# ”What does it refer to?” Casting Reference Resolution as Question Answering

**Marco Concetto Rudilosso** zcabmcr@ucl.ac.uk
**Reemma Muthal Puredath** zcabrmu@ucl.ac.uk
**Sadir Abdul Hadi** zcababd@ucl.ac.uk

## Abstract

Most approaches to coreference resolution which are present in the current literature do not make extensive use of large pre-trained language models representations. We present two new such approaches, by casting coreference resolution as question answering and by using a pre-trained BERT model fine-tuned on the GAP dataset. This dataset was created in response to presence of gender bias in previous datasets, thus it is a gender balanced coreference resolution dataset. We achieve competitive results in this dataset under 2 different variations of the task it proposes. Our best result is 84.7% accuracy.

## 1 Introduction

Pronoun-noun disambiguation has been a long-standing challenge to Natural Language Processing and it was since established that coreference resolution presented a potential approach to this problem. Coreference resolution requires finding all the entities/expressions that refer to a common entity in a given piece of text.

Different studies have highlighted how the existing coreference resolution datasets are plagued by a strong gender bias, skewed by a majority of the data points being about male entities (Rudinger et al., 2018) (Zhao et al., 2018). Furthermore, it has been shown that Wikipedia itself is affected by an inherent gender bias caused by how people write differently about men and women (Graells-Garrido et al., 2015). Due to these reasons, models that are trained over these corpora are inherently doomed to exhibit gender bias, thus limiting their usefulness in downstream tasks.

In order to mitigate such issues, we build coreference resolution models which we train on GAP (Webster et al., 2018), a gendered-balanced labelled corpus. This dataset contains a relatively limited amount of data to learn from, thus we focus our attention on large pre-trained language models, which have been shown to learn a variety of linguistic phenomena from unsupervised learning (Tenney et al., 2019) (Radford et al., 2019). We choose to use a pre-trained BERT model, which we fine-tune on the GAP dataset, as it has been a well performing model in the recent literature.

Furthermore, we cast the task of coreference resolution as question answering, following the work done by Weissenborn et al. (2018) and McCann et al. (2018) which have casted a variety of NLP tasks as QA.

We build two models which cast coreference resolution into one type of question answering each:

- Extractive Question Answering (EQA) (Section 4.2.) The model is not made aware of the two potential names.
- Multiple Choice Question Answering (MCQA) The model is given the reference to two candidate names to which the pronoun might refer (Section 4.3). We call this task the gold-two-mention task.

While we have observed good performance by both models, our EQA has performed particularly well. It has scored higher than Transformer Models for Coreference Resolution and all the baselines based on traditional cues for coreference which were used by Webster et al. (2018).

Both our models exhibit better sample efficiency than the model proposed by Lee et al. (2017a), which achieves poor performance under such a limited amount of data to train on.

## 2 Related Work

### 2.1 Coreference Resolution Models

Lee et al. (2017a) recognised that existing corefer-

ence models don't generalize well when new languages are introduced, because those models make use of syntactic parser. A new model was therefore introduced, and was trained to learn the best way to cluster mentions of entity. It makes use of a model which decides for each span of entity mention, whether any of the existing spans logically precede it.

Webster et al. (2018) introduces a Transformer-based neural model for coreference resolution. The model works by fragmenting the input text into tokens and creates two self attention matrices based on the source and target sentence. From this, we are able to extract the attention values. The main difference in these models depend on how the attention is spread between entities which are used repeatedly.

The End-to-End coreference resolution model proved to perform better using GAP than entity-centric coreference resolution with model stacking (Clark and Manning, 2015), the latter performing better with OntoNotes (Pradhan et al., 2007) test set.

More recently, contextual embeddings which are pre-trained on a large amount of unlabelled text have been shown to be able to accomplish large set of tasks well. Tenney et al. (2019) highlights how these representations contain hints about a variety of linguistic phenomena, including coreference resolution. Moreover, large language models have been shown to perform relatively well in a zero-shot learning scenario on a variety of tasks (Radford et al., 2019). This suggests that we could improve over the previous results by using one of such models.

### 2.2 Casting Problems as Question Answering

McCann et al. (2018) have introduced the Natural Language Decathlon (decaNLP), a challenge that consists in building a single model that solves ten tasks, including Machine Translation, Summarization, and Pronoun Resolution. They cast all these tasks into Question Answering, allowing their model to effectively multitask and generalize, and to perform transfer learning and meta-learning.

The same idea can be seen in the Jack the Reader framework for machine reading (Weissenborn et al., 2018), where Natural Language Inference and Link Prediction have been cast to Question Answering tasks.

We think that we can extend the number of tasks casted as QA by adding coreference resolution to this list.

## 3 Background

### 3.1 GAP dataset

The dataset used in this project is the GAP dataset compiled by Webster et al. (2018). This was created to address the predominant issue with many coreference resolution models that are trained on gender biased datasets that skew towards male. Research into this issue by Rahman and Ng (2012), Webster et al. (2018) has shown that resolving Ambiguous Gendered Pronouns has made very little progress towards being addressed or solved. An example of what Ambiguous Gendered Pronoun Resolution aims to achieve is shown below, where the pronoun in question is 'He':

(1) President Michel Aoun shook hands with Prime Minister Saad Hariri at the Parliament. He wanted to discuss the recent situation in Lebanon.

In order to compile the dataset, the human annotators were presented the job of manually performing coreference resolution. They had to decide whether each of the sentences fit into the following classes:

1. Name A, when the pronoun refers to the first name that the extraction process had found
2. Name B, when the pronoun refers to the second name that the extraction process had found
3. Neither, when the pronoun does not refer to the first name (Name A), nor the second name (Name B)
4. Both, when the pronoun refers to both name A and name B
5. Not Sure, when the annotator was not sure whether the pronoun belongs to Name A or Name B

In order to filter the results to focus on the pronouns that could actually be resolved, the examples where the annotators marked the case as 'Both' or 'Not sure' were removed from the final

version of the dataset. In order ensure gender balance, the authors made sure that there remained an equal counter of references to male and female entities.

The models which aim to solve this dataset can have a degree of access to the labelled data. In the following work we test 2 such ways:

1. The model has access to Name A and Name B and their position in the sentence
2. The model only has access only to the Pronoun

The dataset in focus is a corpus of Gendered Ambiguous Pronouns, containing 8,908 pronoun-pairs extracted from text found in Wikipedia (Webster et al., 2018). GAP is a good choice when it comes to a dataset to train a model on, as it is not affected by gender bias as others such as OntoNotes (Pradhan et al., 2007).

### 3.2 BERT

Devlin et al. (2018) presented BERT which is a unique pre-trained context embeddings model. BERT overcomes the challenges faced by unidirectionality of other language models by pre-training the models from right to left and left to right.

With BERT's introduction of bidirectionality, the pre-training step considers the context in both sides of the token such that when a random token is masked, the model is capable of utilizing the context to predict the missing token. This overcoming of limitations imposed by unidirectionality means that BERT is a better model for fine-tuning tasks at a token-level, as well as sentence-level.

BERT has famously superseded state-of-art for a range of NLP tasks from improving GLUE benchmark by 7.6% (Wang et al., 2018), MultiNLI by 5.6% (Devlin et al., 2018) and SQuAD v1.1 by 1.5% (Rajpurkar et al., 2016).

## 4 Methods

In the following section we describe how we can fine-tune a BERT model on the GAP dataset (Webster et al., 2018), while also casting coreference resolution as question answering (QA).

### 4.1 Casting Coreference Resolution as QA

As we have previously mentioned the GAP dataset can be solved with a variable degree of access to label data. When we have access to what Name A and Name B are, we cast this task as multiple choice question answering (MCQA), where the model has to choose between 3 choices: "Name A", "Name B", "Neither".

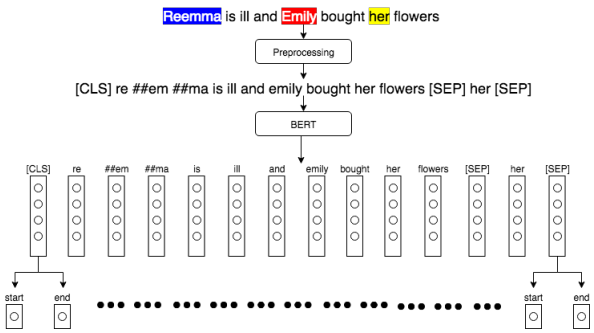


Figure 1: How the BERT EQA model process a datapoint. Name A is shown in blue, Name B in red and the pronoun in yellow.

Conversely, when the model does only have access to the pronoun, we cast coreference resolution as extractive question answering (EQA), where the question is the pronoun.

### 4.2 Using BERT for EQA coreference resolution

In order to adapt BERT for our task we have first converted the GAP data to follow the format used by the SQuAD 2.0 dataset (Rajpurkar et al., 2018). To do so, we use the pronoun as the question, the text as the context used by the model to answer the question and the correct name as the answer. Furthermore, when the correct answer is 'Neither', we mark the sentence as an unanswerable question. Having transformed the GAP data to the SQuAD 2.0 format, we train a pre-trained BERT model to do extractive question answering as described in Devlin et al. (2018). The model will take the sentence and question and the produce the score for each token in the context for being the start and the end of the answer, as shows in Figure 1.

At test time we have to modify the normal SQuAD evaluation pipeline, as the labels we are interested in are the probabilities of the answer being: "Name A", "Name B" or "Neither". In order to do so we use the start/end of answer scores for

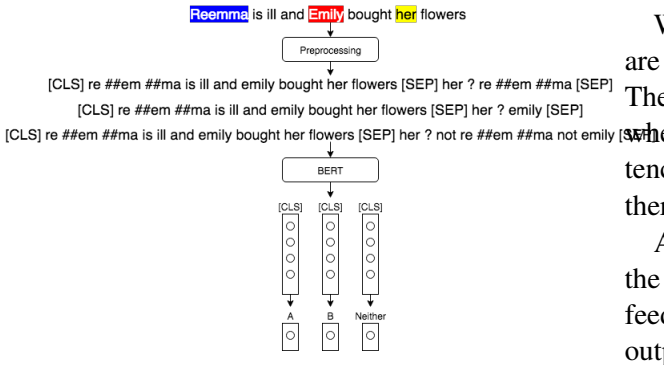


Figure 2: How the BERT MCQA model process a datapoint. Name A is shown in blue, Name B in red and the pronoun in yellow.

each token to calculate the following scores:

$$A\ Score = start^A + end^A \quad (1)$$

$$B\ Score = start^B + end^B \quad (2)$$

$$Neither\ Score = start^{[CLS]} + end^{[CLS]} + \frac{1}{N} \sum_{i \not\subset A\ \&\ i \not\subset B}^{N} start^i + end^i \quad (3)$$

The score of A is calculated by adding the start score of the first token of A and the end score of the last token of A, similarly we calculate the score of B, as shown in Equation 1 and 2. Instead, in order to calculate the score for Neither, we add the start and end score for the [CLS] token, which is used to classify unanswerable questions (the higher the harder the question is to answer), and we also add the mean score for the start and end score for each token which is not contained in Name A nor Name B, as shown in Equation 3.

Finally, we try to give the model a stronger signal as to which pronoun in the phrase it should look at by masking the pronoun in the context and only having it appear in the question.

### 4.3 Using BERT for MCQA coreference resolution

In order to cast coreference resolution as multiple choice question answering using BERT we first take a single datapoint and produce 3 sentences as follows:

1. [CLS] context [SEP] ”pronoun” ? ”Name A”
2. [CLS] context [SEP] ”pronoun” ? Name B”
3. [CLS] context [SEP] ”pronoun” ? not ”Name A” not ”Name B”

Where the context is the sentence which we are trying to perform coreference resolution on. The first sequence should be chosen by our model when the right label is ”Name A”, the second sentence when it’s ”Name B” and the third for ”Neither”.

After this, we feed each sequence to BERT, get the vector for the [CLS] token and pass it through a feed forward layer, independently. This last layer outputs a single scalar. During training we take each of the scores for the sentences, concatenate them and pass the resulting vector through a softmax layer and we optimize the model using cross-entropy loss.

During testing, we similarly get the scores for each of the sentences and choose the one with the highest score as the prediction of our model.

## 5 Experiments

We use the original implementation of BERT[1], which uses tensorflow and we train use a single TPUv2, which are freely available on Colab at the time of writing. We use the pre-trained weights the the original authors of BERT have released.

For our EQA BERT we use the default hyperparameters used by the BERT model used for the SQuAD 2.0 (Rajpurkar et al., 2018) model described by (Devlin et al., 2018), except we modify the learning rate to $1e-5$ and the number of training epochs to $4$.

Similarly, for our MCQA model, we use the default hyperparameters used for text classification, except we modify the learning rate to be $5e-6$, the number of epochs to be $5$ and the training batch size to be $16$.

### 5.1 Baselines

We compare our models against 2 baselines, both of which have access to the gold mentions. One is the End-to-end neural coreference resolution model proposed by Lee et al. (2017b) and trained on the GAP dataset. We have used the openly available Kaggle kernel which implements this model in our experiments[2].

The other baseline is inspired by the work done by Tenney et al. (2019), in which they use just the contextual embeddings of 2 words to predict if one references the other. Similarly, we this baseline

[1] https://github.com/google-research/bert

[2] https://www.kaggle.com/keyit92/end2end-coref-resolution-by-attention-rnn

takes the contextual embeddings of "Name A", "Name B" and the pronoun and pass them to a feed forward layer which does 3-way classification. We use the openly available Kaggle kernel which implements this baseline[3]

## 6 Results and Discussion

We summarize the results on the GAP dataset in Table 1. It is worth highlighting how the BERT baseline which uses the fixed contextual embeddings for Name A, Name B and the pronoun is the best in the class, thus showing how much is learned from unsupervised training and the importance of using large pre-trained models.

To further indicate the importance of pre-training, the results achieved by the baseline which implements the model presented in (Lee et al., 2017a) show the difficulty of learning from a small datasets as the GAP one, which only contains 2000 datapoints, where having good sample efficiency is crucial.

Our 2 models perform well, with EQA achieving the best results when compared to other models which do not have access to Name A and Name B. MCQA has space for improvements as shown by the fixed BERT embeddings baseline.

### 6.1 Performance on Neither

From our analysis of the results both our model suffer particularly when it comes to predict correctly the Neither label, as shown in Table 2.

| Model | Neither Accuracy |
|---|---|
| EQA BERT | 52.42 |
| MCQA BERT | 69.60 |

Table 2: Neither accuracies for EQA and MCQA BERT

This could be caused by the fact that the dataset contains only a small amount of datapoints whose label is Neither. We report the number of datapoints for each label in Table 3.

| | # sentences |
|---|---|
| A | 874 |
| B | 925 |
| Neither | 201 |

Table 3: Number of datapoints for each category

[3]https://www.kaggle.com/mateiionita/taming-the-bert-a-baseline

In order to better understand what kind of challenge the Neither label poses, in Table 4 we show a small sample of them which highlights 2 possible situations when a datapoint can be labelled as 'Neither':

1. The pronoun refers to an entity outside the sentence (Sentence 1)
2. The pronoun refers to an entity inside the sentence, but it does not refer to Name A or Name B (Sentences 2 and 3)

| | |
|---|---|
| 1 | She was born in Paris, the non-marital daughter of the Mexican artist Diego Rivera and his mistress, the Russian-born painter Marie Vorobieff-Stebelska ("Marevna"). **Rivera**, who was married to **Angelina Beloff** at the time, did not accept his daughter. So **she** grew up under the care of her mother. |
| 2 | On June 13 she hosted with Raul Gracia, Felipe Viel, **Elizabeth Lopez** and **Carolina La O**, Bienvenido el Mundial (Welcome to the World Cup), welcoming the FIFA World Cup. On June 28 in North Carolina, Sandoval received an award for being the boldest host. On July 12, **she** was a host on the memorial special for Michael Jackson. |
| 3 | The 2017 Daytona 500 was first time the "stage format" was used in the Daytona 500; it was set up in three stages similar to the all star race (but without the requirement to pit), where drivers are given an chance to pit if they wish to do so. Kyle Busch won the first stage earning **him** ten points in the new points format, **Kevin Harvick** won the second stage and **Kurt Busch** won the final stage in a dramatic last lap pass (passing Kyle Larson who ran out of gas) to win the race. |

Table 4: Interesting sentences from the Neither category which were sampled from the dataset. Name A is in bold and red, Name B is in bold and blue and the pronoun that we need to link is in bold and violet.

### 6.2 Ablation on masking the pronoun for EQA

We hypothesized that by masking the pronoun in the context for our EQA BERT model we could have better results, as it would make less ambiguous which pronoun the model should resolve the

| Model | Accuracy | Recall | F1 | Bias |
|---|---|---|---|---|
| Models that have access to Name A and Name B | | | | |
| BERT Embeddings + MLP | 95.4 | 96.4 | 95.9 | 1.01 |
| BERT MCQA | 84.7 | 84.7 | 84.7 | 0.98 |
| End-to-end (Lee et al., 2017b) | 42.5 | 46.5 | 44.4 | 0.94 |
| Multi-Transformer (Webster et al., 2018) | - | - | 62.3 | 0.98 |
| Models that don't have access to Name A and Name B | | | | |
| BERT EQA | 82.2 | 85.2 | 83.7 | 0.99 |
| Multi-Transformer (Webster et al., 2018) | - | - | 56.2 | 0.89 |

Table 1: Results on GAP dataset

coreference for. In order to prove this hypothesis, we train 2 models with the same hyper-parameters, one with the pronouns masked and the other without. Table 5 shows the result of this ablation study.

| Model | Accuracy |
|---|---|
| EQA BERT with masking | 81.3 |
| EQA BERT without masking | 82.2 |

Table 5: Accuracy for BERT EQA with and without pronoun masking

Given these results, it is unclear if masking the pronouns helps at all, thus our chosen EQA model does not use it.

## 7 Future Work

There are several extensions to this project that could be followed in the future. We could begin by exploring how other pre-trained models such as Elmo (Peters et al., 2018) and GPT-2 (Radford et al., 2019) perform with the GAP dataset; like BERT, these models also produce context embeddings for each token in a document.

Whilst GAP is one of the few options for a gender-balanced dataset, it is quite small in size as it only contains 2000 datapoints in the training set. Therefore, for further experiments, we could train on a larger dataset in addition to GAP.

Due to this, we could improve on EQA BERT, by giving it a better learning signal for the label Neither, which seem to be the hardest to learn, as shown in our analysis. We could do so by penalizing the model if it predicts spans in Name A or Name B when the label is Neither, which could help the model better understand what the Neither label means.

## 8 Conclusion

In this project, we have experimented with a means of reducing gender bias in coreference resolution, predominantly through the use of GAP, a gender-balanced dataset. We cast the task of coreference resolution as a question answering task and compared the accuracy of the model against the current state-of-the-art. We created two versions of the BERT models that had variable access to the GAP dataset; BERT MCQA had access to both names and BERT EQA which has more limited access. Overall, the our models perform better than some existing models such as end-to-end, however, it underperforms in comparison to other BERT model variations.

## References


Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 165–174. ACM.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017a. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017b. End-to-end neural coreference resolution. *CoRR*, abs/1707.07045.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

S. S. Pradhan, E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2007. Ontonotes: A unified relational semantic representation. In *International Conference on Semantic Computing (ICSC 2007)*, pages 517–526.

Sameer S Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in ontonotes. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 446–453. IEEE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

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. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *arXiv preprint arXiv:1810.05201*.

Dirk Weissenborn, Pasquale Minervini, Tim Dettmers, Isabelle Augenstein, Johannes Welbl, Tim Rocktäschel, Matko Bosnjak, Jeff Mitchell, Thomas Demeester, Pontus Stenetorp, and Sebastian Riedel. 2018. Jack the reader - A machine reading framework. *CoRR*, abs/1806.08727.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.