Incorporating Demographic Embeddings into Language Understanding

Justin Garten[1], Brendan Kennedy[1], Joe Hoover[1], Kenji Sagae[2], Morteza Dehghani[1]

[1]Computational Social Science Laboratory,University of Southern California, Los Angeles, CA 90089 USA

[2]Department of Linguistics, University of California, Davis, Davis, CA 95616

Affiliation

To be Published in Cognitive Science

Author Note

This research was sponsored by the Army Research Lab. The content of this publication does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. Correspondence regarding this article should be addressed to Morteza Dehghani, mdehghan@usc.edu, 3620 S. McClintock Ave, Los Angeles, CA 90089-1061.

Abstract

Meaning depends on context. This applies in obvious cases like deictics or sarcasm as well as more subtle situations like framing or persuasion. One key aspect of this is the identity of the participants in an interaction. Our interpretation of an utterance shifts based on a variety of factors including personal history, background knowledge, and our relationship to the source. While obviously an incomplete model of individual differences, demographic factors provide a useful starting point and allow us to capture some of this variance. However, the relevance of specific demographic factors varies between situations—where age might be the key factor in one context, ideology might dominate in another. To address this challenge, we introduce a method for combining demographics and context into situated demographic embeddings—mapping representations into a continuous geometric space appropriate for the given domain, showing the resulting representations to be functional and interpretable. We further demonstrate how to make use of related external data so as to apply this approach in low-resource situations. Finally, we show how these representations can be incorporated to improve modeling of real-world natural language understanding tasks, improving model performance and helping with issues of data sparsity.

Keywords: natural language processing; demographic representation; continuous representations; neural networks; moral reasoning

Incorporating Demographic Embeddings into Language Understanding

The meaning of an utterance depends on context. This is trivially true with deictics (Fillmore, 1997) where a phrase such as "this goes here" tells us little outside of a context. But it is generally true when considering language as a record of the intents, actions, and reactions of real humans. Nonetheless, the current dominant techniques in computational language understanding generally ignore these factors. Semantic representations are trained on large bodies of "neutral" data, yielding results which capture only certain facets of human language understanding (Griffiths, Steyvers, & Tenenbaum, 2007). Supervised models are based on corpora of fragmented text and labels, isolated from their sources. These techniques are extremely powerful and have proved very useful—for tasks and questions where we can ignore local variation as noise. But if we want to go beyond that subset of tasks, context becomes increasingly important.

One critical aspect of this is the identity of the participants in a given interaction. Prior work in fields including linguistics, philosophy, and cognitive science (Carnap, 1947; Fodor, 1981; Frege, 1892) has considered the importance of the interaction between individuals and language as essential to understanding both the intentionality and meaning of utterances in context. At the simplest level, two individuals can interpret the same statement differently—at times very differently—and dismissing these differences as "mistakes" or "noise" only serves to limit our modeling efforts to the most arid of settings. But the choice to focus on statistical regularity was not made from ignorance, but rather due to practical limitations of data, computational resources, and methods. Modeling the variation in human response can easily devolve into a study of exceptions, treating each interaction as novel and losing sight of the patterns and structure which make language possible at all.

These arguments about the importance of individual-level differences in the meaning of language are supported by work in the representation of words. When measuring response time and performance in a variety of retrieval and recognition tasks, the

"Contextual Diversity" — the amount of distinct contexts in which a word has been seen — is a better predictor of performance than the word's overall frequency statistics (Adelman, Brown, & Quesada, 2006; Chen et al., 2017; Johns, Dye, & Jones, 2016; Plummer, Perea, & Rayner, 2014). This comparison, between viewing lexical qualities of language as fixed versus varying across people's experience of them, mirrors the comparison at the semantic level. What language means to an individual is not a result of the inherent qualities of language, but of their unique prior experiences of both language and other stimuli.

Rather than attempt to model the full scope of human variation, we focus on a simplified representation of individuals, using demographics as an obviously incomplete but useful starting point. However, just as a sentence or phrase doesn't have a single meaning across contexts, the impact of identity, especially as summarized by demographic factors, varies depending on the situation. This is clear even from surface factors such as word choice. Whereas regional variation will explain whether we expect someone to use "crawfish" versus "crawdad" or "boot" versus "trunk", age will be far more informative as to the likelihood of observing "hook up" versus "Netflix and chill." Generally, no single measure of concept similarity will capture the range of human or situational variation (Goldstone, Medin, & Gentner, 1991; Tversky, 1977).

In this paper, we explore combining demographics and context into situated demographic representations and demonstrate how these can be applied to modeling language understanding. The goal is to take raw demographic variables and transform them so as to capture situational rather than abstract similarity between individuals. Our approach is to learn a mapping from demographics into a continuous geometric representation where measures of similarity are meaningful for the given task. Just as we don't expect to learn language anew with each interaction, language processing generally depends on representations trained on large quantities of external data. Our method is similarly capable of training situated demographic representations on related, higher-resource, tasks. This allows for wider applicability and opens the potential for

exploring the relationship between the context in which a representation is learned and the domain in which it's applied.

We apply this approach to an area where prior psychological research has found demographic-driven differences in individual responses: moral reasoning (Haidt, 2012). For example, liberals and conservatives have been found to respond differently to a wide range of moral concerns (Graham, Haidt, & Nosek, 2009). We compare modeling participants' responses based on (1) representations of the text by itself (2) text plus demographic factors and (3) text plus situated demographic factors encoded using our approach. We show that the incorporation of situated demographic representations is better able to model participant responses. We follow this by exploring a practical issue where distributed representations have proved useful in other domains: improving modeling with sparse or missing data.

Prior work has looked at how incorporating demographic information about speakers can improve classification performance on a range of tasks (Hovy, 2015; Johannsen, Hovy, & Søgaard, 2015) and has proven particularly valuable in venues such as social media (Hovy & Søgaard, 2015) where group-based linguistic variations can be extreme. Word representations have also been learned on demographically split data (Bamman, Dyer, & Smith, 2014; Garimella, Banea, & Mihalcea, 2017) which has been shown to be useful for community-specific classification in areas such as sentiment analysis (Yang & Eisenstein, 2015). Our work differs from these in modeling how different individuals will respond to and interpret a piece of language rather than focusing the original intent of the source. We further extend prior work by demonstrating a method for generating and applying domain-specific demographic representations to model the situational variability of similarity judgements. Finally, by learning those models on outside sources of information, we demonstrate a method for transferring this type of contextual understanding between models.

The main contributions of this paper are: (1) a novel method for encoding

demographics and context into a combined continuous representation; (2) an exploration of the use of those representations on a concrete language understanding task; (3) a demonstration of the utility of demographic information when faced with missing data; and (4) the release of a new dataset of responses to the moral vignette stimuli developed by (Clifford, Iyengar, Cabeza, & Sinnott-Armstrong, 2015) along with demographic information on the respondents.

Demographic Embeddings

In this section, we introduce a general method for learning situated demographic embeddings. Some of the factors we have already discussed provide useful constraints on the choice of modeling approach. First, since even the trivial examples we have discussed so far involve non-order preserving transforms, the method must be able to learn non-linear mappings. For example, given three individuals (A, B, C), imagine that we had a task focused on regional variation with A from California, B from Kansas, and C from Ohio. In that case we might find that A is most similar to B with C more similar to B than A yielding: the order A, B, C. But for another task where age was most salient, imagine that A was 20, C 25, and B 60. Here we might prefer the order: A, C, B. Given that linear maps can only preserve or reverse order metrics, to allow for these two possibilities we require the ability to learn non-linear transformations.

Second, in order to allow geometric interpretability and facilitate downstream interaction with other semantic representations, we prefer a method which yields a continuous representation. Finally, as learning interaction effects requires multiple times more data than learning first-order effects, we require a method which can handle larger quantities of data. Given these factors, we chose to generate representations using a shallow feedforward neural network model.

The overall approach is to train a feedforward neural network model which learns to predict individual responses to related questions or topics. The structure of simple neural

networks can be best thought of as a series of nested functions (with each layer of the network serving as the next function in the series). We begin with an input x, in this case, the demographic information about the respondent. This is passed to the first layer of the network which outputs a low-dimensional intermediate representation $f(x)$. This is then passed to the next layer which then outputs a prediction as to the response to a particular item, $g(f(x))$. However, after training this network, we then throw away that final layer, instead using the function $f(x)$ (the first layer of the neural network) directly to yield a representation of the user optimized to the domain in question.

The size of the hidden layer determines the size of the resulting embeddings. Effectively, the hidden layer is learning a representation optimized for the output tasks. This is similar to the approach of Mikolov, Chen, Corrado, and Dean (2013) but, where they learned word similarity based on predicting other words in the local context, we learn demographic similarity based on the chosen output tasks. The choice of which tasks to use for this training depends on the domain at hand.

After training, network weights are saved and a new network is initialized consisting of only the input and hidden layers. At this point, arbitrary demographic inputs may be entered with the output of the hidden layer yielding the representation for that input. This is different from situations like word embeddings where training yields representations for a fixed subset of words from the training set. Instead, we learn a mapping function which can be applied to arbitrary future inputs. This flexibility allows the generation of representations not only for observed cases, but also for previously unseen examples or even examples with missing data.

Training Moral Embeddings

In our experiments below, we make use of a version of this architecture trained to learn demographic embeddings adapted for use in the moral reasoning domain. The goal is to leverage a much larger set of related data in order to learn how to map demographics

into a domain-specific geometric space. We train on responses to the Moral Foundations Questionnaire (MFQ) collected on the YourMorals.org website[1] (Graham et al., 2011). This dataset included 133,237 individuals who had both answered the full set of 32 MFQ questions and provided the required demographic information.

As seen in Figure 1, the network is structured as a multi-task network (Collobert & Weston, 2008). Categorical demographic variables are converted to one-hot representations[2] and concatenated to generate the input vector. In terms of the previous description of the overall architechture, this is the input x. This feeds in to a single 20 dimensional hidden layer (the value of 20 was selected by grid search over values ranging from 2 to 50) with rectified linear activation functions (Nair & Hinton, 2010). This is what we described as the function $f(x)$ which learns how to translate raw demographics into a domain-specific latent representation. Finally, the output from the hidden layer is connected to dense output layers for each of the 32 individual MFQ questions ($g(f(x))$). Training makes use of the Adam optimizer (Kingma & Ba, 2014) with models trained to convergence based on a randomly selected 20% validation sample (generally 50-100 epochs). The average mean squared error over the 32 MFQ questions was 1.571. One area for future improvement is further optimizing this network to see how this affects the quality of the resulting representations.

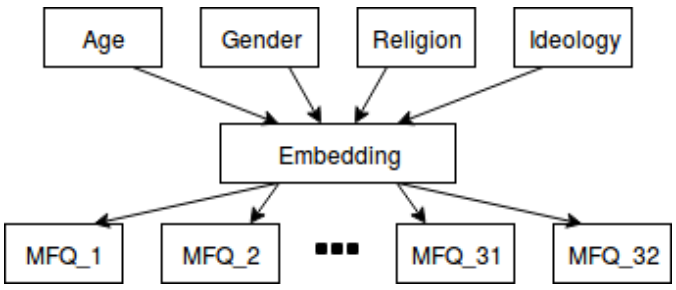

Figure 1. The multi-task training network.

[1] Specific survey questions and collection information available at `http://www.yourmorals.org`

[2] One-hot representations are vectors where each element represents a possible value with the selected value set to one while all other values are set to zero.

After this model is trained, the final prediction layers are discarded. In this case, that was the part of the network focused on predicting responses to MFQ questions. What remains are the inputs for the raw demographic factors and the intermediate representation layer which learned how to translate those raw demographics into a representation optimized for predicting moral responses. What was previously a hidden layer in the middle of the network is now the final output of the network. In the formulation described above, after learning $g(f(x))$, we discard $g(.)$ and directly make use of the output of the intermediate function $f(x)$. The outputs of this ablated network are used as the representation of the input demographic values. Intuitively, the use of 20 dimensional embeddings allows the model to capture interaction effects between the underlying demographic variables (as opposed to dimensionality reduction). However exploring the precise differences in structure between the learned networks is left for future work.

We used the YourMorals dataset to train moral-focused demographic embeddings and then applied the resulting mapping function to the 950 participants in the study discussed below. We found that these representations, in spite of being trained on a different (though clearly related) task, combined to yield the best overall model for predicting participant reactions.

Observations on Moral Space

Before considering experimental results, it is useful to consider the structure of this mapping from raw demographics to moral space. Given that this method generates a geometric space, it is simple to apply standard measures of vector space similarity. The effects of this transform, in particular inversions of ordering between the spaces, proved to be highly intuitive. For one observed example of this, consider three users, U_1: 30 years old religious, conservative, U_2: 30 years old non-religious, liberal, and U_3: 60 years old non-religious, liberal. In demographic space, U_2 is strictly closer to U_1 than U_3. That is, $Dist(U_1, U_2) < Dist(U_1, U_3)$ for any standard distance metric. But, as seen in Figure 2, in

the transformed moral space, U_3 is closer to U_1 than U_2 ($Dist(U_1, U_2) > Dist(U_1, U_3)$)[3]. This makes intuitive sense in the context of contemporary US politics and culture as we would expect, all else being equal, that an older liberal would be slightly more conservative than a young liberal (Truett, 1993). The mapping learned is non-affine (in this case, non-order preserving on the distance metric) but is capturing precisely the sort of relations we would wish to find.

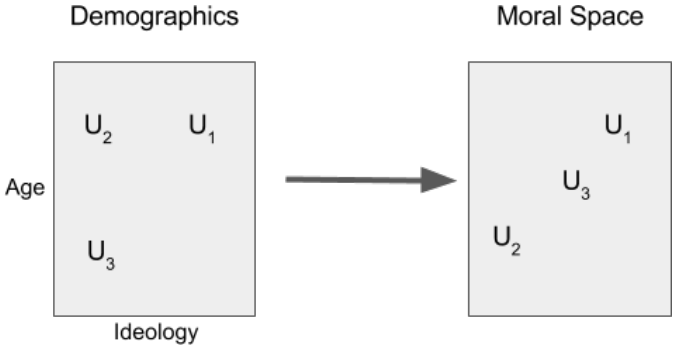

Figure 2. Example transform into moral space.

Capturing this type of relationship opens the potential of applying this method not just for use in predictive modeling, but also for directly exploring the structures of these similarity spaces. In the example of the transformation described above, on another domain (for example, questions relating to media consumption) we wouldn't expect the same inversion (where two thirty-year-olds would likely be more similar regardless of political/religious leanings). This opens the potential of directly exploring the space of transforms.

Study 1: Incorporating Demographic Information into Language Understanding

While we have considered some of the qualitative aspects of these embeddings, in this section we evaluate their usefulness for a concrete natural language understanding task. In particular, we model responses to a set of moral vignettes (Clifford et al., 2015), 14-17

[3] In the model used in the paper, the cosine similarity between U_1 and U_2 is 0.9757 while the similarity between U_1 and U_3 is 0.9796

word stories designed to evoke a particular moral concern which participants evaluated in terms of degree of moral wrongness and level of emotional response. This domain was chosen based on prior work which has found a strong correlation between moral reasoning and demographic factors (Graham et al., 2009; Haidt, 2012).

To evaluate this, we compare the results of predicting responses based on: (1) continuous representations trained from text features only, (2) adding demographic representations to those text representations, (3) adding demographic cluster information to the text representations, and (4) adding domain-adapted demographic embeddings to the text representations.

These allow for several useful evaluations and comparisons. Text features are useful in evaluating how much variance is explained by the story itself, especially in light of work suggesting that such features can explain the majority of variance in many survey studies (Arnulf, Larsen, Martinsen, & Bong, 2014). The addition of demographic features allows us to compare how much general demographics improves our ability to model these responses. The comparison of demographic embeddings with raw demographic information allows evaluation of the impact of situational adaptation. In this case, does mapping general demographic information into morality-focused continuous representations improve modeling? Finally, comparing demographic embeddings with demographic clusters lets us evaluate whether changes in model accuracy are due to simple noise reduction from the embedding or whether the information contained in the outside task was helpful.

However, given that we collected at least 50 responses for each primary demographic group (see Dataset section below), this provides a unique test case. Effectively, this is the best possible case for the raw demographic factors in that there is sufficient information available to directly learn the relationships between factors and interactions for this domain. This is not a real-world situation (for that, see Study 2) as even in the most extensive studies and surveys there are almost always gaps in the data coverage. However, it does allow us to evaluate two key factors. First, we can confirm prior work on the basic

importance of demographic factors on moral judgements. Second, we can compare the results of raw demographics and embeddings in a context where we would expect embeddings to provide minimal improvement. This is critical for setting up further evaluations and verifying the basic functioning of the method.

Dataset

We make use of a collection of "moral vignettes" by Clifford et al. (2015), short (14-17 word) scenarios. These vignettes were designed to capture a scene, or context, which evokes a potential moral violation (for example, in the fairness domain: "You see a runner taking a shortcut on the course during the marathon in order to win."). The vignettes make use of Moral Foundations Theory (Greene & Haidt, 2002; Haidt, 2001, 2003), which classifies moral concerns into five domains: Care/harm, Fairness/cheating, Loyalty/betrayal, Authority/subversion, and Sanctity/degradation (Haidt, Graham, & Joseph, 2009). The original authors (Clifford et al., 2015) generated their final validated set by having approximately 30 annotators evaluate each candidate vignette with the final collection consisting of those which reached a minimum of 70% agreement as to the primary moral domain.

For this study, we selected the five most agreed upon vignettes for each moral foundation, 10 most agreed upon non-moral stories (those lacking any moral content), and the five stories with the highest combination of degree of violation and disagreement between annotators. This yielded a total of 40 vignettes.

Based on a target of at least 50 responses to each vignette for the primary demographic groups, we recruited a total of 950 participants through Amazon Mechanical Turk (allowing margins for platform demographic bias and attention check failures). Each was asked to evaluate a random subset of 10 of the vignettes on the basis of which foundation was involved (or none), the degree of moral wrongness, and their subjective level of emotional response to the story. Additionally, for each participant we collected

demographic data including age, religious affiliation and religiosity, gender, and political affiliation. We limited the demographic categories collected to those that we had sufficient data for in the MFQ dataset. This yielded a total of 9500 individual vignette evaluations with paired demographic data.

Method

Text representations were generated using *InferSent* (Conneau, Kiela, Schwenk, Barrault, & Bordes, 2017), which learns a model for generating continuous sentence representations based on a range of natural language inference data. For this task, we trained a model[4] to yield 256 dimensional text embeddings. We additionally trained a model with 4096 dimensions (the number used by the authors to achieve optimal prediction results), but the difference in performance on our task between 4096 and 256 was negligible. The resulting 256-unit model was used to generate representations for each of the vignettes used in the study.

Models were trained to predict users' responses to the questions "How morally wrong is the action depicted in this scenario?" and "How strong was your emotional response to the behavior depicted in the scenario?". Given the small overall size of the dataset and the fact that our goal was to explore model differences rather than maximize predictive accuracy, linear regression models with elastic net regularization[5] were estimated for each of the following feature sets: (1) text representations; (2) text representations and raw demographic variables; (3) text representations and k-means demographic cluster membership; and (4) text representations and demographic embeddings.

Each model was fit with 10 repetitions of 10-fold cross-validation. Then model

---

[4] Code available from:

`https://github.com/facebookresearch/InferSent`

[5] Elastic net regularization is a regression method that linearly combines the penalties of Lasso (L_1) and Ridge (L_2) regularization methods.

performance was evaluated, in terms of R^2, via a second round of ten repetitions of 10-fold cross-validation, yielding 100 estimates of R^2 for each feature set. Finally, to compare model performance across feature sets, permutation tests with 10,000 iterations were conducted for each pair-wise combination of R^2s in order to evaluate the evidence that a given model explained more variance than a comparison model. Permutation tests were used because R^2 is not normally distributed (Ohtani, 1994) and permutation tests provide non-parametric (Manly, 2006; Menke & Martinez, 2004; Smucker, Allan, & Carterette, 2007) estimates of significance. Specifically, permutation tests test the null hypothesis that two vectors of values (e.g. our vectors of R^2s) are drawn from the same sampling distribution, which is conventionally operationalized as the hypothesis that they have the same measure of central tendency. This is accomplished by randomly sampling two new vectors from the observed vectors, which, per the null hypothesis, are treated as a single sample. For each pair of new vectors, their measures of central tendency are compared and this comparison is stored. A p-value for the null hypothesis can then be calculated as the ratio of the number of times the absolute value of the permuted test statistic exceeded the absolute value of the observed test statistic.

The included demographic features were age, gender (free-response with results mapped to male/female/other), political ideology (liberal, conservative, moderate, other), and religiosity (five levels ranging from highly to non religious). Religious affiliation was dropped for two reasons. First, the dataset was heavily skewed towards Christianity (85% of the total responses) making the feature largely a repetition of the religiosity variable. Second, as a free-response form, the data proved to be both noisy and subject to ambiguous interpretation (e.g. should "Protestant" and "Christian" be separated?).

Clusters were generated based on participant demographics using the *k-nearest neighbors* algorithm (Fix & Hodges Jr, 1951). k was determined based on comparing model performance over values ranging from $k = 2$ to $k = 10$. The best performance was found at $k = 2$, which was used for the results reported below.

As described previously, a mapping function for generating demographic embeddings was learned using a separate dataset of answers to the Moral Foundations Questionnaire. We applied this map to the participants in this study, generating a 20 dimensional demographic embedding for each.

Results and Discussion

Overall, there are several key points. First, demographic information (in any form) improves our ability to predict responses to moral scenarios[6]. This is true in all cases except for non-moral scenarios (for these, due to most participants rating them as generating no emotional response and expressing no moral wrongness, there was simply no variance to explain). Finally, both raw demographics and moral embeddings were significantly better at predicting responses than demographic clusters ($p < 0.0001$).

The low R^2 values when working with only text features is unsurprising given the nature of the data. These vignettes were designed to avoid using similar words or words from the descriptions of the foundations themselves (Clifford et al., 2015), in effect being structured to limit the value of simple textual features. All methods of including demographic information showed significantly improved performance over the text-only models. However, both the use of raw demographic information and moral embeddings did significantly better ($p < 0.0001$) than making use of demographic clusters. As the clustering method reduces multiple demographic factors to a choice between, in this case, two possible clusters, it appears that the potential advantage of noise reduction is in this case outweighed by the loss of information. While the fact that demographic information provided the largest relative improvements for the Authority domain makes intuitive sense in terms of the increasingly wide US political divides around the use and abuse of power

[6] Given the known importance of political ideology, we separately tested by removing this factor to check if it alone was driving these results. While R^2 values were slightly lowered, the overall effects and patterns were the same as with the full set of variables.

(circa 2018), it doesn't match with prior work which found larger gaps in the Sanctity domain (Haidt & Graham, 2007). Exploration of whether this points to a general shift in attitudes or is a feature of the local dataset will require targeted experimental follow-ups.

Table 1

Mean R^2 with standard error by model and domain predicting participants' reported emotional response

	Domain	Text	Text + embedding	Text + demographics	Text + K-mean
1	Authority	0.018 (0.0011)	0.127 (0.0031)	0.120 (0.0031)	0.067 (0.0024)
2	Care	0.026 (0.0014)	0.053 (0.0017)	0.052 (0.0017)	0.026 (0.0013)
3	Fairness	0.139 (0.0033)	0.167 (0.0034)	0.163 (0.0035)	0.137 (0.0033)
4	Loyalty	0.071 (0.0022)	0.108 (0.0028)	0.103 (0.0027)	0.088 (0.0025)
5	Sanctity	0.092 (0.0028)	0.133 (0.0032)	0.133 (0.0032)	0.109 (0.0028)
6	Non-moral	0.011 (7e-04)	0.012 (7e-04)	0.010 (6e-04)	0.014 (8e-04)

The comparison between adapted demographic embeddings and raw demographic representations is both more complex and more interesting. Overall, the performance of the domain-adapted embeddings and the raw demographics was extremely similar. While the results were distributionally different[7], the practical differences between the two were negligible. This fits with the structure of the experiment in that given the large numbers of data points for each condition, there was little missing information for the embeddings to capture. Given the small effect size of the differences, the particulars of these differences should be treated as at most suggestive. Nonetheless, the embeddings provided a small improvement over all categories ($p < 0.0001$ for Authority, Fairness, and Loyalty and

[7] Random permutation tests allow us to determine whether two datasets were drawn from the same underlying distribution.

Table 2

Mean R^2 with standard error by model and domain predicting participants' evaluation of degree of moral wrongness

	Domain	Text	Text + embedding	Text + demographics	Text + K-mean
1	Authority	0.059 (0.0023)	0.144 (0.0034)	0.141 (0.0032)	0.102 (0.0028)
2	Care	0.051 (0.0022)	0.085 (0.0026)	0.084 (0.0026)	0.05 (0.0021)
3	Fairness	0.222 (0.0035)	0.247 (0.0036)	0.245 (0.0037)	0.233 (0.0035)
4	Loyalty	0.108 (0.0029)	0.127 (0.003)	0.125 (0.003)	0.115 (0.0029)
5	Sanctity	0.064 (0.0021)	0.109 (0.0024)	0.108 (0.0025)	0.085 (0.0022)
6	Non-moral	0.009 (5e-04)	0.01 (6e-04)	0.009 (5e-04)	0.01 (7e-04)

$p < 0.05$ for Care) except Sanctity ($p = 0.408$). Results were similar for evaluating model differences in predicting moral wrongness ($p < 0.0001$ for Fairness and Loyalty and $p < 0.05$ for Authority and Care) with the exception of Sanctity ($p = 0.0393$). The improvement observed when predicting emotional response suggests that higher order interactions may be more important there, but more targeted follow-ups would be required to confirm this.

Practically, though, in cases where you have access to 50 examples per case and situation, the additional complexity would not justify the use of embeddings. However, that is almost never the situation in real-world studies and surveys. This motivated study 2 where we explore more real-world conditions.

Study 2: Demographic Embeddings for Missing Data

While embeddings in other contexts have been explored in terms of the abstract representations they encode, they have also proven extremely useful for dealing with practical modeling challenges. One such issue is the general challenge of dealing with

missing and sparse data in high dimensional contexts (Bühlmann & Van De Geer, 2011). The ability to learn similarity metrics between words has allowed for far more generalizable approaches to a range of downstream language modeling tasks (Bengio, Ducharme, Vincent, & Jauvin, 2003). For example, a classifier trained for sentiment might never have seen the word "tedious" but if "boring" and "tiresome" were both in the training set, the fact that "tedious" is distributionally similar would allow for a much better prediction than just treating "tedious" as an unknown word (a common previous approach for dealing with this issue).

In this study, we explore whether similar advantages might hold true for demographic embeddings. If our training data is missing particular groups, can we make use of domain-specific similarity to improve the quality of subsequent predictions? If a model never saw a particular subgroup in training, how well can it predict that group's behavior when encountered later? We compare making use of demographic embeddings with applying the raw demographic features (the two best performing models from Study 1).

Method

The general method is similar to that of Study 1 in that we first train models based on a subset of the data and then test those models against a held out subset. The difference comes in how the data is split. Rather than randomly-sampled cross validation, in this study, we iteratively remove a single group, training on all remaining data and testing solely on that missing group (models are still trained with cross validation to avoid overfitting the training set).

We first iterated through all possible values of the demographic variables (age, gender, religiosity, and ideology) with age bucketed into the groups: 0-30, 31-50, and 50+. A separate test/train split was generated for each of these values by putting all participants with the selected value into the test set and leaving all other participants as the training set. For example, when testing the case of age=31-50, all participants in that

age range (regardless of other demographic factors) were separated and held out as a test group. Thus, during training, the model would never observe anyone with the selected demographic value. This yielded a total of 13 separate train/test splits for the data (one for each possible missing category).

A similar procedure was then repeated for each possible *pair* of factors, with each possible value for each demographic factor combined with all possible values for the other factors. For example, starting with ideology="conservative", we considered each possible combination with values of religiosity, gender, and age. This yielded a total of 84 possible combinations.

For each data split generated either by a single factor or pair of factors, we compare two ways of encoding the situation and generating features for modeling based on the two best performing models from Study 1. The first makes use of the full set of raw demographic values with the sentence representations ("raw demographics"). The second combines the sentence representations with pre-trained demographic embeddings ("embedding").

In all cases we follow the procedures outlined in Study 1: training linear models with elastic net regularization. Models are optimized based on 10-fold cross validation over the training set with pre-processing applied to center, scale, and remove zero-variance variables. Model performance is evaluated based on R^2 of the test set. Model performance is compared via permutation tests with 10,000 iterations. These steps were repeated separately for both the prediction of participants' reported degree of subjective emotional response to a vignette and their reported evaluation of the degree of moral wrongness embodied in the vignette.

Results and Discussion

The use of demographic embeddings did significantly better than the raw demographic model across all moral domains except Care. This was observed for both

Table 3

Mean R^2 with standard error by model and domain predicting participants' reported emotional response over missing data models

	Domain	Raw demographics	Embeddings	p-value
1	authority	0.0617 (0.0149)	0.0825 (0.0146)	0.03
2	care	0.0423 (0.0103)	0.0404 (0.006)	0.82
3	fairness	0.0859 (0.0111)	0.1272 (0.0084)	0.00
4	loyalty	0.0614 (0.0103)	0.0903 (0.0089)	0.00
5	sanctity	0.1124 (0.0161)	0.1335 (0.0168)	0.00
6	non-moral	0.0176 (0.0039)	0.0135 (0.0026)	0.19

participant evaluations of subjective emotional response (Table 3) and evaluations of the degree of moral wrongness of a given vignette (Table 4). The primary difference was that the raw demographic model was trained on values for variables which were not present in the test set. While for some variables (age and religiosity), simple linear interpolation would provide some information, as previously noted, many of the observed transformations into moral space were non-linear. For categorical variables (ideology and gender), the situation was even worse where unseen values were initialized to zero[8].

The two exceptions to this pattern were the Non-moral and Care categories, neither of which showed a significant difference between models. For the non-moral category, this is similar to what was observed in Study 1. If we specifically adapt a model to a particular domain, the better that adaptation, the worse the model is likely to do outside of that domain (especially in a case such as this where examples were selected specifically to avoid overlap with the domain in question). However, the reason for the difference on the Care domain is not immediately obvious. Prior work on moral values has suggested that this domain is connected to the basic patterns of reproduction and attachment (Haidt, 2012)

[8] In initial experiments, those values had been randomly initialized leading to far worse performance

Table 4

Mean R^2 with standard error by model and domain predicting participants' evaluation of degree of moral wrongness over missing data models

	Domain	Raw demographics	Embeddings	p-value
1	authority	0.0774 (0.0165)	0.1244 (0.0147)	0.00
2	care	0.0634 (0.0144)	0.0702 (0.0113)	0.38
3	fairness	0.122 (0.0123)	0.204 (0.0104)	0.00
4	loyalty	0.0567 (0.0074)	0.1082 (0.0086)	0.00
5	sanctity	0.0719 (0.0151)	0.1048 (0.018)	0.00
6	non-moral	0.0216 (0.0061)	0.0203 (0.0046)	0.77

and is one of the two "binding" concerns (along with Fairness) that are thought to show less variance over demographic groups. However, prior work focused on political differences (Graham et al., 2009) found less variation in the "fairness" domain between ideological groups in the US (where we found a significant difference), potentially undercutting this hypothesis. More targeted follow-up work will be required to evaluate the precise reasons for this finding.

The use of domain-adapted demographic embeddings depends on the availability of sufficient quantities of domain-related data. In this case, we were able to make use of a much larger set of questionnaire data on a closely related topic. When available, this allows for highly relevant embeddings to be trained and offers a viable response to the issue of missing data. However, as the amount of related data declines and/or the closeness of the relationship between the domains diminishes, the effectiveness of demographic embeddings is likely to vary more widely. The further exploration of the importance of these tradeoffs in a wider range of domains remains an interesting area for future work.

Discussion and Future Work

Incorporating contextual factors into computational language understanding is important both in terms of advancing practical modeling as well as improving our theoretical understanding of linguistic representation. Just as our understanding of semantics has advanced due to the ability to compare the information captured by different language embedding models, we hope that consideration of context can similarly advance our understanding of pragmatics. One key facet of this is exploring questions of representing language in context—in particular modeling the interactions between individuals with diverse backgrounds, intents, and beliefs.

The present work aims to contribute to this goal by exploring the question of how to combine representations of language and individuals given a particular situation. Our primary contribution is a novel method for learning domain-specific demographic embeddings and validating their potential for inclusion in natural language understanding tasks. Additional contributions include (1) providing a new dataset of responses to moral stories which extends existing work in the domain, (2) confirming prior work on the importance of demographic differences on responses to moral concerns and demonstrating the value of that information in predictive modeling.

The general approach of training situational demographic embeddings has potential applications to a wide range of questions and tasks. Given that there is no single measure of similarity which captures the range of human responses across contexts, focusing instead on domain-specific representation provides a better foundation for computational modeling and better captures human intuitive characterizations.

We were particularly intrigued by the structure of the moral embedding space learned here where the relationships captured by the transformation were highly intuitive. In particular, the non-linearities in the mapping seemed to capture precisely the sort of information we would hope for in such a model. A more formal exploration of the structure of these spaces is a key area for future work.

This also provides a potentially valuable method for extending statistical models to account for differences in individual responses. If we aim to model subjective language understanding, incorporating information about the respondent will be essential. The features required to predict the response to a piece of text go beyond the text itself. This provides an intriguing point of contact between social scientific experimentation and computational language modeling.

In particular, it points to the importance of going beyond averaged responses in our statistical learning methods. Most training corpora used for language processing have focused on providing sets of text and labels. Differences in individual responses are either washed out as average values or discarded as "bad" data. This has both focused efforts on questions where we wouldn't expect high degrees of subjective difference and complicated modeling and learning given that the key information for a given model may not even be present in a training set.

Treating linguistic annotation as records of human responses under particular circumstances is a better frame for these data than "gold standard" truth on a given problem. In the case of our current study, we showed that basic demographic information provided significant improvements in model performance, but the more important general point is that annotators cannot be treated as interchangeable. Just as in any psychological study, what we are recording are human responses where disagreement has as much to tell us as agreement.

In computational language modeling, the shift to continuous representations has provided benefits including improved model performance, better handling of missing data, the facilitation of gradient-based learning, and direct advancements in the understanding of semantic representation. We believe that continuous demographic representations have the potential to provide a similar range of benefits. While we demonstrated the ability of these representations to improve model performance, as with continuous language representations, we expect that these early results will be supplanted as improved training

methods are developed. We have demonstrated here that demographic embeddings have the potential to address similar challenges.

One major question which we leave for future work is the question of how to generate representations. Here we have separately learned linguistic and demographic representations and then combined those as features for further modeling. This carries the assumption that language can be represented in the abstract, effectively filtering those abstract representations through individual differences. Another possible approach would be to generate linguistic representations directly in the context of individual differences, effectively learning personalized semantic spaces. Regardless, just as distributional semantic representations have evolved rapidly over recent years, we expect to see methods for representing individual differences go through a similar transformation.

Finally, while we focused on demographic context here, that is only one aspect of understanding a situation. There are a great many other factors to consider including local discourse context, the relationship of the speaker and listener, and shared background knowledge. While the general task of creating fully context-aware models of language understanding and response is AI-complete (i.e. essentially unsolvable in reasonable period of time), incorporation of this information in more local forms will be essential as we move into increasingly dynamic and rich domains.

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological science*, *17*(9), 814–823.

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2014). Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour. *PloS one*, *9*(9), e106361.

Bamman, D., Dyer, C., & Smith, N. A. (2014). Distributed representations of geographically situated language. In *Proceedings of d annual meeting of the association for computational linguistics (short papers)* (Vol. 828, p. 834).

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137–1155.

Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Carnap, R. (1947). *Meaning and necessity: a study in semantics and modal logic*. University of Chicago Press.

Chen, Q., Huang, X., Bai, L., Xu, X., Yang, Y., & Tanenhaus, M. K. (2017). The effect of contextual diversity on eye movements in chinese sentence reading. *Psychonomic bulletin & review*, *24*(2), 510–518.

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods*, *47*(4), 1178–1198.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on machine learning* (pp. 160–167).

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data.

arXiv preprint arXiv:1705.02364.

Fillmore, C. J. (1997). *Lectures on deixis*. CSLI publications.

Fix, E., & Hodges Jr, J. L. (1951). *Discriminatory analysis-nonparametric discrimination: consistency properties* (Tech. Rep.). Berkeley, CA: California Univ Berkeley.

Fodor, J. A. (1981). *Representations: Philosophical essays on the foundations of cognitive science*. MIT Press.

Frege, G. (1892). Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, *100*(1), 25–50.

Garimella, A., Banea, C., & Mihalcea, R. (2017). Demographic-aware word associations. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2285–2295).

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive psychology*, *23*(2), 222–262.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, *96*(5), 1029.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, *101*(2), 366.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in cognitive sciences*, *6*(12), 517–523.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Haidt, J. (2003). The moral emotions. *Handbook of affective sciences*, *11*, 852–870.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 98–116.

Haidt, J., Graham, J., & Joseph, C. (2009). Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry*, *20*(2-3), 110–119.

Hovy, D. (2015). Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (Vol. 1, pp. 752–762).

Hovy, D., & Søgaard, A. (2015). Tagging performance correlates with author age. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)* (Vol. 2, pp. 483–488).

Johannsen, A., Hovy, D., & Søgaard, A. (2015). Cross-lingual syntactic variation over age and gender. In *Proceedings of the nineteenth conference on computational natural language learning* (pp. 103–112).

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic bulletin & review*, *23*(4), 1214–1220.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Manly, B. F. (2006). *Randomization, bootstrap and monte carlo methods in biology*. Chapman and Hall/CRC.

Menke, J., & Martinez, T. R. (2004). Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons. In *Neural networks, 2004. proceedings. 2004 ieee international joint conference on* (Vol. 2, pp. 1331–1335).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann

machines. In *Proceedings of the 27th international conference on machine learning (icml-10)* (pp. 807–814).

Ohtani, K. (1994). The density functions of r2 and r2, and their risk performance under asymmetric loss in misspecified linear regression models. *Economic Modelling*, *11*(4), 463–471.

Plummer, P., Perea, M., & Rayner, K. (2014). The influence of contextual diversity on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 275.

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on conference on information and knowledge management* (pp. 623–632). New York, NY, USA: ACM.

Truett, K. R. (1993). Age differences in conservatism. *Personality and Individual Differences*, *14*(3), 405–411.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Yang, Y., & Eisenstein, J. (2015). Putting things in context: Community-specific embedding projections for sentiment analysis. *Arxiv-Social Media Intelligence*.