

# An Unbiased Approach to Quantification of Gender Inclination using Interpretable Word Representations

Navid Rekabsaz<sup>1</sup>, Allan Hanbury<sup>2</sup>

<sup>1</sup>Idiap Research Institute / Martigny, Switzerland

<sup>2</sup>TU Wien / Vienna, Austria

navid.rekabsaz@idiap.ch

allan.hanbury@tuwien.ac.at

## Abstract

Recent advances in word embedding provide significant benefit to various information processing tasks. Yet these dense representations and their estimation of word-to-word relatedness remain difficult to interpret and hard to analyze. As an alternative, explicit word representations i.e. vectors with clearly-defined dimensions, which can be words, windows of words, or documents are easily interpretable, and recent methods show competitive performance to the dense vectors. In this work, we propose a method to transfer word2vec SkipGram embedding model to its explicit representation model. The method provides interpretable explicit vectors while keeping the effectiveness of the original model, tested by evaluating the model on several word association collections. Based on the proposed explicit representation, we propose an unbiased method to quantify the degree of the existence of gender bias in the English language (used in Wikipedia) with regard to a set of occupations. By measuring the bias towards explicit Female and Male factors, the work demonstrates a general tendency of the majority of the occupations to male and a strong bias in a few specific occupations (e.g. nurse) to female.

## 1 Introduction

Word embedding models provide significant benefit to information processing tasks. While easy to construct based on raw unannotated corpora, these dense representations and their estimation of term-term relatedness remain difficult to interpret and hard to analyze. In fact, when using word embed-

ding, it remains opaque what the dimensions of the vectors refer to, or in what extent a semantic concept is present in the vector representation of a term.

A natural solution to this problem is using explicit representations of words i.e. vectors with clearly-defined dimensions, where each dimension represents an explicit concept such as a term, window of terms, or document. Such an explicit vector of a word is easily interpretable, as each dimension stands for the degree of relation between the word and the corresponding concept.

As shown by Levy et al. (Levy et al., 2015), the recent explicit representation models such as Shifted Positive Point Mutual Information (SPPMI), show competitive performance in comparison to the state-of-the-art word embeddings on a set of term association test collections. Regarding efficiency, the explicit representations often require much bigger memory space in comparison to the low-dimensional dense vectors. However, in practice the memory issue can be mitigated by suitable data structures if the vectors are highly sparse.

Our first contribution in this chapter is in line with previous studies (Levy and Goldberg, 2014; Levy et al., 2015) on providing fully interpretable vectors by proposing a novel explicit representation for the word2vec SkipGram model. We propose a method to transfer the low-dimensional (dense) vectors of a trained SkipGram model to explicit vector representations in a high-dimensional space. Our approach is in the opposite direction to the methods such as LSI or GloVe, where they start from a high-dimensional matrix and result in low-dimensional embeddings. In contrast, the main objective of our work is to provide an interpretable variation of the SkipGram vectors, enabling error resolution and better causal analysis.

We evaluate our explicit SkipGram model on 6 term-to-term association benchmarks, showing results on par with the SPPMI model as the state of the art of explicit representation vectors. These results support the reliability of our approach to create high quality interpretable vectors of the SkipGram model.

To show an application of our explicit SkipGram representation, in our next contribution, we propose a novel approach based on explicit vectors to quantify the degree of gender bias in a corpus. We particularly focus on the inclination of a set of gender-neutral occupations to male or female in a Wikipedia English corpus.

As a close study to our work, Bolukbasi et al. (Bolukbasi et al., 2016) quantify the gender bias of an occupation by calculating the semantic similarity of the vectors of the terms ‘she’ and ‘he’ ( $v_{she}$  and  $v_{he}$ ), as the representative of female and male, to the vector of the occupation using the SkipGram model. We point out an intrinsic issue in this approach, by arguing that  $v_{she}$  and  $v_{he}$  are not precise representatives of female and male concepts, since due to bias in language they also contain other types of concepts, specially the ones related to occupations. For instance, if ‘nurse’ is biased to female, we expect that  $v_{nurse}$  contains many concepts related to female. However, it also means that  $v_{she}$  contains high relation to the concept ‘nurse’. We refer to this characteristic of word embedding as *circularity*. Considering this trait, given that  $v_{nurse}$  naturally contains the concept ‘nurse’, calculating the semantic similarity between  $v_{she}$  and  $v_{nurse}$  (as the degree of bias of ‘nurse’ to female) is wrongly inclined by the ‘nurse’ concept.

To address the issue caused by circularity, we exploit the interpretability characteristic of the explicit SkipGram representations, by selecting only the gender-related concepts (dimensions) of the gender vectors. In our approach, the bias towards female is quantified by defining a new gender vector  $v_{SHE}$ , where its female-related dimensions are explicitly set to the ones of  $v_{she}$  and the rest to zero (the same process for bias towards male by defining the vector  $v_{HE}$ ).

The proposed gender vectors  $v_{SHE}$  and  $v_{HE}$  therefore only consist of gender-specific concepts which arguably provide a more precise approach to gender bias quantification. These results specially demonstrate the high bias of some specific

jobs to female-specific concepts. This inherent bias in data and therefore word representations can potentially be propagated to information systems, leading to ethically-biased decisions.

## 2 Background and Related Work

### 2.1 Embedding with Negative Sampling

The SG model consists of two sets of vectors: word ( $\mathbf{V}$ ) and context ( $\tilde{\mathbf{V}}$ ) vectors, both of size  $|\mathbb{W}| \times d$ , where  $\mathbb{W}$  is the set of words in the collection and  $d$  is the embedding dimensionality.

The SG model is optimized with Negative Sampling, a descendent of *Noisy Contrastive Estimation (NCE)* (Mnih and Teh, 2012) method. Negative Sampling aims to maximize the difference between  $p(y = 1|w, c)$ , the probability that the co-occurrence of word  $w$  and  $c$  come from a *genuine* distribution, with  $p(y = 1|w, \check{c})$  for  $k$  negative samples  $\check{c}$ .

While the co-occurrence of  $w$  and  $c$  is observed in the given data corpus, the negative samples are drawn from a *noisy* distribution  $\mathcal{N}$ , defined using the unigram distribution of the words in the corpus. In the word2vec framework,  $p(y = 1|w, c)$  is defined as  $\sigma(\mathbf{v}_w \tilde{\mathbf{v}}_c)$  where  $\mathbf{v}_w$  is the vector representation of  $w$ ,  $\tilde{\mathbf{v}}_c$  context vector of  $c$ , and  $\sigma$  denotes the sigmoid function,

### 2.2 Interpretable Representation

A well-known explicit representations is defined based on the Point Mutual Information (PMI) measure. In the PMI word representation, for the word  $w$ , the value of the corresponding dimension to the context word  $c$  is defined as  $\text{PMI}(w, c) = \log \frac{p(w, c)}{p(w)p(c)}$  where probabilities are calculated by counting the number of co-occurrences over the size of the full co-occurrence matrix.

Levy and Goldberg (Levy and Goldberg, 2014) show an interesting relation between PMI and SG representations, i.e. when the dimension of the vectors is very high (as in explicit representations), the optimal solution of SG objective function is equal to PMI shifted by  $\log k$ . Based on this idea, they propose Shifted Positive PMI (SPPMI) representation by subtracting  $\log k$  from PMI vector representations and setting the negative values to zero.

They finally show the competitive performance of the SPPMI model on word association tasks to the SG model. Their definitions of PPMI and SPPMI are the current state-of-the-art in explicit

representations, against which we will compare our method.

Another direction of interpretable word vector representation is explored by [Faruqui et al. \(2015\)](#) and [Sun et al. \(2016\)](#). In these studies, the aim is to increase the sparsity of the dense vectors. The rationale of these approaches is that by having more sparse vectors, it becomes more clear which dimension of the vectors might be referring to which concepts in language.

### 3 Explicit Skip-Gram Representation

In this section, we first explain our approaches to create explicit representations of the SkipGram model, followed by evaluation and comparison of the proposed representations.

#### 3.1 Definition

To define our novel explicit representations, let us first revisit the  $p(y = 1|w, c)$  probability in the word2vec SkipGram model (referred to as SG in the rest of this chapter).  $p(y = 1|w, c)$  measures the probability that the co-occurrence of two terms  $w$  and  $c$  comes from the training corpus and not from a random distribution. The purpose of this probability is in fact related to the conceptual goal of the PMI-based representations i.e. to distinguish a genuine from a random co-occurrence. Indeed, both of these probabilities aim to capture the first-order relationship between two terms, based on the corpus at hand. Based on this idea, an immediate way of defining an explicit representation would be to use SkipGram co-occurrence probability as follows:

$$\text{eSG}(\mathbf{v}_w)_{:c} = p(y = 1|w, c) = \sigma(\mathbf{v}_w \tilde{\mathbf{v}}_c) \quad (1)$$

where the eSG function, standing for *explicit Skip-Gram*, returns an explicit representation of the given vector, and  $\text{eSG}_{:c}$  is the value of its concept  $c$ . The value of each concept in the eSG vector representation is between 0 to 1, reflecting the first-order relation between the word to the corresponding concept.

It is however intuitive to consider that the very low values do not represent a genuine relation and can potentially introduce noise in computation. Such very low values can be seen in the relation of a term to very frequent or completely unrelated terms. We can extend this idea to all the values of eSG, i.e. some portion (or all) of every relation contains noise.

To measure the noise in eSG, we use the definition of noise probabilities in the Negative Sampling approach: the expectation value of  $p(y = 1|w, c)$  where  $c$  (or  $w$ ) is randomly sampled from the dictionary for several times. Based on this idea, we define the Shifted Explicit SkipGram (SeSG) model by subtracting the two expectation values from eSG:

$$\begin{aligned} \text{SeSG}(\mathbf{v}_w)_{:c} = \text{eSG}(\mathbf{v}_w)_{:c} &- \mathbb{E}_{\check{c} \sim \mathcal{N}}[p(y = 1|w, \check{c})] \\ &- \mathbb{E}_{\check{w} \sim \mathcal{N}}[p(y = 1|\check{w}, c)] \end{aligned} \quad (2)$$

where  $\mathbb{E}$  is the expectation value over any  $\check{c}$  term, sampled from the noisy distribution  $\mathcal{N}$ .

Since the expectation values can be calculated off-line, in contrast to Negative Sampling (restricted to a set of  $k$  sampled terms), we compute it over the entire vocabulary:

$$\mathbb{E}_{\check{w} \sim \mathcal{N}}[p(y = 1|\check{w}, c)] = \frac{\sum_{i=1}^{|\mathbb{W}|} \#(\check{w}_i) \cdot \sigma(\mathbf{v}_{\check{w}_i} \tilde{\mathbf{v}}_c)}{\sum_{i=1}^{|\mathbb{W}|} \#(\check{w}_i)} \quad (3)$$

For the sampling of the context term  $\check{c}$ , similar to SG and  $\text{PMI}_\alpha$ , we apply the `cds` method by raising frequency to the power of  $\alpha$ , as follows:

$$\mathbb{E}_{\check{c} \sim \mathcal{N}}[p(y = 1|w, \check{c})] = \frac{\sum_{i=1}^{|\mathbb{W}|} \#(\check{c}_i)^\alpha \cdot \sigma(\mathbf{v}_w \tilde{\mathbf{v}}_{\check{c}_i})}{\sum_{i=1}^{|\mathbb{W}|} \#(\check{c}_i)^\alpha} \quad (4)$$

Similar to PPMI, our last proposed representation removes the negative values. The Positive Shifted Explicit SkipGram (PSeSG) is defined as follows:

$$\text{PSeSG}(\mathbf{v}_w)_{:c} = \max(\text{SeSG}(\mathbf{v}_w)_{:c}, 0) \quad (5)$$

Setting the values to zero in PSeSG facilitates the use of efficient data structures i.e. sparse vectors. We analyze the efficiency and effectiveness of the explicit representations in the next section.

#### 3.2 Evaluation

To analyze the representations, we create a SkipGram model similar to the previous chapters with 300 dimensions on the Wikipedia dump file for August 2015 using the gensim toolkit ([Řehůřek and Sojka, 2010](#)). As suggested by [Levy et al. \(2015\)](#), we use a window of 5 terms, negative sampling of  $k = 10$ , down sampling of  $t = 10^{-5}$ , a `cds` value of  $\alpha = 0.75$ , trained on 20 epochs, and filtering out terms with frequency less than 100. The final model contains 199851 terms.

Table 1: Term association evaluation. Best performing among explicit/all embeddings are shown with bold/underline.

Method	Sparsity	WS Sim.	WS Rel.	MEN	Rare	SCWS	SimLex
PPMI	98.6%	.681	.603	.702	.309	.601	.284
SPPMI	99.6%	<b>.722</b>	<u>.661</u>	.704	.394	.571	<b>.296</b>
eSG	0%	.596	.404	.645	.378	.549	.231
SeSG	0%	.527	.388	.606	.311	.507	.215
PSeSG	94.1%	.697	.626	<b>.711</b>	<b>.406</b>	<b>.614</b>	.272
SG	0%	<u>.770</u>	.620	<u>.750</u>	<u>.488</u>	<u>.648</u>	<u>.367</u>

The same values are used for the common parameters in the PPMI and SPPMI representations.

We conduct our experiments on 6 term association benchmark collections. Each collection contains a set of term pairs where the association between each pair is assessed by several human annotators (*annotation score*). The evaluation is done by calculating the Spearman correlation between the list of pairs scored by similarity values versus by annotation scores. The collections used are: WordSim353 partitioned into Similarity and Relatedness (Agirre et al., 2009); MEN dataset (Bruni et al., 2014); Rare Words dataset (Luong et al., 2013); SCWS (Huang et al., 2012); and SimLex dataset (Hill et al., 2016).

The evaluation results for the explicit representations as well as SG are reported in Table 1. The bold values show the best performing explicit representation and the values with underline refer to the best results among all representations. Based on the results, PSeSG and SPPMI show very similar performance (in 3 benchmarks PSeSG and in the other 3 SPPMI shows the best performance), both considerably outperforming the other explicit representations. As also shown in previous studies (Levy et al., 2015), SG in general performs better than the best performing explicit representations. The results confirm the quality of the PSeSG model as a well-performing representation on term association benchmarks. Also looking at the sparsity ratio of the explicit representations, reported in Table 1, we observe that the PSeSG and SPPMI representations are highly sparse, making them amenable to storage in volatile memory in practical scenarios.

In this section, we introduced the PSeSG model and showed its strong performance in practice. In the next section, we use PSeSG for gender bias quantification, and compare our results to the approach of Bolukbasi et al. (Bolukbasi et al., 2016)

conducted on SkipGram vectors. Using PSeSG—an explicit representation variation of the SkipGram model—enables comparison between the two gender quantification approaches, since the PSeSG representation exploited in our method is created from the SkipGram embedding, used in the approach of Bolukbasi et al..

## 4 Gender Bias Quantification

To study the gender bias in occupations, we prepare a list of 343 occupations, from which 26 are female-specific (e.g. ‘congresswoman’), and 22 male-specific (e.g. ‘congressman’), and the rest are gender neutral (e.g. ‘nurse’, ‘dancer’, ‘book-keeper’), listed in Table X, Table Y, and Table Z respectively. For brevity, we use  $e_w$  as the explicit PSeSG representation of the dense  $v_w$  vector:

$$e_w = \text{PSeSG}(v_w) \quad (6)$$

In the following, we first explain in detail our approach to gender bias quantification using the  $e_w$  vectors as well as the one used in Bolukbasi et al.. We then visualize the degrees of inclinations of the mentioned occupations to female and male by processing a corpus of Wikipedia.

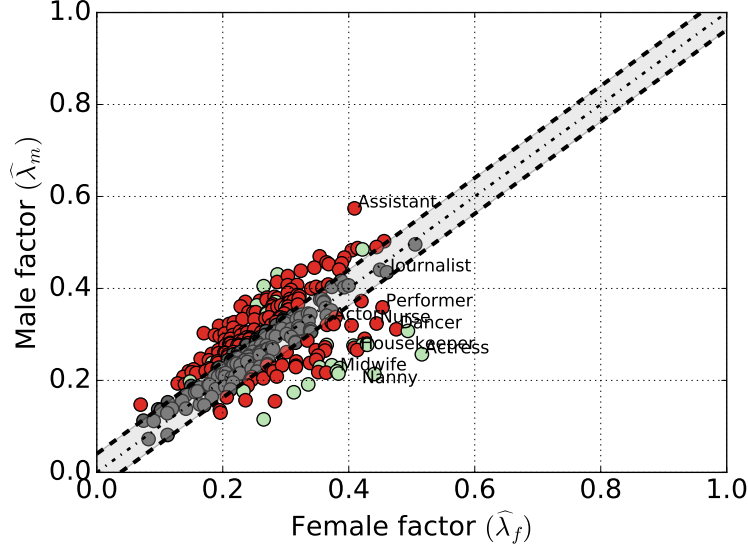
### 4.1 Method

In Bolukbasi et al. (Bolukbasi et al., 2016), the degree of gender bias of a word is measured using the following approach:

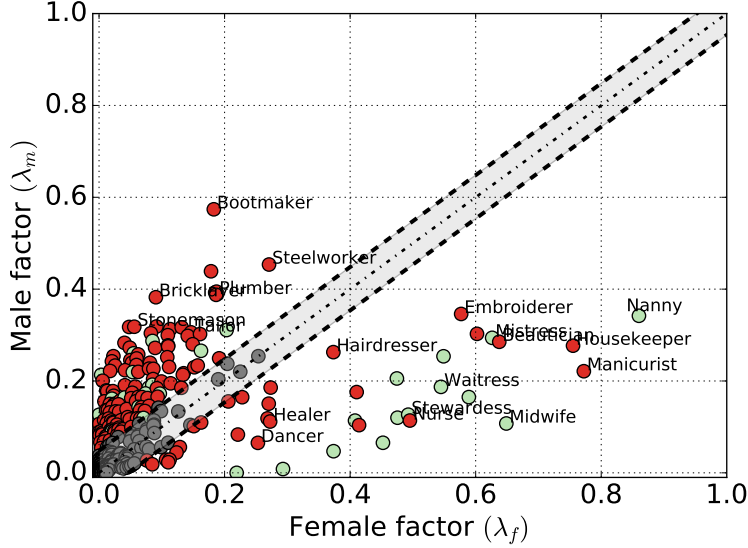
$$\begin{aligned} \hat{\lambda}_f(w) &= \text{cosine}(v_{she}, v_w) \\ \hat{\lambda}_m(w) &= \text{cosine}(v_{he}, v_w) \end{aligned} \quad (7)$$

where  $\hat{\lambda}_f$  ( $\hat{\lambda}_m$ ) denotes the degree of bias of a word  $w$  (occupation in our case) to female (male).

As mentioned in introduction, due to the circularity in word embedding, using  $v_{she}$  and  $v_{he}$  does not provide a precise quantification of bias,



(a) Bolukbasi et al. (Bolukbasi et al., 2016)



(b) Our approach

Figure 1: The inclination of occupations towards male and female genders. Gender-specific occupations are shown in green and gender-neutral ones in red. The gray area indicates gender-neutrality.

as these vectors also contain concepts related to the occupations. To validate the existence of circularity, we can use the explicit variations of  $\mathbf{v}_{she}$  and  $\mathbf{v}_{he}$ , namely  $e_{she}$  and  $e_{he}$  and examine  $e_{she:c}$  and  $e_{he:c}$ , the values regarding each occupation  $c$  (as a concept) in the explicit vector representation of ‘she’ and ‘he’, respectively. Among the 343 occupations, we observe 123, and 168 values higher than zero for  $e_{she}$  and  $e_{he}$  respectively, indicating significant existence of occupation-related concepts in the gender vectors.

To address the issue raised by circularity, we first select a set of terms, representing gender-specific concepts in language. These terms are

shortlisted from the gender-specific list, provided by Bolukbasi et al. after filtering the occupations. The final list contains 32 female-specific terms (e.g. ‘she’, ‘her’, ‘woman’) and 32 equivalent male-specific terms (e.g. ‘he’, ‘his’, ‘man’), denoted as  $\mathbb{G}_f$  and  $\mathbb{G}_m$ , shown in Table XF, and Table XM respectively.

Using these lists gender-concepts, we define the new gender factors as the sum of gender-related concepts of the explicit vector of each word, shown as follows:

$$\lambda_f(w) = \sum_{c \in \mathbb{G}_f} e_{w:c} \quad \lambda_m(w) = \sum_{c \in \mathbb{G}_m} e_{w:c} \quad (8)$$

As the values of  $\lambda$  appear in a different range than the ones of  $\hat{\lambda}$ , to make the approaches comparable, we apply Min-Max normalization on each approach, calculated over the gender factor values of all terms of the corpus.

Another important consideration in our analysis is to distinguish between truly gender-biased terms from low range values of gender factors (which can occur for every random term). To indicate the terms with no considerable inclinations to genders, we define gender-neutrality for a term when the difference between its gender factors is less than a threshold:

$$|\lambda_f - \lambda_m| < \zeta \quad |\hat{\lambda}_f - \hat{\lambda}_m| < \hat{\zeta} \quad (9)$$

To find such a threshold for each approach, since the number of gender-specific terms in English are limited, we assume that a randomly sampled term from the vocabulary is a gender-neutral term. This approach is similar to the one used in the Negative Sampling method. We can repeat this sampling for all the terms and calculate the expected values of  $\zeta$  and  $\hat{\zeta}$  by averaging  $|\lambda_f(w) - \lambda_m(w)|$  and  $|\hat{\lambda}_f(w) - \hat{\lambda}_m(w)|$  respectively over the terms. In our experiments, this results in  $\zeta = 0.046$  and  $\hat{\zeta} = 0.038$ .

## 4.2 Quantification of Gender Bias in Wikipedia

The results of gender bias quantification methods, applied on the Wikipedia corpus are shown in Figure 1. Figures 1a and 1b depict the method used in Bolukbasi et al. and our approach to gender bias quantification method, respectively. In both figures, the gender-specific occupations are colored in green, the gender-neutral ones in red, and the gender-neutrality area in gray.

Comparing the two figures, we observe considerable differences between the gender bias, measured by the two approaches. To compare the approaches, we use WinoBias (Zhao et al., 2018), a recently introduced dataset which reports the degree of gender bias in 40 occupations, using the statistics gathered from the US Department of Labor. The degree of bias of each occupation to female in the dataset is the percent of people in the occupation who are reported as female (e.g. 90% of nurses are women).

We compare the results of the two approaches by calculating the correlation of female bias of these 40 occupations, quantified by each approach,

Table 2: Spearman and Pearson Correlation results of female bias, quantified by our approach using PSeSG representation and Bolukbasi et al.’s approach using SkipGram, to the female bias statistics of 40 occupations, provided by the WinoBias dataset (Zhao et al., 2018).

Method	Correlation	
	Spearman	Pearson
Bolukbasi et al.	0.38	0.39
Our Approach	<b>0.53</b>	<b>0.51</b>

with the values in the WinoBias dataset. The degree of bias to female for occupation  $w$  in our and Bolukbasi et al.’s approach is computed by  $\lambda_f(w) - \lambda_m(w)$ , and  $\hat{\lambda}_f(w) - \hat{\lambda}_m(w)$ , respectively. The evaluation makes the assumption that the bias in the real world is reflected in the text of Wikipedia.

The results of Spearman and Pearson correlations are shown in Table 2. For both Spearman and Pearson correlations, our approach shows higher correlation to the female bias values, provided by the WinoBias dataset. The results show that our approach more accurately resonates the state of gender bias in the real world, and is therefore a more precise method for bias quantification. In fact, our approach corrects the algorithmic bias in Bolukbasi et al.’s method, by addressing the issue of circular effect in word representations using explicit definition of gender-related concepts.

Looking at the results of our approach in Figure 1b, it reveals an interesting pattern in gender bias for the gender-neutral occupations. The majority of these occupations are inclined towards the male factor while in general having weak bias. ‘Bootmaker’, ‘tailor’, and ‘stonemason’ are some of the male-biased occupations. On the other hand, there exist relatively few occupations with inclination to the female factor while some of them have very strong gender bias, for example gender-neutral occupations like ‘housekeeper’, ‘nurse’, and ‘manicurist’. These observations provide a quantification of gender bias in machine learned representations and enable future automated gender debiasing.

## 5 Conclusion

We propose a method to create an explicit representation of the word2vec SkipGram model by capturing the probability of genuine co-occurrence of

the terms. The proposed representation performs on par with the state of the art explicit representations on a set of term association benchmarks, and suggests a novel approach to interpret the vector embeddings of the SkipGram model.

Further on, we propose a method for quantifying gender bias using our explicit SkipGram representation, which addresses the problem of circular effect in word embeddings. We study our method on a set of occupations, observing a general tendency of the majority of jobs to the male factor while there is strong bias in a few specific occupations to the female factor. This study enables further research on algorithmic gender debiasing, especially by using explicit vectors.

## References

- E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proc. of HTL-NAACL*.
- T. Bolukbasi, K. Chang, J. Zou, V. Saligrama, and A. Kalai. 2016. Man is to computer programmer as woman is to homemaker? In *Proc. of NIPS*.
- E. Bruni, N. Tran, and M. Baroni. 2014. Multimodal distributional semantics. *JAIR*.
- M. Faruqui, Y. Tsvetkov, D. Yogatama, C. Dyer, and N. A. Smith. 2015. Sparse overcomplete word vector representations. In *Proc. of ACL*.
- F. Hill, R. Reichart, and A. Korhonen. 2016. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.
- E. Huang, R. Socher, C. Manning, and A. Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proc. of ACL*.
- O. Levy and Y. Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proc. of NIPS*.
- O. Levy, Y. Goldberg, and I. Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *TACL*.
- T. Luong, R. Socher, and C. Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proc. of CoNLL*.
- A. Mnih and Y. Teh. 2012. A fast and simple algorithm for training neural probabilistic language models. In *Proc. of ICML*.
- R. Řehůřek and P. Sojka. 2010. Software framework for topic modelling with large corpora. In *Proc. of LREC W. on New Challenges for NLP*.
- Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2016. Sparse word embeddings using l1 regularized online learning. In *Proc. of IJCAI*.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL-HLT*.