Exploring the Value of Personalized Word Embeddings

Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas and **Rada Mihalcea**
Computer Science & Engineering
University of Michigan
{cfwelch,jkummerf,vrncapr,mihalcea}@umich.edu

Abstract

In this paper, we introduce personalized word embeddings, and examine their value for language modeling. We compare the performance of our proposed prediction model when using personalized versus generic word representations, and study how these representations can be leveraged for improved performance. We provide insight into what types of words can be more accurately predicted when building personalized models. Our results show that a subset of words belonging to specific psycholinguistic categories tend to vary more in their representations across users and that combining generic and personalized word embeddings yields the best performance, with a 4.7% relative reduction in perplexity. Additionally, we show that a language model using personalized word embeddings can be effectively used for authorship attribution.

1 Introduction

Word embeddings have become ubiquitous in natural language processing applications. Usually, embeddings are trained from a large corpus of news or web data that contains writing from many sources and authors (Mikolov et al., 2013; Pennington et al., 2014). These embeddings capture syntactic and semantic properties of the language of all authors who contributed to this corpus.

Multi-source corpora provide large volumes of data, but they may not lead to the ideal representations for individuals. For instance, the word "hometown" may have a different representation for different individuals. For some, it may relate to words such as "hills," "trees," and "family," whereas for others may be more strongly connected to "ocean," "beach," and "friends." These personalized representations differ among individuals, and also differ from a more generic representation that often tends to capture words that are semantically related at concept level, such as "city," "town," or "place."

In this paper, we present the idea of personalized word embeddings. We explore differences in personalized word representations using a corpus of English Reddit posts that contains a large number of posts per author. We use the embeddings to initialize a language model and show that personalization leads to better results than generic embeddings. One potential application of this work is personalized text generation for auto-completion to speed up text entry. Another application is dialog systems that follow the speaking style of certain professionals (e.g., counselors, advisors). Finally, personalized word representations could particularly help users with atypical writing styles that are not currently well served by models trained to suit the majority.

2 Related Work

Prior work has considered *user embeddings*, where one vector is learned for each user in the data (we learn a set of vectors per user, one for each word in the vocabulary). User embeddings have been used for dialog generation (Li et al., 2016), query auto-completion (Jaech and Ostendorf, 2018), authorship attribution (Ebrahimi and Dou, 2016), and sarcasm detection (Kolchinski and Potts, 2018). Amer et al. (2016) learn a set of embeddings from the books that a user adds to their profile. Some approaches also use network information (Zeng et al., 2017; Huang et al., 2016).

Proceedings of the 28th International Conference on Computational Linguistics, pages 6856–6862
Barcelona, Spain (Online), December 8-13, 2020

User	Example Use	Nearest Neighbors
A	doctors think this is bad for her **health** ...	preventative, insurance, reform, medical, education
B	it is usually bad for your **health** ...	professional, mental, conduct, experiences, online
All	N/A	medical, preventative, insurance, safety, healthcare

Table 1: Nearest neighbors of "health" for two personalized embedding spaces and the generic space.

Personalization has been studied for marketing, webpage layout, recommendations, query completion, and dialog (Eirinaki and Vazirgiannis, 2003; Das et al., 2007). Our prior work (Welch et al., 2019a; Welch et al., 2019b) explored predicting response time, common messages, and author relationships from personal conversation data. Zhang et al. (2018) conditioned dialog systems on artificially constructed personas and Madotto et al. (2019) used meta-learning to improve this process. Goal-oriented dialog has used demographics (i.e., age, gender) to condition system response generation, showing that this relatively coarse grained personalization improves system performance (Joshi et al., 2017).

3 Personalized Word Embeddings

Definition. Personalized word embeddings are vector representations of words derived from the text produced by a single author. We use the text produced by a Reddit user s in their posts C_s to create their word embeddings. We apply the method described below to this set and produce an embedding matrix, $C_s \mapsto W_s^{|V| \times k}$, where V is the vocabulary and k represents the embedding dimension.

Joint Learning of Personal and Generic Word Embeddings. We jointly learn a generic embedding matrix and an embedding matrix for each author, inspired by Bamman et al. (2014). Each matrix $W \in \mathbb{R}^{|V| \times k}$ has a row for each vocabulary word and a k-dimensional vector for each embedding. The hidden layer is calculated as $h = w^\intercal W_{generic} + w^\intercal W_s$ where w represents the one-hot encoding of a word and s represents an author. This is a modified skip-gram architecture (Mikolov et al., 2013), which sums two terms so that back-propagation updates the generic matrix and a author-specific matrix. It allows the generic matrix to benefit from all data while learning author-specific deviations in the same space.

Dataset. We use data for the 100 most active users[1] in a corpus collected from Reddit. These users have from 49k to 249k posts, with 73k on average. Posts contain 29 tokens on average and come from 3.6k subreddits. The largest fraction (18.6%) belong to the subreddit *AskReddit*; the next two largest are *blog* and *politics* with 4.8% and 4.7% of the posts respectively.

We use the set of messages from all 100 authors to generate embeddings for all words that occur at least five times (across all users). This yields a vocabulary of 177k words. We learn 100-dimensional embeddings with an initial learning rate of 0.025 and a window size of five, using L2 regularization due to the increased number of parameters (tuned in preliminary experiments). Using this method, we learn 101 embeddings for each word – a generic representation, and a separate representation for each user.

Reddit users have been found to be primarily male, young adults (under 30), located in the USA and primarily identify as christian or atheist (Welch et al., 2020). It is possible that results presented in this paper do not generalize to populations that differ significantly from the population of Reddit users. Future work may consider isolating the effects of topics and style by modeling subreddits for comparison though in this work we consider a personalized embedding as a representation that may capture both.

4 Differences across Individual Word Representations and Usages

Individuals use the same word in different ways in different contexts. Examining these differences can give insight into individual topic and style preferences, or their word associations. To illustrate these differences, in Table 1, we show different ways that two users in our dataset use the word "health." Although these words may be used in similar contexts, the meaning of, and topics associated with these

[1]We excluded users that appear on a `https://www.reddit.com/r/autowikibot/wiki/redditbots`public list of bots or who appear to be automated based on manual inspection.

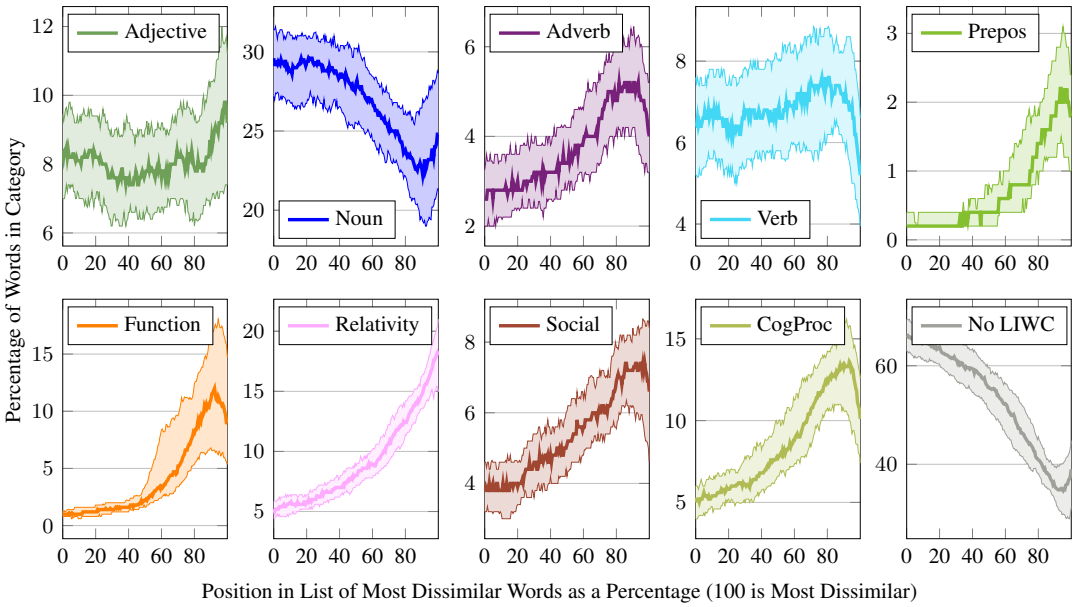

Figure 1: Relationship between embedding similarity and distribution of word categories. Y-axes are scaled separately for each sub-plot and sliding window on x-axis shows word types ordered by average dissimilarity across embedding spaces. Top row groups words by their part-of-speech and bottom groups words by LIWC categories. We show average and interquartile range for values calculated across all users. Categories have more or less personalization, e.g., adverbs and words not in LIWC respectively.

words is often different for each user, which affects the words we would expect to come after it. These preferences are reflected in the top neighbors for the word "health" for each user.

To gain a deeper understanding of these differences, we analyze personal and generic embeddings for specific word groups based on the Linguistic Inquiry and Word Count (LIWC) lexicon (Pennebaker et al., 2001) and part-of-speech tags. This analysis can help us understand what types of words tend to have different representations across users and are therefore more personal in nature. We part-of-speech tag the messages with the Stanford CoreNLP tagger (Toutanova et al., 2003; Manning et al., 2014). For each word in the vocabulary, we assign a tag if the tagger gives the same tag at least 95% of the time, otherwise the word is ignored. LIWC categories are looked up in the lexicon that contains words and word stems and a word may have multiple categories, in which case it counts toward each.

We look at the proportion of word types in the 5k most dissimilar words for each user. We define word similarity as the cosine distance between a word's generic embedding and its author-specific embedding. Note that theses are unique words and that the x-axis is the percentage of the way through the 5k dissimilar words, with the most dissimilar at 100%. We break this into subsets and look at how the distribution changes as we approach the most dissimilar words. A visualization is provided in Figure 1. We find the set of most dissimilar words includes more function words, words relating to space and time (Relativ), cognitive processes (CogProc), and social words, as well as more adjectives and nouns. This suggests that these types of words may tend to have more personal usage than other types. These results are consistent with prior work that has found function words are effective for recognizing style and measuring style similarity (Gonzales et al., 2010) and for authorship attribution (Mosteller and Wallace, 1963; Gamon, 2004; Argamon et al., 2003).

5 Language Modeling

Our first test case for personalized word embeddings is language modeling. We train language models on our data with our embeddings as input. We use AWD-LSTM language model from previous work as our baseline (Merity et al., 2018b; Merity et al., 2018a). It is an autoregressive language model that has state-of-the-art results by combining regularization techniques and has been widely used. Although more recent models can achieve better perplexities on standard benchmarks (Melis et al., 2020; Dai

Model	LM		Attribution	
	MRR	PPL	MRR	Accuracy
Merity et al. (2018b)	0.364	65.53	0.452	34.8%
Single User Vectors	0.361	66.70	0.450	34.9%
Personalized Embeddings	**0.371**	**62.43**	**0.462**	**36.1%**

Table 2: Results for Language Modeling (LM) and Authorship Attribution. Personalized word embeddings significantly improve performance (permutation test, $p < 0.0001$).

et al., 2019), we find that not all models have code available or that they take far more time to run than Merity et al. (2018b)'s model. We use the same hidden layer sizes and drop-out rates as in their original experiments, but untie the weights of the encoder and decoder as that gave better performance in preliminary experiments.

To use our personalized embeddings, we modify the architecture to take as input the concatenation of the personalized user-specific embedding and the generic embedding for each word. The same embedding dropout mask is applied to both word embeddings. The embeddings are trained on all available user data, but the more computationally expensive language models are not. We use a subsample of our dataset with 1,000 posts for each user and an 80/10/10 split for training, validation, and testing. The same splits are used for generic and personalized models, varying only the embedding layer.

To measure the ability of our models to predict the next word, we use two metrics: (1) mean reciprocal rank (MRR), calculated as one divided by the rank of the correct word choice in the descending list of next word probabilities and averaged over all instances, and (2) perplexity.

Single User Embeddings. We also consider an approach in which just one vector is learned for each user (rather than one for each user-word pair). This is an approach widely used in previous work (Kolchinski and Potts, 2018; Li et al., 2016). This user vector is concatenated to the generic word embedding.

The results in Table 2 suggest that using the combined personalized and generic embeddings improves performance significantly over single vector user representations and over generic embeddings. We note that the number of parameters of the LSTM input is not the same when comparing the baseline Merity et al. (2018b) model to the other cases. We ran an additional experiment doubling the size of the embeddings for the baseline and found that the perplexity improved to 64.21, although the personalized embeddings still significantly outperformed this baseline.

We can also analyze the accuracy when predicting words belonging to particular parts of speech and LIWC categories. Our intuition is that a model that uses personalized word embeddings would be better at predicting words belonging to the four LIWC categories whose words are most distant from the generic space. Tables 3a and 3b show that for almost all categories, the personalized word embeddings lead to the best performance, although for relativity words, single user vectors performs slightly better.

6 Authorship Attribution

We also use a language model trained with personalized word embeddings to perform the task of authorship attribution.[2] We build a language model for each author using a sample of 10k posts for training and 1k for validation. We then hold out another sample of 1k posts to use for authorship attribution. The language models for all authors are separately run on the held out set, and the model with the lowest perplexity is then chosen as the assigned author. Table 2 shows there is a statistically significant improvement for our personalized embeddings method. This is a difficult task with 100 classes, so the accuracy is low, but the MRR suggests that the correct author is usually in the top 3 model choices.

7 Limitations and Ethical Considerations

In applying our method to text prediction systems, users may experience unintended negative effects. For instance, embeddings may become unintentionally biased toward language that becomes inappropriate

[2]Note that we do not consider datasets such as Kestemont et al. (2019) because they do not provide the volume of data needed for our approach and our goal is to compare generic and personalized embeddings, not to set a new state-of-the-art.

Model	DT	IN	JJ	NN	PR	RB	VB	PUNCT	OTHER
Merity et al. (2018b)	19.1	30.9	703.6	632.5	23.4	146.6	65.4	10.9	72.6
Single User Vectors	19.4	34.2	708.2	621.4	23.7	148.7	65.8	11.1	73.9
Personalized Word Embeddings	**18.9**	**30.7**	**681.2**	**597.7**	**22.6**	**143.7**	**62.1**	**10.2**	**70.2**

(a) Perplexity results broken down by POS tag (OTHER includes all other tags).

Model	Affect	Bio	CogProc	Drives	Funct	Inform	Percept	Relativ	Social
Merity et al. (2018b)	88.3	93.9	73.4	69.8	68.8	**35.3**	95.9	28.7	48.7
Single User Vectors	85.6	95.7	75.8	75.0	71.2	36.0	90.7	**28.1**	48.9
Personalized Word Embeddings	**82.7**	**92.4**	**73.3**	**70.7**	**67.6**	**35.3**	**85.4**	28.5	**46.7**

(b) Perplexity results broken down by high-level LIWC category.

Table 3: Perplexity results broken down by the type of target word, with the best result in bold.

when later suggested in another context. Additionally, users who are learning a language may bias embeddings toward improper language use, reinforcing errors and making it more difficult for the user to learn the language. It may be appropriate to use our embeddings if users consent and are made aware of the possible consequences of doing so.

It is possible that our method could be used for authorship attribution and surveillance of individuals online (Stamatatos, 2009). Such an application risks potential discrimination, coercion, and threats to intellectual freedom (Richards, 2013). Personalized language models could also be used to develop a tool that tells the user who their writing most resembles, or if their writing resembles their past writing, with the objective of obfuscating the author's identity (Potthast et al., 2018). A tool like this could also be used maliciously to impersonate a particular author. Although we believe the difficulty of this task currently makes these minor risks, we advocate against the use of our methods for these tasks.

Our method requires more computation and memory than the baseline method we compared to. The additional computation is relatively small, as learning the embeddings takes around 3 hours using 30 threads on a machine with 16 Intel Xeon Silver 4108 CPUs. The memory required to store embeddings for N users is $N + 1$ times the amount of storage required by a generic matrix only.

8 Conclusion

In this paper, we explored personalized word embeddings. Using a large corpus of Reddit posts, we generated personalized word embeddings for 100 individuals, and performed analyses of the differences between personalized and generic embeddings for specific groups of words. We showed that using personalized word embeddings to initialize a language model improves perplexity over a model that uses generic word embeddings, or a model that only learns single vectors for each user as has been frequently done in previous work. Further, we showed that the embeddings can be used to improve performance on authorship attribution. We cannot release the data due to licensing restrictions but our code is available online with instructions for how to obtain and process the data in order to support future work on personalization.[3]

Acknowledgements

We would like to thank the anonymous reviewers for their helpful suggestions. This material is based in part on work supported by IBM (Sapphire Project), DARPA (grant #D19AP00079), Bloomberg (Data Science Research Grant), the NSF (grant #1815291), and the John Templeton Foundation (grant #61156). Any opinions, findings, conclusions, or recommendations in this material are those of the authors and do not necessarily reflect the views of IBM, DARPA, Bloomberg, the NSF, or the John Templeton Foundation.

[3] https://lit.eecs.umich.edu/downloads.html

References

Nawal Ould Amer, Philippe Mulhem, and Mathias Géry. 2016. Toward word embedding for personalized information retrieval. In *Neu-IR: The SIGIR 2016 Workshop on Neural Information Retrieval*.

Shlomo Argamon, Marin Šarić, and Sterling S Stein. 2003. Style mining of electronic messages for multiple authorship discrimination: First results. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

David Bamman, Chris Dyer, and Noah A Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July.

Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web*. ACM.

Javid Ebrahimi and Dejing Dou. 2016. Personalized semantic word vectors. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM.

Magdalini Eirinaki and Michalis Vazirgiannis. 2003. Web mining for web personalization. *ACM Transactions on Internet Technology (TOIT)*, 3(1).

Michael Gamon. 2004. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In *Proceedings of the 20th International Conference on Computational Linguistics*. Association for Computational Linguistics.

Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1).

Yu-Yang Huang, Rui Yan, Tsung-Ting Kuo, and Shou-De Lin. 2016. Enriching cold start personalized language model using social network information. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 21, Number 1, June 2016*.

Aaron Jaech and Mari Ostendorf. 2018. Personalized language model for query auto-completion. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia, July.

Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *Advances in Neural Information Processing Systems 2017 Conversational AI Workshop*.

Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019. Overview of the Cross-domain Authorship Attribution Task at PAN 2019. In *CLEF 2019 Labs and Workshops, Notebook Papers*.

Y. Alex Kolchinski and Christopher Potts. 2018. Representing social media users for sarcasm detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, October-November.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*.

Gábor Melis, Tomáš Kočiský, and Phil Blunsom. 2020. Mogrifier LSTM. In *International Conference on Learning Representations (ICLR)*.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018a. An analysis of neural language modeling at multiple scales. *arXiv preprint arXiv:1803.08240*.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018b. Regularizing and optimizing LSTM language models. In *International Conference on Learning Representations (ICLR)*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *International Conference on Learning Representations (ICLR)*.

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302).

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October. Association for Computational Linguistics.

Martin Potthast, Felix Schremmer, Matthias Hagen, and Benno Stein. 2018. Overview of the author obfuscation task at PAN 2018: A new approach to measuring safety. In *Proceedings of the Conference and Labs of the Evaluation Forum*.

Neil M Richards. 2013. The dangers of surveillance. *Harv. L. Rev.*, 126.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3).

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology (NAACL-HLT)*. Association for computational Linguistics.

Charles Welch, Verónica Pérez-Rosas, Jonathan K. Kummerfeld, and Rada Mihalcea. 2019a. Learning from personal longitudinal dialog data. *IEEE Intelligent systems*, 34(4).

Charles Welch, Verónica Pérez-Rosas, Jonathan K. Kummerfeld, and Rada Mihalcea. 2019b. Look who's talking: Inferring speaker attributes from personal longitudinal dialog. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, La Rochelle, France.

Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, November.

Ziqian Zeng, Yichun Yin, Yangqiu Song, and Ming Zhang. 2017. Socialized word embeddings. In *International Joint Conferences on Artificial Intelligence (IJCAI)*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.