Now, It's Personal : The Need for Personalized Word Sense Disambiguation

Milton King
Faculty of Computer Science
University of New Brunswick
milton.king@unb.ca

Paul Cook
Faculty of Computer Science
University of New Brunswick
paul.cook@unb.ca

Abstract

Authors of text tend to predominantly use a single sense for a lemma that can differ among different authors. This might not be captured with an author-agnostic word sense disambiguation (WSD) model that was trained on multiple authors. Our work finds that WordNet's first senses, the predominant senses of our dataset's genre, and the predominant senses of an author can all be different and therefore, author-agnostic models could perform well over the entire dataset, but poorly on individual authors. In this work, we explore methods for personalizing WSD models by tailoring existing state-of-the-art models toward an individual by exploiting the author's sense distributions. We propose a novel WSD dataset and show that personalizing a WSD system with knowledge of an author's sense distributions or predominant senses can greatly increase its performance.

1 Introduction

Authors of text tend to predominantly use a single sense for a lemma that can differ among different authors (Gella et al., 2014). This might not be captured with an author-agnostic word sense disambiguation (WSD) model that is trained on multiple authors. Our work finds that WordNet's first senses, the predominant senses of our dataset's genre, and the predominant senses of an author can all be different and therefore, author-agnostic models could perform well over an entire dataset, but poorly on individual authors. Ideally, each author would have access to a personalized WSD model, which is a model that was tailored toward that individual. In this work, we explore methods for personalizing WSD models by tailoring existing state-of-the-art models toward an individual by exploiting the author's sense distributions. We evaluate our models on our proposed dataset which contains 1586 sense-annotated instances for 11 lemmas across 36 authors. Our evaluation includes metrics that focus on the performance of models with respect to the authors that they perform poorly on, which highlights the potential gain for individual authors.

The most similar work to this is presented in Gella et al. (2014), which created a dataset that contains sense-annotated instances of tweets for a list of authors. We differ from them by including more annotated instances from a single author and by using text from blog posts, which do not have a limitation on the length of text, unlike text from tweets. We evaluate different WSD models, including a state-of-the-art model (*SensEmBERT*) (Scarlini et al., 2020), which we extend with personalization techniques to achieve our best scores. We personalize *SensEmBERT* with knowledge of an author's sense distributions or predominant senses, which outperform author-agnostic WSD systems. Our work also shows that the use of author-specific sense distributions outperforms the use of genre-specific senses. In this first work on personalized WSD, we do not automatically learn the sense distributions of an author, but instead, we use the author's true sense distributions to demonstrate the importance of learning author-level sense distributions when considering personalized WSD.

2 Related Work

Two common frameworks for WSD systems include knowledge-based, which utilizes information contained in a sense inventory, and supervised, which involves training on annotated instances. In this work, we focus on knowledge-based models because they do not require annotated instances and they are able to work well on less frequent lemmas and senses (Scarlini et al., 2020). Many modern knowledge-based methods extend the simplified Lesk method (Kilgarriff and Rosenzweig, 2000), which involves classifying a token with the sense that contains the most overlapping words in its defi-

Proceedings of Recent Advances in Natural Language Processing, pages 692–700
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_079

nition with the context of the word being classified. WSD methods that involved a Lesk-style approach are Banerjee and Pedersen (2002), which included looking at the definition of words that are similar to the token being classified according to WordNet and Basile et al. (2014), which looked at the definition of similar words but added a vector representation of the target token — generated by a topic model — into their similarity calculations. Topic modeling is a probabilistic model that views documents as a distribution over topics and topics as a distribution over types (Blei et al., 2003). Topic modeling was also applied to WSD in Boyd-Graber et al. (2007) and Li et al. (2010).

The current knowledge-based model, that achieves state-of-the-art performance on the unified WSD evaluation framework (Raganato et al., 2017), creates an embedding for each sense of a word by using BERT (Devlin et al., 2019) to embed both the sentences in Wikipedia articles related to that sense — determined by words in the sense's WordNet synset — and the gloss of the sense (Scarlini et al., 2020). The two vectors are then concatenated to make the final sense embedding. At runtime, the context of the target word is embedded using BERT and is compared to the previously described sense embeddings to determine which sense is assigned. *SensEmBERT* works well on rare lemmas and senses (Scarlini et al., 2020), which is important to consider for our experiments that contain author-specific text since authors tend to favour a single sense for a word (Gella et al., 2014), which might not be the predominant sense of a domain. Therefore, if a model performs poorly on rare senses according to the sense inventory, then it might perform poorly on the favoured sense of the author. Likewise, if an author frequently uses rare words — according to the sense inventory — then a model that performs poorly on rare words would perform poorly for the author. The ability to perform well on all authors and demographics is a way to evaluate the fairness of a model (Ethayarajh and Jurafsky, 2020; Hashimoto et al., 2018). In this work, we consider the fairness of our models by focusing on authors that they perform poorly on.

3 Data Statement

In this section, we discuss the properties of our dataset, while following the proposed schema from Bender and Friedman (2018).

3.1 Data selection

We collect all text from blogs of the top 50 authors that contain the most tokens from the corpus that was originally presented in Schler et al. (2006). The original corpus consists of English blog posts from 19,320 authors. There were some authors who have blog posts that are copied from other authors and therefore, we do not consider these authors in the top 50. We consider the top 50 authors to ensure that each author possesses enough text to allow the ability to study the potential benefits of using text from the author for personalizing a WSD model. For selecting lemmas, we first consider all 20 nouns from Gella et al. (2014). We consider the top 10 authors — authors that have the 10^{th} highest frequency of a lemma — to ensure that there is text from multiple authors for each lemma to study the effects of personalization on a per-lemma basis. From this group of lemmas, we retain all lemmas that have been used 20 or more times by the author who has used this lemma the 10^{th} most frequently. We selected the cutoff of 20, because there is not a large possibility that all instances from an author will be usable due to them not being a noun or not representing a sense from our chosen sense inventory. The group of lemmas that we include moving forward will be referred to as *SHORT_LIST*.

For each lemma in *SHORT_LIST*, we randomly sample approximately 10 sentences that contain the lemma from 10 authors and manually assign the lemma a sense ourselves.[1] We then use this annotated subset to perform two different analyses that focus on quantifying the diversity of the senses for a lemma. The first analysis involves calculating the predominant sense of each lemma for each author and then finding the most frequent sense among the author-level predominant senses, which we call the grand sense. We then calculate the percent of author-level predominant senses for a given lemma that are not the grand sense. The second analysis calculates the number of assigned senses that are not the grand sense. Both types of analysis assist in showing which lemmas have senses that vary among authors and therefore, they might benefit from a personalized model which is the main focus of this work. Lemmas that score low on these metrics could be ideal for models that predict the predominant sense, but would most likely not benefit from an author-level model. We originally wanted the top 10 lemmas that scored

[1] We are native English speakers.

Lemma	10^{th}MF	Predom	Token	# Senses
form	54	0.60	0.69	16
position	40	0.60	0.74	16
degree	27	0.56	0.57	7
sign	77	0.50	0.62	11
track	36	0.44	0.67	12
paper	54	0.44	0.57	7
deal	75	0.40	0.44	9
field	30	0.30	0.43	17
case	97	0.22	0.47	19
charge	36	0.22	0.34	15
rule	43	0.22	0.27	12

Table 1: List of lemmas and their frequency for the author who uses the lemma the 10^{th} most frequently (10^{th}MF). The percent of author-level predominant senses and token senses that are not the grand sense, represented by *Predom* and *Token*, respectively (higher values indicate more diversity). The number of WordNet senses for each lemma under the label *#Senses* is also shown.

the highest when comparing predominant senses with the grand sense, but we received a tie for the lemmas that scored 9^{th}, 10^{th}, and 11^{th}. We remove all lemmas from *SHORT_LIST* that scored less than 0.22 on the percent of predominant senses that are not the grand sense, which was the score for the lemmas with the 9^{th}, 10^{th}, and 11^{th} highest values. Table 1 shows our final 11 lemmas with their frequency for the top 10 authors, and their two diversity metrics.

Additional preprocessing was applied to the text from authors and the annotated instances, including the replacing of tokens that contain URL identifiers (*www*, *html*, *https*, etc.) with the token *urlLink* and the removal of what appear to be artifacts of text encoding, such as *\xx\xx\xx*.

For each of the 11 lemmas in the dataset, we gather the top 10 authors that used that lemma most frequently and gather all sentences from them that contain that lemma tagged as a noun by a part-of-speech tagger (Qi et al., 2020). This results in 36 authors and a total of 1607 instances across all lemmas. We manually scan through all instances ourselves and remove all instances that were incorrectly tagged as nouns.

3.2 Annotation

In this work, we use WordNet (Miller, 1995) as the sense inventory due to its popularity among WSD tasks (Raganato et al., 2017). We show the number of WordNet senses for each lemma in Table 1, which ranges from 7 for *degree* to 19 for *case*.

We used Amazon Mechanical Turk — a common crowd-sourcing site — to annotate the instances of each lemma. We divided the instances into groups of 5 — known as HITs — and ensure all 5 instances belong to the same lemma. Each instance in a HIT is presented to the annotator, known as a turker, with a piece of text that contains a single target token written in bold for each text. Each instance contains up to 20 tokens before the target token and 20 tokens after the target token, which can cross sentences but does not cross blog posts. This provides the turker with more context than only looking at a single sentence, which can assist their annotations. The turkers are asked to select the sense that best applies to the target token from a list of the WordNet senses for the lemma of the target token or they can select *I cannot assign a sense*. There was space available for feedback for each instance, where turkers can write the reason that they cannot assign a sense or provide general feedback. It is possible that a token can exhibit multiple senses (Erk et al., 2009), but our dataset will only consider one sense as the ground truth for each instance. Therefore, following Chklovski and Mihalcea (2002) and Pradhan et al. (2007), a turker can only select one sense for any instance. Each HIT was annotated by 10 turkers. An example of the task assigned to turkers is seen in Figure 1.

3.2.1 Annotators

For a turker to be eligible to annotate the HITs, they need to be 19 years of age, speak English as a first language, live in Canada or United States, and have a previous HIT acceptance rate of 98%. We paid the turkers between \$0.05 and \$0.10 per HIT, which is competitive with other sense annotation tasks (Akkaya et al., 2010; Hong and Baker, 2011; Rumshisky, 2011; Passonneau and Carpenter, 2014). Our work consists of 185 annotators producing a total of 14,607 annotations and 137 instances of feedback.

Some turkers may provide poor annotations and therefore an initial pass over the annotations can help identify these turkers (Gella et al., 2014). Gella et al. (2014) included a gold-standard instance within each HIT and disregarded all annotations from turkers that performed poorly on these gold-standard instances. Instead of providing a gold-standard instance within each HIT, we compare each turker's annotations against a majority vote. We avoid the use of a gold standard because we did not want to mix text types, i.e., blog posts and WordNet, and because not all senses in WordNet have an example, and those that do have

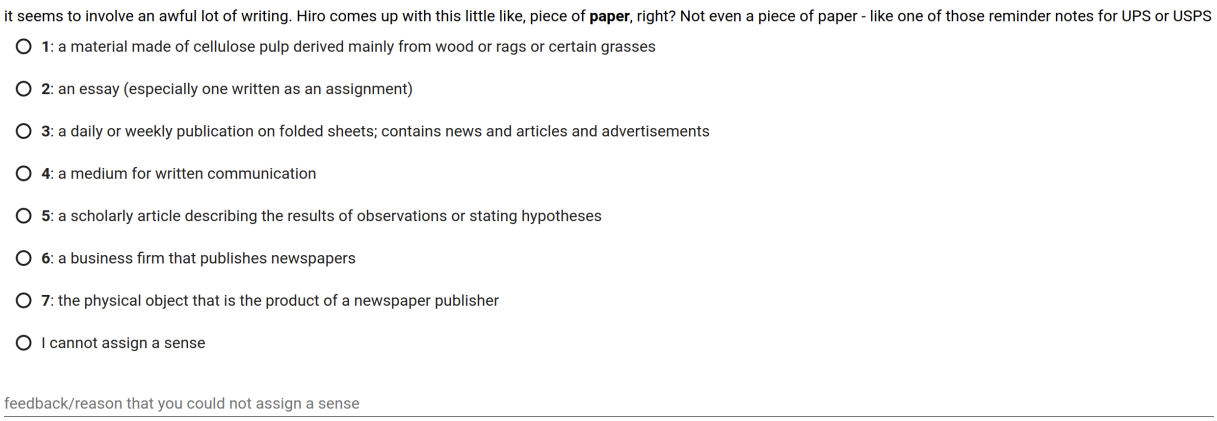

Figure 1: Example of the task assigned to turkers.

examples do not always contain the target lemma. We do this by performing an initial pass over the annotations to calculate the majority vote for each instance and then calculate each annotator's accuracy with the majority vote. Annotations from any turker that scored less than 50% agreement with the majority vote are removed from consideration, leaving 162 turkers in the dataset. We perform a second pass through the dataset and calculate the majority vote for each instance and remove 7 instances, which were assigned *I cannot assign a sense* as the most frequent label and 2 instances where there was a tie for the most frequent label. We randomly annotated 86 instances and our annotations agreed with the turkers' majority vote 85% of the time.

3.3 Speech Situation

All text was originally obtained from downloading all accessible blogs from `blogger.com` on a single day in August of 2004 (Schler et al., 2006). The number of instances per author ranges from 3 to 152 with a mean of 44 and a median of 31 instances per author. The age of the authors range from 17 to 48 years with a mean of 30 and a median of 27. The sex of the authors is disproportionate, with 10 females and 26 males.

3.4 Speaker Demographic

All text in the original corpus was English, although there was non-English text in the blogs, which was removed by Schler et al. (2006).

3.5 Dataset Analysis

The final dataset consists of 11 lemmas and 1586 annotated instances.[2] The dataset consists of multi-sentence instances that were annotated by the turkers. We did this to maintain consistency with the text being annotated by turkers and the text being used by WSD models. Table 2 shows the number of instances per author for each lemma ranges from 93 for *charge* to 192 for *paper* with an average of 144. The number of assigned senses ranges from 4 for *deal* to 13 for *field* with an average number of 8.5, known as sense ambiguity (Jurgens, 2014). The sense ambiguity is a metric that can be used to measure the difficulty of a WSD dataset. The dataset's sense ambiguity of 8.5 is among the higher values of the datasets in the collection from Raganato et al. (2017) and the dataset from Gella et al. (2014), which range from 4.9 to 8.9

Figure 2 shows the sense distributions for four authors with respect to the lemma *deal* and shows how authors can use the same lemma differently and the potential benefits of tailoring models toward an individual author. Specifically, each author in Figure 2 has a different predominant sense for the lemma *deal* and only *Author_0* shares their predominant sense with the predominant sense across all authors.

4 Methods

In this section, we discuss the WSD methods that we evaluate on the dataset. This includes predominant sense baselines, a state-of-the-art method (*SensEmBERT*), and our proposed personalized models.

4.1 Baselines

We apply three WSD baselines, which include always predicting the predominant sense for each lemma. The predominant sense is calculated via

[2]Code that generates our dataset from the corpus of Schler et al. (2006) is located at `https://github.com/Mordecaffe/Personalized_WSD_Dataset`.

Lemma	# instances	# senses assigned
Paper	192	7
Position	176	12
Sign	163	9
Form	156	10
Case	154	10
Degree	146	6
Track	146	8
Deal	140	4
Field	121	13
Rule	99	6
Charge	93	8
Average	144	8.5

Table 2: The number of instances and assigned senses for each lemma.

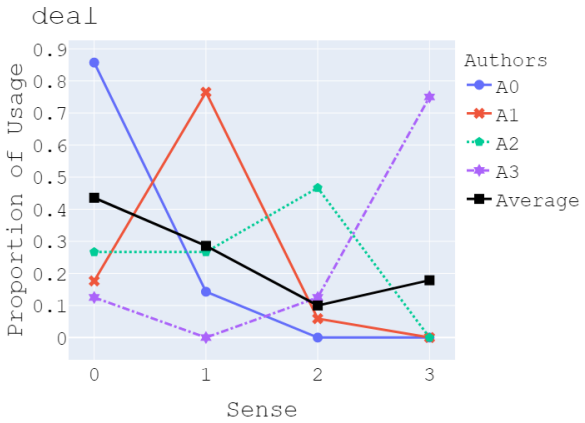

Figure 2: Sense distributions for four authors for the lemma *deal*. The average distribution across all authors is in black.

WordNet (*WORDNET*), the predominant sense of the dataset (*DATASET_PREDOM*), and the predominant sense for each author (*AUTHOR_PREDOM*).

4.2 SensEmBERT

This method is an unsupervised method that has achieved state-of-the-art results on the English datasets from Raganato et al. (2017) and can outperform supervised WSD models on less frequent lemmas and senses (Scarlini et al., 2020). It uses sense embeddings of a noun by embedding text from Wikipedia articles related to the noun and concatenating it with an embedding of text from the noun's BabelNet entry. The embedding process of a token is done by averaging BERT embeddings of each considered token. A BERT embedding is calculated by summing the last four layers of BERT after using the target token in context as input.

SensEmBERT assigns a token's sense by embedding the token with BERT and concatenating this embedding onto itself to double the vector length. Cosine similarity is calculated between this vector and all sense embeddings for all possible senses for the lemma of the target token. The sense that has the highest similarity is assigned as the target token's sense.

4.3 Personalizing SensEmBERT

We extend *SensEmBERT* by using text from the author to tailor the model to them by assuming knowledge about the author's sense distributions. These methods are used to explore the benefits of personalized WSD systems and the potential gains of learning the sense distributions of an author. We discuss these types of methods in this subsection.

4.3.1 SEBERT_PERS

In this method, we exploit the Zipfian distribution that sense frequencies tend to exhibit (Kilgarriff, 2004). Specifically, the weights for each sense that are outputted by *SensEmBERT* are ranked and the final score for a sense is calculated by the inverse rank of the sense multiplied by the probability of this sense given an author (calculated by the author's gold standard sense distributions) as seen in Equation 1. The sense that results in the highest score is assigned as the sense of the target token.

$$weight(sense) = \frac{1}{rank} * p(sense|author) \quad (1)$$

4.3.2 SEBERT_PERS_PREDOM

The author-level sense distribution of a lemma could be difficult to automatically estimate and it might be easier to estimate the author-level predominant sense. Therefore, we assume knowledge of only the author-level predominant sense of a lemma for this method by using the author's gold standard predominant sense. Specifically, we assign the sense that was given the most weight according to *SensEmBERT* if the predominant sense of the author is not among the top k ranked senses. If the predominant sense is in the top k ranked senses, we assign the predominant sense. We refer to this k value as the override rank and it is a hyperparameter that needs to be tuned. We also explore the use of the predominant senses from the dataset and WordNet; we refer to these methods as *SEBERT_DATASET_PREDOM* and *SEBERT_WORDNET_PREDOM*, respectively.

5 Experimental Results

In this section, we evaluate our different models. We first explore the tuning of the override rank for *PREDOM*-based methods. We then evaluate our

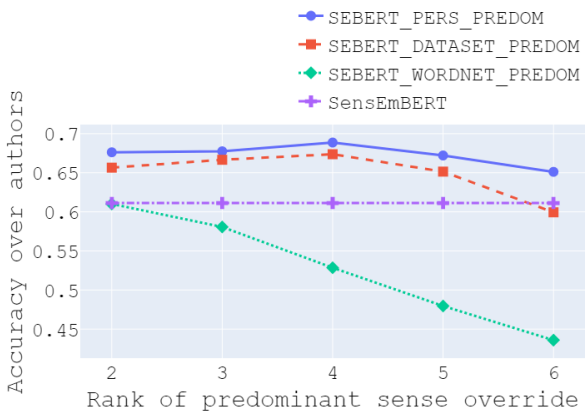

Figure 3: Average accuracy across authors for the *PREDOM*-based methods compared to *SensEmBERT*.

models using average accuracy across all authors and for individual authors.

5.1 Tuning SEBERT_PERS_PREDOM

Unfortunately, due to the relatively small size of our dataset, we are unable to use a held-out subset of the data for model tuning. Therefore, we show the performance of the *PREDOM*-based models using different override ranks. Figure 3 shows that *SEBERT_PERS_PREDOM* and *SEBERT_DATASET_PREDOM* outperform the *SensEmBERT* approach for any of the tested override rank values except for *SEBERT_DATASET_PREDOM* with an override rank of 6. This finding eliminates the need to fine-tune this model, since any value between 2 and 5, inclusively, works reasonably well. Increasing the override rank results in the model that uses WordNet first-senses to perform worse, which suggests that the authors' predominant senses do not align with WordNet's first senses. We do not consider *SEBERT_WORDNET_PREDOM* for the remaining experiments due to its poor performance and we use an override rank value of 4 for *SEBERT_PERS_PREDOM* and *SEBERT_DATASET_PREDOM*.

5.2 Overall Accuracy

In this subsection, we discuss the results of each method on the entire dataset. We evaluated each model using accuracy across instances, average accuracy across lemmas, and average accuracy across authors. We evaluated using these three metrics to eliminate the issue of having non-uniform distributions of instances in the dataset. For example, the number of instances per author ranges from 3 to 152 and, therefore, we would like to weigh each author equally in the case of accuracy across authors, instead of favouring models that only perform well on authors with more instances in the dataset.

Method	Inst	Lem	Auth
WordNet First-Sense	0.204	0.208	0.193
Dataset Predominant sense	0.384	0.395	0.396
Author Predominant sense	0.552	0.560	0.569
SensEmBERT	0.574	0.585	0.611
SEBERT_PERS	0.712	0.716	0.738
SEBERT_PERS_PREDOM	0.656	0.661	0.689
SEBERT_DATASET	0.660	0.655	0.664
SEBERT_DATASET_PREDOM	0.625	0.627	0.674

Table 3: Average accuracy across instances (Inst), lemmas (Lem), and authors (Auth).

Table 3 shows the scores for each method across the different evaluations. The three predominant sense baselines' performances scored in the expected order, such that using the predominant sense of the author outperforms using the predominant sense of the dataset, which outperforms using the first sense from WordNet. *SensEmBERT* outperformed all other baselines. The inclusion of the author's sense distribution in *SEBERT_PERS* and their predominant sense in *SEBERT_PERS_PREDOM* both outperform *SensEmBERT*. *SEBERT_PERS* achieves the highest score of 0.738 in terms of average accuracy across all authors, which is an absolute improvement of 0.127 above SensEmBERT. The use of the dataset-level sense distributions in *SEBERT_DATASET* and predominant senses in *SEBERT_DATASET_PREDOM* outperform *SensEmBERT* but does not outperform *SEBERT_PERS* and *SEBERT_PERS_PREDOM*, which supports the importance of using author-specific data. Interestingly, *SEBERT_DATASET_PREDOM* outperforms *SEBERT_DATASET* for average accuracy across authors. These findings indicate that SensEmBERT can be improved through personalization by incorporating information about author-level sense distributions or predominant senses.

5.3 Author-level Performance

In this subsection, we consider the models' performances on individual authors and observe the lower bound of each model's score with respect to the authors. By observing the lower bound of each model's performance with respect to the authors, we can observe the fairness of the models.

Figure 4 shows the performance of our two personalized models (*SEBERT_PERS* and *SEBERT_PERS_PREDOM*) and the *SensEmBERT*

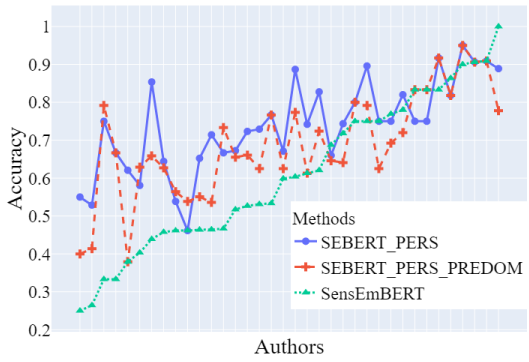

Figure 4: Author-level performances for the two personalized methods and *SensEmBERT*.

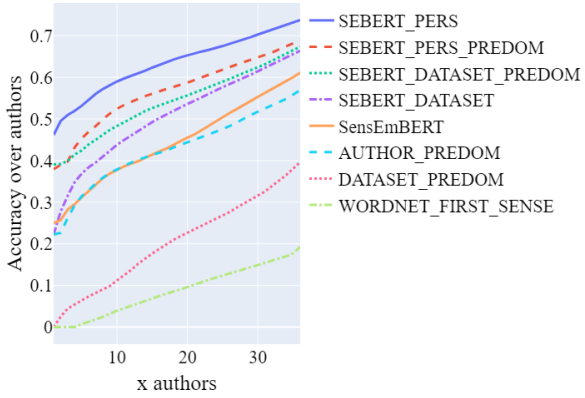

Figure 5: Performance of methods on the x authors that they achieve the lowest accuracy on.

baseline for each author in the dataset. The authors are sorted in ascending order with respect to the accuracy of *SensEmBERT*. It shows that authors that *SensEmBERT* performs below average on with respect to accuracy across authors (i.e. 0.611) often receive the largest boost in performance from personalized models. This could be due to those individuals having different writing styles as compared with text that *SensEmBERT* was trained on, which is an interesting topic for further exploration. The authors that achieve approximately 0.70 or greater for *SensEmBERT* can have their performance hindered by personalization, although, often not by a large amount. *SEBERT_PERS* usually outperforms *SEBERT_PERS_PREDOM* for a given author.

One of the main pillars of our work is to provide personalized models that work well for authors that scored poorly with conventional non-personalized models. Therefore, we would ideally want models that do not perform poorly on any single author, which we can evaluate by averaging the accuracy over the authors that each model achieves the lowest accuracy on. In Figure 5, we evaluate our models on the x authors that each model achieves their lowest accuracy on for values of x ranging from 1 author to all 36 authors. It shows that *SensEmBERT* achieves 0.25 on its worst author, while *SEBERT_PERS* achieves 0.46 on its worst author. Furthermore, *SEBERT_PERS*, *SEBERT_PERS_PREDOM*, and *SEBERT_DATASET_PREDOM* always outperform *SensEmBERT* for every value of x in the x worst authors evaluation, with *SEBERT_PERS* always achieving the highest score. This finding demonstrates that Personalized WSD models such as *SEBERT_PERS* and *SEBERT_PERS_PREDOM* are more fair than non-personalized models (*SensEmBERT*).

6 Conclusions

In this work, we proposed a novel dataset for personalized WSD and showed that sense distributions and predominant senses of an author can be used to personalize an existing knowledge-based WSD model (*SensEmBERT*). Our experiments consistently show that models that consider author-specific sense distributions (*SEBERT_PERS*) or predominant senses (*SEBERT_PERS_PREDOM*) can outperform models that do not consider any knowledge of sense distributions (*SensEmBERT*). Furthermore, we show that models that use author-level sense distributions or predominant senses outperform models that use genre-level sense distributions (*SEBERT_DATASET*) or predominant senses (*SEBERT_DATASET_PREDOM*). *SensEmBERT* achieved the highest accuracy across all authors with an absolute improvement of 0.127 above *SensEmBERT*. We further explore the fairness of our models by evaluating their accuracy on authors that they perform poorly on and showed that the lowest accuracy achieved by *SEBERT_PERS* on a single author is 0.21 above *SensEmBERT's* lowest accuracy. This finding demonstrates that *SEBERT_PERS* is more fair than *SensEmBERT*, which indicates that personalization can produce more fair WSD systems. Our work shows the importance of learning sense distributions of individual authors for WSD and therefore, we plan on developing methods for learning an author's sense distributions in future work similar to Pasini et al. (2020) and Bennett et al. (2016). Our personalized models could be learning topic-related content from the author to assist with their classification, therefore, an extension of this work could further explore this dataset with a focus on topic-related features.

7 Ethical Considerations

The involvement of turkers as annotators was reviewed and approved by the University of New Brunswick's ethics committee. We selected the authors based on their amount of text available and therefore the distributions over sexes is not equal — 26 males and 10 females — and therefore groups should consider this when working with this dataset.

Acknowledgments

This work is financially supported by the Natural Sciences and Engineering Research Council of Canada and the University of New Brunswick.

References

Cem Akkaya, Alexander Conrad, Janyce Wiebe, and Rada Mihalcea. 2010. Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, pages 195–203.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 136–145, Mexico City, Mexico.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1591–1600, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Andrew Bennett, Timothy Baldwin, Jey Han Lau, Diana McCarthy, and Francis Bond. 2016. LexSemTm: A semantic dataset based on all-words unsupervised sense distribution learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1513–1524, Berlin, Germany. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jordan L. Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *EMNLP-CoNLL*.

Timothy Chklovski and Rada Mihalcea. 2002. Building a sense tagged corpus with open mind word expert. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions*, pages 116–122.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18, Suntec, Singapore. Association for Computational Linguistics.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboard design. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853.

Spandana Gella, Paul Cook, and Timothy Baldwin. 2014. One sense per tweeter ... and other lexical semantic tales of Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 215–220, Gothenburg, Sweden.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR.

Jisup Hong and Collin F Baker. 2011. How good is the crowd at "real" wsd? In *Proceedings of the 5th linguistic annotation workshop*, pages 30–37.

David Jurgens. 2014. An analysis of ambiguity in word sense annotations. In *LREC*, pages 3006–3012. Citeseer.

Adam Kilgarriff. 2004. How dominant is the commonest sense of a word? In *International conference on text, speech and dialogue*, pages 103–111. Springer.

Adam Kilgarriff and Joseph Rosenzweig. 2000. English senseval: Report and results. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Linlin Li, Benjamin Roth, and Caroline Sporleder. 2010. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1138–1147, Uppsala, Sweden.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Tommaso Pasini, Federico Scozzafava, and Bianca Scarlini. 2020. Clubert: A cluster-based approach for learning sense distributions in multiple languages.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110. Association for Computational Linguistics.

Anna Rumshisky. 2011. Crowdsourcing word sense definition. In *Proceedings of the 5th linguistic annotation workshop*, pages 74–81.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8758–8765.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.