Demographic-Aware Word Associations

Aparna Garimella, Carmen Banea, and Rada Mihalcea
University of Michigan
Ann Arbor, MI
{gaparna,carmennb,mihalcea}@umich.edu

Abstract

Variations of word associations across different groups of people can provide insights into people's psychologies and their world views. To capture these variations, we introduce the task of demographic-aware word associations. We build a new gold standard dataset consisting of word association responses for approximately 300 stimulus words, collected from more than 800 respondents of different gender (male/female) and from different locations (India/United States), and show that there are significant variations in the word associations made by these groups. We also introduce a new demographic-aware word association model based on a neural net skip-gram architecture, and show how computational methods for measuring word associations that specifically account for writer demographics can outperform generic methods that are agnostic to such information.

1 Introduction

Understanding the associations that are formed in the mind is paramount to understanding the way humans acquire language throughout a lifetime of learning (Elman et al., 1997; Rogers and McClelland, 2004). Furthermore, word associations are believed to mirror the mental model of the conceptual connections in a human mind, and constitute a direct path to assessing one's semantic knowledge (Nelson et al., 2004; Mollin, 2009).

Word associations start forming early in life, as language is acquired and one learns based on the environment where concepts lie in relation to each other. For example, we may learn to associate "mother" with "warmth," or "fire" with "burn." Yet, this mental model is not static but highly dynamic, and is shaped by new experiences over a lifetime. For instance, (Tresselt and Mayzner, 1964) showed that word associations change with time, and that for respondents in younger age groups their variability is lower, while for those in older age groups the variability is higher, as their life experiences modify the commonality between respondents from the same group.

Computational linguistics has traditionally taken the "one-size-fits-all" approach, with most models being agnostic to the language of the speakers behind the language. With the introduction and adoption of Web 2.0, there has been an exponential increase in the availability of digital user-centric data in the form of blogs, microblogs and other forms of online participation. Such data often times can be augmented with demographic or other user-focused attributes, whether these are user-provided (e.g., from a user's online profile) or labeled using an automatic system. This enables computational linguists to go beyond generic corpus-based metrics of word associations, and attempt to extract associations that pertain to given demographic groups that would not have been possible without administering time consuming and resource intensive word association surveys.

While current NLP methods generally deal with more advanced tasks (relation extraction, text similarity, etc.), at their very core many of these tasks assume some way of drawing connections (or associations) between words. Therefore, as a step toward demographic-aware NLP, we choose to work on the core task of "word association." The algorithms we introduce can be immediately applied to demographic-aware word similarity, and with some minor changes to demographic-aware text similarity. Future stages could also include demographic-aware labeled associations, and more advanced applications such as information retrieval (which relies heavily on word associations/similarity), demographic-aware keyword extraction, dialogue personalization, and so forth.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2285–2295
Copenhagen, Denmark, September 7–11, 2017.

Note that a few other researchers have explored demographic-aware NLP models with promising results, primarily focusing on the use of demographics for various forms of text classification (Hovy, 2015) or sentiment and subjectivity classification (Volkova et al., 2013).

The paper makes several main contributions. First, we create a novel dataset of demographic-aware word associations, consisting of approximately 300 stimulus words along with 800 responses per word collected from a demographically-diverse group of respondents, for a total of 228,800 responses. Removing spam responses resulted in 176,097 responses. Analyses that we perform on this dataset demonstrate that indeed word associations vary across user dimensions.[1] Second, we show that the associations we obtained follow the same pattern as those elicited during traditional classroom surveys. Third, we propose an evaluation metric suited for the free association norms task. Fourth, we introduce a demographic-aware model based on a skip-gram architecture and through several comparative experiments, we show that we are able to surpass the performance attainable on demographic agnostic models.

We specifically focus on two demographic dimensions: location and gender. For location, we consider India and United States (US), choice made primarily because these two countries have a large English-speaking population, represented both on social media and on crowdsourcing platforms.

2 Related Work

Word associations have captured the attention of psychologists since at least the early 1900. In (1910), Kent and Rosanoff proposed the use of a set of 100 emotionally neutral words for word associations surveys. A psycholinguistics study that looked at the impact that the nationality of respondents may have on formed word associations was carried out by Rosenzweig (1961), employing the stimulus word list proposed by Kent and Rosanoff (1910) manually translated into several West European languages. Based on the primary responses coming from native speakers of English, French, German and Italian, which were mapped back into English, the author concludes that the associations formed by speakers of the four languages are very similar, with "almost half the comparisons in any pair of languages yielding agreements," where the most frequent responses are encountered across pairs of languages. Given that the primary responses were compared across languages and people with a relatively common origin (West European), our work seeks to investigate whether similar results are encountered when looking at different locations (namely US versus India). Furthermore, our study is conducted in English from the beginning, to eliminate a third party's subjectivity in mapping primary responses from one language to another.

There have also been attempts in computational linguistics to derive associations not based on survey results (which are static and resource intensive), but based on statistics derived from large corpora (Church et al., 1989; Wettler and Rapp, 1989; Church and Hanks, 1990). Research in semantic similarity can also be used to model associations based on several directions: (1) co-occurrence metrics that rely on large corpora such as PMI (Church and Hanks, 1990), second order PMI (Islam and Inkpen, 2008), or Dice (Dice, 1945); (2) distributional similarity-based measures, that characterize a word by its surrounding context such as LSA (Landauer and Dumais, 1997), ESA (Gabrilovich and Markovitch, 2007), or SSA (Hassan and Mihalcea, 2011); and (3) knowledge-based metrics that rely on resources such as lexica or thesauri (Leacock and Chodorow, 1998; Lesk, 1986; Jarmasz and Szpakowics, 2003; Hughes and Ramag, 2007). However, most of these metrics have so far been applied to model the relatedness between two words, namely given a word pair, to score how similar the two words are; as such, they have not been used to predict free association norms, namely given a word, to attempt to determine the most likely word that a human would associate with that stimulus.

Large word association databases exist, such as the one collected by Deyne et al. (2013), who used a set of 12,000 stimulus words and surveyed 70,000 participants. Yet to our knowledge, no concerted attempt has been made to gather word associations jointly with the demographic characteristics of the people behind them.

While not directly seeking to extract word associations but rather trying to represent language meaning through a locality lens, (Bamman et al., 2014) have proposed using distributed representations to model words employed by social media

[1] This work is not centered around comparing different word forms, as one would encounter for example in British English and American English, but rather around different word associations that people with a particular demographic characteristic are inclined to make, e.g., "health" in India is more strongly associated with "wealth", while in the United States it is more strongly associated with "sick."

users from different US states. They were able to show that the regional meaning of words can successfully be carried by word embeddings, for example the word "wicked" was most similar to the word "evil" in Kansas, while in Massachusetts, it was most similar to "super" (based on the cosine similarity of the words' vectorial representation). In contrast, our rationale in this article is to explore if word associations can be automatically derived from large corpora annotated with user-centered attributes such as location or gender.

3 Word Associations Dataset

Word association data collection typically consists of providing participants with a list of words, also known in the psycholinguistics literature as *stimulus words*, and asking them to provide the first word that comes to mind in response to each stimulus. For instance, given a stimulus word such as *cat*, one would expect answers such as *dog* or *mouse*. Earlier work on word associations administered the tests in classroom settings, with 100 words per survey, and the results were compiled into tables of norms of word associations (Kent and Rosanoff, 1910; Nelson et al., 2004).

Since our goal is to explore the effect of demographics on word associations, we created a task on Amazon Mechanical Turk (AMT) able to reach a wide and demographically diverse audience. The survey was structured into two sections: the word association part, followed by a demographic survey. Given the online nature of the survey, and since we aimed for a high quality dataset, each participant was presented with a set of 50 stimulus words at a time (instead of 100). The demographic section consisted of seven questions covering gender, age, location, occupation, ethnicity, education, and income.

Stimuli. The stimulus list consists of a set of approximately 300 words. Among these, 99 words are sourced from the word list proposed by Kent and Rosanoff (1910) (*standard* list).[2] The remaining words are identified using the method for finding word-usage differences between two groups introduced in (Garimella et al., 2016), which relies on large collections of texts authored by the two groups to identify words that can be accurately classified by an automatic classifier as belonging to one group versus another. Using their method, we obtain 100 words as the top most different words between US and India (*culture* list), and another set of 100 words as the top most different words between male and female (*gender* list). The reunion of these three lists results in 286 stimulus words for which we collect word associations. Examples are shown in Table 1.

Responses. The task was published separately for respondents from US and India, as AMT has an option of only presenting the survey to people from a preselected geographical location. Six different surveys, each including approximately 50 stimulus words, were administered for each region. The survey was conducted in English for both countries, noting that one of the official languages of India is English (alongside Hindi). Each survey also included four spam-checking questions with previously known answers (e.g., *What is the color of the sky?*, with five options *blue*, *red*, *pink*, *green*, *yellow*), which were used to filter out respondents who were filling out the survey without reading the questions.

For each set, we gathered 400 responses per region, resulting in 800 responses for both US and India. After removing the respondents who did not pass the spam-checking questions, we were left with an average of 752 responses per word, which we then balanced by gender, to retain an equal number of Indian women, Indian men, US women, and US men. This resulted in 492 and 480 responses for the two sets of 50 *standard* stimulus words, 436 and 468 for the *culture* words, and 440 and 432 for the *gender* words. Similar to (Rosenzweig, 1961), all the responses were normalized (i.e. plural was mapped to singular, gerund to infinitive, etc.); in our case we used the Stanford CoreNLP Lemmatizer (Manning et al., 2014), ultimately aggregating the responses into a gold standard.

Table 1 shows the top associations for a few sample stimuli, as collected from India and US, and males and females. Finer-grained qualitative analyses also reveal interesting distinctions. For instance *bath* is overwhelmingly associated by men with *water*, while US women associate it with *bubble*, and Indian women with *soap*. Interestingly, US men seem to provide responses based on collocations, e.g., they answer *Kane* for *citizen* (citizen Kane), *weight* for *heavy* (heavyweight), or *lion* for *mountain* (mountain lion); on the contrary, women more often provide responses that consist of synonym or antonym words, e.g., *person* for *citizen*, *health* for *sick*, or *light* for *heavy*.

For further insight, Table 2 shows the average

[2]Note that this list originally included 100 words. The word "foot" was however misspelled in our survey, and instead we gathered answers for "food."

Word	Gender		Location	
	Male	Female	India	US
beautiful	girl, woman, pretty	pretty, girl, ugly	girl, nature, flower	pretty, girl, ugly
cheese	pizza, bread, milk	butter, mouse, pizza	pizza, butter, bread	cracker, swiss, cheddar
hard	soft, rock, work	soft, work, rock	work, stone, rock	soft, rock, time
health	good, wealth, care	good, wealth, sick	wealth, good, fitness	good, sick, care
range	distance, gun, shooting	gun, rover, mountain	price, rover, wide	gun, distance, rover
admit	hospital, guilt, card	hospital, confess, one	hospital, card, accept	guilt, one, confess
mix	tape, match, juice	cake, tape, stir	juice, tape, match	stir, tape, cake
organize	clean, arrange, party	clean, arrange, meeting	arrange, meeting, party	clean, sort, neat
stack	pile, book, box	book, pile, hay	book, queue, pile	pile, book, pancake

Table 1: Top three most frequent responses for sample stimulus words.

number of different responses obtained for a given stimulus word, with the lowest variability word, and the highest variability word.[3] The second column lists the correlations between the frequency of the primary response and the number of different responses, as also reported by (Jenkins and Palermo, 1965). This correlation is negative, as the more people agree on the primary response, the fewer overall unique answers for a stimulus word are provided. Additionally, Figure 1 shows the Zipfian distribution of average norm frequency; the most frequent response is given on average by 24% of the respondents, while the third most frequent response is given by 7% of them.

Demographic	Avg.	Correlation	Lowest Variability	Highest Variability
Standard				
India	60.88	-0.52	stove	city
US	51.19	-0.53	bath	trouble
Male	61.63	-0.45	stove	city
Female	56.75	-0.55	stove	city
All				
India	72.27	-0.59	stove	regardless
US	57.03	-0.56	east	basically
Male	70.33	-0.52	stove	regardless
Female	66.54	-0.59	east	respectively

Table 2: Average number of responses obtained for a given stimulus word, correlation between frequency of primary response and number of different responses, words exhibiting the lowest variability, and words with the highest variability.

Analyses of Demographic Variations. To model norm strength within a given demographic group or across groups, we tabulate how often respondents from a group match the most frequent answer (*Primary*) or one of the most frequent ten answers for that group (*Top10*). That is, given the response for one stimulus word as provided by one held-out survey respondent at a time, we determine whether that response matches the most frequent association of the *remaining* members of the same group (Table 3, *Primary* columns), or one of the top 10 associations pertaining to that same group (Table 3, *Top10* columns). Similarly, we measure the match with the most frequent or the top 10 responses from the other group, as shown in Table 4. As expected, the intra-group similarities are significantly higher than the inter-group similarities, which supports our hypothesis that different groups make different word associations, which tend to be more coherent within a group than across groups. While males and females have similar ranges for their agreement figures, we notice that on average US respondents have stronger intra-group agreements. Note also that inter-group similarities are asymmetrical, as multiple words may have the same association frequency for one group, yet for the complementary group that may not be the case.

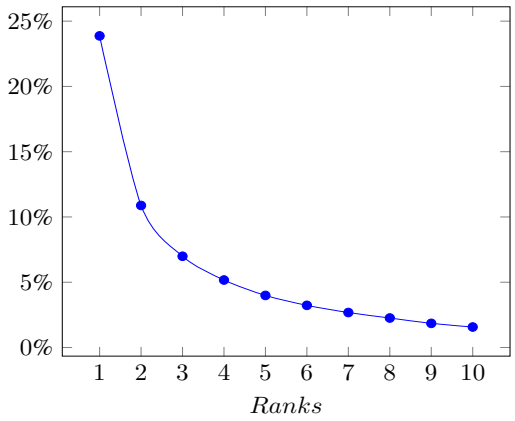

Figure 1: Primary response frequency (in percent) versus rank for the Standard word list.

[3] In several of our data analyses, in order to allow for a direct comparison with the word list from (Kent and Rosanoff, 1910), in addition to showing statistics for the entire dataset (*All*), we also show statistics separately compiled for the list from (Kent and Rosanoff, 1910) (*Standard*).

As an additional analysis of demographic variations in the responses received, for each respondent, we predict his / her demographic group using a majority vote conducted across all the user's responses using a simple rule-based system that assigns each response to the group having the highest frequency for that particular association. For instance, given the response *sun* obtained from a respondent for the stimulus *yellow*, we assign the respondent to either India or US depending on the highest normalized frequency of the response *sun* for the same stimulus in each of those groups. A similar rule-based assignment is also used for gender. Thus, we compute the response words and their normalized frequencies based on the responses from 80% of the users chosen randomly, and accordingly predict the demographic group for the remaining 20% of the users based on a decision across the entire set of a user's responses. Table 5 shows the results of these predictions, which indicate high location variability (i.e., we can predict with high accuracy the location of a respondent), and medium gender variability.

Demographic	Standard		All	
	Primary	Top10	Primary	Top10
India-India	0.23	0.77	0.18	0.78
US-US	0.29	0.82	0.25	0.81
Male-Male	0.23	0.79	0.19	0.79
Female-Female	0.25	0.80	0.21	0.81

Table 3: Intra-group similarities (the higher the similarity, the more cohesive the group is).

Demographic	Standard		All	
	Primary	Top10	Primary	Top10
India-US	0.18	0.55	0.14	0.50
US-India	0.20	0.60	0.16	0.56
Male-Female	0.22	0.63	0.17	0.59
Female-Male	0.24	0.66	0.19	0.61

Table 4: Inter-group similarities (the higher the similarity, the less distinct the groups are).

Demographic	Standard	All
Gender	0.60	0.56
Location	0.94	0.94

Table 5: Predictions based on similarity to group.

4 Computational Models of Word Associations

We first introduce a new model for measuring word associations that leverages a shallow neural net architecture to embed demographically-enriched words. We then compare the performance of the predicted associations to those resulting from other approaches, including traditional corpus-based measures such as mutual information or vector-space models, as well as a recent distributed learning model with word embeddings. For each of these methods, we predict, evaluate, and compare generic associations (devoid of any demographic information), as well as demographic-aware associations.

4.1 Composite Skip-gram Models

We introduce a new word association model, which relies on the skip-gram neural net architecture (Mikolov et al., 2013), and leverages its efficiency and ability to deal with less frequent words.

The skip-gram model tries to predict the context given a word, that is, for each word w_i in the input sequence $w_1, \ldots, w_T$, the model tries to predict w_{i-2}, w_{i-1}, w_{i+1} and w_{i+2}, assuming, for example, a sliding window of five words. Mathematically, the model maximizes the objective function

$$J = \frac{1}{T} \sum_{i=1}^{T} \sum_{j=-c, j\neq 0}^{c} \log P(w_{i+j}|w_i) \quad (1)$$

where T is the number of tokens in the data set, c is the number of context words on each side of the target word w_i and $P(w_{i+j}|w_i)$ is the probability to observe word w_{i+j} in the context of word w_i.

To make this model demographic-aware, we propose two variations, which we refer to as composite skip-gram models ($C-SGM$). In the first one ($EMB1$), the target word w_i is tagged with a demographic label L (e.g., gender). For example, for the target word "formula$^{L=\text{female}}$" we try to predict a high probability for "baby" and "milk" occurring in the neighboring context. The underlying reasoning is that tagged words that appear in similar contexts will be nudged toward each other, while those that do not, will further distance themselves. This allows discrepancies to emerge between how the words are embedded given a demographic dimension.

In the second variation ($EMB2$), we also include the demographic label in the context. That is, for each skip-gram $(c_{i,\text{left}}, w_i, c_{i,\text{right}})$ we generate three skip-grams

$$\begin{aligned} &(c^{\text{label}}_{i,\text{left}}, \quad w_i, \quad c_{i,\text{right}}) \\ &(c_{i,\text{left}}, \quad w^{\text{label}}_i, \quad c_{i,\text{right}}) \\ &(c_{i,\text{left}}, \quad w_i, \quad c^{\text{label}}_{i,\text{right}}) \end{aligned} \quad (2)$$

The two models seek to capture different scenarios. In the first model, where we only add

the demographic label to the target word, the embedding of the labeled word is optimized with respect to the generic embedding of the context. In the second model, the optimization is rather symmetric, allowing tagged and generic embeddings to influence each other. Thus, the optimization function seeks to predict both tagged and untagged words in the vicinity given a target word, instead of only focusing on predicting untagged words like EMB1. The embeddings resulting from such a model should allow for more accurate representations across the tagged and untagged vocabulary, where for example the word "mother" uttered by a female would be close to the word "mother" (regardless of author gender). In both scenarios, the embeddings space accommodates both tagged and untagged words at the same time, being very computationally robust, and allowing comparisons across the tagged version of words, as well as between generic words and their tagged surrogates. For both variations, we compute the cosine similarity between the stimulus word and each of the vocabulary words (whether generic or demographic-enhanced), and retain the closest unique candidates (after dropping their demographic tag).

4.2 Other Word Association Models

Mutual Information (MI). We implement the information theoretic measure proposed by Church and Hill (1990). It is defined as follows:

$$I(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)} \quad (3)$$

This measure compares the probability of observing words x and y together (the joint probability) with the probabilities of observing x and y independently. The joint probability, $P(x,y)$, is generally estimated by counting the number of times x is followed by y in a window of w words, and normalizing this count with the size of the corpus. We follow Church and Hill and set the window size w to five, as it is large enough to capture verb-argument constraints, and not so large to restrict to strict adjacency. For a given stimulus word, (1) we use the entire corpus and compute the generic MI word association with the rest of the vocabulary, and get the top associations according to their MI scores; and (2) we use the section of the corpus obtained for a given demographic, and determine the top demographic-aware MI word associations.

Vector-Space Model (VSM). We also implement the traditional vector-space model, where each word is represented by a $tf.idf$ weighted vector inside the term-document matrix (representing term occurrences inside the documents in the corpus), with a length equal to the number of documents D in the corpus (Salton and McGill, 1986). For a given stimulus word, cosine similarities are computed with all the remaining word vectors in the vocabulary, and those words having the highest similarity are considered as the top responses. Similar to MI, we use all the documents in the corpus to produce generic word associations, while only those documents pertaining to a specific demographic value are utilized to derive demographic-aware associations.

Skip-gram Language Model. We also use the distributional representation technique of word embeddings ($SGLM$) proposed by Bamman et al. (2014). Specifically, information about the speaker (geography, in their case) is used while learning the vector-space representations of word meanings from textual data that is supplemented with metadata about the authors. In addition to the *global* embedding matrix W_{main} that contains low-dimensional representations for every word in the vocabulary (Mikolov et al., 2013), this approach has an additional $|C|$ matrices $\{W_c\}$ of the same size as W_{main}, where $|C|$ denotes the number of values the demographic variable has in the data (e.g., if gender is the demographic variable, $C = \{female, male\}$ and $|C| = 2$). Each of these $|C|$ matrices captures the effect that each demographic variable value has on each word in the vocabulary. To index the embedding of a stimulus word $w \in \mathbb{R}^{|V| \times k}$, the hidden layer h is computed as the sum of the matrix multiplications with each of the independent embeddings:

$$h = w^T W_{main} + \Sigma_{c \in C} w^T W_c \quad (4)$$

It then predicts the value of the context word y using another parameter matrix $X \in \mathbb{R}^{|V| \times k}$ based on a softmax function $o = softmax(Xh)$, where $o \in \mathbb{R}^{|V| \times k}$. Backpropagation using (input x, output y) word tuples learns the values of the various embedding matrices W and parameter matrix X, which maximize the likelihood of context words y conditioned on the stimulus word x.

We use this approach in its original implementation provided by (Bamman et al., 2014) to compute the word embedding vectors for all the words in the vocabulary. Given a stimulus word, the closest vocabulary words with the highest cosine similarity are retained as the top association predictions for the given stimulus word.

5 Experiments

All our models require textual data with demographic information. We introduce below the data we used and the metrics we adopted for evaluation.

Data. Given the requirement of having gender and location information associated with the data, we resort to blogs, and collect from Google Blogger[4] a large set of blog posts authored between 1999 and 2016. Table 6 shows the breakdown of the raw blog counts per demographic category. From these, we retain only those posts with non-empty content, and preprocess the data by removing HTML tags, converting all the tokens to their lemmatized forms,[5] and discarding those lemmas with a frequency less than 10, in order to avoid misspellings and other noise characteristic to social media content.

Demographic	Raw		Balanced		
	Profiles	Posts	Profiles	Posts	Tokens
India	1,520	339,624	1,520	34,987	16,884K
US	3,273	825,093	1,520	32,782	11,706K
Male	2,031	597,935	1,818	44,299	21,971K
Female	1,818	321,779	1,818	45,980	17,070K

Table 6: Raw and balanced blog dataset statistics.

From the above pool of blog posts, we create two datasets with complementary demographic classes (1) location: India-US and (2) gender: male-female.

We process each of these datasets so that they are profile-balanced with no peaks for any specific years, by applying several heuristics: **(1)** Compute the minimum number of users n over all the classes (e.g., Indian and US authors in the case of the location dataset). **(2)** From each class, select the top n users based on the number of years they were blogging and the number of posts they wrote.[6] This ensures that the maximum amount of data will be available for the selected users. **(3)** For each of these n users, pick at most 50 posts in a round-robin fashion from the years in which they blogged. **(4)** Let M be the total number of posts collected in this manner from all the classes. In order to avoid having most of the posts coming from a small number of years, set a cutoff X as a fraction of M. For each year, a maximum of X posts will be chosen from the set of M posts ($X = 0.1M$). **(5)** To ensure that all the users get to contribute posts, and that the contribution of prolific writers is kept in check, maintain user participation scores:

$$p(user) = \frac{\text{posts collected from user}}{\text{total number of posts collected}} \quad (5)$$

These scores are updated after every year is processed, as explained further. **(6)** Sort the years in increasing order of number of posts and iterate through them; identify the lowest number of posts contributed by the least prolific writer, then collect the minimum number of posts from all users who published in that year in a round-robin manner. Then, select additional posts from users in increasing order of participation scores, until the number of posts for the year reaches the cutoff X. **(7)** After each year, update the user participation scores. Table 6 shows the number of users and posts retained after balancing. This particular composition is used in our *location* data set (consisting of India and US posts) and *gender* data set (consisting of females and males posts).

Metrics. Given that the word association task is relatively similar to the lexical substitution task, in terms of open vocabulary and lack of a "right" answer, we decided to borrow the *best* and out-of-ten (*oo10*) evaluation metrics traditionally used for the latter (McCarthy and Navigli, 2009), yet corrected for weight (Jabbari et al., 2010). Briefly, these measures take the best (or top ten) responses from a system, and compare them against the gold standard, while accounting for the frequencies of the responses in the gold standard. In addition, since Figure 1 shows that the top three ranking norms are provided as answers by approximately 42% of the respondents, with the remaining norms following a long Zipfian distribution in terms of frequency of appearance, we also compute out-of-three (*oo3*), which represents a more focused approximation of our ability to predict human associations (note that out-of-ten covers 62% of the responses). Several recent papers on word associations evaluated their models indirectly via Pearson or Spearman correlation performance on a word similarity task (Chaudhari et al., 2011; Deyne et al., 2016); we choose instead to evaluate word associations directly, by using metrics that more closely align with the evaluations performed in the field of psychology where the best output of a system is compared against the most frequent human response (Bel-Enguix, 2014; Mohammad, 2011).

For a given stimulus word w with human responses H_w, suppose a system returns a set of answers S_w. We estimate how well this system

[4] www.blogger.com

[5] We normalize the word forms using the Stanford CoreNLP lemmatizer (Manning et al., 2014).

[6] Prolific users will be chosen first. For a class with exactly n users, all users will be chosen.

can find a *best* substitute for w using Equation 6, where the function $freq_w(s)$ returns the count of a system response s in H_w, and $maxfreq_w$ returns the maximum count of any response in H_w.

$$best(w) = \frac{\Sigma_{s \in S_w} freq_w(s)}{maxfreq_w \times |S_w|} \quad (6)$$

$$oon(w) = \frac{\Sigma_{s \in S_w^n} freq_w(s)}{|H_w|} \quad (7)$$

Equation 7 measures the coverage of a system by allowing it to offer a set S_w^n of *n* responses for w, where each response s is weighted by its frequency $freq_w(s)$ in H_w.

6 Evaluations and Discussions

We conduct evaluations using all the word association models described in Section 4. The results using the *best*, *out-of-three*, and *out-of-ten* evaluation metrics are listed in Table 7. For all the embeddings experiments, we use 300 latent dimensions. The Gen variation uses the demographic-blind dataset, whereas the DA variation uses the demographic-aware dataset.[7]

The MI and VSM models do not perform well in the word association prediction task, whether considering the generic or the demographic-aware data. We should emphasize, however, that the generic version of these models is able to consider co-occurrences across the entire generic datasets, while the demographic-aware co-occurrences can only be computed from the section of the dataset that matches a particular demographic; as such, these latter models are placed at a disadvantage.

Perhaps not surprisingly, the neural network skip-gram-based architectures, whether SGLM or our C-SGM, always achieve better results when compared to MI or VSM. The demographic-aware variation proposed by (Bamman et al., 2014) uses an extended skip-gram architecture that encodes a generic embedding, and several demographic-based filters per class, which in our case translates into three matrices of 300 dimensions each, the first for the generic words, and the subsequent ones for skews to be applied to the generic words in order to render the embedding through the lens of a given demographic. $SGLM - Gen$ in our case are the predictions based on the generic matrix, while $SGLM - DA$ are the predictions modified along the lines of a particular demographic.

[7]To place the results in this table in perspective, it is important to note that results for this task are traditionally low. Given that the most frequent response is selected on average by 24% of respondents (see Figure 1), we can see that even for humans, the highest score would be around 0.24.

Our composite skip-gram models encode a single matrix that contains a mix of demographic-aware and generic words expressed as 300 latent dimensions. For both gender and location, our gender-aware models ($EMB1$ and $EMB2$) surpass the SGLM gender-aware model. Surprisingly, while SGLM was never meant to be generic, the predictions based on its generic embedding matrix prove to be a difficult baseline to surpass, similar to C-SGM generic. Nonetheless, the composite skip-gram models ($EMB1$ and $EMB2$) do achieve best and second best rankings in the vast majority of cases (when compared to the best among all the other methods), with $EMB1$ being the more robust variation performing well both for gender and for location. Focusing on the performance of $EMB1$, the highest gains are observed for India-based predictions, for best (from 0.05 to 0.08) and out-of-three (from 0.07 to 0.12); for male-based predictions increasing from 0.11 to 0.13 for best, and from 0.17 to 0.20 for out-of-three; and for female-based predictions, increasing from 0.13 to 0.14 for best, and from 0.17 to 0.20 for out-of-three. US-based associations are the hardest to predict, probably because of the diverse makeup of society; additional evaluations are needed to pinpoint the exact cause.

To determine how susceptible the embedding model is to skewed, but larger training data, we also run a separate experiment on the entire raw set of blogs we collected (described on the left of Table 6), where we re-generate the $EMB1$ and $EMB2$ models. While the entire dataset is significantly larger than the balanced set, it is also significantly skewed: the data in the India:US dataset was skewed in a proportion of 1:0.48 tokens, while for Female:Male the proportion was 1:0.41 tokens. As was the case for the balanced dataset, the $EMB1$ model is still the most robust (see the bottom section in Table 7), and it achieves significant gains when compared to its balanced counterpart, in particular for *best* (for the US demographic from 0.03 to 0.13, and for India from 0.08 to 0.11), and for *out-of-three* (for US from 0.07 to 0.15, and for India from 0.12 to 0.17), which suggests that as an avenue for future research, we can explore the use of significantly larger even if unbalanced datasets to train our models.

7 Conclusion

In this paper, we introduced the task of demographic-aware word associations. To understand the various ways in which people associate words, we collected a new large demographics-

Method	Type	best		oo3		oo10		best		oo3		oo10	
		IN	US	IN	US	IN	US	M	F	M	F	M	F
MI	Gen	0.00	0.00	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	DA	0.00	0.00	0.01	0.00	0.02	0.02	0.00	0.01	0.01	0.01	0.01	0.02
VSM	Gen	0.00	0.00	0.01	0.01	0.03	0.03	0.00	0.00	0.02	0.02	0.05	0.05
	DA	0.00	0.00	0.02	0.01	0.04	0.02	0.00	0.01	0.02	0.01	0.04	0.06
SGLM	Gen	0.02	0.02	0.03	0.03	0.06	0.05	**0.13**	0.13	**0.18**	*0.18*	0.20	0.21
	DA	0.05	0.01	0.07	0.02	0.11	0.03	0.10	0.13	0.16	*0.18*	0.18	0.20
C-SGM	Gen	0.05	**0.04**	0.07	**0.07**	0.11	**0.10**	*0.11*	0.13	*0.17*	0.17	0.20	0.21
	EMB1	*0.08*	*0.03*	*0.12*	**0.07**	*0.18*	**0.10**	**0.13**	*0.14*	**0.20**	**0.20**	**0.25**	**0.26**
	EMB2	**0.09**	0.02	**0.14**	*0.04*	**0.19**	*0.06*	0.10	**0.16**	*0.17*	**0.20**	*0.23*	*0.25*
C-SGM-raw	EMB1	0.11	0.13	0.17	0.15	0.21	0.17	0.09	0.16	0.17	0.18	0.21	0.23
	EMB2	0.10	0.08	0.15	0.12	0.19	0.15	0.09	0.14	0.15	0.16	0.18	0.20

Table 7: Best, out-of-three (oo3), and out-of-ten (oo10) scores across the various methods. IN: India, US: United States, M: Male, F: Female. The numbers in bold mark the highest scores, those in italics, the second highest.

enhanced dataset of approximately 300 stimulus words and their associated norms compiled from 800 respondents for a total of 176,097 non-spam responses, and show that for people of different demographics, associations do differ with gender and location.

We proposed a new demographic-aware word association method based on composite skip-gram models that are able to jointly embed generic and gender tagged words. We showed that this method improves over its generic counterpart, and also outperforms previously proposed models of word association, thus demonstrating that it is useful to account for the demographics of the people behind the language when performing the task of automatic word association. We regard this as a first step toward demographic-aware NLP, and in future work we plan to address other more advanced NLP tasks while accounting for demographics.

The word association dataset introduced in this paper is publicly available from `http://lit.eecs.umich.edu/downloads.html`.

Acknowledgements

This material is based in part upon work supported by the Michigan Institute for Data Science, by the National Science Foundation (grant #1344257), and by the John Templeton Foundation (grant #48503). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the Michigan Institute for Data Science, the National Science Foundation, or the John Templeton Foundation.

References

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) (ACL 2014)*. pages 828–834.

Gemma Bel-Enguix. 2014. Retrieving word associations with a simple neighborhood algorithm in a graph-based resource. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon*. Dublin, Ireland, pages 60–63.

Dipak L. Chaudhari, Om P. Damani, and Srivatsan Laxman. 2011. Lexical co-occurrence, statistical significance, and word association. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*. Edinburgh, Scotland, UK, pages 1058–1068.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1989. Parsing, word associations and typical predicate-argument relations. In *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, pages 75–81.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16(1):22–29.

Simon De Deyne, Daniel J. Navarro, and Gert Storms. 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods* 45(2):480–498.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING 2016)*. Osaka, Japan, pages 1861–1870.

Lee R. Dice. 1945. Measures of the amount of ecologic association between species. *Ecology* 26(3):297–302.

Jeffrey L. Elman, Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. 1997. *Rethinking innateness: a connectionist perspective on development*. The MIT Press.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th AAAI International Conference on Artificial Intelligence (AAAI 2007)*. Hyderabad, India, pages 1606–1611.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING 2016)*. Osaka, Japan, pages 674–683.

Samer Hassan and Rada Mihalcea. 2011. Measuring semantic relatedness using salient encyclopedic concepts. *Artificial Intelligence, Special Issue* xx(xx).

Dirk Hovy. 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 (ACL 2015)*. Beijing, China, pages 752–762.

Thad Hughes and Daniel Ramag. 2007. Lexical semantic relatedness with random graph walks. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP 2007)*. Association for Computational Linguistics, Prague, Czech Republic, pages 581–589.

Aminul Islam and Diana Inkpen. 2008. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data* 2(2):10:1–10:25.

Sanaz Jabbari, Mark Hepple, and Louise Guthrie. 2010. Evaluation metrics for the lexical substitution task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010)*. pages 289–292.

Mario Jarmasz and Stan Szpakowics. 2003. Rogets thesaurus and semantic similarity. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2003)*. Borovetz, Bulgaria, pages 111–120.

James J. Jenkins and David S. Palermo. 1965. Further data on changes in word-association norms. *Journal of Personality and Social Psychology* 1(4):303–309.

Grace H. Kent and Aaron J. Rosanoff. 1910. A study of association in insanity. *American Journal of Psychiatry* 67(1):37–96.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2):211–240.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In *WordNet: An Electronic Lexical Database*, pages 305–332.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries. In *Proceedings of the 5th annual international conference on Systems documentation (SIGDOC 1986)*. Toronto, Ontario, pages 24–26.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics System Demonstrations (ACL 2014)*. pages 55–60.

Diana McCarthy and Roberto Navigli. 2009. The english lexical substitution task. *Language Resources and Evaluation* 43(2):139–159.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Neural Information Processing Systems Conference (NIPS 2013)*. pages 3111–3119.

Saif Mohammad. 2011. Colourful language: Measuring word-colour associations. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*. pages 97–106.

Sandra Mollin. 2009. Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory* 5(2):175–200.

Douglas L. Nelson, McEvoy Cathy L., and Schreiber Thomas A. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc* 36(3):402–407.

Timothy R. Rogers and James L. McClelland. 2004. *Semantic cognition: a parallel distributed processing approach*. The MIT Press.

Mark R. Rosenzweig. 1961. Comparisons among word-association responses in English, French, German, and Italian. *The American Journal of Psychology* 74(3):347–360.

Gerard Salton and Michael J. McGill. 1986. *Introduction to modern information retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Margaret E. Tresselt and Mark S. Mayzner. 1964. The Kent-Rosanoff word association: Word association norms as a function of age. *Psychonomic Science* 1(1-12):65–66.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*. Seattle, WA, USA, October, pages 1815–1827.

Manfred Wettler and Reinhard Rapp. 1989. A connectionist system to simulate lexical decisions in information retrieval. *Connectionism in Perspective. Amsterdam: Elsevier* pages 463–469.