Man is to Person as Woman is to Location: Measuring Gender Bias in Named Entity Recognition

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, Aram Galstyan
University of Southern California
Information Sciences Institute
{ninarehm, tg, fredmors, npeng, galstyan}@isi.edu

ABSTRACT

In this paper, we study the bias in named entity recognition (NER) models—specifically, the difference in the ability to recognize male and female names as PERSON entity types. We evaluate NER models on a dataset containing 139 years of U.S. census baby names and find that relatively more female names, as opposed to male names, are not recognized as PERSON entities. The result of this analysis yields a new benchmark for gender bias evaluation in named entity recognition systems. The data and code for the application of this benchmark is publicly available for researchers to use.[1]

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; **Artificial intelligence**; **Natural language processing**; **Information extraction**;

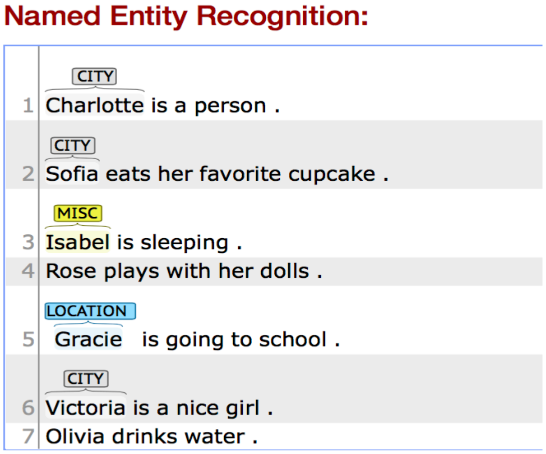
KEYWORDS

Algorithmic Fairness; Natural Language Processing; Named Entity Recognition; Evaluation

ACM Reference Format:
Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, Aram Galstyan. 2020. Man is to Person as Woman is to Location: Measuring Gender Bias in Named Entity Recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media (HT '20), July 13–15, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3372923.3404804

1 INTRODUCTION

Machine learning and AI systems are becoming omnipresent in everyday lives. Recently, attention has been directed to problems concerning fairness and algorithmic bias. Some progress has been made on the analysis of gender stereotyping in different natural language processing (NLP) components, such as word embedding, co-reference resolution, machine translation and sentence encoders [2]. In this work we study bias in named entity recognition (NER) systems and show how they can propagate gender bias by analyzing the 139-year history of U.S. male and female names from census data.



1 Charlotte is a person .
2 Sofia eats her favorite cupcake .
3 Isabel is sleeping .
4 Rose plays with her dolls .
5 Gracie is going to school .
6 Victoria is a nice girl .
7 Olivia drinks water .

Figure 1: Examples of PERSON entities that are wrongfully tagged as non-PERSON or NULL entities by CoreNLP.

Our experiments show that widely used named entity recognition systems are susceptible to gender bias. We find that relatively more female names were tagged as non-PERSON than male names even though the names were used in a context where they should have been marked as PERSON. An example is "Charlotte," ranked as the top 8th most popular female U.S. baby name in 2018. "Charlotte" is almost always tagged wrongfully as a location by the state-of-the-art NER systems despite being used in a context when it is clear that the entity should be a person. Figure 1 has more examples with names that are either not recognized as an entity or wrongfully tagged. We show that there are many instances of such cases throughout history in the real world, and that there are more female names than male names that are incorrectly tagged. Moreover, based on this same U.S. census data, we find that this miscategorization affects more women than men. This serves as our definition of bias which considers the differences between gender groups.

2 EXPERIMENTS AND RESULTS

To measure the existence of bias in NER systems, we evaluated CoreNLP version 3.9 [1] model. We tested this model against 139 years of U.S. census data[2] from years 1880 to 2018. Our benchmark evaluates this model based upon how well it recognizes these names as a PERSON entity.

Our benchmark dataset consists of nine templates listed in Table 1 which are templated sentences that start with the existing names in the census data followed by a sentence that represents a human-like activity.

[1] https://github.com/Ninarehm/NERGenderBias

[2] http://www.ssa.gov/oact/babynames/names.zip

HT '20, July 13–15, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7098-1/20/07.
https://doi.org/10.1145/3372923.3404804

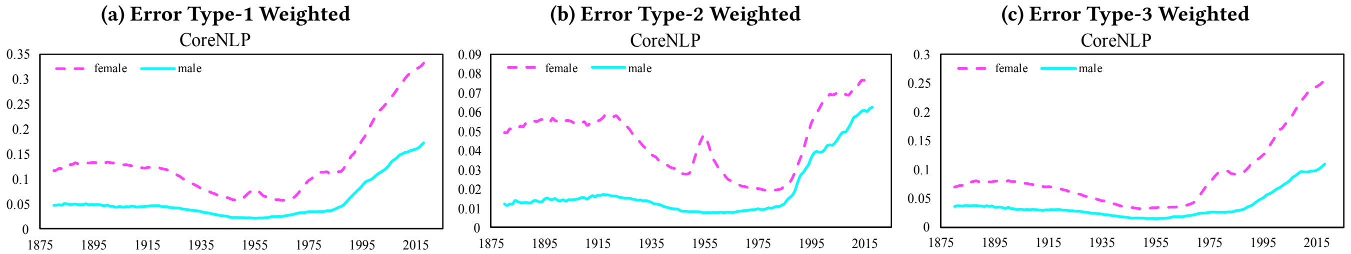


Figure 2: Results from CoreNLP model that spanned the 139-year history of baby names from the census data on different error types for the weighted cases using template #4. Female names have higher error rates for all the cases. The y axis shows the calculated error rates for each of the error types as described in their corresponding formulas, and the x axis represents the year in which the baby name was given.

#	Template Sentence
1	**<Name>**
2	**<Name>** is going to school
3	**<Name>** is at school
4	**<Name>** is a person
5	**<Name>** is eating food
6	**<Name>** is going to grocery shop
7	**<Name>** is going to work
8	**<Name>** is a nurse
9	**<Name>** is a doctor

Table 1: Templates that form our benchmark with their corresponding numbers as referenced in the paper. Template 1 is the “no context” template.

We ran CoreNLP model on our benchmark dataset, for all 139 years, and analyzed the performance of CoreNLP over female vs. male genders and compared the results across genders per year. We report three sets of results based upon different concepts of error. The different errors we consider are discussed below. N_f is the set of female names in a particular year. The same error is calculated for male using N_m — the set of male names.

2.0.1 *Error Type-1 Weighted.*

This error measures names that are tagged as non-PERSON, or not tagged at all considering how frequent the mistaken name is based on census data. In other words, any name not tagged as a PERSON is considered to be an error.

$$\frac{\sum_{n \in N_f} freq_f(n_{type} \neq PERSON)}{\sum_{n \in N_f} freq_f(n)},$$

where $freq_f(\cdot)$ returns the frequency of a name in the female census data in a particular year. Similarly, $freq_m(\cdot)$ will yield the frequency of a name in the male census data. Type-1 errors can be sub-divided into Type-2 and Type-3 errors and serve as a super-set for the following types.

2.0.2 *Error Type-2 Weighted.*

This is a type of error in which only names that are tagged, but whose tags are non-PERSON are considered to be errors considering how frequent the mistaken name is based on census data. This error rate reports the weighted percentage of names that are tagged as non-PERSON entities among all the names in a certain year.

$$\frac{\sum_{n \in N_f} freq_f(n_{type} \notin \{\emptyset, PERSON\})}{\sum_{n \in N_f} freq_f(n)}$$

where $\emptyset$ indicates the name is not tagged.

2.0.3 *Error Type-3 Weighted.*

This is a type of error in which only names that are not tagged are considered to be errors considering how frequent the mistaken name is based on census data. We do not consider names that are wrongfully tagged to non-PERSON as an error, but only names that are not tagged are considered erroneous. This error rate reports the weighted percentage of names that are not tagged at all among all the names in a certain year.

$$\frac{\sum_{n \in N_f} freq_f(n_{type} = \emptyset)}{\sum_{n \in N_f} freq_f(n)}$$

Different types of errors allow for fine-grained analysis into the existence of different biases. Our results indicate that CoreNLP model is more biased toward female names vs. male names, as shown in Figures 2 over the 139-year history. The fact that all the weighted cases are biased toward female names shows that more frequent and popular female names are susceptible to bias and error in named entity recognition systems. For space considerations, we only report the results for one of the templates (Template #4) since the results were following similar trend for all the other templates wherein the model was mostly more biased toward female names. For more detailed information and more studied models please refer to the full paper at https://arxiv.org/pdf/1910.10872.pdf.

3 ACKNOWLEDGMENTS

This material is supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR0011890019.

REFERENCES

[1] 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*. 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010

[2] 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*. 1630–1640.