Disambiguation of morpho-syntactic features of African American English – the case of habitual *be*

Harrison Santiago
Department of Computer and Information Science and Engineering, University of Florida
harrison.santiag@ufl.edu

Joshua L. Martin
Department of Linguistics, University of Florida
joshua.martin@ufl.edu

Sarah Moeller
Department of Linguistics, University of Florida
smoeller@ufl.edu

Kevin Tang
Department of English and American Studies, Heinrich-Heine-University, Düsseldorf
kevin.tang@hhu.de

Abstract

Recent research has highlighted that natural language processing (NLP) systems exhibit a bias against African American speakers. The bias errors are often caused by poor representation of linguistic features unique to African American English (AAE), due to the relatively low probability of occurrence of many such features in training data. We present a workflow to overcome such bias in the case of habitual "be". Habitual "be" is isomorphic, and therefore ambiguous, with other forms of "be" found in both AAE and other varieties of English. This creates a clear challenge for bias in NLP technologies. To overcome the scarcity, we employ a combination of rule-based filters and data augmentation that generate a corpus balanced between habitual and non-habitual instances. With this balanced corpus, we train unbiased machine learning classifiers, as demonstrated on a corpus of AAE transcribed texts, achieving .65 F_1 score disambiguating habitual "be".

1 Introduction

Linguistic discrimination has adversely affected the lives of marginalized populations for centuries, including racially marginalized groups in the United States. In spite of extensive research on linguistic discrimination (Baugh, 2008), many NLP systems inherit the linguistic biases that exist between humans. For example, preliminary studies into the performance of automatic speech recognition (ASR) systems uncovered a performance bias against African American speakers (Tatman and Kasten, 2017; Dorn, 2019). This problem was confirmed most recently by Koenecke et al. (2020) who found that the average word error rate (WER) for white American speakers was significantly lower as compared to the average WER for African American speakers among five prominent ASR systems from such companies as Google, Amazon, and Apple.

This performance gap is rooted in two related issues. First, the linguistic differences between African American English (AAE) and General American English (GAE) include distinctive features in their morphosyntactic structures. Second, incorrect inferences in NLP systems are often caused by the scarcity of certain linguistic features when training, and the many unique features in AAE have a relatively low probability of occurrence.

This paper describes work that overcomes the data scarcity issue for a specific feature unique to AAE: the habitual "be". As the name suggests, this morphologically invariant form of "be" communicates habitual action. Disambiguating habitual "be" from non-habitual "be" is difficult for two prominent reasons. First, the form is isomorphic with the other uses of "be", such as the infinite use in "I want to be...". Second, habitual "be" is relatively rare even in corpora of AAE. Our work addresses both these issues. It uses a rule-based method that capitalizes on morphosyntactic differences to eliminate a portion of non-habitual "be" instances and it uses a method of data augmentation that increases the ratio of habitual "be" instances. The resulting balanced data can then be used to train classifiers to tag "be" instances as habitual or non-habitual.[1]

2 Related work

Distinguishing habitual "be" and non-habitual "be" usage is a word sense disambiguation (WSD) prob-

[1] https://github.com/HarrisonSantiago/Habitual_be_classifier

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 70 - 75

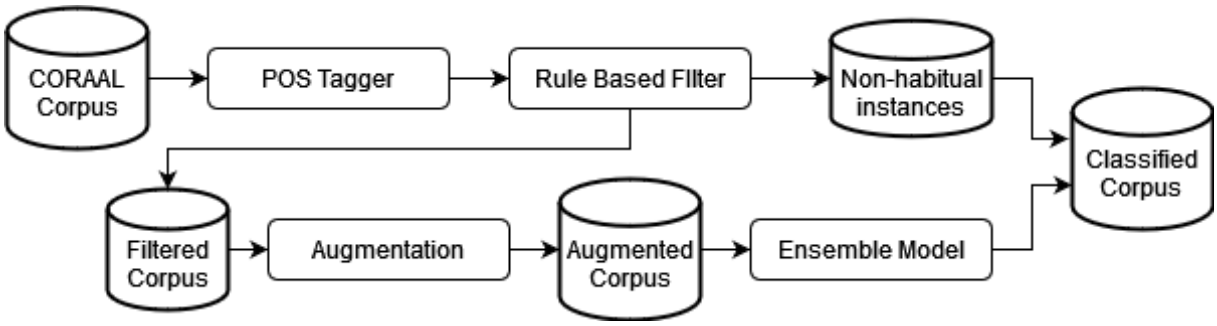

Figure 1: The disambiguation pipeline: the input corpus goes through a Part-of-Speech tagger, after which non-habitual instances are separated by a rule-based filter. Any indeterminate "be" instances are balanced by augmentation and tagged by classification models.

lem because it involves identifying the meaning of words in context (Navigli, 2009). Most successful WSD algorithms make use of contextual embeddings (Melamud et al., 2016; Peters et al., 2018), but some feature extraction algorithms, such as the IMS algorithm by Zhong and Ng (2010), have a comparable level of performance although comparatively much simpler. The IMS algorithm uses a support-vector-machine (SVM) with simple contextual features, such as word form or part-of-speech (POS) tags, and weighted average of embeddings. Similarly, our disambiguation pipeline makes use of the POS tags of the surrounding words. This helps avoid the limited amount of annotated AAE data which could lead to sparse word vectors and unreliable embeddings.

Data augmentation techniques that generate synthetic, or artificial, language in the training data often improve NLP applications when the training corpus is small or when a certain feature occurs rarely (Chen et al., 2021). Our approach follows previously successful examples of data augmentation methods that combine a language model (Fadaee et al., 2017) with a thesaurus (Zhang et al., 2015) or word embeddings (Wu et al., 2019). These methods identify substitutes for words in the data and insert them into synthetic strings that include the target feature.

3 Habitual "be"

The "be" verb has various functions. This includes several types of non-habitual use, as shown in Appendix A. The use of habitual "be" is a prominent, distinct, and well-researched morphosyntactic feature in AAE. Habitual "be" is a morphologically invariant form of the verb that encodes the habitual aspect, as shown below (Green, 2002).

1. I **be** in my office by 7:30. (habitual: AAE)

2. I **am usually** in my office by 7:30. (habitual: GAE)

Syntactic contexts serve as important cues for disambiguating "be" as habitual or non-habitual. Martin and Tang (2020) show that ASR systems not only fail to recognize habitual "be" more often than non-habitual "be" but, when habitual "be" is present in an utterance, the surrounding words are also incorrectly recognized, particularly preceding words. These findings reveal a strong dependency between habitual "be" and its syntactic context. Failure to reflect this dependency in a language model could lead to a less accurate and biased system.

Even in an AAE corpus, habitual "be" is relatively rare. This imbalanced distribution poses a challenge for designing a non-biased NLP system because most classifiers tend to be biased towards the majority class.

The ambiguity and scarcity of habitual "be" presents two obvious approaches to a solution: (i) incorporate more habitual "be" instances in the data, (ii) manually disambiguate habitual and non-habitual "be" before training. Each approach poses a challenge. For (i), simply collecting more data is extremely impractical, as the habitual "be" is naturally rare. For (ii), hand-coding is unsuitable for the scale of the data needed.

Our study addresses these challenges with a rule-based filter based on syntactic cues and with a data augmentation technique. Together the filter and data augmentation increase the ratio of habitual "be", providing a more balanced training set for the model and allowing for a more fine-grained language model.

4 Methodology

The first novel task towards training classifiers to disambiguate habitual "be" is to address the ambiguity of the invariant form by eliminating as many non-habitual "be" instances as possible. The second task is to increase the proportional occurrence of habitual "be" in the training data.

We undertake these two tasks and incorporate them into a pipeline, shown in Figure 1. First, the entire corpus is run through a pre-trained `NLTK` tokenizer and POS tagger trained using the Penn Treebank Project. To eliminate as many non-habitual "be" instances as possible, a rule-based filter identifies determinate instances of non-habitual "be". With these removed, we increase the proportional occurrence of habitual "be" by augmenting the proportion of habitual "be". Finally, we combine the filtered habitual "be" instances back into a now balanced dataset and use that dataset to train an ensemble model for classification. As discussed in section 5.1, the habituality of each instance is known and allows accurately creating rules and training the classifiers.

4.1 Preprocessing

The data is formatted using WordSmith Tools (Scott, 2020) so that each instance of "be" is centered in a 102-character string, the length being determined by the software default. To simplify the task, no breaks between speakers or texts were included, meaning these text segments combine speech from multiple speakers and texts if necessary, with no indication as to where this occurs. If multiple instances of "be" fall within 102 characters, each instance is treated as separate instance that becomes the center of another string slightly offset from the overlapping example. Also, all punctuation, marks made by transcribers (e.g., "/??/"), corpus-specific codes (e.g., "/RD-NAME-3/") and other non-speech text are removed as part of the preprocessing.

4.2 Rule-based filter of non-habitual "be"

In AAE, there are certain syntactic patterns that strongly correlate to occurences of the habitual "be" (Green, 2002; Fasold, 1972). Most patterns are based on the part-of-speech immediately surrounding "be". Two example patterns are a pronoun immediately preceding "be" (e.g., "...*they* **be** like, what you finna do?") and a verb ending in -ing immediately following "be" (e.g., "But LeBron **be** *passing* though").

Following from this, we invert some patterns and create filters that capture a large number of non-habitual instances. For example, if the word that precedes "be" is not a pronoun and the word after it is not a verb ending in -ing, then we can say that instance is non-habitual.

The vast majority of non-habitual "be" instances are caught by these syntactic rules. In addition, we created some ad-hoc rules that showed success at eliminating remaining non-habitual "be", although they generally capture a smaller number. A full list of our rules we can be found in Appendix B.

The goal of the rule-based filter is not to identify instances of habitual "be". Rather, it is used to remove non-habitual "be" instances for which more advanced disambiguation techniques are not needed. This is a step towards creating a more balanced corpus. It serves to narrow the scope of our classifier to those instances which much more difficult to be automatically disambiguated.

4.3 Augmenting habitual "be"

To counter the relative rarity of habitual "be", the dataset needs to be balanced, but without excluding the remaining non-habitual instances after the rule-based filter is applied. Instead, the amount of habitual "be" can be increased. To accomplish this, we use data augmentation to create new, synthetic examples of habitual "be".

We found that the Python library `nlpaug` (Ma, 2019) provides easy synthetic text generation. Focusing on text augmentation, we used the Word2Vec (Mikolov et al., 2013)[2] and WordNet (Fellbaum, 1998) implementations for substituting and inserting words in surrounding examples of habitual "be" instances from our corpus. The Word2Vec implementation both substitutes and inserts new words at random by finding similar words using the cosine distance from pre-trained embeddings. The WordNet augmentation leverages a database of semantic relations to substitute synonyms at random. These methods can occasionally lead to ungrammatical outputs, as seen in Appendix C. We did not remove such occurrences, as the inclusion of all generated perturbations in our data set strengthened the robustness of our model. Combined, these methods inserted or replaced words with a new part of speech in over 90% of the augmentations.

[2]https://github.com/dav/word2vec

4.4 Classifiers

After filtering trivial instances of non-habitual "be" and balancing the remaining data by augmenting instances of habitual "be", we train a logistic regression classifier, a multi-layer perceptron (MLP), and a linear Support Vector Machine (SVM) to disambiguate instances of "be". All are implemented with the `scikit-learn` library. All models set the max-iteration to 10,000 steps to allow for convergence on a regular basis. The MLP was changed to use a limited-memory BFGS algorithm solver, and set to have two hidden layers, the first with five nodes and the second with two. These hyperparameters were set after a non-exhaustive search of looking for the optimal settings. All other default parameters were kept unchanged. We compared these against a majority-rules ensemble model that uses the logistic regression, MLP, and SVM voting algorithms. The votes are equally weighted between all three.

The input to all of the classifiers consists of vectors which contain the number of times each POS occurs within a window around each instance of "be". We treated the size of this window as a hyperparameter, and found that defining our window to start at the 9th word in the string and end at the 5th-from-last word produced optimal results.

5 Experiment

Unbiased NLP systems should successfully disambiguate instances of habitual "be". We implemented our system on a corpus of AAE speakers after training it our filtered and balanced corpus.

5.1 Data

The data comes from the Corpus of Regional African American Language (CORAAL) (Kendall and Farrington, 2018) which contains transcriptions of over 150 sociolinguistic interviews with African American speakers, totaling more than 127 hours of audio and including a rich variety of interviewees by age, socio-economic background, gender identity, and urban/rural origin.

From this corpus, 5,133 instances of "be" were manually annotated as habitual/non-habitual. This resulted in 477 instances of habitual "be" and, 4,656 instances of non-habitual "be", which is to say that non-habitual instances were approximately ten times more frequent. The rule-based filter and augmentation were applied to this data with the resulting statistics shown in Table 1. The rule-based filter incorrectly eliminated 61 instances of habitual "be", reducing the total from 477 to 416. This means the filter has an error rate of about 13% that might be improved with additional ad-hoc rules.

	Orig.	Filter	Augment
Non-hab "be" total	4,656	994	944
Hab "be" total	477	416	963
Hab 'be" %	9%	30%	50%

Table 1: The distribution of habitual "be" in the training corpus: original, rule-based filtered, and augmented. The top two rows show the change in the raw number of "be" instances; the bottom shows the proportion of habitual "be" to non-habitual "be".

When analyzing our classifiers, we used a 70/30 training/test split, with the test set having a ratio of non-habitual to habitual occurrences similar to that of the original corpus. Importantly, the dataset was split before any augmentation occurred to help our results be more transferable to the original corpus. To get a better understanding of the consistency in results that the augmentation methods would lead to, we re-performed our augmentation procedure for each trial. In total, 10 trials were performed.

5.2 Results

Based on our results on the CORAAL corpus, classifying habitual "be" is a feasible task even with a limited supply of natural AAE speech for training. Each algorithm and the ensemble model were tested after being trained on the filtered and the augmented data and on the original corpus. Table 2 shows F_1-scores displays the comparison, showing means and standard deviations over 10 trials. The best results were achieved by the ensemble classifier after both filtering and augmenting. Over 10 trials the ensemble model classified instances of habitual "be" with an average score of 0.65.

All four classifiers' performance rose dramatically when using our filtering and augmentation methods. In addition, the variability in classifier performance decreased after filtering and augmentation, as evident by the lower standard deviations. The lower variability indicates that balancing a data set allowed the classifiers to find a more definitive decision boundary.

6 Conclusion

Our goal was to develop a pipeline which aids the creation of models unbiased against African American English. We proposed and tested a combi-

	Augmented	Not Augmented
Logistic regression	0.648 (0.048)	0.416 (0.039)
SVM	0.628 (0.114)	0.542 (0.206)
MLP	0.627 (0.038)	0.498 (0.058)
Ensemble	**0.652** (0.049)	0.439 (0.084)

Table 2: F_1-scores for different classification algorithms (Logistic regression, Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Ensemble of all three). The mean over 10 trials are reported, with the standard deviation in parentheses.

nation of hand-crafted rules, data augmentation, and machine learning to disambiguate instances of habitual "be" which is a distinct, if relatively infrequent, morphosyntactic feature in AAE. The results show this combination to be a promising pipeline, with each step contributing to success at increasing classification scores and reducing bias.

The hand-crafted rules we used took into consideration morphosyntactic patterns that are unique to AAE and correlate with habitual "be" usage. This allowed us to filter out most non-habitual "be" instances. We then found that Word2Vec and WordNet augmentation methods were able to adequately imitate AAE structure and balance the proportion of habitual "be" instances. Together the filtering and the augmentation resulted in more balanced data with which to train the classifiers.

In the future, with an increased amount of natural speech and more advanced classification algorithms, it is possible that the classification performance could be even higher. However, due to limited data, we treated the entire CORAAL corpus without regard to several interesting factors that should be considered. For example, we did not regard the geographic location or origin of the speaker. Further analysis of our model's performance with respect to regional sub-varieties of AAE would be an interesting avenue to explore. This exploration might refine the hand-crafted rules. Also, our pipeline makes use of the POS tags of the surrounding words, similar to (Zhong and Ng, 2010), but it does not include the surrounding words themselves or their embeddings as features because the limited data would have led to sparse word vectors and unreliable embeddings.

We feel it should be easy to adapt our pipeline to other unique AAE features such as the completive "done" (Green, 2002). Although we expect feature-based models to tend to perform better at low-resource settings than deep learning, we plan to compare our results against state-of-the-art neural models such as the Transformer (Vaswani et al., 2017).

The increase in scores we were able to achieve with these simple methods serves as a proof-of-concept that systems based on similar syntactic filtering and data augmentation approaches have the potential to improve the performance of other AAE-focused NLP systems and provide enough data for more advanced feature representations.

References

John Baugh. 2008. Linguistic discrimination. In *Kontaktlinguistik*, pages 709–714. De Gruyter Mouton.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2021. An empirical survey of data augmentation for limited data learning in nlp.

Rachel Dorn. 2019. Dialect-specific models for automatic speech recognition of african american vernacular english. In *Student Research Workshop*, pages 16–20.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Ralph W. Fasold. 1972. *Tense Marking in Black English: A linguistic and social analysis*. Number 8 in Urban Language Series. Center for Applied Linguistics, Arlington, VA.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Lisa J Green. 2002. *African American English: a linguistic introduction*. Cambridge University Press.

Tyler Kendall and Charlie Farrington. 2018. The corpus of regional African American language. *Version*, 6:1.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Joshua L. Martin and Kevin Tang. 2020. Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be". In *Proc. Interspeech 2020*, pages 626–630.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pages 51–61.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

M. Scott. 2020. Wordsmith tools version 8. Stroud: Lexical Analysis Software.

Rachael Tatman and Conner Kasten. 2017. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *INTERSPEECH*, pages 934–938.

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 *Advances in neural information processing systems*, pages 5998–6008.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Computational Science – ICCS 2019*, pages 84–95, Cham. Springer International Publishing.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden. Association for Computational Linguistics.

A Appendix: Types of non-habitual "be"

- auxiliary "be" in progressive constructions (e.g., "I will **be** going there tomorrow.")
- auxiliary "be" in passive constructions (e.g., "She should **be** given an award.")
- copula or auxiliary "be" preceded by verbal complements (e.g., "He wanted to **be** a lawyer.")
- copula or auxiliary "be" preceded by a modal (e.g., "They might **be** in the house.")
- imperative "be" (e.g., "**Be** quiet!")

B Rules to filter non-habitual "be"

- If the word immediately preceding "be" is a modal, adjective, or "to".
- If the word immediately following "be" is a verbal noun, while the word immediately preceding is not a personal pronoun nor a noun.
- If the word immediately following "be" is an adjective, while the word immediately preceding "be" is not a personal pronoun nor a noun.
- If the word immediately following "be" is a preposition or subordinating conjunction, while the word immediately preceding "be" is a singular present verb.
- If the word immediately preceding "be" is a noun, and the word immediately preceding that noun is an adjective
- If the word immediately preceding "be" is an adverb, and the word immediately following "be" is either a personal pronoun or determiner.
- If the word immediately preceding "be" is an adverb, and either the word immediately preceding the adverb is a verb, or modal

C Examples of augmenting occurrences of the habitual "be"

- "they were like you should totally come here we be having so much fun So I tell my mom about it and" becomes "they were like you should totally come hither we be have got so much fun So I tell my mom astir it and"
- "mixed up all kinds a way everybody just just be there having a good time That s Mm hm that s" becomes "mixed up all dizzying array a way everybody yeah just be happen having a heckuva time That s hm that s"