Squib

Fair Is Better than Sensational: Man Is to Doctor as Woman Is to Doctor

Malvina Nissim
University of Groningen
Center for Language and Cognition
m.nissim@rug.nl

Rik van Noord
University of Groningen
Center for Language and Cognition
r.i.k.van.noord@rug.nl

Rob van der Goot
IT University of Copenhagen
Computer Science Department
robv@itu.dk

Analogies such as man is to king as woman is to X *are often used to illustrate the amazing power of word embeddings. Concurrently, they have also been used to expose how strongly human biases are encoded in vector spaces trained on natural language, with examples like* man is to computer programmer as woman is to homemaker. *Recent work has shown that analogies are in fact not an accurate diagnostic for bias, but this does not mean that they are not used anymore, or that their legacy is fading. Instead of focusing on the intrinsic problems of the analogy task as a bias detection tool, we discuss a series of issues involving implementation as well as subjective choices that might have yielded a distorted picture of bias in word embeddings. We stand by the truth that human biases* are *present in word embeddings, and, of course, the need to address them. But analogies are not an accurate tool to do so, and the way they have been most often used has exacerbated some possibly non-existing biases and perhaps hidden others. Because they are still widely popular, and some of them have become classics within and outside the NLP community, we deem it important to provide a series of clarifications that should put well-known, and potentially new analogies, into the right perspective.*

1. Introduction

Word embeddings are distributed representations of texts that capture similarities between words. Besides improving a wide variety of NLP tasks, their power is often also tested intrinsically. Mikolov et al. (2013) introduced the idea of testing the soundness of embedding spaces via the analogy task. Analogies are equations of the form $A : B :: C : D$, or simply *A is to B as C is to D*. Given the terms A, B, C, the model must return the

Submission received: 28 May 2019; revised version received: 09 December 2019; accepted for publication: 19 January 2020.

https://doi.org/10.1162/COLI_a_00379

word that correctly stands for D in the given analogy. A most classic example is *man is to king as woman is to X*, where the model is expected to return *queen*, by subtracting "manness" from the concept of king to obtain some general royalty, and then re-adding some "womanness" to obtain the concept of queen ($king - man + woman = queen$).

Besides showing this kind of seemingly magical power, analogies have been extensively used to show that embeddings carry worrying biases present in our society and thus encoded in language. This bias is often demonstrated by using the analogy task to find stereotypical relations, such as the classic *man is to doctor as woman is to nurse* or *man is to computer programmer as woman is to homemaker*.

The potential of the analogy task has been recently questioned, though. It has been argued that what is observed through the analogy task might be mainly due to irrelevant neighborhood structure rather than to the vector offset that supposedly captures the analogy itself (Linzen 2016; Rogers, Drozd, and Li 2017). Also, Drozd, Gladkova, and Matsuoka (2016) have shown that the original and classically used 3COSADD method (Mikolov et al. 2013) is not able to capture all linguistic regularities present in the embeddings. With the recently proposed contextual embeddings (Peters et al. 2018; Devlin et al. 2019), it is non-trivial to evaluate on the analogy task, and is thus not commonly used. Recent research has shown that analogies are also not an accurate diagnostic to detect bias in word embeddings (Gonen and Goldberg 2019). Nevertheless, analogies are not only still widely used, but have also left a strong footprint, with some by-now-classic examples often brought up as proof of human bias in language models. A case in point is the opening speech by the ACL President at ACL 2019 in Florence, Italy, where the issue of bias in embeddings is brought up showing biased analogies from a 2019 paper (Manzini et al. 2019b).[1]

This contribution thus aims at providing some clarifications over the past use of analogies to hopefully raise further and broader awareness of their potential and their limitations, and put well-known and possibly new analogies in the right perspective.[2]

First, we take a closer look at the concept of analogy together with requirements and expectations. We look at how the original analogy structure was used to query embeddings, and some misconceptions that a simple implementation choice has caused. In the original proportional analogy implementation, all terms of the equation $A : B :: C : D$ are distinct (Mikolov et al. 2013), that is, the model is *forced* to return a *different* concept than any of the input ones. Given an analogy of the form $A : B :: C : D$, the code explicitly prevents yielding any term D such that $D == B$, $D == A$, or $D == C$. Although this constraint is helpful when all terms are expected to be different, it becomes a problem, and even a dangerous artifact, when the terms *could* or *should* be the same.

Second, we discuss different analogy detection strategies/measures that have been proposed, namely, the original 3COSADD measure, the 3COSMUL measure (Levy and Goldberg 2014), and the Bolukbasi et al. (2016) formula, which introduces a different take on the analogy construction, reducing the impact of subjective choices (Section 4.3).

Third, we highlight the role played by human biases in choosing which analogies to search for, and which results to report. We also show that even when subjective

[1] `https://www.microsoft.com/en-us/research/uploads/prod/2019/08/ACL-MingZhou-50min-ming.v9-5d5104dcbe73c.pdf`, slide 29.

[2] This work does not mean at all to downplay the presence and danger of human biases in word embeddings. On the contrary: Embeddings *do* encode human biases (Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018; Kozlowski, Taddy, and Evans 2019; Gonen and Goldberg 2019), and we agree that this issue deserves the full attention of the field (Hovy and Spruit 2016).

choices are minimized in input (as in Bolukbasi et al. 2016), parameter tuning might have consequences on the results, which should not go unnoticed or underestimated.

2. What Counts as *Analogy*?

In linguistics, analogies of the form $A : B :: C : D$ can be conceived on two main levels of analysis (Fischer 2019). The first one is morphological (so-called strict **proportional analogies**), and they account for systematic language regularities. The second one is more at the lexico-semantic level, and similarities can get looser and more subject to interpretation (e.g., *traffic is to street as water is to riverbed* [Turney 2012]). The original, widely used, analogy test set introduced by Mikolov et al. (2013) consists indeed of two main categories: morphosyntactic analogies (*car is to cars as table is to tables*) and semantic analogies (*Paris is to France as Tokyo is to Japan*). Within these, examples are classified in more specific sub-categories.

There are two important aspects that must be considered following the above. First, analogies are (traditionally) mostly conceived as featuring four *distinct* terms. Second, we need to distinguish between cases where there is one specific, expected, correct fourth term, and cases where there is not. Both aspects bear important methodological consequences in the way we query and analyze (biased) analogies in word embeddings.

2.1 Should All Terms Be Different?

Two of the four constraints introduced by Turney in formally defining analogies indirectly force the terms B and D to be different (Turney 2012, p. 540). Also, all the examples of the original analogy test (Mikolov et al. 2013) expect four different terms. Is this always the case? Are expressions featuring the same term twice non-analogies?

Because most out-of-the-box word embeddings have no notion of senses, homographs are modeled as one unit. For example, the infinitive form and the past tense of the verb *to read*, will be represented by one single vector for the word *read*. A consequence of this is that for certain examples, two terms would be identical, though they would be conceptually different. In strong verbs, infinitive and simple past can be homographs (e.g., split/split), and countries or regions can be homographs with their capitals (e.g., Singapore/Singapore). Other cases where all terms are not necessarily distinct include "is-a" relations (hypernyms, *cat:animal :: dog:animal*), and ordered concepts (*silver:gold :: bronze:silver*). Moreover, the extended analogy test set created by Gladkova, Drozd, and Matsuoka (2016) also includes examples where B is the correct answer, for example *country:language* and *thing:color*. While these examples might not be conceived as standard analogies, the issue with homographs remains.

2.2 Is There a Correct Answer?

In Mikolov's analogy set, all the examples are such that given the first three terms, there is one specific, correct (expected) fourth term. We can call such analogies "factual." While morphosyntactic analogies are in general indeed factual (but there are exceptions due to homographical ambiguities), the picture is rather different for the semantic ones. If we take *man:computer_programmer :: woman:X* as a semantic analogy, what is the "correct" answer? Is there an expected, unbiased completion to this query? Compare it to the case of *he:actor :: she:X*—it seems straightforward to assume that X should be resolved to *actress*. However, such resolution easily rescales the analogy to a morphosyntactic rather than semantic level, thereby also ensuring a factual, unbiased answer.

The morphosyntactic and semantic levels are indeed not always distinct. When querying *man:doctor :: woman:X*, is one after a morphosyntactic or a semantic answer? Morphosyntactically, we should resolve to *doctor*, thereby violating the all-terms-different constraint. If we take the semantic interpretation, there is no single predefined term that "correctly" completes the analogy (or perhaps *doctor* does here too).[3]

In such nonfactual, more creative analogies, various terms could be used for completion depending on the implied underlying relation (Turney 2012), which could be unclear or unspecified in the query. For the analogies used by Manzini et al. (2019b) (see Table 2 later in the article), for example, it is rather unclear what one would expect to find. Some of the returned terms might be biased, but in order to claim bias, one should also conceive the expected unbiased term. So, if *doctor* is not eligible by violating the distinction constraint, what would the unbiased answer be?

When posing queries, all such aspects should be considered, and one should be aware of what analogy algorithms and implementations are designed to detect. If the correct or unbiased answer to *man:woman :: doctor:X* is expected to be *doctor* and the model is not allowed to return any of the input terms as it would otherwise not abide to the definition of analogy, then such a query should not be asked. If asked anyway under such conditions, the model should not be charged with bias for not returning *doctor*.

3. Algorithms

We consider three strategies that have been used to capture analogies. We use the standard 3COSADD function Equation (1) from Mikolov et al. (2013), and 3COSMUL, introduced by Levy and Goldberg (2014) to overcome some of the shortcomings of 3COSADD, mainly ensuring that a single large term cannot dominate the expression Equation (2):

$$\underset{d}{\operatorname{argmax}} \; (\cos(d,c) - \cos(d,a) + \cos(d,b)) \tag{1}$$

$$\underset{d}{\operatorname{argmax}} \; \frac{\cos(d,c)\cos(d,b)}{\cos(d,a) + 0.001} \tag{2}$$

Bolukbasi et al. (2016) designed another formula, specifically focused on finding pairs $B : D$ with a similar direction as $A : C$:

$$S_{(a,c)}(b,d) = \begin{cases} \cos(a-c, b-d) & \text{if } ||b-d|| \leq \delta \\ 0 & \textit{otherwise} \end{cases} \tag{3}$$

They do not assume that B is known beforehand, and generate a ranked list of $B : D$ pairs, with the advantage of introducing less subjective bias in the input query (see Section 4.2). To ensure that B and D are related, the threshold δ is introduced, and set to 1.0 in Bolukbasi et al. (2016). This corresponds to $\pi/3$ and in practice means that B and D have to be closer together than two random embedding vectors. For convenience, because B is known beforehand in our setup and we are interested in examining the

[3] In this sense, it is admirable that Caliskan, Bryson, and Narayanan (2017) try to better understand their results by checking them against actual job distributions between the two genders.

top-N output, we rewrite Equation (3) as Equation (4) (note that they yield the exact same scores).

$$\underset{d}{\text{argmax}} \begin{cases} \cos(a-c, b-d) & \text{if } ||b-d|| \leq \delta \\ 0 & \textit{otherwise} \end{cases} \tag{4}$$

Even though it is not part of the equations, in practice most implementations of these optimization functions specifically ignore one or more input vectors. Most likely, this is because the traditional definition of analogies expects all terms to be different (see Section 2), and the original analogy test set reflects this. Without this constraint, 3COSADD for example would return B in absence of close neighbors. However, we have seen that this is a strong constraint, both in morphosyntactic and semantic analogies. Moreover, even though this constraint is mentioned in the original paper (Mikolov et al. 2013) and in follow-up work (Linzen 2016; Bolukbasi et al. 2016; Rogers, Drozd, and Li 2017; Goldberg 2017; Schluter 2018), we believe this is not common knowledge in the field (analogy examples are still widely used), and even more so outside the field.[4]

4. Is the Bias in the Models, in the Implementation, or in the Queries?

In addition to preventing input vectors from being returned, other types of implementation choices (such as punctuation, capitalization, or word frequency cutoffs), and subjective decisions play a substantial role. So, what is the actual influence of such choices on obtaining biased responses? In what follows, unless otherwise specified, we run all queries on the standard GoogleNews embeddings.[5] All code to reproduce our experiments is available: `https://bitbucket.org/robvanderg/w2v`.

4.1 Ignoring or Allowing the Input Words

In the default implementation of word2vec (Mikolov et al. 2013), gensim (Řehůřek and Sojka 2010) as well as the code from Bolukbasi et al. (2016), the input terms of the analogy query are not allowed to be returned.[6] We adapted all these code-bases to allow for the input words to be returned.[7]

We evaluated all three methods on the test set from Mikolov et al. (2013), in their constrained and unconstrained versions. We observe a large drop in macro-accuracy for 3COSADD and 3COSMUL in the unconstrained setting (from 0.71 to 0.21 and 0.73 to 0.45, respectively). In most cases, this is because the second term is returned as answer (*man is to king as woman is to king*, *D* == *B*), which happens if no close neighbor is found, but in some cases it is the third term that gets returned (*short is to shorter as new is to new*, *D* == *C*). A similar drop in performance was observed before by Linzen (2016) and Schluter (2018). The Bolukbasi et al. (2016) method shows very low scores (0.06 constrained, 0.11 unconstrained), but this was to be expected, since their formula was not specifically designed to capture factual analogies. But what is so different between factual and biased analogies?

[4] This was confirmed by the response we obtained when we uploaded a first version of the paper.

[5] `https://code.google.com/archive/p/word2vec/`.

[6] In Equation (3), in practice B will almost never be returned, as it will always be assigned a score of 0.0, making it the last ranked candidate.

[7] The 3COSADD unconstrained setting can be tested in an online demo: `www.robvandergoot.com/embs`.

Table 1
Example output of the three algorithms for their constrained (const.) and unconstrained (unconst.) implementations for three well-known gender bias analogies.

3COSADD		3COSMUL		BOLUKBASI	
const.	**unconst.**	**const.**	**unconst.**	**const.**	**unconst.**
man is to doctor as woman is to X					
gynecologist	doctor	gynecologist	doctor	midwife	gynecologist
he is to doctor as she is to X					
nurse	doctor	nurse	doctor	nurse	nurse
man is to computer_programmer as woman is to X					
homemaker	computer_programmer	homemaker	computer_programmer	–	schoolteacher

In Table 1, we report the results using the same settings for a small selection of mainstream examples from the literature on embedding bias. It directly becomes clear that removing constraints leads to different (and arguably less biased) results.[8] More precisely, for 3COSADD and 3COSMUL we obtain word B as answer, and using the method described by Bolukbasi et al. (2016) we obtain different results because with the vocabulary cutoff they used (50,000 most frequent words, see Section 4.3), *gynecologist* (51,839) and *computer_programmer* (57,255) were excluded.[9]

The analogy *man is to doctor as woman is to nurse* is a classic showcase of human bias in word embeddings, reflecting gendered stereotypes in our society. This is meaningful, however, only if the system were allowed to yield *doctor* (arguably the expected answer in absence of bias, see Section 2) instead of *nurse*, and it does not. But using the original analogy code, it is impossible to obtain *man is to doctor as woman is to doctor* (where $D == B$). Under such settings, it is not exactly fair to claim that the embedding space is biased because it does not return *doctor*.

4.2 Subjective Factors

Let us take a step back though, and ask: Why do people query *man is to doctor as woman is to X?* In fairness, one should wonder how much bias leaks in from our own views, preconceptions, and expectations. In this section we aim to show how these affect the queries we pose and the results we get, and how the inferences we can draw depend strongly on the choices we make in formulating queries and in reporting the outcome.

To start with, the large majority of the queries posed and found in the literature imply human bias. People usually query for *man:doctor :: woman:X*, which in 3COSADD and 3COSMUL is different from querying for *woman:doctor :: man:X*, both in results and in assumptions. This issue also raises the major, usually unaddressed question as to what would the unbiased, desired, D term be? Such bias-searching queries do not pose factual, one-correct-answer, analogies, unless interpreted morphosyntactically (Section 2).

Another subjective decision has to do with reporting results. One would think that the top returned term should always be reported, or possibly the top five, if willing

[8] This was noticed before: `https://www.youtube.com/watch?v=25nC0n9ERq4`, and `https://medium.com/artists-and-machine-intelligence/ami-residency-part-1-exploring-word-space-andprojecting-meaning-onto-noise-98af7252f749`.

[9] Though *man is to computer programmer as woman is to homemaker* is used in the title of Bolukbasi et al. (2016), this analogy is obtained using 3COSADD.

Table 2
Overview of reported biased analogies in Manzini et al. (2019b) and Manzini et al. (2019c), obtained with 3COSADD without constraints, but their embeddings as they are. "Idx" refers to the average position of the reported biased word as we find it in their five embedding sets trained with different seeds (i.e., the same they used.)

	Analogy	**Reported**	**Idx**	**Top-5 answers (averaged)**
Manzini et al. (2019b)	caucasian lawful black	criminal	2.0	lawful criminal defamation libel vigilante
	asian yuppie caucasian	hillbilly	5.0	yuppie knighting pasty hipster hillbilly
	black killer asian	engineer	5.2	addict aspie impostor killer engineer
	jew liberal christian	conservative	2.0	liberal conservative progressive heterodox secular
	jew journalist muslim	terrorist	1.6	terrorist purportedly journalist watchdog cia
	muslim regressive christian	conservative	9.2	regressive progressive milquetoast liberal neoliberal
Manzini et al. (2019c)	black homeless caucasian	servicemen	211.6	homeless somalis unemployed bangladeshi nigerians
	caucasian hillbilly asian	suburban	60.6	hillbilly hippy hick redneck hippie
	asian laborer black	landowner	3.0	laborer landowner fugitive worker millionaire
	jew greedy muslim	powerless	8.8	greedy corrupt rich marginalized complacent
	christian familial muslim	warzone	7172	familial domestic marital bilateral mutual
	muslim uneducated christian	intellectually	16.6	uneducated uninformed idealistic elitist arrogant

to provide a broader picture. However, subjective biases and result expectation might lead to discard returned terms that are not viewed as biased, and report biased terms that are appearing further down in the list, however. This causes a degree of arbitrariness in reporting results that can be substantially misleading. As a case in point, we discuss here the recent Manzini et al. paper, which is the work from which the examples used in the opening presidential speech of ACL 2019 were taken (see footnote 1). This paper was published in three subsequent versions, differing only in the analogy queries used and the results reported. We discuss this to show how subjective the types of choices above can be, and that transparency about methodology and implementation are necessary.

Initially Manzini et al. (2019a), the authors accidentally searched for the inverse of the intended query: instead of *A is to B as C is to X* (*black is to criminal as caucasian is to X*), they queried *C is to B as A is to X* (*caucasian is to criminal as black is to X*).[10] Surprisingly, they still managed to find biased examples by inspecting the top-N returned *D* terms. In other words, they reported the analogy *black is to criminal as caucasian is to police* to support the hypothesis that there is cultural bias against the black, but they had in fact found *caucasian is to criminal as black is to police*, so the complete opposite.

This mistake was fixed in subsequent versions (Manzini etal. 2019b,c), but it is unclear which algorithm is used to obtain the analogies. We tried the three algorithms in Section 3, and in Table 2 we show the results of 3COSADD, for which we could most closely reproduce their results. For their second version, in five out of their six examples the input word *B* would actually be returned before the reported answer *D*. For three of the six analogies, they pick a term from the returned top-10 rather than the top one. In their third version Manzini et al. (2019c), the authors changed the list of tested analogies,

[10] We confirmed this with the authors.

especially regarding the B terms. It is unclear under which assumption some of these "new" terms were chosen (*greedy* associated to *jew*, for example: what is one expecting to get—biased or non-biased—considering this is a negative stereotype to start with, and the C term is *muslim*?). However, for each of the analogy algorithms, we cannot reasonably reproduce four out of six analogies, even when inspecting the top 10 results.

Although qualitatively observing and weighing the bias of a large set of returned answers can make sense, it can be misleading to cherry-pick and report very biased terms in sensitive analogies. At the very least, when reporting term-N, one should report the top-N terms to provide a more complete picture.

4.3 Other Constraints

Using the BOLUKBASI formula is much less prone to subjective choices. It takes as input only two terms (A and C, like *man* and *woman*), thus reducing the bias present in the query itself, and consequently the impact of human-induced bias expectation. At the same time, though, starting with $A : C$, the formula requires some parameter tuning in order to obtain (a) meaningful $B : D$ pair(s). Such parameter values also affect the outcome, possibly substantially, and must be weighed in when assessing bias.

As shown in Equation (3), Bolukbasi et al. (2016) introduce a threshold δ to ensure that B and D are semantically similar. In their work, δ is set to 1 to ensure that B and D are closer than two random vectors (see Section 3). Choosing alternative values for δ will however yield quite different results, and it is not a straightforward parameter to tune, since it cannot be done against some gold standard, "correct" examples.

Another common constraint that can have a substantial impact on the results is limiting the embedding set to the top-N most frequent words. Both Bolukbasi et al. (2016) and Manzini et al. (2019a) filter the embeddings to only the 50,000 most frequent words, though no motivation for this need or this specific value is provided. Setting such an arbitrary value might result in the exclusion of valid alternatives. Further processing can also rule out potentially valid strings. For example, Manzini et al. (2019a) lowercase all words before training, and remove words containing punctuation after training, whereas Bolukbasi et al. (2016) keep only words that are shorter than 20 characters and do not contain punctuation or capital letters (after training the embeddings).

To briefly illustrate the impact of varying the values of the threshold and the vocabulary size when using the BOLUKBASI formula, in Table 3 we show the results when changing them for the query *man is to doctor as woman is to X* (given that B is known, we

Table 3
Influence of vocabulary size and threshold value for the method of Bolukbasi et al. (2016). With extreme values for the threshold, and allowing to return query words, the answer becomes "doctor" (≤0.5) and "she" (≥1.5). Italics: original settings.

	Threshold (δ)				
Voc. size	**0**.8	**0**.9	**1**.0	**1**.1	**1**.2
10,000	doctors	nurse	nurse	nurse	woman
50,000	doctors	nurse	*midwife*	midwife	woman
100,000	gynecologist	gynecologist	gynecologist	gynecologist	gynecologist
500,000	gynecologist	gynecologist	gynecologist	nurse_midwife	nurse_midwife
full vocab.	gynecologist	gynecologist	gynecologist	nurse_midwife	nurse_midwife

use Equation (4)). The variety of answers, ranging from what can be considered to be biased (*nurse*) to not biased at all (*doctors*), illustrates how important it is to be aware of the influence of choices concerning implementation and parameter values.

5. Final Remarks

If analogies might not be the most appropriate tool to capture certain relations, surely matters have been made worse by the way that consciously or not they have been used (Gonen and Goldberg [2019] have rightly dubbed them sensational "party tricks"). This is harmful for at least two reasons. One is that they get easily propagated both in science itself (Jha and Mamidi 2017; Gebru et al. 2018; Mohammad et al. 2018; Hall Maudslay et al. 2019), also outside NLP and artificial intelligence (McQuillan 2018) and in popularized articles (Zou and Schiebinger 2018), where readers are usually in no position to verify the reliability or significance of such examples. The other is that they might mislead the search for bias and the application of debiasing strategies. And although it is debatable whether we should aim at debiasing or rather at transparency and awareness (Caliskan, Bryson, and Narayanan 2017; Gonen and Goldberg 2019), it is crucial that we are clear and transparent about what analogies can and cannot do as a diagnostic for embeddings bias, and about all the implications of subjective and implementation choices. This is a strict prerequisite to truly understand how and to what extent embeddings encode and reflect biases of our society, and how to cope with this, both socially and computationally.

Acknowledgments

We would like to thank Hessel Haagsma for providing us with interesting and challenging examples, Gertjan van Noord for reading previous versions of this paper, and Stéphan Tulkens for the discussions on the analogy algorithms. We also like to thank the three anonymous reviewers for their comments. Moreover, we are grateful that we could exchange ideas with the authors of the two papers that we discuss mostly in this article, in particular Thomas Manzini, Kai-Wei Chang, and Adam Kalai. The opinions here expressed remain obviously our own.

References

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems 29*, pages 4349–4357, Curran Associates, Inc.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN.

Drozd, Aleksandr, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, Osaka.

Fischer, Olga. 2019. *Analogy in Language and Linguistics*, Oxford Bibliographies in Linguistics, Oxford University Press.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010 V4*.

Gladkova, Anna, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, CA.

Goldberg, Yoav. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

Gonen, Hila and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, MN.

Hall Maudslay, Rowan, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. Its all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5270–5278, Hong Kong.

Hovy, Dirk and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Jha, Akshita and Radhika Mamidi. 2017. When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Levy, Omer and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, MI.

Linzen, Tal. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin.

Manzini, Thomas, Yao Chong Lim, Yulia Tsvetkov, and Alan W. Black. 2019b. Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047 V2*.

Manzini, Thomas, Yao Chong Lim, Yulia Tsvetkov, and Alan W. Black. 2019c. Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047 V3*.

Manzini, Thomas, Lim Yao Chong, Alan W. Black, and Yulia Tsvetkov. 2019a. Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, MN.

McQuillan, Dan. 2018. People's councils for ethical machine learning. *Social Media+ Society*, 4(2):1–10. SAGE Publications, London, England.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, Scottsdale, AZ.

Mohammad, Saif, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, LA.

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, LA.

Řehůřek, Radim and Petr Sojka. 2010. Software framework for topic modeling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta.

Rogers, Anna, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148, Vancouver.

Schluter, Natalie. 2018. The word analogy testing caveat. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 242–246, New Orleans, LA.

Turney, Peter D. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, 44:533–585.

Zou, James and Londa Schiebinger. 2018. AI can be sexist and racist—it's time to make it fair. *Nature*, 559(7714):324.