Simple dynamic word embeddings for mapping perceptions in the public sphere

Nabeel Gillani
MIT
ngillani@mit.edu

Roger Levy
MIT
rplevy@mit.edu

Abstract

Word embeddings trained on large-scale historical corpora can illuminate human biases and stereotypes that perpetuate social inequalities. These embeddings are often trained in separate vector space models defined according to different attributes of interest. In this paper, we develop a single, unified dynamic embedding model that learns attribute-specific word embeddings and apply it to a novel dataset—talk radio shows from around the US—to analyze perceptions about refugees. We validate our model on a benchmark dataset and apply it to two corpora of talk radio shows averaging 117 million words produced over one month across 83 stations and 64 cities. Our findings suggest that dynamic word embeddings are capable of identifying nuanced differences in public discourse about contentious topics, suggesting their usefulness as a tool for better understanding how the public perceives and engages with different issues across time, geography, and other dimensions.

1 Introduction

Language has long been described as both a cause and reflection of our psycho-social contexts (Lewis and Lupyan, 2018). Recent work using word embeddings—low-dimensional vector representations of words trained on large datasets to capture key semantic information—has demonstrated that language encodes several gender, racial, and other common contemporary biases that correlate with both implicit biases (Caliskan et al., 2017) and macro-scale historical trends (Garg et al., 2018).

These studies have validated the use of word embeddings to measure a range of psychological and social contexts, yet in most cases, they have failed to leverage the full power of available datasets. For example, the historical biases presented in (Garg et al., 2018) are computed using decade-specific word embeddings produced by training different Word2Vec (Mikolov et al., 2013) models on a large corpus of historical text from that decade. The authors then use a Procrustes alignment to project embeddings from different models into the same vector space so they can be compared across decades (Hamilton et al., 2016). While this approach is reasonable when there are large-scale datasets available for a given attribute of interest (e.g. decade), it requires an additional optimization step and also disregards valuable training data that could be pooled and leveraged across attribute values to help with both training and regularization. This latter property is particularly appealing—and necessary—in the context of limited data.

In this paper, we use a simple, unified dynamic word embedding model that jointly trains linguistic information alongside any categorical variable of interest—e.g. year, geography, income bracket, etc.—that describes the context in which a particular word was used. We apply this model to a novel data corpus—talk radio transcripts from stations located in over 64 US cities—to explore the evolution of perceptions about refugees during a one-month period in late 2018. The results from our model suggest the potential to use dynamic word embeddings to obtain a granular, near real-time pulse on how people feel about different issues in the public sphere.

2 Model

2.1 Overview

Our dynamic embedding for word w is defined as

$$E(w, A) = \gamma_w + \Sigma_{a \in A}\, \beta_w^a \qquad (1)$$

where γ_w is an attribute-invariant embedding of w computed across the entire corpus, β_w^a is the off-

Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science, pages 94–99
Minneapolis, Minnesota, June 6, 2019. ©2019 Association for Computational Linguistics

set for w with respect to attribute a across the set of attributes A we are interested in computing the word embedding with respect to. For example, if we wish to compute the embedding for the word “refugee” as it was used on the 25th day of a particular 30-day corpus of talk radio transcripts, we would set $w = refugee$ and $A = \{25\}$. This approach, as formalized in Equation 1 above, is identical to one introduced by (Bamman et al., 2014), though finer details of our model and training differ slightly, as described below.

To learn γ_w and β_w^a, we train a neural network. Our model is a simple extension to the distributed memory (DM) model for learning paragraph vectors originally introduced in (Le and Mikolov, 2014). The DM model uses a continuous bag-of-words architecture to jointly train a paragraph ID with a sequence of words sampled from that paragraph to predict a particular word given the words that surround it. The output of this model includes a semantic vector representation of a) each paragraph, and b) each word in the vocabulary.

Our model extends the DM model by adding an additional dimension to the paragraph vector to learn specific *paragraph-by-word*—or, in our context, *attribute-by-word*—embeddings (i.e., β_w^a). The penultimate layer (before word prediction) is computed as an average of the dynamic embeddings for each context word, i.e., $X = \frac{1}{N}\Sigma_{i=1}^{N} E(w_i, S, A)$, where N is the size of our context window. This average embedding is then multiplied by the output layer parameters and fed through the final layer for word prediction. Figure 1 depicts our model architecture.

2.2 Implementation

We build on an existing PyTorch implementation of paragraph vectors[1] to implement our model, setting the dimensionality of γ_w and β_w^a to be 100. We use the Adam optimization algorithm with a batch size of 128, word context window size of 8 (sampling four words to the left and right of a target prediction word), learning rate of 0.001, and L2 penalty to regularize all model parameters. We only train embeddings for words that occur at least 10 times in the corpus. For training, we use the negative sampling loss function, described in (Mikolov et al., 2013) to be much more efficient than the hierarchical softmax and yield competitive results[2]. We train for 1 to 3 epochs and select the model with the lowest loss.

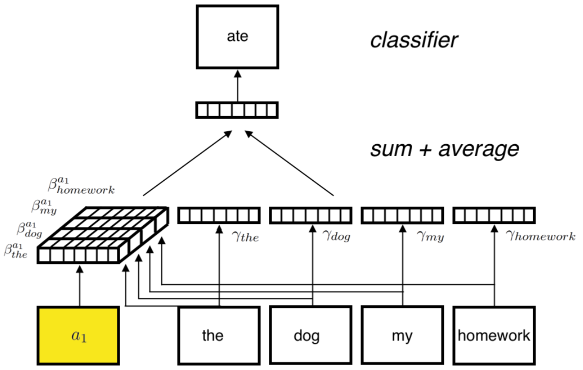

Figure 1: Our dynamic embedding model learns an attribute invariant embedding for each training word w (i.e., γ_w), along with an attribute-specific offset for attribute $A = \{a_1\}$ (i.e., $\beta_w^{a_1}$). The γ_w and $\beta_w^{a_1}$ terms are summed to compute $E(w, A)$ for each context word and averaged across words before classification. Figure inspired by (Le and Mikolov, 2014).

2.3 Validation

To validate our model, we compare our results to those produced via the decade-by-decade models trained in (Garg et al., 2018) using the Corpus of Historical American English (Davies, 2010). We use the same metric and word lists as the authors to compute bias scores. In particular, we compute linguistic bias scores for two analyses presented in (Garg et al., 2018): the extent to which female versus male words are semantically similar to occupation-related words, and the extent to which Asian vs. White last names are semantically similar to the same, from 1910 through 1990. We then compute correlations between changes in these scores and the actual changes in female and Asian workforce participation rates (relative to men and Whites, respectively) over the same time period.

Figure 2 depicts these results. The correlation between our scores and changes in workforce participation rates are similar to the correlation between the scores from (Garg et al., 2018) and the same ($r = 0.8, p = 0.01$ and $r = 0.81, p < 0.01$, respectively, for gender occupation bias; $r = 0.84, p < 0.01$ and $r = 0.79, p = 0.01$, respectively, for Asian/White occupation bias). Qualitative inspection of Figure 2 suggests that our model also produces smoother decade-by-decade scores, suggesting that it not only identifies attribute-

[1] Available at: https://github.com/inejc/paragraph-vectors.

[2] We include three noise words when computing the loss.

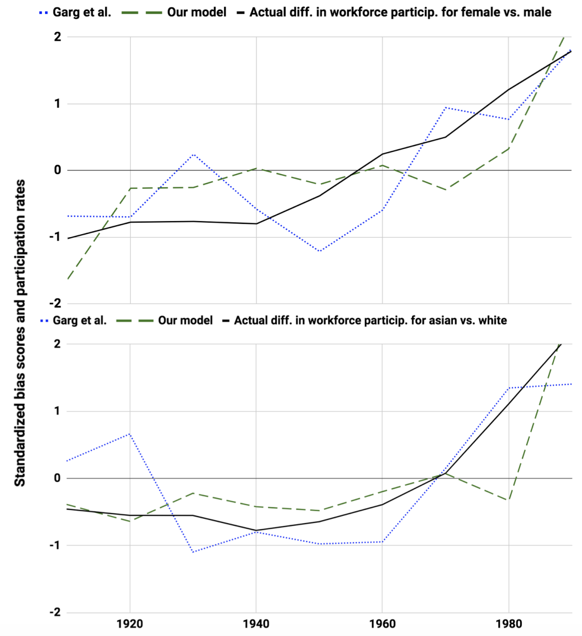

Figure 2: Scores produced by (Garg et al., 2018) and our model (blue dotted and green dashed lines, respectively) compared to actual workforce participation rates (solid lines) for gender (top) and Asian/White (bottom) linguistic biases. To compare all values on a single y-axis, we standardize both sets of bias scores and workforce participation rates by subtracting the mean and dividing by the standard deviation across decades.

specific fluctuations in word semantics, but also, may provide a more general, regularized model for learning attribute-conditioned word embeddings. Future research should include a comparison of our model's outputs to the outputs of other dynamic word embedding models that treat time as a continuously-valued attribute, e.g. (Bamler and Mandt, 2017; Rudolph and Blei, 2018; Yao et al., 2018).

3 Case study: refugee bias on talk radio

We are interested in applying our dynamic embedding model to better-understand talk radio-show biases towards refugees. Talk radio is a significant source of news for a large fraction of Americans: In 2017, over 90% of Americans over the age of 12 listened to some type of broadcast radio during the course of a given week, with news/talk radio serving as one of the most popular types (Pew, 2018). With listener call-ins and live dialog, talk radio provides an interesting source of information, commentary, and discussion that distinguishes it from discourse found in both print and social media. Given the proliferation of refugees and displaced peoples in recent years (totalling nearly 66 million individuals in 2016 (UNHCR, 2017))—coupled with the rise of talk radio as a particularly popular media channel for conservative political discourse (Mort, 2012)—analyzing bias towards refugees across talk radio stations may provide a unique window into a large portion of the American population's views on the issue.

3.1 Dataset and analyses

Our data is sourced from talk radio audio data collected by the media analytics nonprofit Cortico[3]. Audio data is ingested from nearly 170 different radio stations and automatically transcribed to text. The data is further processed to identify different speaker turns into "snippets"; infer the gender of the speaker; and compute other useful metrics (more details on the radio data pipeline can be found in (Beeferman and Roy, 2018)).

We train our dynamic embedding model on two talk radio datasets sourced from 83 stations located in 64 cities across the US. Dataset 1 includes 4.4 million snippets comprised of 114 million words produced by 390 shows between September 1 and 30, 2018. Dataset 2 includes over 4.8 million snippets comprised of 119 million total words produced by 433 shows between August 15, and September 15, 2018[4]. These datasets are used for analyses 1 and 2, respectively, described below.

Finally, we define bias towards refugees similar to how the authors of (Garg et al., 2018) define bias against Asians during the 20th century, measuring to what extent radio shows associate "outsider" adjectives like "aggressive", "frightening", "illegal", etc. with refugee and immigrant-related terms in comparison to all other adjectives. To compute refugee bias scores with respect to the attribute set A, we use the relative norm distance metric from (Garg et al., 2018):

$$bias_A = \Sigma_{r \in R}||E(r,A) - \overline{all}||_2 - ||E(r,A) - \overline{out}||_2$$

Where $E(r, A)$ is the dynamic embedding for a given word refugee word r in the set of all refugee-related words R (e.g. "refugee", ”immigrant”, ”asylum”, etc); $\overline{all}$ is the average dynamic embedding computed for each w in the set of all adjectives with respect to A; $\overline{out}$ is analogously defined for outsider adjectives; and $|| \cdot ||_2$ is the L2 norm.

[3] http://cortico.ai.

[4] As a rough proxy for removing syndicated content, we include only those snippets produced by a talk radio shows that air on one station.

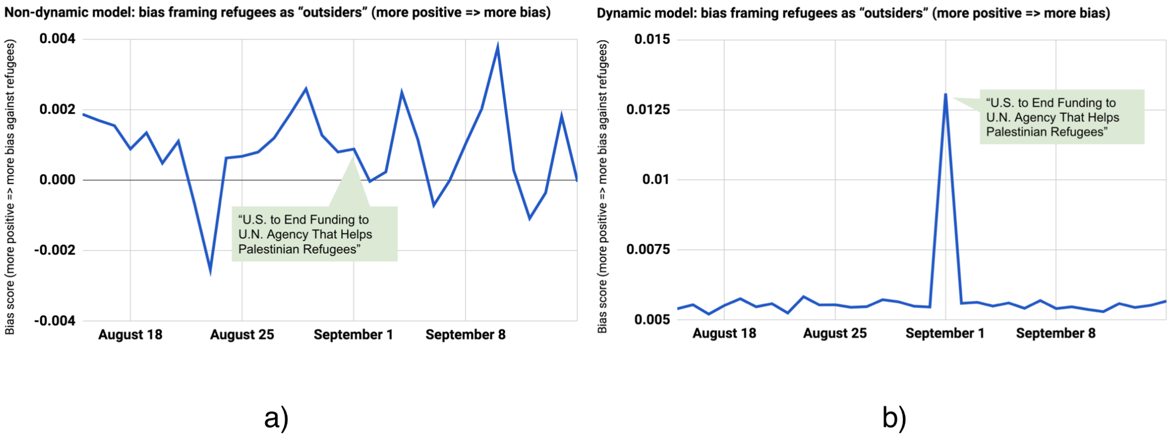

Figure 3: Bias towards refugees as outsiders across talk radio shows from mid-August to mid-September 2018: (a) depicts bias scores computed using a "non-dynamic model", i.e., training multiple Word2Vec models (one per day of data) and then projecting these models into the same vector space using orthogonal Procrustes alignment, and (b) depicts bias scores computed using our dynamic model. From qualitative inspection, the dynamic model appears to regularize scores across days during which refugee-related news is likely less-salient in public discourse.

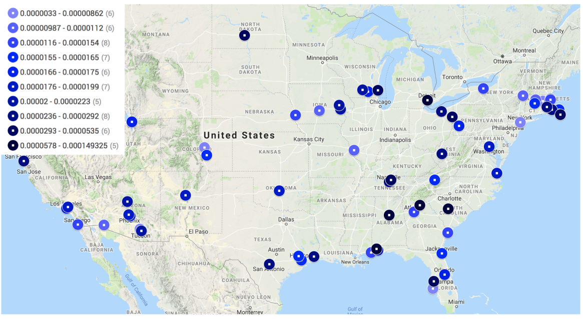

Figure 4: Bias towards refugees as outsiders computed across cities for radio shows aired between September 1 and 30, 2018 (darker means more biased).

3.2 Analysis 1: refugee bias over time

We analyze how refugee biases on talk radio vary by day between August 15 and September 15, 2018. We choose this interval to center on the August 31, 2018 news story regarding the Trump administration's contentious decision to pull funding from a UN agency that supports Palestinian refugees[5]. Our attribute of interest is the day in which a particular snippet occurred. Figure 3(b) illustrates the temporal variation in bias scores, highlighting a notable shift towards *greater bias* against refugees in response to the news story. Interestingly, bias towards refugees returns to pre-event levels very quickly after the spike. Computing the correlation between daily bias scores and the number of mentions of the keyword "refugee" across stations yields $r = 0.56, p < 0.001$, suggesting that additional discourse about refugees tends to be biased against them.

As a comparison, we also compute bias scores by training one Word2Vec model per day and projecting all day-by-day models into the same vector space using orthogonal Procrustes alignment[6] similar to (Hamilton et al., 2016). The resulting scores from this non-dynamic model are depicted in 3(a). From qualitative inspection, the day-by-day scores produced by the non-dynamic model appear much less smooth, and hence, fail to show the relative shift in discourse that likely occurred in response to a major refugee-related news event. One possible reason for this is that the median number of words for each day in the talk radio corpus is 4 million—over 5x fewer than a median of 22 million words per decade used to train each decade-specific model in (Garg et al., 2018). These results suggest that using our dynamic embedding approach is particularly valuable when data is sparse for any given attribute.

3.3 Analysis 2: refugee bias by city

Next, we analyze how bias towards refugees varies by city for talk radio produced between September 1 and 30, 2018. We first train our model to learn a city-specific embedding for each word

[5] For historical coverage of different refugee-related news events, please see https://www.nytimes.com/topic/subject/refugees-and-displaced-people.

[6] We use the Gensim implementations of Word2Vec and orthogonal Procrustes alignment, aligning hyperparameters as closely as possible to our dynamic model.

and then use these embeddings to compute corresponding bias scores, which are depicted in figure 4. Qualitatively, cities in the Southeastern US, those closer to the US-Mexico border, and some that have suffered from economic decline in recent years (e.g. Detroit, MI; Youngstown, OH) tend to have talk radio coverage that is more biased towards refugees, though the trends are quite varied. Interestingly, there is a weak negative, though marginally insignificant, correlation between the level of bias per city and the number of refugees the city admitted in 2017[7] ($r = -0.21, p = 0.1$). This relationship persists even after controlling for state fixed effects. A more thorough analysis with additional cities and other city-level covariates may reveal meaningful patterns and perhaps even help illuminate which geographies are particularly welcoming towards refugees.

4 Conclusion

In this paper, we present a unified dynamic word embedding model mirroring the earlier work of (Bamman et al., 2014) to learn attribute-specific embeddings. We validated our model by replicating gender and ethnic stereotypes produced in (Garg et al., 2018) by training multiple word embedding models and applied it to a novel corpus of talk radio data to analyze how perceptions of refugees as "outsiders" vary by geography and over time. Our results illustrate that dynamic word embeddings capture salient shifts in public discourse around specific topics, suggesting their potential usefulness as a tool for obtaining a granular understanding of how the media and members of the public perceive different issues, especially when data is sparse.

Opportunities for future work include a) comparing the results of our model to other existing dynamic embedding models, particularly when the attribute of interest is temporal in nature, b) exploring embeddings defined with respect to other attributes of interest, perhaps in combination with other contextual embedding models like (Peters et al., 2018), c) exploring alternative definitions of bias towards refugees and other groups, and d) learning a dynamic embedding model for continuous attributes in order to limit the need to impose (perhaps arbitrary) discretizations. We believe these approaches hold promise in helping us illuminate evolving attitudes and perceptions towards different issues and groups across a rapidly expanding digital public sphere.

Acknowledgments

We thank Doug Beeferman, Prashanth Vijayaraghavan and David McClure for their valuable input on this project. We also thank Nikhil Garg for sharing prior results to help us benchmark our model.

References

R. Bamler and S. Mandt. 2017. Dynamic Word Embeddings. In *Proceedings of the International Conference on Machine Learning (ICML).*

D. Bamman, C. Dyer, and N. 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).*

D. Beeferman and B. Roy. 2018. Making radio searchable. `https://medium.com/cortico/making-radio-searchable-f337de9fa325`. Accessed: March 10, 2019.

A. Caliskan, J. J. Bryson, and A. Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

M Davies. 2010. The 400 million word corpus of historical American English (1810 2009). In *Selected Papers from the Sixteenth International Conference on English Historical Linguistics (ICEHL 16).*

N. Garg, L. Schiebinger, D. Jurafsky, and J. Zhou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

W. Hamilton, J. Leskovec, and D. Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.

Q.V. Le and T. Mikolov. 2014. Distributed Representations of Sentences and Documents. *arXiv: 1405.4053.*

M. Lewis and G. Lupyan. 2018. Language use shapes cultural norms: Large scale evidence from gender. In *The Annual Meeting of the Cognitive Science Society*, pages 2041–2046.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. *NIPS*.

[7] We sourced per-city 2017 refugee arrival numbers from the Refugee Processing Center's interactive reporting webpage: `http://ireports.wrapsnet.org/`.

S. Mort. 2012. Tailoring Dissent on the Airwaves: The Role of Conservative Talk Radio in the Right-Wing Resurgence of 2010. *New Political Science*, 34(4):485–505.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. *arXiv: 1802.05365*.

Pew. 2018. Audio and podcasting fact sheet. http://www.journalism.org/fact-sheet/audio-and-podcasting/. Accessed: March 10, 2019.

M. Rudolph and D. Blei. 2018. Dynamic Embeddings for Language Evolution. In *WWW 2018: The 2018 Web Conference*, pages 1003–1011.

UNHCR. 2017. Forced displacement in 2016. Global Trends Report.

Z. Yao, Y. Sun, W. Ding, N. Rao, and H. Xiong. 2018. Dynamic Word Embeddings for Evolving Semantic Discovery. In *Proceedings of the The Eleventh ACM International Conference on Web Search and Data Mining (WSDM)*.