the reference entity by freezing the hidden layers.

3.2 Pooling Approach

Since the entity spans consist of various tokens, the contextual representation is re-computed to maintain better correspondence. This is done using several pooling methods:

- Meanpooling
- Maxpooling
- Minpooling

The pooling is computed as the average, maximum or minimum of embeddings in the span corresponding to a particular layer.

3.3 Pooling with Attention Mechanism

An attention mechanism is used in which, instead of weighting each token equally, the attention mechanism is used to weight the tokens (Zili Wang, 2019). This method is different from commonly used attention functions in that it has no parameters and thus, is more space efficient.

The weights are learned automatically from the contextual similarity between pronoun and the token within the span.

4. Experiments

Several models were trained to evaluate the performance on the task. These models were

trained on Google Colab with GPU, using PyTorch as the deep learning framework. Due to the limited size of the dataset, 5-fold cross validation technique was adopted.

4.1 Datasets

4.1.1 GAP

The gender balanced GAP Coreference dataset, published by Google AI, is used for this task. The dataset has 4454 records and is split into development set (2000 records), validation set (454 records) and test set (2000 records). Table 1 shows the data distribution. The training, test and validation datasets are balanced. The GAP corpus is publicly available.

	MALE	FEMALE	TOTAL
GAP-DEVELOPMENT	1000	1000	2000
GAP-VALIDATION	227	227	454
GAP-TEST	1000	1000	2000
NEITHER A NOR B	129	124	253

Table 1: GAP corpus statistics

This dataset was generated using Wikipedia text and consists of two mention pronoun labels in each record. A single example contains a text passage, pronoun, entity A, entity B; location offsets of pronoun, entity A and entity B, the URL of the passage and a True/False label of whether the pronoun refers to A, B or Neither.

4.1.2 Definite Pronoun Resolution (DPR)

The DPR corpus (Rahman and Ng, 2012) is a collection of problems that resemble the Winograd Schema Challenge. The criteria for this dataset have been relaxed, and it contains examples that might not require common-sense reasoning or examples where the special word is actually a whole phrase. The dataset was constructed manually and consists of 564 training, 282 validation and 282 test samples.

4.2 Data Preprocessing

To ensure the token counts are less than 300 after tokenizing, the head or tail is truncated in a few examples based on where the target pronoun and entities reside.

4.3 Evaluation Method

We predict the probability that candidate A, candidate B or neither co-refers with the pronoun. The performance is evaluated using accuracy and multiclass logarithmic loss as defined below, where N is the number of examples and M=3, the number of classes (A, B, Neither).

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log\left(p_{ij}\right)$$

GAP has a particular focus on the balance of male and female pronouns and allows for gender-specific evaluation.

4.4 Dimension of Similarity Vector

As a similarity vector is used to represent the triple-wise semantic similarity among the pronoun and the entities, its dimension can influence the performance. A smaller dimension might lead to losing information while a bigger dimension could overfit and have generalization problems. The similarity dimension is treated as a hyperparameter.

4.5 Optimal Hidden Layers

Hidden layers in language models encode different linguistic knowledge. As pointed out by Tenney et al. (2019), basic syntactic

information appears earlier in the network, while high-level semantic information appears at higher layers.

Thus, instead of the common practice of taking the last hidden layer of BERT, intermediate hidden layers 16-20 were used based on the layer-wise metrics on BERT Large given by Tenney et al. (2019) for the task of coref-erence resolution. The top 8 layers, second to last hidden layer and combination of top & bottom layers were also used to experiment with hidden layers.

4.6 Hyperparameters

The models were trained using a dropout rate of 0.6 for BERT and 0.4 for RoBERTa models with Adam optimizer and a batch size of 32. Learning rate of 3e-4 gave the best performance with maximum epochs set to 30. Similarity vector dimension of 8 was best for BERT and dimension of 32 and 256 were best for RoBERTa models.

4.7 Testing on unbalanced dataset

To test the hypothesis that a balanced dataset eliminates gender bias, the best model is tested on highly unbalanced data.

The original GAP test data is filtered to subsets containing only male pronouns and only female pronouns. The performance of the strongest model is evaluated against each of these subsets. Tables 2 and 3 show the data distribution for pronouns by gender.

	TRAIN	VAL	TEST	TOTAL
SHE	396	87	428	911
HER	603	140	572	1315
HERS	1	0	0	1
TOTAL	1000	227	1000	-

Table 2: Female pronoun distribution

	TRAIN	VAL	TEST	TOTAL
HE	348	93	373	814
HIM	96	26	98	150
HIS	556	108	529	1193
TOTAL	1000	227	1000	-

Table 3: Male pronoun distribution

To further test the above hypothesis, the best performing model is also evaluated on DPR dataset.

5. Results & Analysis

Performance results on GAP and DPR datasets are shown in Table 4 and 5. The best performing model gives an accuracy score of 88.6% on GAP and 92.9% on DPR. Results on DPR dataset replicate the benchmark set by Deming Ye et al. (2020), but using RoBERTa model with a simple mention score architecture instead of the complex pre-training of CorefRoBERTa.

5.1 Detailed Analysis

The large versions of pretrained models consistently outperformed the base versions. This is expected as a more complex model with a greater number of layers can capture the contextual representations better.

An interesting observation was that when using log loss as the evaluation metric, BERT outperformed RoBERTa consistently. However, when using accuracy score as the metric, RoBERTa outperformed BERT. It is important to note that accuracy and cross-entropy loss measure two different things. Cross-entropy loss awards lower loss to predictions which are closer to the class label and is continuous whereas accuracy is a binary (true/false) measure for a particular sample. The official GAP challenge uses log

loss as the metric for measuring performance. However, since we care more about the model's performance in an interpretable way than optimizing the models, the models are primarily evaluated based on accuracy score.

As noted by Kocigan et al. (2019), the unlabeled test samples are known in advance to the models trained on GAP because BERT has been pre-trained on the entire English Wikipedia and has thus seen the text in the GAP dataset at pre-training time. This can explain the excellent performance of BERT based models on GAP dataset.

RoBERTa builds on BERT's language masking strategy and was trained with much larger mini-batches and on an order of magnitude more data than BERT, for a longer amount of time. Thus, RoBERTa generalizes and performs better on the task compared to BERT. CorefRoBERTa performs even better than RoBERTa as expected, because of its pre-training for mention reference prediction task, making it more suitable for coreference resolution.

Using these off-the-shelf transformer models, a pooling approach was employed to re-compute the span representation. Meanpooling performed the best among all pooling methods. Average or meanpooling method smooths out the representations and is better suited for coreference resolution instead of picking sharp features with max or minpooling.

Attention mechanism for entities A and B in combination with meanpooling for pronoun further improved the model performance.

Models trained on the top 8 hidden layers gave far better results than using only the top layer. Further probing revealed that layers 16 to 20 produced the best results, and this is in line with the work done by Tenney et al.

Last 8 layers, second to last hidden layer and combining top & bottom layers produced worse results compared to layers 16 to 20.

Similarity vector dimension varied with different models of BERT showing that the dimension of the similarity vector has a slight effect on the performance.

Model #	Pre-trained model	Hidden Layers	Similarity vector Dimension	Span	Accuracy	Log loss
1	BERT Base	Last layer	256	No Pooling	80.25	0.4838
2	RoBERTa Base	Last layer	256	No Pooling	76.85	0.5444
3	CorefRoBERTa Base	Last layer	256	No Pooling	80.6	0.488
4	BERT Large	Last layer	256	No Pooling	83.7	0.4513
5	RoBERTa Large	Last layer	256	No Pooling	79.7	0.521
6	CorefRoBERTa Large	Last layer	256	No Pooling	78.4	0.538
7	BERT Large	Last 8	8	Meanpooling	87.8	0.3428
8	RoBERTa Large	Last 8	16	Meanpooling	88.0	0.3561
9	CorefRoBERTa Large	Last 8	16	Meanpooling	87.6	0.3617
10	BERT Large	16 - 20	8	Meanpooling with Attention	88.1	0.3279
11	RoBERTa Large	16 - 20	32	Meanpooling with Attention	88.2	0.3352
12	CorefRoBERTa Large	16 - 20	32	Meanpooling with Attention	**88.6**	0.3327

Table 4: Performance of different models on GAP dataset

#	Train Set	Test Set	Model	Accuracy	Logloss
1	GAP	GAP	BERT	88.1	0.3279
			RoBERTa	88.2	0.3352
			CorefRoBERTa	**88.6**	0.3327
2	DPR	DPR	BERT	86.8	0.4214
			RoBERTa	**92.9**	0.3179
			CorefRoBERTa	91.8	0.3652
3	GAP	DPR (only gendered pronouns)	RoBERTa	**68.0**	0.831
			CorefRoBERTa	64.6	0.8544
4	DPR	GAP	RoBERTa	62.0	0.8977
			CorefRoBERTa	**62.9**	0.8747
5	GAP + DPR	GAP	RoBERTa	87.5	0.3505
			CorefRoBERTa	**87.8**	0.3478
6	GAP + DPR	DPR	RoBERTa	83.3	0.3421
			CorefRoBERTa	**83.6**	0.385
7	Balanced GAP	Only Male GAP	CorefRoBERTa	88.9	0.3179
		Only Female GAP	CorefRoBERTa	87.7	0.3494

Table 5: Accuracy score and multiclass logarithmic loss of the best performing models

5.2 Testing on Unbalanced Data

The best performing model (CorefRoBERTa using attention mechanism with pooling trained on hidden layers16-20) is trained using a similarity vector dimension of 32 and dropout of 0.4. This model is used to test the hypothesis that a strong model trained on a gender balanced dataset will not exhibit gender bias when tested on a gender unbalanced data (Table 5).

The models perform much better when using training and test set belonging to a single dataset (GAP / DPR) rather than being trained on one and tested on the other. This observation has also been made by Kocijan et al. (2019), that training on data from the target distribution improves the performance the most. Models trained on GAP training data usually show more than a 20% increase in their F1-score on GAP test data.

However, we do see from Table 5 that training on a balanced dataset and testing on an unbalanced dataset gives better results than the other way around. These results do come with certain limitations of the datasets. GAP dataset is much bigger than DPR and GAP contains only gendered pronouns whereas DPR contains gender neutral pronouns such as ‘it’, ‘they’, ‘them’.

Combining the GAP and DPR training sets gives better performance. The model performs better on GAP test set than DPR test set. This is probably because of 2454 records in GAP training data vs. only 848 records in DPR training data.

We do see more promising results when the model is trained on balanced GAP training set and tested on unbalanced GAP test set containing only male and only female pronouns.

6. Conclusion & Future Work

The work in this paper supports the claim that a gender balanced training dataset can

prevent the model from exhibiting gender bias on unseen data. Since this paper only experimented with GAP & DPR datasets, one area for future work would be to run similar experiments on other larger gender imbalanced datasets to further strengthen the claim.

7. References

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In EMNLP, pages 7170–7186.

Hector J. Levesque. 2013. On our best behaviour. In Proc. AAAI.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. In ACL, Volume 6, pages 605–617.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In EMNLP, pages 188–197.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The winograd schema challenge. In EMNLP, pages 777–789.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints. In EMNLP, pages 2979–2989.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In NAACL, Volume 2, pages 15-20.

Abbas Ghaddar, Phillippe Langlais. WikiCoref: An English Coreference-annotated Corpus of Wikipedia Articles. 2016. In LREC, pages 136-142.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015b. Solving hard coreference problems. In NAACL, pages 809-819.

Vid Kocijan, Oana-Maria Camburu, Ana-Maria Cretu, Yordan Yordanov, Phil Blunsom, Thomas Lukasiewicz. WikiCREM: A Large Unsupervised Corpus for Coreference Resolution. 2019. In EMNLP, pages 4303-4312.

Zili Wang. MSnet: A BERT-based Network for Gendered Pronoun Resolution. Proceedings of the First Workshop on Gender Bias in Natural Language Processing 2019, pages 89-95.

Ian Tenney, Dipanjan Das, Ellie Pavlick. BERT Rediscovers the Classical NLP Pipeline. 2019. In ACL, pages 4593-4610.