Gender-Inclusive Pronoun Coreference Resolution

Patrick Sonnenberg
Brandeis University
psonnenberg@brandeis.edu

1 Introduction

Natural language processing as a field has yet to properly take into account how language intersects with sociocultural realities, ideologies, and identities, especially with regards to gender, sexuality, race, and power (see Cao and Daumé III, 2020; Bramsen et al., 2011; Field et al., 2021; Hutchinson et al., 2020; Lauscher et al., 2022; Talat et al., 2022). Without proper consideration of macro, meso, and micro-level sociocultural realities and how they vary across time and space, NLP runs the risk of perpetuating harm and violence against marginalized communities (see de Gibert et al., 2018; Fourcade and Johns, 2020; Ray, 2019; Rosa and Flores, 2017; Tomasev et al., 2021).

I focus on the incorporation of singular gender-neutral "they/them" and neopronouns in coreference resolution. First, I define and examine gender in coreference resolution research and adapt sentence templates to create a new dataset. I run the new data through two existing coreference resolution models and then run experiments to evaluate performance on four of my own models using various training sets, features, and pre-trained embeddings. Last, I highlight key insights and takeaways.

1.1 Gender

Gender is a complex, multifaceted sociocultural construct that encompasses a broad range of cultural, linguistic, and embodied practices intimately tied to sexuality, race, class, and other identities. And gender encompasses a spectrum of identities, not simply a binary, that are continually constructed and performed through discourse and embodied expression (see Butler, 2011; Muñoz, 1999).

Sociocultural linguists Bucholtz and Hall (2005) argue that "the constant iteration of particular practices cumulatively produce...gender itself as a socially meaningful system." This includes discourse and interaction, which constitute identities as socially real. Linguistic resources such as labels, implicatures, stances, and styles all indexically produce identity (see Kiesling, 2018; Zimman, 2019).

Queer linguists examining these intersections draw on queer theory to problematize identity categories and critique heteronormativity (Motschenbacher, 2011). For this paper, I use gender in the sense of these queer and sociocultural linguists, rather than grammatical gender, for instance (Ackerman, 2019). Pronouns often index gender itself, along with sociocultural ideologies beyond gender (Lauscher et al., 2022). Thus, using incorrect pronouns to refer to a person is considered misgendering, as this is an explicit refusal to acknowledge one's identity, which also has broader implications (Cameron and Kulick, 2003; Dev et al., 2021).

1.2 Gender in Coreference Resolution

NLP tasks, datasets, and evaluation metrics can all perpetuate gender bias (Sun et al., 2019). Unfortunately, very little has been done to address inclusive pronoun coreference resolution. There exist a few studies drawing on existing Winograd schemas that test for bias. Baumler and Rudinger (2022) create their own "WinoNB schemas" dataset to test singular they/them, finding that five coreference resolution systems all perform incredibly poorly on all forms of singular they/them. In addition, Zhao et al. (2018) use them to create a new benchmark called WinoBias, focused specifically on gender bias. Rudinger et al. (2018) utilize these schemata to measure occupational gender bias. Webster et al. (2018) similarly focus on occupational gender bias.

Much of the research addressing gender bias and inclusion in coreference resolution has unfortunately, and ironically, focused on the gender binary. Zhao et al. (2018) relies on a binary, as does Webster et al. (2018), who focus on pronoun ambiguity. Some do acknowledge pronouns beyond the binary, such as Cao and Daumé III (2020), who use naturally occurring data and carefully tease apart grammatical, referential, lexical, and social gender. Lauscher et al. (2022) explore gender diversity and neopronouns, pointing out that people can use multiple pronouns. Zhang et al. (2019) argue that external knowledge must be incorporated in these models. These studies demonstrate that much work is still needed to make this task more inclusive.

2 Methods

I first updated the Winograd schema templates from Rudinger et al. (2018) to include examples with neopronouns. As the dataset already contained singular they/them pronouns, I added three additional neopronouns: ze, xe, and ey. The nominative, accusative, and possessive forms of these pronouns are displayed in Table 1. I wrote a script to take their sentence templates and insert the appropriate occupations, participants, and pronouns. Each template will thus generate six sentences as there is one for each pronoun. I save these results along with other metadata to make evaluation easier. As there are 120 sentence templates, there are a total of 720 sentences. In the last column, I add a tuple that contains two items: one to indicate whether the pronoun corefers to the first person or second person, and one to indicate the gender corresponding to the pronoun. I chose to do this for a few reasons - it ensures there are enough examples for each tuple for cross validation, it captures whether the relation is for the first or second entity, and it captures the pronoun/gender. And I explicitly include the pronoun/gender since neopronouns are often deemed out-of-vocabulary.

Pronoun	Nom	Acc	Poss
masculine	he	him	his
feminine	she	her	her
neutral they	they	them	their
neutral ze	ze	zir	zir
neutral xe	xe	xem	xir
neutral ey	ey	em	eir

Table 1: Pronouns and their nominative, accusative, and possessive forms

I planned to use these sentence templates to evaluate and re-train existing coreference resolution models. However, after downloading each model, I had significant troubles even getting them to run in the first place, never mind reformatting the data to the necessary format for the model. I tested Stanford CoreNLP's neural-network-based model by using the CoreNLP server online. This is of course not the most efficient method, but it became my only choice as I struggled to get the downloaded version to run. I iterated over the output to obtain the predicted coreferents, and I compared those to the actual gold coreferent pairs.

I next used Coreferee, a perceptron model that uses spaCy. I had to again process the text to get the model's predictions to match a format to which I could compare it with the gold labels.

Next, I generated train and development files from the updated sentence templates dataset, as there were too few data to also create a test set. I created four train files - one with only binary gender pronouns; one with binary and they/them; one with binary, they/them, and ze; and one with all six. There are two dev files - one with binary pronouns and they/them, and the other with all six. As a note, each training set did contain at least one example with each pronoun. If one sentence template was in train, it does not appear in the dev set.

I used four different models to compare performance - LSTM, BiLSTM, RNN, and CNN. I also wrote the file to handle multiple models, where each model is trained separately and the predictions are passed through a voting classifier. I also wrote the code to handle random, GloVe[1], and FastText[2] embeddings. There are additional features for adding entity tags around each entity, adding positional encoding, adding parts of speech tags, and truncation, along with other tunable hyperparameters such as batch size and learning rate.

There are five main experiments. The first tests each model against each set of training data. The next tests each model against the additional features to see if any improve performance. The third tests each model against various embedding types. The fourth tests each model being combined with other models to make the final predictions. Finally, the fifth tests the highest scoring models being combined with all features that led to increased accuracies. All experiments and trainings are run for 20 epochs and using GPU and Google Colab. All scores reported are accuracy scores on the dev set.

3 Initial Experiments

I did a baseline evaluation with the output from Stanford CoreNLP's model[3]. As expected, it per-

[1] GloVe
[2] FastText
[3] Stanford CoreNLP

formed poorly on neopronouns, and it only even predicted them for one sentence. The output specified the neopronoun had a coreferent of “it”, which was incorrect. It is notable, though, that it predicted all three of these neopronouns as coreferents to “it”. As expected, the model performed best for masculine pronouns, followed by gender-neutral they/them, and then feminine pronouns. I find this interesting, especially as NLP models typically perform poorly on gender-neutral they/them, though the original dataset was created to detect encoded gender bias based on occupation. The results for the non-neopronouns can be found in Table 2, as the neopronouns were never correctly predicted.

	He	She	They
CoreNLP	40.8	10.8	25.0
Coreferee spaCy sm	49.2	49.2	46.7
Coreferee spaCy lg	47.5	50.0	50.0
Coreferee spaCy trf	50.0	50.0	50.0

Table 2: Models’ predictions accuracy

I next did the same with Coreferee[4], which uses five multilayer perceptrons and utilizes spaCy. I used three different sized pipelines that spaCy provides[5], including a small, a large, and a transformer (RoBERTa-base) pipeline. The small had slightly lower percentages than the large size and transformer pipeline. Interestingly, the large model worked slightly worse for masculine pronouns than the small. The transformer pipeline accuracies were all even for masculine, feminine, and they/them pronouns. As with CoreNLP, none of the Coreferee spaCy models recognized any of the neopronouns. Finally, I had attempted to use NeuralCoref[6] from HuggingFace, but could never get it to function on my laptop, on Google colab, or through post requests.

4 Experiments

The following experiments are based off using the gold labels, which include a 0 or 1 and the corresponding pronoun. This means there are 12 possible categories - six pronouns times the two coreferent entity possibilities.

I calculate accuracies based off the six pronouns, but I only consider a result correct if both the pronoun and entity number correspond. Thus, if the prediction and gold are both ‘masculine’, it would not be counted as correct if it predicted entity 1 and the gold is entity 0. Regardless, the count for the total number of ‘masculine’ would still increase.

Table 3 reflects the first experiment, showing how adding more pronouns to the training data increases accuracy on the development set. Unsurprisingly, any pronoun not seen in training almost always had a score of 0% in the dev data. The neopronouns were being predicted accurately around the same levels as the other pronouns when training included all six pronouns. This means in training, the model is better able to pick up on these pronoun relations. Adding only one neopronoun in training, however, did not help improve accuracy for the other neopronouns in the dev data. For all remaining experiments, I only use the training data the contains all six pronouns.

	Bin	Bin/Th	Bin/Th/Ze	All 6
LSTM	8.33	22.9	35.4	42.4
BiLSTM	25.0	31.3	34.7	48.6
RNN	9.7	12.5	23.6	8.3
CNN	15.3	23.6	31.9	35.4

Table 3: Testing Models by Training Sets: Binary, Binary/They, Binary/They/Ze, and All 6 in Training

Table 4 shows that, for the most part, truncation helped the models the most. Across each test, the LSTM appears to outperform all the other models, except for truncation with a BiLSTM. The RNN model continues to perform poorly. I do not add any of these features to the next experiment, as the next experiment is testing embeddings.

	EntTag	Position	POS	Trunc
LSTM	44.4	43.1	46.5	48.6
BiLSTM	43.1	40.3	40.3	50.7
RNN	11.1	11.8	8.3	11.1
CNN	29.9	15.3	18.8	48.6

Table 4: Testing Models by Additional Features: Entity Tags, Positional Encoding, POS Tags, and Truncation

The results from the third experiment are found in Table 5. Using pretrained embeddings increased performance for LSTM and CNN pretty significantly, but less so for RNN. The highest performing was the LSTM using FastText, though it was very similar to both GloVe results. Because the LSTM performed the best, I ran an additional test with the LSTM where I train only on binary pronouns,

[4] Coreferee
[5] spaCy English Pipelines
[6] NeuralCoref

as opposed to all six pronouns. All embeddings performed very poorly, which again shows that more inclusive training data would lead to higher accuracies for testing data.

	Random	G-100	G-300	FT
LSTM	42.4	49.3	50.0	50.7
BiLSTM	48.6	50.0	49.3	45.8
RNN	8.3	8.3	9.0	9.7
CNN	35.4	50.7	49.3	42.4
Binary	8.33	10.4	17.4	16.7

Table 5: Testing Models by Embedding Type

I also tested combinations of models, where multiple models are used to determine the final prediction. As the LSTM and BiLSTM cannot really be combined, each of them is combined with the other models. Table 6 shows these results, with the best score coming from combining the BiLSTM with the CNN. The best score for the LSTM was combining it with both the RNN and CNN. However, the LSTM, BiLSTM, and CNN all had higher accuracies when simply run as standalone models.

	LSTM	BiLSTM
RNN	25.7	18.8
CNN	43.8	49.3
RNN + CNN	45.8	44.4

Table 6: Testing (Bi)LSTM with Other Models, Using GloVe 300d Embeddings

The highest score throughout all experiments has been 50.7%, which the LSTM, BiLSTM, and CNN all reached. For the next test, shown in Table 7, I combine each of those models with every feature that increased its accuracy across all experiments. Thus, I trained LSTM with FastText embeddings, added entity tags, positional encoding, POS tags, and truncation; BiLSTM with GloVe 100d embeddings and truncation; and CNN with GloVe 100d embeddings and truncation.

	Accuracy
LSTM + FastText + 4 Features	50.0
BiLSTM + G-100 + Trunc	51.4
CNN + G-100 + Trunc	47.9

Table 7: Combining Models with Parameters that Increased Accuracy

The only model that increased accuracy with Experiment 5 was the BiLSTM, being combined with GloVe 100d embeddings and truncation. Given the small size of the dataset, I tried a smaller learning rate of 0.0025, which slightly increased accuracy to 52.1%. I additionally decreased the batch size to 32, which increased accuracy even more to 56.3%. This model's accuracy is broken down by pronoun in Table 8, demonstrating that it performs relatively uniformly across pronouns.

He	She	They	Ze	Xe	Ey
54.17	58.33	54.17	58.33	58.33	54.17

Table 8: Pronoun Accuracies of Highest Scoring Model

I wrote a script to use OpenAI's API[7]. I tried various prompts, including recognizing the pronoun and its coreferent, as well as simply returning the pronoun. However, all of the results were poor and often did not return anything of use, including with davinci, curie, ada, and babbage models. This could be because none of them were trained specifically to do this task or coreference resolution in general.

5 Conclusion

These experiments show that more inclusive training data leads to better results on test data, lowering the risk of erasing certain gender identities. The existing coreference models I tested could not even detect neopronouns. The pretrained embeddings performed especially poorly when only trained on binary pronouns, showing that those embeddings are also erasing certain identities.

The (Bi)LSTM consistently scored the highest, likely because they are designed to handle long-term dependencies across sequential data (see Hochreiter and Schmidhuber, 1997). The CNN often performed similarly, perhaps because it can capture local features and capture patterns by using convolutional filters (see Kim, 2014). The RNN likely performed poorly because RNNs suffer from vanishing gradients and are less effective at capturing local features (see Bengio et al., 1994). The best model had an accuracy of only 56.3%, meaning that there is significantly more research needed.

Gender has been understood as beyond a binary for decades, yet NLP has alarmingly and overwhelmingly ignored such basic concepts, treating gender as a static category. More inclusive work must not solely fall on the shoulders of those most

[7] OpenAI API

affected (Ahmed, 2012; Berenstain, 2016). We must proactively and explicitly take into account political and sociocultural realities regarding gender, sexuality, race, and class in order to reduce bias, harm, and violence, even if that comes at the expense of "accuracy" or capital (Zuboff, 2019; see also Burrell and Fourcade, 2021; Carter and Hurtado, 2007; Monk Jr, 2022; Nguyen et al., 2016; O'Neil, 2017; Talat et al., 2022).

Limitations

There are several limitations for this project. First, time and scope limited my abilities. Consequently, the dataset was pretty small. The dataset also was individual sentences, rather than more naturally occurring data. Additionally, the dataset did not take into account the fact that people can use more than one pronoun, and even switch the pronoun in the same conversation. Gender fluidity is important to take into consideration. Pronouns do not always correspond to gender identity, meaning using "he/him" pronouns does not entail identifying as a 'man' or 'masculine.' While I have tried to make this distinction, it is very tricky given pronoun coreference resolution as a task tends to assume pronouns directly index gender.

This project is also English-centric and relying on sociocultural knowledge of present-day United States, and still does not reflect all contemporary realities across the US. More inclusive studies will take into account other languages and sociocultural contexts.

Another limitation is that this project assumes that the two entities and the pronoun in the sentence have been pre-defined. This is of course not how naturally ocurring data would be structured, so it could be necessary to add some sort of entity or occupation recognition before the coreference resolution if not using naturally-occuring data.

Ethics Statement

The ethical issues relate back to the identity-related limitations of the study. The project was limited in scope of neopronouns, did not contain in-context data, and did not account for fluidity and the usage of multiple pronouns. There can also be privacy concerns given gender identity and pronoun usage. These all need to be included in future research in order to truly reduce harm and violence. Given the scope of the project, other factors were not taken into consideration, such as race or class, which would necessarily need to be included in all future research.

There are a few other ethical concerns. No financial compensation was provided to those who compiled the original dataset of sentence templates. Additionally, the pre-trained word embeddings I used certainly contributed to carbon emissions in their original trainings, as did the running of these experiments. Using these embeddings also further reinforced large tech companies' concentrated power, as they are the only ones who can afford to train these models and embeddings.

While completing on this project, I resided on land that belongs to the Wampanoag, and I want to acknowledge the painful history, violence, and forced removal of indigenous peoples from this area.

References

Lauren M Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa*.

Sara Ahmed. 2012. *On being included: Racism and diversity in institutional life*. Duke University Press.

Connor Baumler and Rachel Rudinger. 2022. Recognition of they/them as singular personal pronouns in coreference resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3426–3432, Seattle, United States. Association for Computational Linguistics.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Nora Berenstain. 2016. Epistemic exploitation. *Ergo*, 3(22):569–590.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782, Portland, Oregon, USA. Association for Computational Linguistics.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: a sociocultural linguistic approach. *Discourse Studies*, 7(4-5):585–614.

Jenna Burrell and Marion Fourcade. 2021. The society of algorithms. *Annual Review of Sociology*, 47:213–237.

Judith Butler. 2011. *Gender trouble: Feminism and the subversion of identity*. Routledge.

Deborah Cameron and Don Kulick. 2003. *Language and sexuality*. Cambridge University Press.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Deborah Faye Carter and Sylvia Hurtado. 2007. Bridging key research dilemmas: Quantitative research using a critical eye. *New Directions for Institutional Research*, 133:25–35.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Marion Fourcade and Fleur Johns. 2020. Loops, ladders and links: the recursivity of social and machine learning. *Theory and Society*, 49:803–832.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Scott F. Kiesling. 2018. Masculine stances and the linguistics of affect: on masculine ease. *NORMA*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ellis P Monk Jr. 2022. Inequality without groups: Contemporary theories of categories, intersectional typicality, and the disaggregation of difference. *Sociological Theory*, 40(1):3–27.

Heiko Motschenbacher. 2011. Taking queer linguistics further: Sociolinguistics and critical heteronormativity research. *International Journal of the Sociology of Language*, 2011.

José Esteban Muñoz. 1999. *Disidentifications: Queers of color and the performance of politics*, volume 2. U of Minnesota Press.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Survey: Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Cathy O'Neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Victor Ray. 2019. A theory of racialized organizations. *American Sociological Review*, 84(1):26–53.

Jonathan Rosa and Nelson Flores. 2017. Unsettling race and language: Toward a raciolinguistic perspective. *Language in society*, 46(5):621–647.

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.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Nenad Tomasev, Kevin R McKee, Jackie Kay, and Shakir Mohamed. 2021. Fairness for unobserved characteristics: Insights from technological impacts on queer communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 254–265.

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.

Hongming Zhang, Yan Song, Yangqiu Song, and Dong Yu. 2019. Knowledge-aware pronoun coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 867–876, 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.

Lal Zimman. 2019. Trans self-identification and the language of neoliberal selfhood: Agency, power, and the limits of monologic discourse. *International Journal of the Sociology of Language*, 2019(256):147–175.

Shoshana Zuboff. 2019. *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile books.