

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

We study the bias in several state-of-the-art named entity recognition (NER) models—specifically, a 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. We study the extent of this bias in several NER systems that are used prominently in industry and academia. In addition, we also report a bias in the datasets on which these models were trained. 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 will be publicly available for researchers to use.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 (Bolukbasi et al. 2016; Zhao et al. 2018b), co-reference resolution (Zhao et al. 2018a), machine translation (Font and Costa-jussà 2019) and sentence encoders (May et al. 2019). 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.

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

Named Entity Recognition:

<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;">1</td><td style="padding: 2px;">CITY</td></tr> <tr> <td colspan="2" style="padding: 2px;">Charlotte is a person .</td></tr> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;">2</td><td style="padding: 2px;">CITY</td></tr> <tr> <td colspan="2" style="padding: 2px;">Sofia eats her favorite cupcake .</td></tr> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;">3</td><td style="padding: 2px;">MISC</td></tr> <tr> <td colspan="2" style="padding: 2px;">Isabel is sleeping .</td></tr> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;">4</td><td style="padding: 2px;">MISC</td></tr> <tr> <td colspan="2" style="padding: 2px;">Rose plays with her dolls .</td></tr> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;">5</td><td style="padding: 2px;">LOCATION</td></tr> <tr> <td colspan="2" style="padding: 2px;">Gracie is going to school .</td></tr> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;">6</td><td style="padding: 2px;">CITY</td></tr> <tr> <td colspan="2" style="padding: 2px;">Victoria is a nice girl .</td></tr> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;">7</td><td style="padding: 2px;">CITY</td></tr> <tr> <td colspan="2" style="padding: 2px;">Olivia drinks water .</td></tr> </table>	1	CITY	Charlotte is a person .		2	CITY	Sofia eats her favorite cupcake .		3	MISC	Isabel is sleeping .		4	MISC	Rose plays with her dolls .		5	LOCATION	Gracie is going to school .		6	CITY	Victoria is a nice girl .		7	CITY	Olivia drinks water .		<h3>Named Entity Recognition:</h3>
1	CITY																												
Charlotte is a person .																													
2	CITY																												
Sofia eats her favorite cupcake .																													
3	MISC																												
Isabel is sleeping .																													
4	MISC																												
Rose plays with her dolls .																													
5	LOCATION																												
Gracie is going to school .																													
6	CITY																												
Victoria is a nice girl .																													
7	CITY																												
Olivia drinks water .																													

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

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 following the statistical parity notion of fairness (Kusner et al. 2017; Dwork et al. 2012).

The contributions of this paper are fourfold:

1. We introduced a benchmark dataset that tests NER models for gender bias. This benchmark can be run on any blackbox system.
2. We studied the existence and extent of gender bias in current NER systems by analysing a 139-year history of names according to the statistical parity notion of fairness (Kusner et al. 2017; Dwork et al. 2012).
3. We compared different versions of models and found that newer versions are amplifying gender bias in an effort to boost performance. Based on our observations, we defined a new source of bias that arises from version updates in the systems.

¹<https://github.com/Ninarehm/NERGenderBias>

#	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.

4. Finally, we analyzed datasets currently used for training many NER models and found the extent of gender bias in these datasets.

Models and Experiments

To measure the existence of bias in NER systems, we evaluated five named entity models used in research and industry. We used Flair (Akbik, Blythe, and Vollgraf 2018; Akbik et al. 2019), CoreNLP version 3.9 (Manning et al. 2014; Finkel, Grenager, and Manning 2005), and Spacy version 2.1 with small, medium, and large models.² We test these models against 139 years of U.S. census data³ from years 1880 to 2018. Our benchmark evaluates these models based upon how well they recognize these names as a PERSON entity.

Benchmark

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. The aim was to test the performance of the NER from different perspectives. Template 1, containing just the name, purely tests the name itself and reveals something about the distribution of the training data. Template 4 is designed to direct the model to tag the name as a person. Template 3 may reveal more subtle gender bias that stems from society as, historically, men have received greater levels of education from women. It is possible that the error rate may be even higher for female names under Template 3.

Experimental Design and Results

We ran each NER model on our created benchmark dataset, for all 139 years, and analyzed the performance of each template over female vs. male genders and compared the results of these models across genders per year. We report six 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.

Error Type-1 Unweighted. This is a type of error that measures names that are tagged as non-PERSON, or not tagged at all. In other words, any name not tagged as a PERSON is considered to be an error.

$$\frac{\sum_{n \in N_f} I(n_{type} \neq PERSON)}{|N_f|}$$

Error Type-1 Weighted. This type of error is similar to Error Type-1 Unweighted; however, we considered how frequent the mistaken name is based on census data while calculating the 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.

Error Type-2 Unweighted. This is a type of error in which only names that are tagged, but whose tags are non-PERSON are considered to be errors. This measure is similar to precision, where the model is only punished for incorrect retrieval. This error rate reports the percentage of names that are tagged as non-PERSON entities among all the names in a certain year.

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

where \emptyset indicates the name is not tagged.

Error Type-2 Weighted. This type of error is similar to Error Type-2 Unweighted; however, here we considered how frequent the mistaken name was while calculating the error.

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

Error Type-3 Unweighted. This is a type of error in which only names that are not tagged are considered to be errors. 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 percentage of names that are not tagged at all among all the names in a certain year.

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

Error Type-3 Weighted. This type of error is similar to Error Type-3 Unweighted; however, we considered how frequent the mistaken name was while calculating the error.

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

²<https://spacy.io>

³<http://www.ssa.gov/oact/babynames/names.zip>

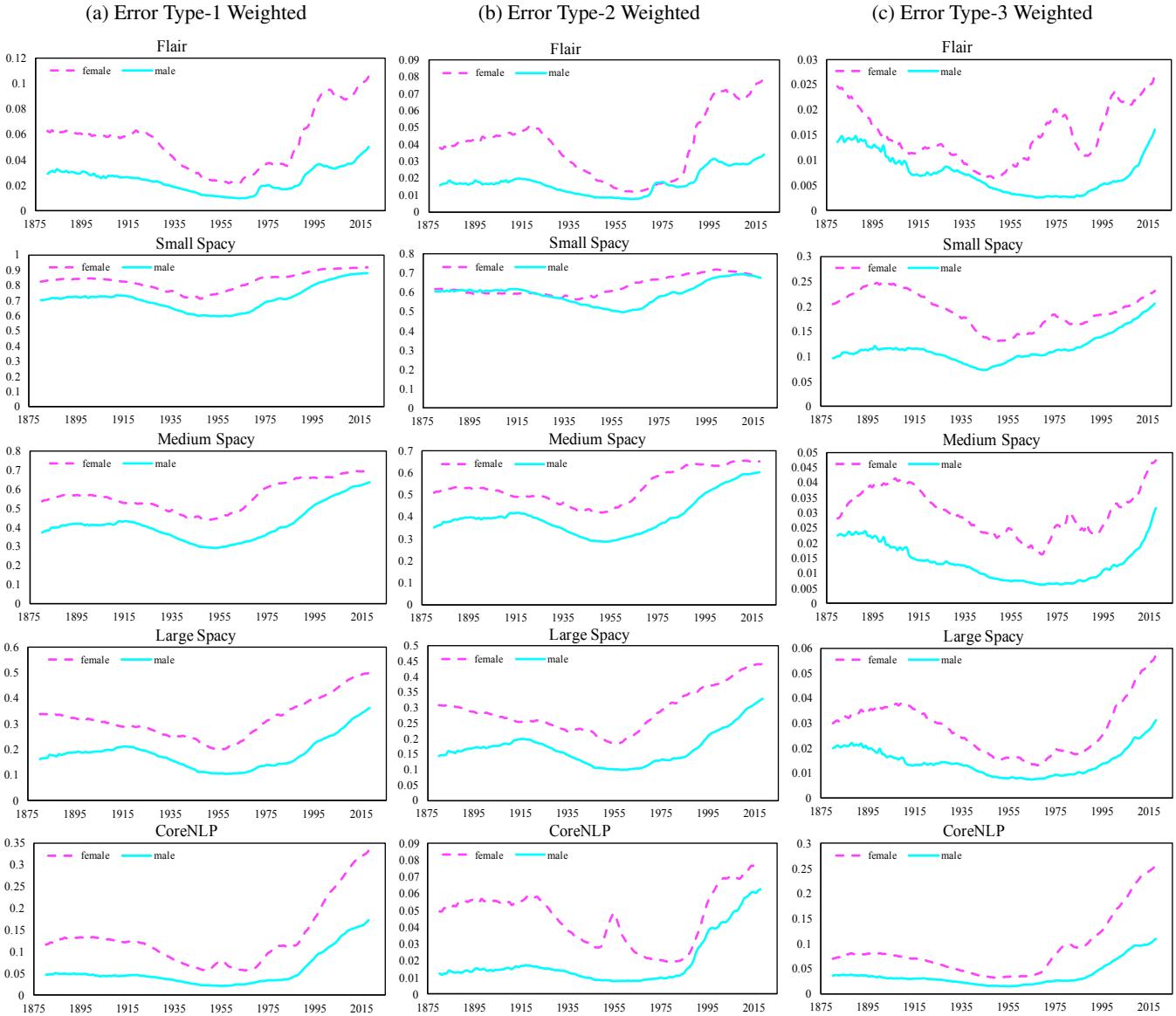


Figure 2: Results from different models that spanned the 139-year history of baby names from the census data on different error types for the weighted cases using template #4. Female names have higher error rates for all the cases. The y axis shows the calculated error rates for each of the error types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

Different types of errors allow for fine-grained analysis into the existence of different biases. Our results indicate that all models are mostly more biased toward female names vs. male names, as shown in Figures 2 and 3 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—which is a more serious type of error to consider. 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 models were mostly more biased toward female

names. We have included results from other templates in our supplementary material to demonstrate that other templates also follow a similar pattern. That being said, in Figure 4 we showed how all the models perform on all the templates, for Error Type-1 Weighted (the super-set error) case, which on its own is showing an interesting trend. This result shows how context helps some models over others by bringing down the error rates when sentences are added to the names (templates #2 through #9). Other models perform better on template #1, showing that context, in fact confuses the model. We observe that in contextual-based models such as Flair, context indeed helps the model make the right de-

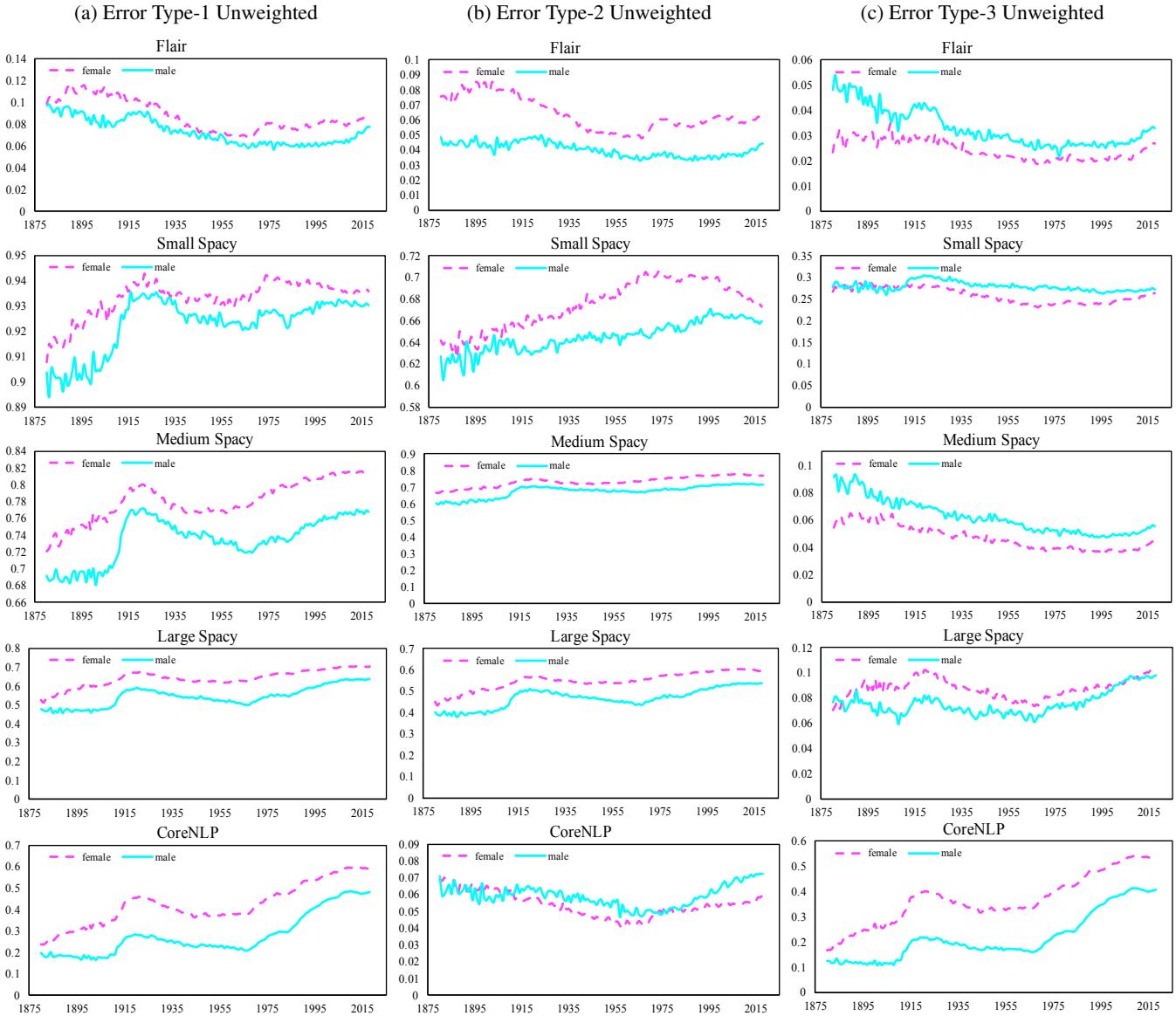


Figure 3: Results from different models over the 139-year history of baby names from the census data on different error types for the unweighted cases using template #4. Female names have higher error rates for all cases except four marginal cases. The y axis shows the calculated error rates for each of the error types as described in their corresponding formulas, and the x axis represents the year in which the baby name was given.

cisions by it having less error rates for templates #2 through #9 compared to template #1. Other models do not necessarily follow this pattern. As an example, we provide the types of names and errors that can happen in these models. We list the top six most frequent male and female names which were tagged erroneously by the Flair model in year 2018 from our benchmark evaluated on template #4 in Table 2.

Model Version Evaluation and Comparison

Updates to models often lead to superior performance; however, this may come with a detriment to fairness and bias. In this section we want to see how much the version updates in

models will be robust toward fairness constraints and errors defined in the previous section. Thus, we will first define this source of bias. Then we will show the results for the models from the previous section that have various versions.

Definition (Version Bias). This is a type of bias that arises from updates in the systems.

In order to report the results of our analysis, we used four of our models mentioned in the previous section that have version updates—namely small, medium, and large models from Spacy (versions 2.0 and 2.1) and versions 3.8 and 3.9 from CoreNLP. We then repeated the experiments from the previous section to report the results for Error Type-1,

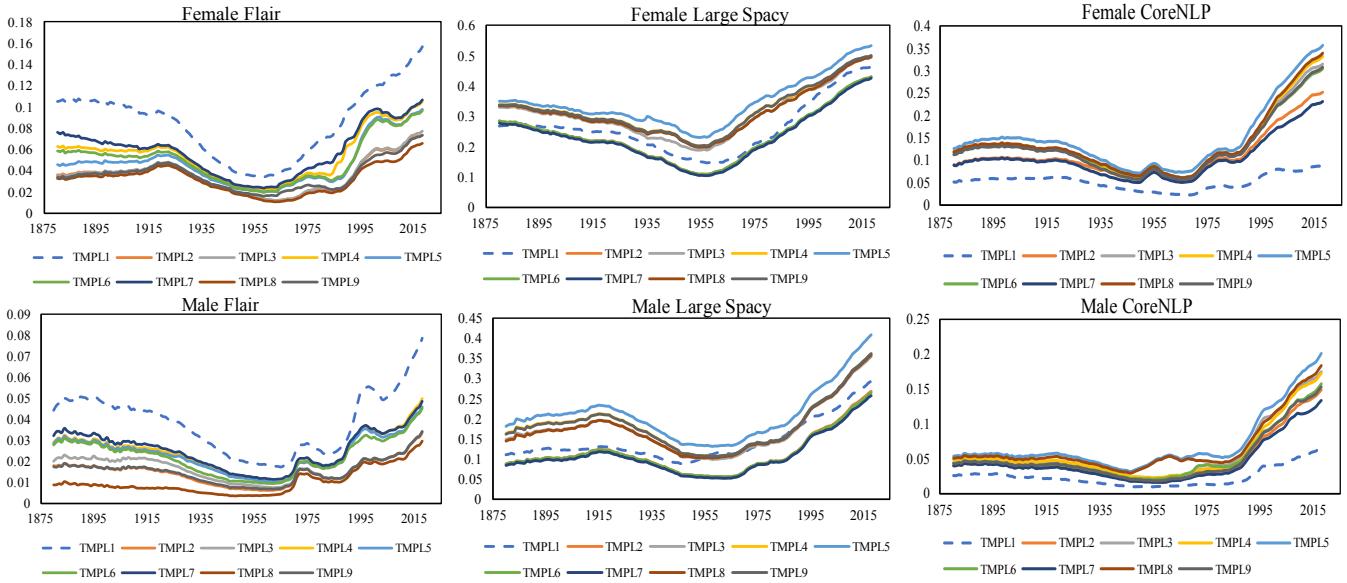


Figure 4: Performance of different models on different templates from our benchmark for female and male names collected for over 139 years. Notice how context in some of the templates helped some models, but showed negative effects on other models.

2, and 3 (Weighted and Unweighted) cases using template #4 from our benchmark. The results for Spacy show that although in not all cases were updated versions worse than that of the previous versions, there were some cases where version updates had serious fairness-related issues. For instance, as shown in Figure 6 for the Spacy medium model, the newer version is more erroneous when it comes to the Error Type-1 Weighted case which is a superset of all the error types discussed in this paper. Not only that, the average increase rate of this error is twice that of female names as opposed to male names when updating the model version. These newer versions may try to boost their performance by trying to tag more entities, which indeed would boost their performance when having other non-PERSON entities in the test set. However, these newer tags may be tagging PERSON entities in the relevant context to the non-PERSON counterpart, which is not desirable when evaluated on our benchmark. The reason that we separated the types of errors and created the benchmark is to show the fine-grained and sensitive issues in these systems. Similarly, for CoreNLP more entities tried to be tagged—which resulted in a slight improvement in Error Type-3, as shown in Figure 5. However, these tags would not correctly assign PERSON tags to PERSON entities, which resulted in a degraded performance in Error Type-2, as shown in Figure 5 and resulted in no change in the overall performance based on Error Type-1 in the newer version update of the CoreNLP model. The results were identical for the unweighted case, and we can see how that degrade in performance was more severe for female names on average. As examples, some of the changes from version 3.8 to version 3.9 of the CoreNLP model are shown in Table 3.

Female Name	Frequency	Error Type
Charlotte	12,940	Tagged as LOC
Sofia	7,621	Tagged as LOC
Victoria	7,089	Tagged as LOC
Madison	7,036	Tagged as LOC
Aurora	4,785	Tagged as LOC
Male Name	Frequency	Error Type
Christian	6,509	Tagged as MISC
Jordan	4,646	Tagged as LOC
Roman	4,364	Tagged as MISC
Kaiden	2,832	Tagged as LOC
King	2,579	Not Tagged

Table 2: Top 5 mistagged examples from the Flair model on Template #4 of female and male names from our benchmark.

Bias in Data

Since data plays an important role in the outcome of the model and can directly affect the fairness constraints if it contains any biases, we decided to analyze some of the datasets that are widely used in the training of NER models to determine whether they show any biases toward a specific group that could result in the biased behavior observed in those results discussed in previous sections. We used the train, test, and development sets from two widely known CoNLL-2003⁴ (Sang and De Meulder 2003) and

⁴<https://www.clips.uantwerpen.be/conll2003/ner/>

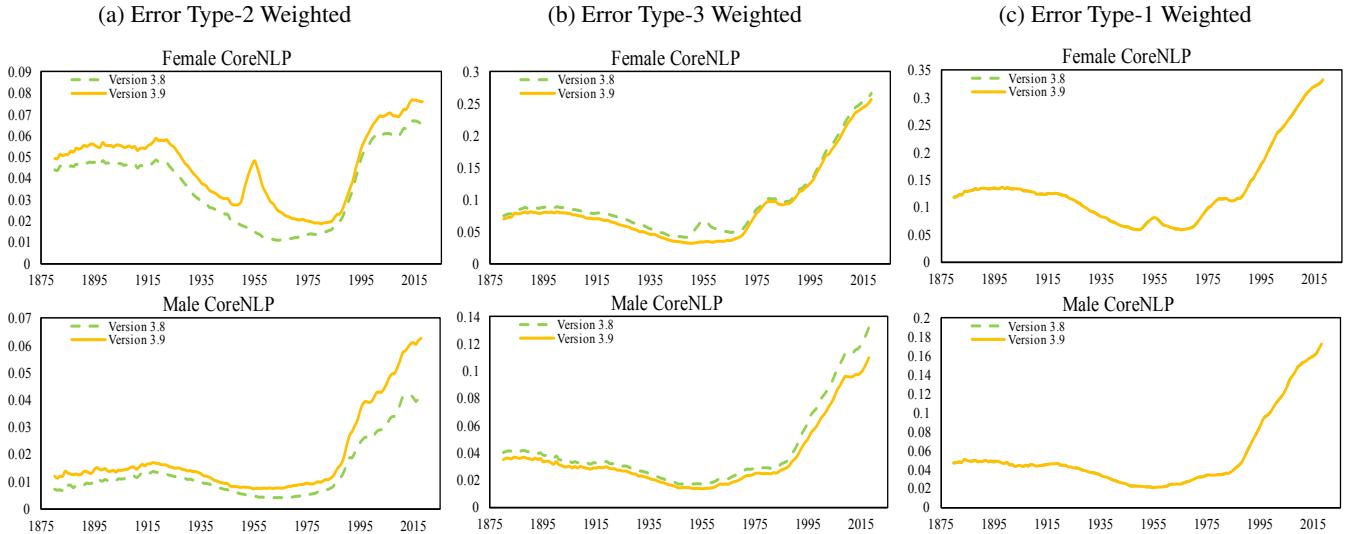


Figure 5: Version update in the CoreNLP model tried to tag more entities and thus a subtle boost in performance with regard to Error Type-3. However, this resulted in more PERSON entities being tagged as non-PERSON entities. This then degraded performance with regard to Error Type-2 in the newer version, and overall no change in the Error Type-1 case which is considered to be the super-set. We observe how the degrade in Error Type-2 affected more female names than males on average.

OntoNotes-5⁵ (Weischedel et al. 2012) datasets which were used in the training and testing of Flair, Spacy, and many other models. The split of the OntoNotes-5 dataset into train, development, and test sets was performed according to (Pradhan et al. 2013). We reported the percentages of male vs. female names from the census data that appeared in train, test, and development sets in each of the datasets and compared this to the percentages of male vs. female names in reality from the census data to see how much these datasets are reflective of the reality or if they pertain to any bias toward a specific gender group.

Our results shown in Table 4 indicate that the datasets used do not reflect the real world, but rather exactly the opposite of that. Unlike the census data, which is representative of real-world statistics, wherein female names have more versatility—62% unique names vs. 38% unique male names—datasets used in training the NER models contain 42% female names vs. 58% male names from the census data. Not only do the datasets not contain more versatile female names to reflect the reality, but instead have even less variety which can itself bias the models by not covering enough female names. Similar patterns are observable in test and development sets of datasets used in the NER systems.

Related Work

We have seen large amounts of attention and work regarding fairness in machine learning and natural language processing models and methods. Recent papers observed a type of stereotyping bias in word embedding methods and tried to mitigate this type of bias by proposing a method that respects the embeddings for gender-specific words but de-

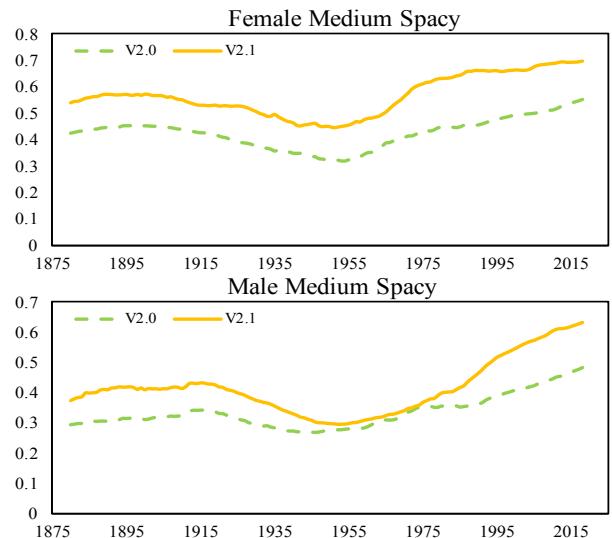


Figure 6: Biased performance of version 2.1 over 2.0 in Spacy medium. The bias against female names was on average twice that of male names.

bias embeddings for gender-neutral words (Bolukbasi et al. 2016), or by generating a gender-neutral version of Glove (called GN-Glove) that aims to preserve gender information in some directions of word vectors, while setting other dimensions free from gender influence (Zhao et al. 2018b) or other data augmentation techniques (Brunet et al. 2019; Zhao et al. 2019). Other work tried to show and address bias in co-reference resolution (Zhao et al. 2018a), semantic role labeling (Zhao et al. 2017), machine translation (Font and Costa-jussà 2019), language models (Bordia and Bowman

⁵<https://catalog.ldc.upenn.edu/LDC2013T19>

Female Name	CoreNLP version 3.8	CoreNLP version 3.9
Isabel	Not Tagged	Tagged as MISC
Angelina	Not Tagged	Tagged as MISC
June	Not Tagged	Tagged as DATE
Charlotte	Tagged as LOCATION	Tagged as CITY
Victoria	Tagged as LOCATION	Tagged as CITY
Sydney	Tagged as LOCATION	Tagged as CITY
Male Name	CoreNLP version 3.8	CoreNLP version 3.9
Christian	Not Tagged	Tagged as RELIGION
Logan	Tagged as LOCATION	Tagged as CITY
Roman	Not Tagged	Tagged as MISC
Santiago	Tagged as LOCATION	Tagged as CITY
Jordan	Tagged as LOCATION	Tagged as CITY
Messiah	Not Tagged	Tagged as TITLE

Table 3: Some examples on how tagging changed during version update of the CoreNLP model. Note how the original problem of tagging PERSON entities correctly has not been addressed.

2019), and sentence embedding (May et al. 2019).

Addressing fairness and bias, not only in NLP but also in general machine learning, has lately gained much attention. In Mehrabi et al. (2019b), the authors created a taxonomy on fairness and bias that discusses how researchers have addressed fairness related issues in different fields. From representation learning (Moyer et al. 2018) to graph embedding (Bose and Hamilton 2019) to community detection (Mehrabi et al. 2019a) and clustering (Backurs et al. 2019), researchers have studied biases in these areas and tried to address them by pointing out the observed problems and proposing new directions and ideas. In Buolamwini and Gebru (2018) authors show and analyze the existing gender bias in facial recognition systems, such as those used by IBM, Microsoft, and Face++, and created a benchmark for better evaluation of bias in facial recognition systems. This is considered a significant contribution as it opens many future research questions and related papers. Paying attention to different AI applications and pointing out their issues in terms of fairness is an important issue that needs serious attention for significant future improvements to these systems.

Conclusion and Future Work

In this work we not only performed a historical analysis of named entity recognition systems and showed the existence of bias, but we also introduced a benchmark that can help future models evaluate the extent of gender bias in their systems. We then performed a cross version analysis of models and showed that model updates can sometimes amplify the existing bias in previous versions. We also analyzed some datasets widely used in current state-of-the-art models and showed the existence of bias in these datasets as well which

Dataset	Female Count	Male Count	Female Pct	Male Pct
Census	67,698	41,475	62%	38%
Train	CoNLL 2003	1,810	2,506	42% 58%
	OntoNotes5	2,758	3,832	42% 58%
Dev	CoNLL 2003	962	1,311	42% 58%
	OntoNotes5	1,159	1,524	43% 57%
Test	CoNLL 2003	879	1,228	42% 58%
	OntoNotes5	828	1,068	44% 56%

Table 4: Percentage of female and male names from the census data appearing in CoNLL 2003 and OntoNotes datasets with their corresponding counts. Both datasets fail to reflect the variety of female names.

can directly affect the biased performance of models. Named entity recognition systems are extensively used in different downstream tasks and having biased NER systems can have implications beyond just the NER task. We believe that using our benchmark for evaluation of future named entity recognition systems can help mitigate the gender bias issue in these applications.

This work identifies an important problem with the current state-of-the-art in named entity recognition. Nevertheless, this measure is a glimpse into the many possible biases that NER may contain, and there are some key limitations that we plan to address in future work. First, the nine templates used to test the models are not necessarily representative of real-world text. There is a limitless supply of sentences that could be fed to the model. Moving forward, we seek to generate a sentence corpora that is based on real-world text. Second, our approach is based upon names taken from United States census data. This work can be extended to different languages to demonstrate the biases they pertain.

Through our analysis, we provide some suggestions for avoiding gender bias in NER systems, as listed below:

1. Using benchmarks designed for evaluation of bias in future named entity recognition systems can help mitigate the gender bias issue in these applications.
2. Using contextual-based models as shown in our results can help reduce the error rates.
3. We encourage introduction of new models to overcome the observed gender bias in these systems. As future work, these benchmarks can be used as loss functions to train the NER system.
4. A collection of new datasets reflecting reality, such as following name distributions observed in the real-world, would be helpful to the community to build models that avoid these biases. As a step towards this, we will provide our code and data upon acceptance.

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under

Agreement No. HR0011890019. We would want to thank Kai-Wei Chang and Jieyu Zhao for their constructive feedbacks.

References

- Akbik, A.; Bergmann, T.; Blythe, D.; Rasul, K.; Schweter, S.; and Vollgraf, R. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59. Minneapolis, Minnesota: Association for Computational Linguistics.
- Akbik, A.; Blythe, D.; and Vollgraf, R. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, 1638–1649.
- Backurs, A.; Indyk, P.; Onak, K.; Schieber, B.; Vakilian, A.; and Wagner, T. 2019. Scalable fair clustering. In Chaudhuri, K., and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 405–413. Long Beach, California, USA: PMLR.
- Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, 4349–4357.
- Bordia, S., and Bowman, S. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 7–15.
- Bose, A., and Hamilton, W. 2019. Compositional fairness constraints for graph embeddings. In *International Conference on Machine Learning*, 715–724.
- Brunet, M.-E.; Alkalay-Houlihan, C.; Anderson, A.; and Zemel, R. 2019. Understanding the origins of bias in word embeddings. In Chaudhuri, K., and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 803–811. Long Beach, California, USA: PMLR.
- Buolamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91.
- Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS ’12, 214–226. New York, NY, USA: ACM.
- Finkel, J. R.; Grenager, T.; and Manning, C. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL ’05, 363–370. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Font, J. E., and Costa-jussà, M. R. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.
- Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. 4066–4076.
- Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S. J.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, 55–60.
- May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.
- Mehrabi, N.; Morstatter, F.; Peng, N.; and Galstyan, A. 2019a. Debiasing community detection: The importance of lowly-connected nodes. *arXiv preprint arXiv:1903.08136*.
- Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2019b. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.
- Moyer, D.; Gao, S.; Brekelmans, R.; Galstyan, A.; and Ver Steeg, G. 2018. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, 9084–9093.
- Pradhan, S.; Moschitti, A.; Xue, N.; Ng, H. T.; Björkelund, A.; Uryupina, O.; Zhang, Y.; and Zhong, Z. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 143–152.
- Sang, E. F., and De Meulder, F. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.
- Weischedel, R.; Pradhan, S.; Ramshaw, L.; Kaufman, J.; Franchini, M.; El-Bachouti, M.; Xue, N.; Palmer, M.; Hwang, J. D.; Bonial, C.; et al. 2012. Ontonotes release 5.0.
- Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.
- Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods.
- Zhao, J.; Zhou, Y.; Li, Z.; Wang, W.; and Chang, K.-W. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4847–4853.
- Zhao, J.; Wang, T.; Yatskar, M.; Cotterell, R.; Ordonez, V.; and Chang, K.-W. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 629–634.

Appendices

Weighted Results from Template # 5

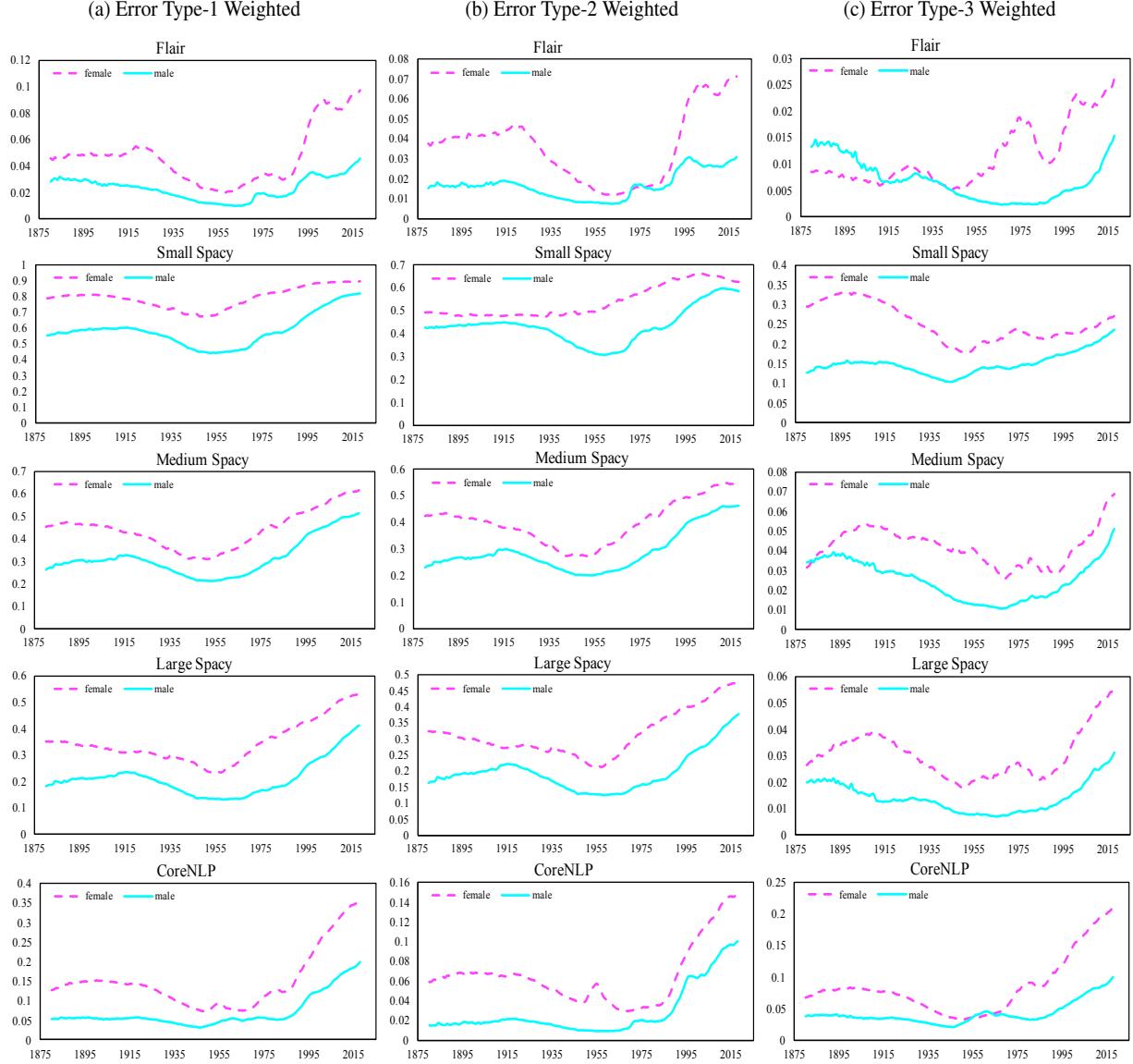


Figure 8: Results from different models ran over 139-year history of baby names from the census data on different error types for the weighted case using template # 5. Female names have higher error rates for all the cases. The y axis shows the calculated error rates for each of the error types as described in their corresponding formulas, and the x axis represents the year in which the baby name was given.

Unweighted Results from Template # 5

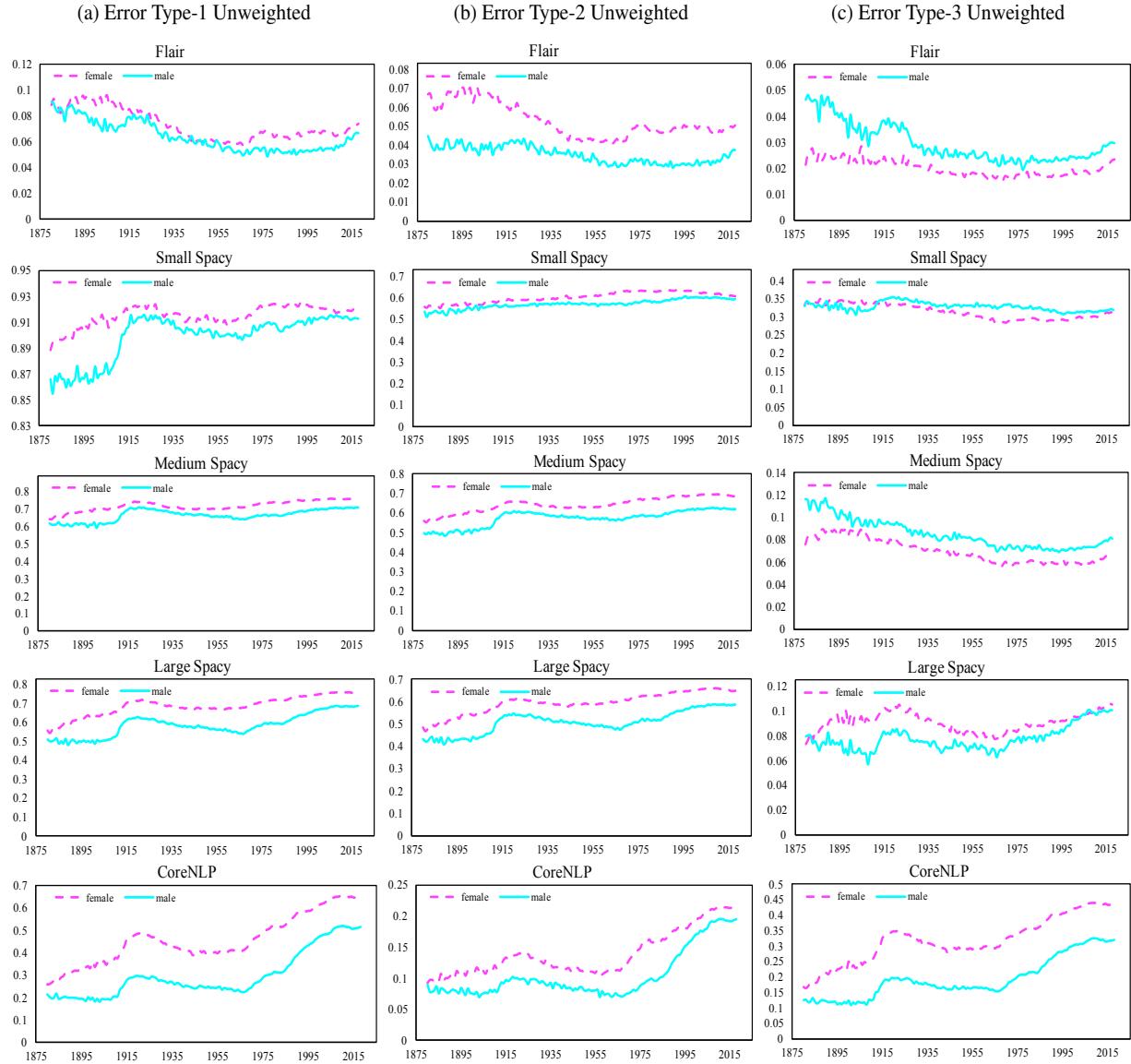


Figure 9: Results from different models ran over 139-year history of baby names from the census data on different error types for the unweighted case using template #5. Female names have higher error rates for all the cases except three marginal cases. Overall performance is more biased towards female names. 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.