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

Navid Rekabsaz[1], **Allan Hanbury**[2]
[1]Idiap Research Institue / Martigny, Switzerland
[2]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 embedding, it remains opaque what the dimensions of the vectors refer to, or in what extend 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' ($\boldsymbol{v}_{she}$ and $\boldsymbol{v}_{he}$), 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 $\boldsymbol{v}_{she}$ and $\boldsymbol{v}_{he}$ 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 $\boldsymbol{v}_{nurse}$ contains many concepts related to female. However, it also means that $\boldsymbol{v}_{she}$ contains high relation to the concept 'nurse'. We refer to this characteristic of word embedding as *circularity*. Considering this trait, given that $\boldsymbol{v}_{nurse}$ naturally contains the concept 'nurse', calculating the semantic similarity between $\boldsymbol{v}_{she}$ and $\boldsymbol{v}_{nurse}$ (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 $\boldsymbol{v}_{SHE}$, where its female-related dimensions are explicitly set to the ones of $\boldsymbol{v}_{she}$ and the rest to zero (the same process for bias towards male by defining the vector $\boldsymbol{v}_{HE}$).

The proposed gender vectors $\boldsymbol{v}_{SHE}$ and $\boldsymbol{v}_{HE}$ 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 ($\boldsymbol{V}$) and context ($\widetilde{\boldsymbol{V}}$) 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 drwan 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 $\sigma(\boldsymbol{v}_w \widetilde{\boldsymbol{v}}_c)$ where $\boldsymbol{v}_w$ is the vector representation of of w, $\widetilde{\boldsymbol{v}}_c$ context vector of c, and σ 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