

Resolving Gendered Ambiguous Pronouns with BERT

Matei Ionita

University of Pennsylvania
matei@sas.upenn.edu

Yury Kashnitsky

Moscow Institute of Physics and Technology
yury.kashnitsky@phystech.edu

Ken Krige

Kitsong
kenkrige@gmail.com

Vladimir Larin

PJSC Sberbank
vlarine@gmail.com

Denis Logvinenko

Jet Infosystems
dennis_l@rambler.ru

Atanas Atanasov

Sofia University
a.atanasov1@ocado.com

Abstract

Pronoun resolution is part of coreference resolution, the task of pairing an expression to its referring entity. This is an important task for natural language understanding and a necessary component of machine translation systems, chat bots and assistants. Neural machine learning systems perform far from ideally in this task, reaching as low as 73% F1 scores on modern benchmark datasets. Moreover, they tend to perform better for masculine pronouns than for feminine ones. Thus, the problem is both challenging and important for NLP researchers and practitioners. In this project, we describe our BERT-based approach to solving the problem of gender-balanced pronoun resolution. We are able to reach 92% F1 score and a much lower gender bias on the benchmark dataset shared by Google AI Language team.

1 Introduction

In this work, we are dealing with gender bias in pronoun resolution. A more general task of coreference resolution is reviewed in Sec. 2. In Sec. 3, we give an overview of a related Kaggle competition. Then, Sec. 4 describes the GAP dataset and Google AI's heuristics to resolve pronominal coreference in a gender-agnostic way, so that pronoun resolution is done equally well in cases of masculine and feminine pronouns. In Sec. 5, we provide the details of our BERT-based solution while in Sec. 6 we analyze pleasantly low gender

bias specific for our system (our code is shared on GitHub¹). Lastly, in Sec. 7, we draw conclusions and express some ideas for further research.

2 Related work

Among popular approaches to coreference resolution are:² rule-based, mention pair, mention ranking, and clustering. As for rule-based approaches, they describe naïve Hobbs algorithm (Hobbs, 1986) which, in spite of being naïve, has shown state-of-the-art performance on the OntoNotes dataset³ up to 2010.

Recent state-of-the-art approaches (Lee et al., 2018, 2017; Peters et al., 2018a) are pretty complex examples of mention ranking systems. The 2017 version is the first end-to-end coreference resolution model that didn't utilize syntactic parsers or hand-engineered mention detectors. Instead, it used LSTMs and attention mechanism to improve over previous NN-based solutions.

Some more state-of-the-art coreference resolution systems are reviewed in (Webster et al., 2018) as well as popular datasets with ambiguous pronouns: Winograd schemas (Levesque et al., 2012), WikiCoref (Ghaddar and Langlais, 2016), and

¹<https://github.com/Yorko/gender-unbiased-BERT-based-pronoun-resolution>

²<https://bit.ly/2JbKxv1>

³<https://catalog.ldc.upenn.edu/LDC2013T19>

The Definite Pronoun Resolution Dataset (Pradhan et al., 2007). We also refer to the GAP paper for a brief review of gender bias in machine learning.

We further outline that e2e-coref model (Lee et al., 2018), in spite of being state-of-the-art in coreference resolution, didn't show good results in the pronoun resolution task that we tackled, so we only used e2e-coref predictions as an additional feature.

3 Kaggle competition “Gendered Pronoun Resolution”

Following Kaggle competition “Gendered Pronoun Resolution”,⁴ for each abstract from Wikipedia pages we are given a pronoun, and we try to predict the right coreference for it, i.e. to which named entity (A or B) it refers. Let's take a look at this simple example:

“John entered the room and saw [A] Julia. [Pronoun] She was talking to [B] Mary Hendriks and looked so extremely gorgeous that John was stunned and couldn't say a word.”

Here “Julia” is marked as entity A, “Mary Hendriks” – as entity B, and pronoun “She” is marked as Pronoun. In this particular case the task is to correctly identify to which entity the given pronoun refers.

If we feed this sentence into a coreference resolution system (see Fig. 1 and online demo⁵), we see that it correctly identifies that “she” refers to Julia, it also correctly clusters together two mentions of “John” and detects that Mary Hendriks is a two-word span.

For instance, if you take an abstract like this it's pretty hard to resolve coreference.

“Roxanne, a poet who now lives in France. Isabel believes that she is there to help Roxanne during her pregnancy with her toddler infant, but later realizes that her father and step-mother sent her there so that Roxanne would help the shiftless Isabel gain some direction in life. Shortly after she (pronoun) arrives, Roxanne confides in Isabel that her French husband, Claude-Henri has left her.”

Google AI and Kaggle (organizers of this competition) provided the GAP dataset (Webster et al., 2018) with 4454 snippets from Wikipedia articles, in each of them named entities A and B are labeled

along with a pronoun. The dataset is labeled, i.e. for each sentence a correct coreference is specified, one of three mutually-exclusive classes: either A or B or “Neither”. Thus, the prediction task is actually that of multiclass classification type.

Moreover, the dataset is balanced w.r.t. masculine and feminine pronouns. Thus, the competition was supposed to address the problem of building a coreference resolution system which is not susceptible to gender bias, i.e. works equally well for masculine and feminine pronouns.

These are the columns provided in the dataset (Webster et al., 2018):

- ID - Unique identifier for an example (matches to Id in output file format)
- Text - Text containing the ambiguous pronoun and two candidate names (about a paragraph in length)
- Pronoun - target pronoun (text)
- Pronoun-offset - character offset of Pronoun in Text
- A - first name candidate (text)
- A-offset - character offset of name A in Text
- B - second name candidate
- B-offset - character offset of name B in Text
- URL - URL of the source Wikipedia page for the example

Evaluation metric chosen for the competition⁶ is multiclass logarithmic loss. Each pronoun has been labeled with whether it refers to A, B, or “Neither”. For each pronoun, a set of predicted probabilities (one for each class) is submitted. The formula is then

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

where N is the number of samples in the test set, M is 3, \log is the natural logarithm, y_{ij} is 1 if observation i belongs to class j and 0 otherwise, and p_{ij} is the predicted probability that observation i belongs to class j .

⁴<https://www.kaggle.com/c/gendered-pronoun-resolution>

⁵<https://bit.ly/2I4tECI>

⁶<https://www.kaggle.com/c/gendered-pronoun-resolution/overview/evaluation>

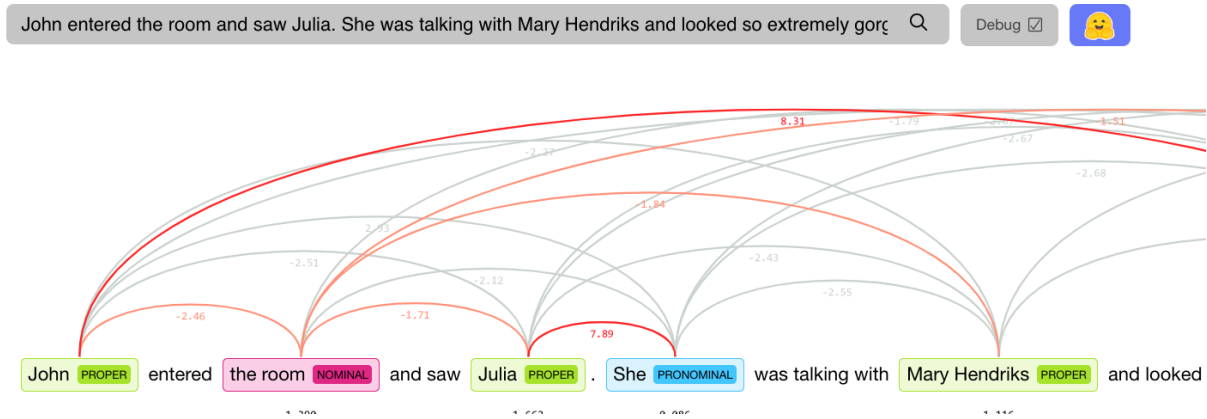


Figure 1: Coreference resolution visualized with HuggingFace demo <https://huggingface.co/coref/>.

Unfortunately, the chosen evaluation metric does not reflect the mentioned above goal of building a gender-unbiased coreference resolution algorithm, i.e. the metric does not account for gender imbalance - logarithmic loss may not reflect the fact that e.g. predicted pronoun coreference is much worse for masculine pronouns than for feminine ones. Therefore, we explore gender bias separately in Sec. 6 and compare our results with those published by the Google AI Language team (reviewed in Sec. 4).

4 Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns

Google AI Language team addresses the problem of gender bias in pronoun resolution (when systems favor masculine entities) and a gender-balanced labeled corpus of 8,908 ambiguous pronoun-name pairs sampled to provide diverse coverage of challenges posed by real-world text (Webster et al., 2018) (further referred to as the GAP dataset). They run 4 state-of-the-art coreference resolution models (Lee et al., 2013; Clark and Manning, 2015; Wiseman et al., 2016; Lee et al., 2017) on the OntoNotes and GAP datasets reporting F1 scores separately for masculine and feminine pronoun-named entity pairs (metrics **M** and **F** in the paper). Also they measure “gender bias” defined as $\mathbf{B} = \mathbf{F} / \mathbf{M}$. In general, they conclude, these models perform better for masculine pronoun-named entity pairs, but still pronoun resolution is challenging - all achieved F1 scores are less than 0.7 for both datasets.

Further, they propose simple heuristics (called surface, structural and Wikipedia *cues*). The best reported cues are “Parallelism” (if the pronoun is a subject or direct object, select the closest can-

didate with the same grammatical argument) and “URL” (select the syntactically closest candidate which has a token overlap with the page title). They compare the performance of “Parallelism + URL” cue with e2e-coref (Lee et al., 2017) on the GAP dataset and, surprisingly enough, conclude that heuristics work better achieving better F1 scores (0.742 for **M** and 0.716 for **F**) at the same time being less gender-biased (some of heuristics are totally gender-unbiased, for “Parallelism + URL” $\mathbf{B} = \mathbf{F} / \mathbf{M} = 0.96$).

Finally, they explored Transformer architecture (Vaswani et al., 2017) for this task and observed that the coreference signal is localized on specific heads and that these heads are in the deep layers of the network. In Sec. 5 we confirm this observation. Actually, they select the candidate which attends most to the pronoun (“Transformer heuristic” in the paper). Even though they conclude that Transformer models implicitly learn language understanding relevant to coreference resolution, as for F1 scores, they didn’t make it work better than e2e-coref or Parallelism cues (F1 scores lower than 0.63). More to that, proposed Transformers heuristics are a bit biased towards masculine pronouns with **B** from 0.95 to 0.98.

Further we report a much stronger gender-unbiased BERT-based (Devlin et al., 2018) pronoun resolution system.

5 System

BERT (Devlin et al., 2018) is a transformer architecture, pre-trained on a large corpus (Wikipedia + BookCorpus), with 12 to 24 transformer layers. Each layer learns a 1024-dimensional representation of the input token, with layer 1 being similar to a standard word embedding, layer 24 special-

ized for the task of predicting missing words from context. At the same time BERT embeddings are learned for a second auxiliary task of resolving whether two consequent sentences are connected to each other or not.

In general, motivated by (Tenney et al., 2019), we found that BERT provides very good token embeddings for the task in hand.

Our proposed pipeline is built upon solutions by teams “Ken Krige” and “[ods.ai] five zeros” (placed 5 and 22 in the final leaderboard⁷ correspondingly). The way these two teams approached the competition task are described in two Kaggle posts.⁸⁹ The combined pipeline includes several subroutines:

- Extracting BERT-embeddings for named entities A, B, and pronouns
- Fine-tuning BERT classifier
- Hand-crafted features
- Neural network architectures
- Correcting mislabeled instances

5.1 Extracting BERT-embeddings for named entities A, B, and pronouns

We concatenated embeddings for entities A, B, and Pronoun taken from Cased and Uncased large BERT “frozen” (not fine-tuned) models.¹⁰ We noticed that extracting embeddings from intermediate layers (from -4 to -6) worked best for the task. Also we added pointwise products of embeddings for Pronoun and entity A, Pronoun and entity B as well as AB - PP. First of these embedding vectors expresses similarity between pronoun and A, the second one expresses similarity between pronoun and B, the third vector is supposed to represent the extent to which entities A and B are similar to each other but differ from the Pronoun.

⁷<https://www.kaggle.com/c/gendered-pronoun-resolution/leaderboard>

⁸<https://www.kaggle.com/c/gendered-pronoun-resolution/discussion/90668>

⁹<https://www.kaggle.com/c/gendered-pronoun-resolution/discussion/90431>

¹⁰<https://github.com/google-research/bert>

5.2 Fine-tuning BERT classifier

Apart from extracting embeddings from original BERT models, we also fine-tuned BERT classifier for the task in hand. We made appropriate changes to the “run_classifier.py” script from Google’s repository.¹¹ Preprocessing input data for the BERT input layer included stripping text to 64 symbols, then into 4 segments, running BERT Wordpiece for each segment, adding start and end tokens (with truncation if needed) and concatenating segments back together. The whole preprocessing is reproduced in a Kaggle Kernel¹² as well as in our final code on GitHub.¹³

5.3 Hand-crafted features

Apart from BERT embeddings, we also added 69 features which can be grouped into several categories:

- Neuralcoref,¹⁴ Stanford CoreNLP (Manning et al., 2014) and e2e-coref (Lee et al., 2017) model predictions. It turned out that these models performed not really well in the task in hand, but their predictions worked well as additional features.
- Predictions of a Multi-Layered Perceptron trained with ELMo (Peters et al., 2018b) embeddings
- Syntactic roles of entities A, B, and Pronoun (subject, direct object, attribute etc.) extracted with SpaCy¹⁵.
- Positional and frequency-based (distances between A, B, Pronoun and derivations, whether they all are in the same sentence or Pronoun is in the following one etc.). Many of these features we motivated by the Hobbs algorithm (Hobbs, 1986) for coreference resolution.
- Named entities predicted for A and B with SpaCy

¹¹<https://github.com/google-research/bert>

¹²<https://www.kaggle.com/kenkrige/bert-example-prep>

¹³https://github.com/Yorko/gender-unbiased_BERT-based_pronoun_resolution

¹⁴<https://github.com/huggingface/neuralcoref>

¹⁵<https://spacy.io/>

- GAP heuristics outlined in the corresponding paper (Webster et al., 2018) and briefly discussed in Sec. 4

We need to mention that adding all these features had only minor effect on the quality of pronoun resolution (resulted in a 0.01 decrease in logarithmic loss when measured on the Kaggle test dataset) as compared to e.g. fine-tuning BERT classifier.

5.4 Neural network architectures

Final setup includes:

- 6 independently trained fine-tuned BERT classifiers with preprocessing described in Subsec. 5.2. In Tables 1, 2, and 3, we refer to their averaged prediction as to that of a “fine-tuned” model (🔥)
- 5 multi-layered perceptrons trained with different combinations of BERT embeddings for A, B, Pronoun (see Subsec. 5.1) and hand-crafted features (see Subsec. 5.3), all together referred to as “frozen” in Tables 1, 2, and 3 (❄️). Using MLPs with pre-trained BERT embeddings is motivated by (Tenney et al., 2019). Two MLPs- separate for Cased and Uncased BERT models - both taking 9216-d input and outputting 112-d vectors. Two Siamese networks were trained on top of distances between Pronoun and A-embeddings, Pronoun and B-embeddings as inputs. One more MLP took only 69-dimensional feature vectors as an input. Finally, a single dense layer mapped outputs from the mentioned 5 models into 3 classes corresponding to named entities A, B or “Neither”.
- Blending (🌀) involves taking predicted probabilities for A, B and “Neither” with weight 0.65 for the “fine-tuned” model and summing the result with 0.35 times corresponding probabilities output by the “frozen” model.

In the next Section, we perform the analysis identical to the one done in (Webster et al., 2018) to measure the quality of pronoun resolution and the severity of gender bias in the task in hand.

	GAP test	Kaggle test
🔥 fine-tuned	0.29	0.192
❄️ frozen	0.299	0.226
🌀 blend	0.257	0.185

Table 1: Logarithmic loss reported for the GAP test set, and Kaggle test (Stage 2) data for the model with fine-tuned BERT classifier (🔥), MLPs with pre-trained BERT embeddings and hand-crafted features (❄️) and a blend of the previous two (🌀). There are 66 corrections done for GAP test labels as described in Subsec. 5.5.

5.5 Correcting mislabeled instances

During the competition, 158 label corrections were proposed for the GAP dataset¹⁶ - when Pronoun is said to mention A but actually mentions B and vice versa. For the GAP test set, this resulted in 66 pronoun coreferences being corrected. It’s important to mention that the observed mislabeling is a bit biased against female pronouns (39 mislabeled feminine pronouns versus 27 mislabeled masculine ones), and it turned out that most of the gender bias for F1 score and accuracy comes from these mislabeled examples.

6 Results

In Table 1, we report logarithmic loss that we got on GAP test (“gap-test.tsv”), and Kaggle test (Stage 2) datasets. Kaggle competition results can also be seen on the final competition leaderboard.¹⁷ We report GAP test results as well to further compare with the results reported in the GAP paper: measured are logarithmic loss, F1 score and accuracy for masculine and feminine pronouns (Table 2). Logarithmic loss and accuracy are computed for a 3-class classification problem (A, B, or Neither) while F1 is computed for a 2-class problem (A or B) to compare with results reported by the Google AI Language team in (Webster et al., 2018).

We also incorporated 66 label corrections as described in 5.5 and, interestingly enough, this lead to a conclusion that with corrected labels, models are less susceptible to gender bias. Table 3 reports the same metric in case of corrected labeling, and we see that in this case the proposed models are

¹⁶<https://www.kaggle.com/c/gendered-pronoun-resolution/discussion/81331>

¹⁷<https://www.kaggle.com/c/gendered-pronoun-resolution/leaderboard>




	Logarithmic loss				Accuracy				F1 score			
	M	F	O	B	M	F	O	B	M	F	O	B
 fine-tuned	0.294	0.398	0.346	0.738	0.908	0.884	0.896	0.974	0.927	0.9	0.914	0.971
 frozen	0.308	0.368	0.338	0.837	0.883	0.866	0.874	0.981	0.904	0.882	0.893	0.976
 blend	0.259	0.338	0.299	0.766	0.907	0.883	0.895	0.974	0.923	0.898	0.911	0.973

Table 2: Performance of the proposed two models and their blending on the GAP test set, split by **Masculine**, **Feminine** (Bias shows F/M in case of F1 and accuracy, and M/F in case of logarithmic loss), and **Overall**.

almost gender-unbiased.

These results imply that:

- Overall, in terms of F1 score, the proposed solution compares very favorably with the results reported in the GAP paper, achieving as high as 0.911 overall F1 score, compared to 0.729 for “Parallelism + URL” heuristic from (Webster et al., 2018);
- Blending model predictions improves logarithmic loss pretty well but does not impact F1 score and accuracy that much. It can be explained: logarithmic loss is high for confident and at the same time incorrect predictions. Blending averages predicted probabilities so that they end up less extreme (not so close to 0 or 1);
- With original labeling, all models are somewhat susceptible to gender bias, especially in terms of logarithmic loss. However, in terms of F1 score, gender bias is still less than for e2e-coref and “Parallelism + URL” heuristic reported in (Webster et al., 2018);
- Fixing some incorrect labels almost eliminates gender bias, when we talk about F1 score and accuracy of pronoun resolution.

7 Conclusions and further work

We conclude that we managed to propose a BERT-based approach to pronoun resolution which results in considerably better quality (as measured in terms of F1 score and accuracy) than in case of pronoun resolution done with heuristics described in the GAP paper. Moreover, the proposed solution is almost gender-unbiased - pronoun resolution is done almost equally well for masculine and feminine pronouns.

Further we plan to investigate which semantic and syntactic information is carried by different BERT layers and how it refers to coreference resolution. We are also going to benchmark our system on OntoNotes, Winograd, and DPR datasets.

Acknowledgments

Authors would like to thank Open Data Science¹⁸ community for all insightful discussions related to Natural Language Processing and, more generally, to Deep Learning. Authors are also grateful to Kaggle and Google AI Language teams for organizing the Gendered Pronoun Resolution challenge.

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, Beijing, China. 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](#). Cite arxiv:1810.04805Comment: 13 pages.
- Abbas Ghaddar and Phillippe Langlais. 2016. Wikicoref: An english coreference-annotated corpus of wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).
- J Hobbs. 1986. [Readings in natural language processing](#). chapter Resolving Pronoun References, pages 339–352. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. [Deterministic coreference resolution based on entity-centric, precision-ranked rules](#). *Comput. Linguist.*, 39(4):885–916.
- Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. [End-to-end neural coreference resolution](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

¹⁸<https://ods.ai>




	Logarithmic loss				Accuracy				F1 score			
	M	F	O	B	M	F	O	B	M	F	O	B
 fine-tuned	0.268	0.311	0.29	0.863	0.914	0.905	0.91	0.99	0.932	0.919	0.926	0.987
 frozen	0.292	0.306	0.299	0.954	0.886	0.89	0.888	1.005	0.908	0.906	0.907	0.997
 blend	0.241	0.273	0.257	0.882	0.913	0.908	0.91	0.995	0.928	0.921	0.924	0.992

Table 3: Performance of the proposed two models and their blending on the GAP test set with 66 corrected labels, split by Masculine, Feminine (Bias shows F/M in case of F1 and accuracy, and M/F in case of logarithmic loss), and **Overall**.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. [Higher-order coreference resolution with coarse-to-fine inference](#). 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 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. [The winograd schema challenge](#). In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, KR’12*, pages 552–561. AAAI Press.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. [The Stanford CoreNLP natural language processing toolkit](#). In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. [Deep contextualized word representations](#). Cite arxiv:1802.05365Comment: NAACL 2018. Originally posted to openreview 27 Oct 2017. v2 updated for NAACL camera ready.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018b. [Deep contextualized word representations](#). *CoRR*, abs/1802.05365.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica Macbride, and Linnea Micciulla. 2007. [Unrestricted coreference: Identifying entities and events in ontonotes](#). *International Conference on Semantic Computing*, 0:446–453.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. 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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In I. Guyon, U. V. Luxburg, S. Bengio,

H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. [Learning global features for coreference resolution](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, San Diego.