Showing 1–30 of 30 results for all: Word embeddings bias

1. arXiv:1904.08783 [pdf, other] cs.CL cs.LG

Evaluating the Underlying Gender Bias in Contextualized Word Embeddings

Authors: Christine Basta, Marta R. Costa-jussà, Noe Casas

Abstract: Gender **bias** is highly impacting natural language processing applications. **Word embeddings** have clearly been proven both to keep and amplify gender biases that are present in current data sources. Recently, contextualized **word embeddings** have enhanced previous **word embedding** techniques by computing **word** vector representations dependent on the sentence they appear in. In this paper, we study the impact of this conceptual change in the **word embedding** computation in relation with gender **bias**. Our analysis includes different measures previously applied in the literature to standard **word embeddings**. Our findings suggest that contextualized **word embeddings** are less biased than standard ones even when the latter are debiased. △ Less

Submitted 18 April, 2019; **originally announced** April 2019.

2. arXiv:1904.05233 [pdf, other] cs.LG cs.CL stat.ML

What's in a Name? Reducing Bias in Bios without Access to Protected Attributes

Authors: Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, Adam Tauman Kalai

Abstract: There is a growing body of work that proposes methods for mitigating **bias** in machine learning systems. These methods typically rely on access to protected attributes such as race, gender, or age. However, this raises two significant challenges: (1) protected attributes may not be available or it may not be legal to use them, and (2) it is often desirable to simultaneously consider multiple protected attributes, as well as their intersections. In the context of mitigating **bias** in occupation classification, we propose a method for discouraging correlation between the predicted probability of an individual's true occupation and a **word embedding** of their name. This method leverages the societal biases that are encoded in **word embeddings**, eliminating the need for access to protected attributes. Crucially, it only requires access to individuals' names at training time and not at deployment time. We evaluate two variations of our proposed method using a large-scale dataset of online biographies. We find that both variations simultaneously reduce race and gender biases, with almost no reduction in the classifier's overall true positive rate. △ Less

Submitted 10 April, 2019; **originally announced** April 2019.

Comments: Accepted at NAACL 2019; Best Thematic Paper

3. arXiv:1904.04047 [pdf, other] cs.CL cs.LG stat.ML

Black is to Criminal as Caucasian is to Police:Detecting and Removing Multiclass Bias in Word Embeddings

Authors: Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, Alan W Black

Abstract: Online texts -- across genres, registers, domains, and styles -- are riddled with human stereotypes, expressed in overt or subtle ways. **Word embeddings**, trained on these texts, perpetuate and amplify these stereotypes, and propagate biases to machine learning models that use **word embeddings** as features. In this work, we propose a method to debias **word embeddings** in multiclass settings such as race and religion, extending the work of (Bolukbasi et al., 2016) from the binary setting, such as binary gender. Next, we propose a novel methodology for the evaluation of multiclass debiasing. We demonstrate that our multiclass debiasing is robust and maintains the efficacy in standard NLP tasks. △ Less

Submitted 3 April, 2019; **originally announced** April 2019.

Comments: Accepted as a conference paper at NAACL. 5 Pages excluding references, additional page for appendix

4. arXiv:1904.03310 [pdf, other] cs.CL

Gender Bias in Contextualized Word Embeddings

Authors: Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, Kai-Wei Chang

Abstract: In this paper, we quantify, analyze and mitigate gender **bias** exhibited in ELMo's contextualized **word** vectors. First, we conduct several intrinsic analyses and find that (1) training data for ELMo contains significantly more male than female entities, (2) the trained ELMo **embeddings** systematically encode gender information and (3) ELMo unequally encodes gender information about male and female entities. Then, we show that a state-of-the-art coreference system that depends on ELMo inherits its **bias** and demonstrates significant **bias** on the WinoBias probing corpus. Finally, we explore two methods to mitigate such gender **bias** and show that the **bias** demonstrated on WinoBias can be eliminated. △ Less

Submitted 5 April, 2019; **originally announced** April 2019.

5. arXiv:1904.01628 [pdf, other] cs.CL stat.AP

Identification, Interpretability, and Bayesian Word Embeddings

Authors: Adam M. Lauretig

Abstract: Social scientists have recently turned to analyzing text using tools from natural language processing like **word embeddings** to measure concepts like ideology, **bias**, and affinity. However, **word embeddings** are difficult to use in the regression framework familiar to social scientists: **embeddings** are are neither identified, nor directly interpretable. I offer two advances on standard **embedding** models to remedy these problems. First, I develop Bayesian **Word Embeddings** with Automatic Relevance Determination priors, relaxing the assumption that all **embedding** dimensions have equal weight. Second, I apply work identifying latent variable models to anchor the dimensions of the resulting **embeddings**, identifying them, and making them interpretable and usable in a regression. I then apply this model and anchoring approach to two cases, the shift in internationalist rhetoric in the American presidents' inaugural addresses, and the relationship between bellicosity in American foreign policy decision-makers' deliberations. I find that inaugural addresses became less internationalist after 1945, which goes against the conventional wisdom, and that an increase in bellicosity is associated with an increase in hostile actions by the United States, showing that elite deliberations are not cheap talk, and helping confirm the validity of the model. △ Less

Submitted 2 April, 2019; **originally announced** April 2019.

Comments: Accepted to the Third Workshop on Natural Language Processing and Computational Social Science at NAACL-HLT 2019

6. arXiv:1903.10561 [pdf, other] cs.CL cs.CY

On Measuring Social Biases in Sentence Encoders

Authors: Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, Rachel Rudinger

Abstract: The **Word Embedding** Association Test shows that GloVe and word2vec **word embeddings** exhibit human-like implicit biases based on gender, race, and other social constructs (Caliskan et al., 2017). Meanwhile, research on learning reusable text representations has begun to explore sentence-level texts, with some sentence encoders seeing enthusiastic adoption. Accordingly, we extend the **Word Embedding** Association Test to measure **bias** in sentence encoders. We then test several sentence encoders, including state-of-the-art methods such as ELMo and BERT, for the social biases studied in prior work and two important biases that are difficult or impossible to test at the **word** level. We observe mixed results including suspicious patterns of sensitivity that suggest the test's assumptions may not hold in general. We conclude by proposing directions for future work on measuring **bias** in sentence encoders. △ Less

Submitted 25 March, 2019; **originally announced** March 2019.

Comments: NAACL 2019

7. arXiv:1903.03862 [pdf, other] cs.CL

Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

Authors: Hila Gonen, Yoav Goldberg

Abstract: **Word embeddings** are widely used in NLP for a vast range of tasks. It was shown that **word embeddings** derived from text corpora reflect gender biases in society. This phenomenon is pervasive and consistent across different **word embedding** models, causing serious concern. Several recent works tackle this problem, and propose methods for significantly reducing this gender **bias** in **word embeddings**, demonstrating convincing results. However, we argue that this removal is superficial. While the **bias** is indeed substantially reduced according to the provided **bias** definition, the actual effect is mostly hiding the **bias**, not removing it. The gender **bias** information is still reflected in the distances between "gender-neutralized" **words** in the debiased **embeddings**, and can be recovered from them. We present a series of experiments to support this claim, for two debiasing methods. We conclude that existing **bias** removal techniques are insufficient, and should not be trusted for providing gender-neutral modeling. △ Less

Submitted 9 March, 2019; **originally announced** March 2019.

Comments: Accepted to NAACL 2019

8. arXiv:1902.00496 [pdf, ps, other] cs.CL doi 10.13140/RG.2.2.15426.02240

Examining the Presence of Gender Bias in Customer Reviews Using Word Embedding

Authors: A. Mishra, H. Mishra, S. Rathee

Abstract: Humans have entered the age of algorithms. Each minute, algorithms shape countless preferences from suggesting a product to a potential life partner. In the marketplace algorithms are trained to learn consumer preferences from customer reviews because user-generated reviews are considered the voice of customers and a valuable source of information to firms. Insights mined from reviews play an indispensable role in several business activities ranging from product recommendation, targeted advertising, promotions, segmentation etc. In this research, we question whether reviews might hold stereotypic gender **bias** that algorithms learn and propagate Utilizing data from millions of observations and a **word embedding** approach, GloVe, we show that algorithms designed to learn from human language output also learn gender **bias**. We also examine why such biases occur: whether the **bias** is caused because of a negative **bias** against females or a positive **bias** for males. We examine the impact of gender **bias** in reviews on choice and conclude with policy implications for female consumers, especially when they are unaware of the **bias**, and the ethical implications for firms. △ Less

Submitted 1 February, 2019; **originally announced** February 2019.

9. arXiv:1901.07656 [pdf, other] cs.CL

Attenuating Bias in Word Vectors

Authors: Sunipa Dev, Jeff Phillips

Abstract: **Word** vector representations are well developed tools for various NLP and Machine Learning tasks and are known to retain significant semantic and syntactic structure of languages. But they are prone to carrying and amplifying... ▽ More

Submitted 22 January, 2019; **originally announced** January 2019.

Comments: To appear in AIStats 2019

10. arXiv:1901.07002 [pdf, other] cs.LG cs.CL stat.ML

Error-Correcting Neural Sequence Prediction

Authors: James O' Neill, Danushka Bollegala

Abstract: In this paper we propose a novel neural language modelling (NLM) method based on \textit{error-correcting output codes} (ECOC), abbreviated as ECOC-NLM. This latent variable based approach provides a principled way to choose a varying amount of latent output codes and avoids exact softmax normalization. Instead of minimizing measures between the predicted probability distribution and true distribution, we use error-correcting codes to represent both predictions and outputs. Secondly, we propose multiple ways to improve accuracy and convergence rates by maximizing the separability between codes that correspond to classes proportional to **word embedding** similarities. Lastly, we introduce a novel method called \textit{Latent Mixture Sampling}, a technique that is used to mitigate exposure **bias** and can be integrated into training latent-based neural language models. This involves mixing the latent codes (i.e variables) of past predictions and past targets in one of two ways: (1) according to a predefined sampling schedule or (2) a differentiable sampling procedure whereby the mixing probability is learned throughout training by replacing the greedy argmax operation with a smooth approximation. In evaluating Codeword Mixture Sampling for ECOC-NLM, we also baseline it against CWMS in a closely related Hierarhical Softmax-based NLM. △ Less

Submitted 21 January, 2019; **originally announced** January 2019.

11. arXiv:1901.03116 [pdf, other] cs.CL

Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques

Authors: Joel Escudé Font, Marta R. Costa-jussà

Abstract: Neural machine translation has significantly pushed forward the quality of the field. However, there are remaining big issues with the translations and one of them is fairness. Neural models are trained on large text corpora which contains biases and stereotypes. As a consequence, models inherit these social biases. Recent methods have shown results in reducing gender **bias** in other natural language processing applications such as **word embeddings**. We take advantage of the fact that **word embeddings** are used in neural machine translation to propose the first debiased machine translation system. Specifically, we propose, experiment and analyze the integration of two debiasing techniques over GloVe **embeddings** in the Transformer translation architecture. We evaluate our proposed system on a generic English-Spanish task, showing gains up to one BLEU point. As for the gender **bias** evaluation, we generate a test set of occupations and we show that our proposed system learns to equalize existing biases from the baseline system. △ Less

Submitted 10 January, 2019; **originally announced** January 2019.

12. arXiv:1812.10424 [pdf, other] cs.CL cs.LG stat.ML

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

Authors: Navid Rekabsaz, Allan Hanbury

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 a novel 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. △ Less

Submitted 13 December, 2018; **originally announced** December 2018.

Comments: arXiv admin note: text overlap with arXiv:1707.06598

13. arXiv:1812.08769 [pdf, other] cs.CL cs.LG

What are the biases in my word embedding?

Authors: Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, Adam Tauman Kalai

Abstract: This paper presents an algorithm for enumerating biases in **word embeddings**. The algorithm exposes a large number of offensive associations related to sensitive features such as race and gender on publicly available **embeddings**, including a supposedly "debiased" **embedding**. These **embedded** biases are concerning in light of the widespread use of **word embeddings**. The associations are identified by geometric patterns in **word embeddings** that run parallel between people's names and common lower-case **words** and phrases. The algorithm is highly unsupervised: it does not even require the sensitive groups (such as gender or race) to be pre-specified. This is desirable because it may not always be easy to identify all vulnerable groups a priori, and because it makes it easier to identify biases against intersectional groups, which depend on combinations of sensitive features. The inputs to our algorithm are a list of target tokens, e.g. names, and a **word embedding**, and the outputs are a number of **Word Embedding** Association Tests (WEATs) that capture various biases present in the data. We illustrate the utility of our approach on publicly available **word embeddings** and lists of names, and evaluate its output using crowdsourcing. We also show how removing names may not remove potential proxy **bias**. △ Less

Submitted 22 December, 2018; **v1** submitted 20 December, 2018; **originally announced** December 2018.

Comments: At AIES 2019: the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society

14. arXiv:1812.04224 [pdf, other] cs.LG cs.CL stat.ML

On the Dimensionality of Word Embedding

Authors: Zi Yin, Yuanyuan Shen

Abstract: In this paper, we provide a theoretical understanding of **word embedding** and its dimensionality. Motivated by the unitary-invariance of **word embedding**, we propose the Pairwise Inner Product (PIP) loss, a novel metric on the dissimilarity between **word embeddings**. Using techniques from matrix perturbation theory, we reveal a fundamental **bias**-variance trade-off in dimensionality selection for **word embeddings**. This **bias**-variance trade-off sheds light on many empirical observations which were previously unexplained, for example the existence of an optimal dimensionality. Moreover, new insights and discoveries, like when and how **word embeddings** are robust to over-fitting, are revealed. By optimizing over the **bias**-variance trade-off of the PIP loss, we can explicitly answer the open question of dimensionality selection for **word embedding**. △ Less

Submitted 11 December, 2018; **originally announced** December 2018.

Comments: 18 pages, Advances in Neural Information Processing Systems 31 (NeurIPS 2018, Oral Presentation)

15. arXiv:1811.11002 [pdf, other] cs.CL cs.LG stat.ML

Correcting the Common Discourse Bias in Linear Representation of Sentences using Conceptors

Authors: Tianlin Liu, João Sedoc, Lyle Ungar

Abstract: Distributed representations of **words**, better known as **word embeddings**, have become important building blocks for natural language processing tasks. Numerous studies are devoted to transferring the success of unsupervised **word embeddings** to sentence **embeddings**. In this paper, we introduce a simple representation of sentences in which a sentence **embedding** is represented as a weighted average of **word** vectors followed by a soft projection. We demonstrate the effectiveness of this proposed method on the clinical semantic textual similarity task of the BioCreative/OHNLP Challenge 2018. △ Less

Submitted 17 November, 2018; **originally announced** November 2018.

Comments: Accepted by the BioCreative/OHNLP workshop of ACM-BCB 2018

16. arXiv:1810.13407 [pdf, other] eess.AS cs.CL cs.LG cs.SD

On The Inductive Bias of Words in Acoustics-to-Word Models

Authors: Hao Tang, James Glass

Abstract: Acoustics-to-**word** models are end-to-end speech recognizers that use **words** as targets without relying on pronunciation dictionaries or graphemes. These models are notoriously difficult to train due to the lack of linguistic knowledge. It is also unclear how the amount of training data impacts the optimization and generalization of such models. In this work, we study the optimization and generalization of acoustics-to-**word** models under different amounts of training data. In addition, we study three types of inductive **bias**, leveraging a pronunciation dictionary, **word** boundary annotations, and constraints on **word** durations. We find that constraining **word** durations leads to the most improvement. Finally, we analyze the **word embedding** space learned by the model, and find that the space has a structure dominated by the pronunciation of **words**. This suggests that the contexts of **words**, instead of their phonetic structure, should be the future focus of inductive **bias** in acoustics-to-**word** models. △ Less

Submitted 12 November, 2018; **v1** submitted 31 October, 2018; **originally announced** October 2018.

17. arXiv:1810.04528 [pdf, other] cs.CL cs.AI

Is there Gender bias and stereotype in Portuguese Word Embeddings?

Authors: Brenda Salenave Santana, Vinicius Woloszyn, Leandro Krug Wives

Abstract: In this work, we propose an analysis of the presence of gender **bias** associated with professions in Portuguese **word embeddings**. The objective of this work is to study gender implications related to stereotyped professions for women and men in the context of the Portuguese language...

Submitted 10 October, 2018; **originally announced** October 2018.

Journal ref: The 13th edition of the International Conference on the Computational Processing of Portuguese (PROPOR 2018)

18. arXiv:1810.03611 [pdf, other] cs.LG cs.CY stat.ML

Understanding the Origins of Bias in Word Embeddings

Authors: Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, Richard Zemel

Abstract: The power of machine learning systems not only promises great technical progress, but risks societal harm. As a recent example, researchers have shown that popular **word embedding** algorithms exhibit stereotypical biases, such as gender **bias**. The widespread use of these algorithms in machine learning systems, from automated translation services to curriculum vitae scanners, can amplify stereotypes in important contexts. Although methods have been developed to measure these biases and alter **word embeddings** to mitigate their biased representations, there is a lack of understanding in how **word embedding bias** depends on the training data. In this work, we develop a technique for understanding the origins of **bias** in **word embeddings**. Given a **word embedding** trained on a corpus, our method identifies how perturbing the corpus will affect the **bias** of the resulting **embedding**. This can be used to trace the origins of **word embedding bias** back to the original training documents. Using our method, one can investigate trends in the **bias** of the underlying corpus and identify subsets of documents whose removal would most reduce **bias**. We demonstrate our techniques on both a New York Times and Wikipedia corpus and find that our influence function-based approximations are extremely accurate. △ Less

Submitted 8 October, 2018; **originally announced** October 2018.

19. arXiv:1808.07231 [pdf, ps, other] cs.CL

Reducing Gender Bias in Abusive Language Detection

Authors: Ji Ho Park, Jamin Shin, Pascale Fung

Abstract: Abusive language detection models tend to have a problem of being biased toward identity **words** of a certain group of people because of imbalanced training datasets. For example, "You are a good woman" was considered "sexist" when trained on an existing dataset. Such model **bias** is an obstacle for models to be robust enough for practical use. In this work, we measure gender biases on models trained with different abusive language datasets, while analyzing the effect of different pre-trained **word embeddings** and model architectures. We also experiment with three **bias** mitigation methods: (1) debiased **word embeddings**, (2) gender swap data augmentation, and (3) fine-tuning with a larger corpus. These methods can effectively reduce gender **bias** by 90-98% and can be extended to correct model **bias** in other scenarios. △ Less

Submitted 22 August, 2018; **originally announced** August 2018.

Comments: 6 pages. Accepted at the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018

20. arXiv:1807.11714 [pdf, other] cs.CL

Gender Bias in Neural Natural Language Processing

Authors: Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, Anupam Datta

Abstract: We examine whether neural natural language processing (NLP) systems reflect historical biases in training data. We define a general benchmark to quantify gender **bias** in a variety of neural NLP tasks. Our empirical evaluation with state-of-the-art neural coreference resolution and textbook RNN-based language models trained on benchmark datasets finds significant gender **bias** in how models view occupations. We then mitigate **bias** with CDA: a generic methodology for corpus augmentation via causal interventions that breaks associations between gendered and gender-neutral **words**. We empirically show that CDA effectively decreases gender **bias** while preserving accuracy. We also explore the space of mitigation strategies with CDA, a prior approach to **word embedding** debiasing (WED), and their compositions. We show that CDA outperforms WED, drastically so when **word embeddings** are trained. For pre-trained **embeddings**, the two methods can be effectively composed. We also find that as training proceeds on the original data set with gradient descent the gender **bias** grows as the loss reduces, indicating that the optimization encourages **bias**; CDA mitigates this behavior. △ Less

Submitted 31 July, 2018; **originally announced** July 2018.

21. arXiv:1806.06301 [pdf, ps, other] cs.CL cs.AI stat.ML

Biased Embeddings from Wild Data: Measuring, Understanding and Removing

Authors: Adam Sutton, Thomas Lansdall-Welfare, Nello Cristianini

Abstract: Many modern Artificial Intelligence (AI) systems make use of data **embeddings**, particularly in the domain of Natural Language Processing (NLP). These **embeddings** are learnt from data that has been gathered "from the wild" and have been found to contain unwanted biases. In this paper we make three contributions towards measuring, understanding and removing this problem. We present a rigorous way to measure some of these biases, based on the use of **word** lists created for social psychology applications; we observe how gender **bias** in occupations reflects actual gender **bias** in the same occupations in the real world; and finally we demonstrate how a simple projection can significantly reduce the effects of **embedding bias**. All this is part of an ongoing effort to understand how trust can be built into AI systems. △ Less

Submitted 16 June, 2018; **originally announced** June 2018.

Comments: Author's original version

22. arXiv:1805.11295 [pdf] cs.CL

Unsupervised detection of diachronic word sense evolution

Authors: Jean-François Delpech

Abstract: Most **words** have several senses and connotations which evolve in time due to semantic shift, so that closely related **words** may gain different or even opposite meanings over the years. This evolution is very relevant to the study of language and of cultural changes, but the tools currently available for diachronic semantic analysis have significant, inherent limitations and are not suitable for real-time analysis. In this article, we demonstrate how the linearity of random vectors techniques enables building time series of congruent **word embeddings** (or semantic spaces) which can then be compared and combined linearly without loss of precision over any time period to detect diachronic semantic shifts. We show how this approach yields time trajectories of polysemous **words** such as amazon or apple, enables following semantic drifts and gender **bias** across time, reveals the shifting instantiations of stable concepts such as hurricane or president. This very fast, linear approach can easily be distributed over many processors to follow in real time streams of social media such as Twitter or Facebook; the resulting, time-dependent semantic spaces can then be combined at will by simple additions or subtractions. △ Less

Submitted 30 May, 2018; **v1** submitted 29 May, 2018; **originally announced** May 2018.

Comments: 10 pages, 1 figure, 10 tables

23. arXiv:1804.06876 [pdf, other] cs.CL cs.AI

Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods

Authors: Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang

Abstract: We introduce a new benchmark, WinoBias, for coreference resolution focused on gender **bias**. Our corpus contains Winograd-schema style sentences with entities corresponding to people referred by their occupation (e.g. the nurse, the doctor, the carpenter). We demonstrate that a rule-based, a feature-rich, and a neural coreference system all link gendered pronouns to pro-stereotypical entities with higher accuracy than anti-stereotypical entities, by an average difference of 21.1 in F1 score. Finally, we demonstrate a data-augmentation approach that, in combination with existing **word-embedding** debiasing techniques, removes the **bias** demonstrated by these systems in WinoBias without significantly affecting their performance on existing coreference benchmark datasets. Our dataset and code are available at http://winobias.org. △ Less

Submitted 18 April, 2018; **originally announced** April 2018.

Comments: NAACL '18 Camera Ready

24. arXiv:1803.11175 [pdf, other] cs.CL

Universal Sentence Encoder

Authors: Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil

Abstract: We present models for encoding sentences into **embedding** vectors that specifically target transfer learning to other NLP tasks. The models are efficient and result in accurate performance on diverse transfer tasks. Two variants of the encoding models allow for trade-offs between accuracy and compute resources. For both variants, we investigate and report the relationship between model complexity, resource consumption, the availability of transfer task training data, and task performance. Comparisons are made with baselines that use **word** level transfer learning via pretrained **word embeddings** as well as baselines do not use any transfer learning. We find that transfer learning using sentence **embeddings** tends to outperform **word** level transfer. With transfer learning via sentence **embeddings**, we observe surprisingly good performance with minimal amounts of supervised training data for a transfer task. We obtain encouraging results on **Word Embedding** Association Tests (WEAT) targeted at detecting model **bias**. Our pre-trained sentence encoding models are made freely available for download and on TF Hub. △ Less

Submitted 12 April, 2018; **v1** submitted 29 March, 2018; **originally announced** March 2018.

Comments: 7 pages; fixed module URL in Listing 1

25. arXiv:1710.07045 [pdf, other] cs.CL

Unsupervised Context-Sensitive Spelling Correction of English and Dutch Clinical Free-Text with Word and Character N-Gram Embeddings

Authors: Pieter Fivez, Simon Šuster, Walter Daelemans

Abstract: We present an unsupervised context-sensitive spelling correction method for clinical free-text that uses **word** and character n-gram **embeddings**. Our method generates misspelling replacement candidates and ranks them according to their semantic fit, by calculating a weighted cosine similarity between the vectorized representation of a candidate and the misspelling context. To tune the parameters of this model, we generate self-induced spelling error corpora. We perform our experiments for two languages. For English, we greatly outperform off-the-shelf spelling correction tools on a manually annotated MIMIC-III test set, and counter the frequency **bias** of a noisy channel model, showing that neural **embeddings** can be successfully exploited to improve upon the state-of-the-art. For Dutch, we also outperform an off-the-shelf spelling correction tool on manually annotated clinical records from the Antwerp University Hospital, but can offer no empirical evidence that our method counters the frequency **bias** of a noisy channel model in this case as well. However, both our context-sensitive model and our implementation of the noisy channel model obtain high scores on the test set, establishing a state-of-the-art for Dutch clinical spelling correction with the noisy channel model. △ Less

Submitted 19 October, 2017; **originally announced** October 2017.

Comments: Appears in volume 7 of the CLIN Journal, http://www.clinjournal.org/biblio/volume

Journal ref: CLIN Journal, Volume 7, 2017

26. arXiv:1704.06380 [pdf, other] cs.CL

Improving Context Aware Language Models

Authors: Aaron Jaech, Mari Ostendorf

Abstract: Increased adaptability of RNN language models leads to improved predictions that benefit many applications. However, current methods do not take full advantage of the RNN structure. We show that the most widely-used approach to adaptation (concatenating the context with the **word embedding** at the input to the recurrent layer) is outperformed by a model that has some low-cost improvements: adaptation of both the hidden and output layers. and a feature hashing **bias** term to capture context idiosyncrasies. Experiments on language modeling and classification tasks using three different corpora demonstrate the advantages of the proposed techniques. △ Less

Submitted 20 April, 2017; **originally announced** April 2017.

27. arXiv:1608.07187 [pdf, other] cs.AI cs.CL cs.CY cs.LG doi 10.1126/science.aal4230

Semantics derived automatically from language corpora contain human-like biases

Authors: Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan

Abstract: Artificial intelligence and machine learning are in a period of astounding growth. However, there are concerns that these technologies may be used, either with or without intention, to perpetuate the prejudice and unfairness that unfortunately characterizes many human institutions. Here we show for the first time that human-like semantic biases result from the application of standard machine learning to ordinary language---the same sort of language humans are exposed to every day. We replicate a spectrum of standard human biases as exposed by the Implicit Association Test and other well-known psychological studies. We replicate these using a widely used, purely statistical machine-learning model---namely, the GloVe **word embedding**---trained on a corpus of text from the Web. Our results indicate that language itself contains recoverable and accurate imprints of our historic biases, whether these are morally neutral as towards insects or flowers, problematic as towards race or gender, or even simply veridical, reflecting the {\em status quo} for the distribution of gender with respect to careers or first names. These regularities are captured by machine learning along with the rest of semantics. In addition to our empirical findings concerning language, we also contribute new methods for evaluating **bias** in text, the **Word Embedding** Association Test (WEAT) and the **Word Embedding** Factual Association Test (WEFAT). Our results have implications not only for AI and machine learning, but also for the fields of psychology, sociology, and human ethics, since they raise the possibility that mere exposure to everyday language can account for the biases we replicate here. △ Less

Submitted 25 May, 2017; **v1** submitted 25 August, 2016; **originally announced** August 2016.

Comments: 14 pages, 3 figures

28. arXiv:1607.06520 [pdf, other] cs.CL cs.AI cs.LG stat.ML

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Authors: Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai

Abstract: The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with **word embedding**, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even **word embeddings** trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender **bias** is first shown to be captured by a direction in the **word embedding**. Second, gender neutral **words** are shown to be linearly separable from gender definition **words** in the **word embedding**. Using these properties, we provide a methodology for modifying an **embedding** to remove gender stereotypes, such as the association between between the **words** receptionist and female, while maintaining desired associations such as between the **words** queen and female. We define metrics to quantify both direct and indirect gender biases in **embeddings**, and develop algorithms to "debias" the **embedding**. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender **bias** in **embeddings** while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting **embeddings** can be used in applications without amplifying gender **bias**. △ Less

Submitted 21 July, 2016; **originally announced** July 2016.

29. arXiv:1606.06121 [pdf, other] cs.CL cs.LG stat.ML

Quantifying and Reducing Stereotypes in Word Embeddings

Authors: Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai

Abstract: Machine learning algorithms are optimized to model statistical properties of the training data. If the input data reflects stereotypes and biases of the broader society, then the output of the learning algorithm also captures these stereotypes. In this paper, we initiate the study of gender stereotypes in {\em **word embedding**}, a popular framework to represent text data. As their use becomes increasingly common, applications can inadvertently amplify unwanted stereotypes. We show across multiple datasets that the **embeddings** contain significant gender stereotypes, especially with regard to professions. We created a novel gender analogy task and combined it with crowdsourcing to systematically quantify the gender **bias** in a given **embedding**. We developed an efficient algorithm that reduces gender stereotype using just a handful of training examples while preserving the useful geometric properties of the **embedding**. We evaluated our algorithm on several metrics. While we focus on male/female stereotypes, our framework may be applicable to other types of **embedding** biases. △ Less

Submitted 20 June, 2016; **originally announced** June 2016.

Comments: presented at 2016 ICML Workshop on #Data4Good: Machine Learning in Social Good Applications, New York, NY

30. arXiv:1505.07931 [pdf, ps, other] cs.CL

Supervised Fine Tuning for Word Embedding with Integrated Knowledge

Authors: Xuefeng Yang, Kezhi Mao

Abstract: Learning vector representation for **words** is an important research field which may benefit many natural language processing tasks. Two limitations exist in nearly all available models, which are the **bias** caused by the context definition and the lack of knowledge utilization. They are difficult to tackle because these algorithms are essentially unsupervised learning approaches. Inspired by deep learning, the authors propose a supervised framework for learning vector representation of **words** to provide additional supervised fine tuning after unsupervised learning. The framework is knowledge rich approacher and compatible with any numerical vectors **word** representation. The authors perform both intrinsic evaluation like attributional and relational similarity prediction and extrinsic evaluations like the sentence completion and sentiment analysis. Experiments results on 6 **embeddings** and 4 tasks with 10 datasets show that the proposed fine tuning framework may significantly improve the quality of the vector representation of **words**. △ Less

Submitted 29 May, 2015; **originally announced** May 2015.