Part of Speech Tagging of Neopronouns:
An Exploration of Gender Inclusion in Artificial Intelligence

Lain Nelson

August 2023

Introduction:

The field of natural language processing (NLP), explores the intersection of human languages and artificial intelligence (AI). While incredible progress has been made in this field, such as the advent of ChatGPT, human biases are often replicated and even exacerbated in language models. Language models are the result of utilizing machine learning with instances of human language such as newspaper articles. Gender bias is also present in these language models, especially as it relates to individuals who use neopronouns, or “new pronouns” (e.g. ey/em). These pronouns often fail to even be identified as such.

Research Journey:

I began with the hope of exploring how inclusive language can be further incorporated into artificial intelligence. While this is an area that I had been interested in for some time, I had no idea what to expect when formulating my project. I had initially hoped to perform a comparative analysis on the difficulties of incorporating gender inclusive language into AI systems in English, which is generally less gendered, versus Spanish, which implements almost ubiquitous binary grammatical gender. However, as I began to work I quickly discovered that even working solely in English would likely extend far beyond the six weeks of my Laidlaw research experience. With the support of the Laidlaw Leadership and Research Program, I connected with my mentor, Professor Cardie, who specializes in natural language processing (NLP). I then decided to focus on gender inclusive language and began with a broad literature review to help me hone a specific research project.

The majority of my time this summer was spent on that literature review. One article I examined was "Learning Gender-Neutral Word Embeddings" (Zhao et al. 2018). I learned that words are often translated into mathematical vectors, or embeddings in NLP (Ibid). In more recent work, these embeddings can change over time, but in this article the embeddings are static. This article utilized a technique to restrict information about gender largely to specific dimensions of the vector, thus allowing it to be ignored when looking at word embeddings to determine the similarity of words (Ibid). However, the work only recognized the gender binary and explored bias solely from that standpoint.

Reading "How Conservative are Language Models? Adapting to the Introduction of Gender-Neutral Pronouns", I realized that when processing text automatically, current technologies have difficulty with gender-neutral pronouns (Brandl, Cui and Søgaard 2022).

Specifically, higher perplexity, more dispersed attention patterns and worse downstream performance is evident when utilizing gender-neutral pronouns (Ibid). Furthermore, this work recognized not only they/them pronouns as gender-neutral pronouns, but also neopronouns (Ibid). Neopronouns, or "new" pronouns, are pronouns that are not as widely recognized (e.g. ze/zir).

"Towards the Necessity for Debiasing Natural Language Inference Datasets" introduced the technique of delexicalization (Panenghat et al. 2020). Specifically, the article explored replacing entities with the super sense tag corresponding to them (Ibid). In addition, whenever the same entity appears again, the work investigated giving it the same tag (as differentiated by a number), and whenever a different entity appears it is giving it a different tag (Ibid). Training on this delexicalized data resulted in improved performance in natural language inference tasks when compared to not training on delexicalized data (Ibid). This perhaps occurred by preventing patterns irrelevant to the task at hand from being learned (Ibid). This technique was interesting to me, especially as it relates to coreference resolution. Coreference resolution is the task of determining what entity a word, such as a pronoun, is referring to. I was curious whether replacing a pronoun with a tag could remove some of the gender bias in coreference resolution. However, I was not sure what information should be included in this tag. Thus, I decided to keep investigating other possibilities for my project.

I was most inspired by "Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender" (Lauscher, Crowley and Hovy 2022). This paper describes how pronouns are actually an open word class, not a fixed list of words, a fact which has largely been ignored by NLP (Ibid). For instance, pronouns can be formed from a variety of words (e.g. sun/suns/sunself). There are some types of pronouns that are only expressible in

certain media, such as emojiself pronouns (e.g. 🍓/🍓s / 🍓self) (Ibid). Furthermore, there are a variety of language phenomena that tackle gender inclusion, and it is unclear which will persist (Ibid). With this in mind as well as an additional investigation, the paper proposed five desiderata for handling pronouns:

1. "Refrain from assuming an individual's identity and pronouns."
2. "Allow for the existing sets of pronouns as well as for neopronouns."
3. "Allow for novel pronouns at any point in time."
4. "Allow for multiple, alternating, and changing pronouns."
5. "Provide an option for individuals to define their sets of pronouns." (Ibid)

This paper led me to believe that handling of pronouns is extremely important for ensuring gender equity in technology. Furthermore, the relative absence of scholarship on neopronouns encouraged me to focus on them.

After settling on the subject of neopronouns, I considered ways in which they could be examined in the context of NLP. I realized that because neopronouns have largely been ignored by the field, natural language models which have been used to determine the part of speech (e.g. verb or adjective) of words may not even be able to identify them as pronouns. Looking at part of speech tagging is also an effective way to analyze the treatment of neopronouns, whereas with other, more complex, natural language inference tasks it is more difficult to determine how the presence of neopronouns is affecting performance on the task.

Once I decided to focus on part of speech tagging on neopronouns, I began by searching for textual examples of neopronouns. I found that such writings were found most easily on Archive of Our Own (AO3). Using the boolean search "neopronoun OR neopronouns", I found 1,751 works. I initially wrote code to scrape fanfictions from AO3, but it stopped working after

the website experienced a distributed denial-of-service (DDoS) attack. Thus, in order to perform some analysis, I manually gathered the text of some works and then ran part of speech taggers supplied by SpaCy on this text. My initial explorations revealed that neopronouns are often misidentified, commonly as either nouns or proper nouns. I created a file for each fanfiction that I was exploring and stored it in my Google Drive. I then uploaded these files to Google Colab, where I imported different part of speech taggers from SpaCy. I utilized these part of speech taggers to create an object with the text of the works, as well as the tags associated with the works. While my analysis is still in progress and I have many functions left to tag, I will continue my work this fall. I plan to produce more concrete statistics on the treatment of neopronouns and explore how part of speech tagging of neopronouns can be improved.

Citations:

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.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Mithun Paul Panenghat, Sandeep Suntwal, Faiz Rafique, Rebecca Sharp, and Mihai Surdeanu. 2020. Towards the Necessity for Debiasing Natural Language Inference Datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6883–6888, Marseille, France. European Language Resources Association.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. How Conservative are Language Models? Adapting to the Introduction of Gender-Neutral Pronouns. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3624–3630, Seattle, United States. Association for Computational Linguistics.