BERT fined-tuned on GAP dataset for Pronoun Resolution

Isabelle Bouchard[1*], **Carolyne Pelletier**[2] and **Pierre-Alexis Nault**[3]
{isabelle.bouchard, carolyne.pelletier, pierre-alexis.nault }@polymtl.ca

Abstract

This project was realised in the context of the INF8225 AI course. In this paper, we aim to reduce gender bias in pronoun resolution by creating a coreference resolver that performs well on a gender-balanced pronoun dataset, called the Gendered Ambiguous Pronouns (GAP) dataset. We leverage BERT's strong pre-training tasks on large unsupervised datasets and transfer these contextual representations to the fine-tuning stage.

We present an adaptation of the BERT fine-tuning task called SWAG to the coreference resolution problem. We used data from the **Gendered Pronoun Resolution** Kaggle competition. The aim is to train the BERT_{base} model to identify which person the pronoun refers to in the input sequence. We show that using BERT_{base} with a very simple fine-tuning architecture outperforms the state-of-the-art results by a large margin on the GAP score.

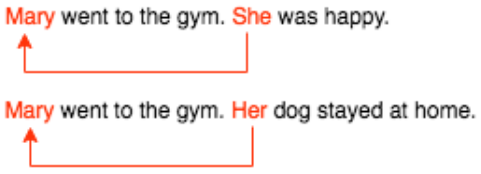

Figure 1: Pronoun resolution examples.

1 Introduction

Coreference resolution is a task aimed at pairing a phrase to its referring entity. Pronoun resolution is related to coreference resolution, but aims to pair pronouns with names as seen in Figure 1. Recent studies [4] [9], however suggest that state-of-the-art resolvers are gender biased, partly due to the biased data they have been trained on. For example, Zhao et al. reported that OntoNotes 5.0, which is a dataset used in the training of coreference systems, contains a gender imbalance. One such example of these imbalances are in the frequency of gendered mentions related to job titles: "Male gendered mentions are more than twice as likely to contain a job title as female mentions" [9]. When coreference resolution decisions are used to process text in automatic systems, any bias present in these decisions will be passed on to downstream applications. This has caused underrepresented groups to be treated unfairly by these biased downstream applications [2]. The Gendered Ambiguous Pronouns (GAP) [6] dataset was released by Google AI Language to improve gender-fairness in coreference resolvers.

In the context of our project, we use the GAP dataset to train an unbiased pronoun resolver. Our hypothesis is that given a gender-balanced dataset, we can utilize the learnings of a pretrained language understanding model to resolve the task of coreference resolution in an unbiased manner.

The main contribution of this paper is the adaptation of BERT to the pronoun resolution task, more specifically gender balanced pronoun resolution. We indeed show that we can reuse a potentially gender biased pretrained model, finetune it on an balanced dataset to correct the model on such bias and achieve state-of-the-art results on the $\text{GAP}_{\text{scorer}}$ metric.

We will first present the BERT model and show how it's been used in previous work. Then we will present the GAP dataset. Finally, we will show how we adapted the fine-tuning of BERT to this dataset and we will compare our results to SOTA results on this dataset.

Our code is available here.

2 Related Work

2.1. BERT model

BERT has been widely used since its release for several NLP tasks and has contributed significantly to improving the performance of these tasks. BERT is a language understanding model pretrained on a very large corpus of unlabelled data, namely Wikipedia (2.5G words) and BookCorpus (800M words). It learns to represent the language and can be fine-tuned to solve a specific task. In the paper introducing BERT [1], they have shown how BERT outperforms state-of-the-art models on many NLP tasks, most of them from the GLUE benchmark.

We have re-used the idea of the fine-tuning procedure on the SWAG dataset. SWAG is a *Large-Scale Adversarial*

*Contact Author

Dataset for Grounded Commonsense Inference [7]. The SWAG dataset has the format of a multiple choice test. Given the beginning of a sentence, the model tries to predict which ending fits better among the given choices. The example provided in the BERT paper [1] is the following:

Context	*A girl is going across a set of monkey bars. She*
Option 1	*jumps up across the monkey bars.*
Option 2	*struggles onto the bars to grab her head.*
Option 3	*jumps up and does a back flip.*

Table 1: SWAG finetuning task example for BERT

The fine-tuning procedure described in the BERT paper [1] for SWAG training involves a multiple choice task for end of sentence prediction. We will show how the pronoun resolution task of GAP has a similar structure and how we adapted the SWAG approach to design a procedure to fine-tune GAP.

2.2. GAP dataset

The GAP dataset is a gender-balanced labeled corpus of 8,908 ambigious pronoun-name pairs (ambiguous pronoun, antecedent name), sampled from Wikipedia [1]. This dataset has been specifically built for pronoun resolution, which is a type of the more broader coreference resolution task. In the dataset, two candidates are presented as possible reference for a given pronoun in a context paragraph or sentence. These two candidates compete to get the best matching score which leads to an adversarial training setup. The examples presented in the dataset are gender-balanced (fifty percent contain feminine pronouns, and fifty percent contain masculine pronouns), so that pronoun resolver models trained on this data will not favor one gender over the other. The hope is this will in turn reduce the bias present in current pronoun resolver models.

Table 2 shows a sample taken from the dataset. Given a context and a pronoun, the task is to predict the right candidate among two candidates (**Option 1** and **Option 2**) or neither of them (**Option 3**). In the following example, the pronoun of interest is "*His*" (last sentence) :

In this particular example, the correct answer would be *Dehner*.

2.3. BERT fine-tuning procedure for GAP dataset

We justify the usage of a pretrained language model such as BERT, that's been trained with no supervision on a very large corpus of data, to finetune a very small labelled GAP dataset (2K training examples only). We indeed leverage the rich language understanding of BERT and only need to finetune a very simple MLP at the output of BERT to get a prediction.

The similarity between GAP and SWAG task relies in the adversarial learning, where multiple choices compete to get the best score. However in GAP, in order to make a prediction, the model needs to know both the context and a given pronoun, while in SWAG, only the context is necessary.

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

Context	*Upon their acceptance into the Kontinental Hockey League,* ***Dehner*** *left Finland to sign a contract in Germany with EHC Mnchen of the DEL on June 18, 2014. After capturing the German championship with the Mnchen team in 2016, he left the club and was picked up by fellow DEL side EHC Wolfsburg in July 2016. Former NHLer Gary Suter and Olympic-medalist* ***Bob Suter*** *are* ***Dehner****'s uncles.* ***His*** *cousin is Minnesota Wild's alternate captain Ryan Suter.*
Pronoun	*His*
Option 1	*Bob Suter*
Option 2	*Dehner*
Option 3	Neither

Table 2: Co-reference resolution GAP dataset example

3 Experiments

3.1. Training procedure

As mentioned before, we adapted BERT to the GAP dataset following a similar learning procedure they used to finetune BERT for the SWAG dataset. We construct an input for each candidate A, B or neither. These inputs are the result of a concatenation of the context sentence, the pronoun and the candidate separated by the appropriate tokens. The task is very similar to SWAG, where multiple choices are competing to best match the beginning of the sentence. However, in our implementation, the elements passed after the separator token refer directly to words of the context sentence as seen in Table 3.

Input 1	[CLS] context sentence [SEP] pronoun [SEP] candidate A
Input 2	[CLS] context sentence [SEP] pronoun [SEP] candidate B
Input 3	[CLS] context sentence [SEP] pronoun [SEP] neither

Table 3: BERT inputs in co-reference resolution task

We use the embedding of the word "neither" directly and assume the network will learn to recognize it as the fallback option when candidate A and candidate B are not relevant.

We then introduce a task-specific MLP at the output of BERT, specifically at the aggregate representation C_i. The output of this MLP will be passed though a Softmax layer to

output the probability distribution over the three candidates (A, B, neither). Using the labels from the dataset, the neural network will be fine-tuned along with the BERT model.

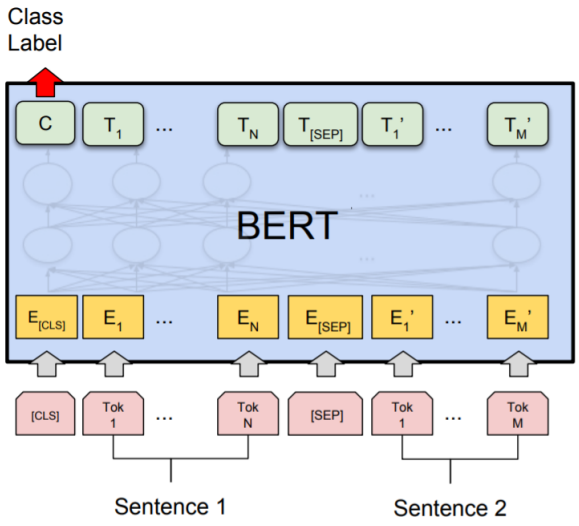

Figure 2: BERT fine-tuning setup [1].

Since the context sentences of GAP are very long (in average 430 words sentences in the training set), we used the maximum length of 512 for the input size to fine-tune the model. This constraint results directly from the BERT pre-trained model. Since the GAP dataset has long context sentences of over 512 words, some sentences were truncated. The presence of these truncated context sentences probably diminished the efficiency of the learning.

This also brought training resource difficulties. Indeed, we were constrained to use a batch size of 1 on 6 GPUs (Tesla P100) using BERT_{base} and were unable to train on BERT_{large} due to lack of resources.

We used bert-base-cased pretrained model since upper-cases are useful information for pronoun resolution. We used Adam to optimize with initial learning rate 5e-5. Early stopping with a patience of five epochs was used to fine-tune the model on GAP dataset. Figures 3 and 4 show respectively the evolution of the training loss and accuracy that resulted from the experiment .

3.2. Other experiments

We also experimented using only two choices as input (A and B) and removing the "neither" input choice. Our hypothesis is that the injection of this third input was not providing any additional information and therefore removing it should not hurt the model performance. The inputs took the form of Table 4.

We however had to adapt the two channels output of BERT to back propagate on the three choices target (A, B, neither). To do so, we added a 1D convolution with a kernel of size 1 with a ReLU activation to the output, before the MLP classifier, in order to convert our two channels output to three channels.

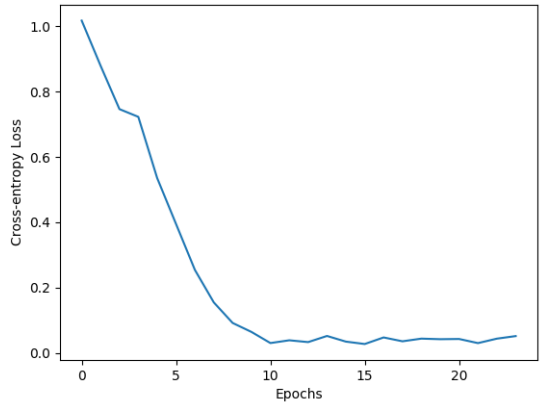

Figure 3: Training Loss

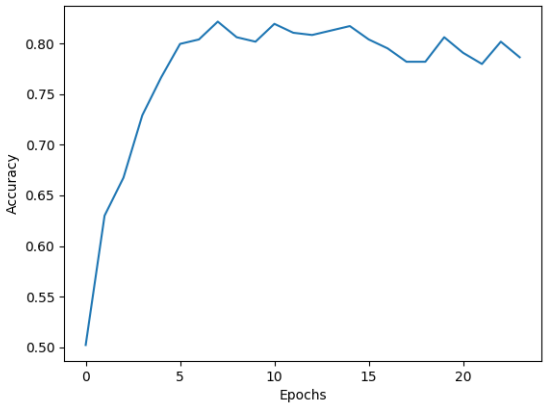

Figure 4: Validation Accuracy

Input 1	[CLS] context sentence [SEP] pronoun [SEP] candidate A
Input 2	[CLS] context sentence [SEP] pronoun [SEP] candidate B

Table 4: Alternative BERT inputs in co-reference resolution task

This would allows us to significantly reduce our memory usage (0.33 less) and we were hoping to be able to use BERT large with this input.

4 Results

We achieve an accuracy of 0.8215 on the test set for GAP pronoun resolution using the 3 input training method (table 5).

We also rank 229 on the Gender Pronoun Resolution Kaggle competition over 263 submissions (under the team name

	accuracy
our model (2 candidates input)	61.67
our model (3 candidates input)	**82.15**

Table 5: Comparison of best test accuracy for two and three candidates input

Secret). Submissions were evaluated using a multi-class log loss.

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

where N is the number of samples in the test set, M is 3 for the number of candidates (A, B, neither), yij is the target (0 or 1), and pij is the predicted probability that observation i belongs to class j.

	log loss
Best score	0.13667
our model	1.60924

Table 6: Gender Pronoun Resolution Kaggle competition best scores

This metric however does not measure the bias in the result, it simply assume that a model that performs well on balanced test set is de facto unbiased. We therefore also evaluate our model on the GAP scorer, which was introduced to give a measure of the bias of a resolver. It calculates the F1 score **O**verall as well as by the gender of the pronoun (**M**asculine and **F**eminine). The **B**ias is then calculated by taking the ratio of **F**eminine to **M**asculine F1 scores where a perfect score would be 1.

	M	F	B	O
Lee et al. (2017)	67.7	60.0	0.89	64.0
Parallelism	69.4	64.4	0.93	66.9
Parallelism+URL	72.3	68.8	0.95	70.6
our model	**87.9**	**85.1**	**0.97**	**86.5**

Table 7: Baselines on the GAP challenge test set and our model. The baselines are from Table 10 in the *Mind the GAP* paper [6]

Although BERT has been trained on a dataset that has been proved to be gender biased, namely Wikipedia and BookCorpus, the resolver we trained remains unbiased. Indeed, given a very small training set (2000 examples) on the fine-tuning task, we are able to get well balanced results. In table 7, we show that our model outperforms the model that's been presented in Mind the GAP on both F1 scores and Bias measure. Note that this paper was published before the introduction of BERT.

5 Future work

5.1 Train on BERT large

In this projet, even though we had access to great computational resources, we were only able to use BERT$_{large}$ in the reduced input scenario (4). We got worst results, due to the training format that was less efficient. However, if we had access to the necessary resources, we could run BERT$_{large}$ with the 3 inputs scenario and we would then expect better results. In this context, we could also vary the batch size, which has been clamped at 1 during the whole project. We expect a varying batch size to produced better performances.

5.2 Data augmentation

One of the biggest challenges in NLP is the shortage of training data. The labelled datasets are typically smaller because NLP is a diversified field with many distinct tasks. A way to increase a labelled dataset is by performing data augmentation. One approach for text data augmentation is to replace words with their synonyms selected from WordNet [8] or a word similarity calculation [5]. Contextual text data augmentation by words with paradigmatic relations can also be used [3]. Instead of using synonyms, Kobayashi [3] proposes to use words that are predicted by a Language Model given the context surrounding of the original words to be augmented. The hope is that by creating more training examples, the generalization error of our model will decrease.

5.3 Non-binary gender

Non-binary gender describes any gender identity which does not fit the male and female binary. The grammatical use of they/them as a singular pronoun could be used in the GAP dataset as a data augmentation technique that will allow for a more inclusive pronoun resolver.

6 Conclusion

Gender Biased coreference resolution decisions have the potential to be passed along to downstream tasks that rely on these decisions to operate. The GAP dataset was therefore created by Google AI Language to challenge the machine learning community to build robust resolvers that perform well on pronoun resolution no matter what the gender is. Using pre-trained BERT$_{base}$ fine-tuned in a SWAG-like manner on the GAP dataset, we report SOTA results on the baselines from the GAP challenge test set reported in Table 10 of the Mind the GAP paper [6]. These results speak to the strong transfer learning capabilities of BERT$_{base}$, allowing us to leverage a large amount of unlabelled text data in its pre-training phase to then better train a supervised model on a relatively minuscule labelled dataset.

References

[1] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[2] Moritz Hardt. “How big data is unfair - Understanding unintended sources of unfairness in data driven decision making”. In: *Medium* abs/1804.06876 (2014). URL: https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de.

[3] Sosuke Kobayashi. “Contextual Augmentation: Data Augmentation byWords with Paradigmatic Relations”. In: *Proceedings of NAACL* (2018). arXiv: 1805.06201. URL: https://arxiv.org/abs/1805.06201.

[4] Rachel Rudinger et al. “Gender Bias in Coreference Resolution”. In: *CoRR* abs/1804.09301 (2018). arXiv: 1804.09301. URL: http://arxiv.org/abs/1804.09301.

[5] William Yang Wang and Diyi Yang. “That’s So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using petpeeve Tweets”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2557-2563* (2015). URL: http://www.emnlp2015.org/proceedings/EMNLP/pdf/EMNLP306.pdf.

[6] Kellie Webster et al. “Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns”. In: *Transactions of the ACL*. 2018.

[7] Rowan Zellers et al. “SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference”. In: *arXiv e-prints*, arXiv:1808.05326 (Aug. 2018), arXiv:1808.05326. arXiv: 1808.05326 [cs.CL].

[8] Xiang Zhang, Junbo Zhao, and Yann LeCun. “Character-level Convolutional Networks for Text Classification”. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS’15. Montreal, Canada: MIT Press, 2015, pp. 649–657. URL: http://dl.acm.org/citation.cfm?id=2969239.2969312.

[9] Jieyu Zhao et al. “Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods”. In: *Proceedings of NAACL* abs/1804.06876 (2018). arXiv: 1804.06876. URL: https://arxiv.org/pdf/1804.06876.pdf.