

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

Malvina Nissim
University of Groningen

Rik van Noord
University of Groningen

Rob van der Goot
University of Groningen

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 built on natural language, like man is to computer programmer as woman is to homemaker.

Recent work has shown that analogies are in fact not such a diagnostic for bias, and other methods have been proven to be more apt to the task. However, beside the intrinsic problems with the analogy task as a bias detection tool, in this paper we show that a series of issues related to how analogies have been implemented and used might have yielded a distorted picture of bias in word embeddings. Human biases are present in word embeddings and need to be addressed. Analogies, though, are probably not the right tool to do so. Also, the way they have been most often used has exacerbated some possibly non-existing biases and perhaps hid 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 cases into the right perspective.

1. Introduction

Word embeddings are distributed representations of texts which capture similarities between words. Besides improving a wide variety of NLP tasks, the power of word embeddings 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 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 ($king - man + woman = queen$).

Beside showing this kind of 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. And regarding bias, Gonen and Goldberg (2019) have shown that analogies are not a good diagnostic for bias in embeddings, and have in a way misled bias identification, with consequences on debiasing efforts. Indeed, on the recent contextualised embeddings (Peters et al. 2018; Devlin et al. 2019), the analogy task is not used anymore to either evaluate their soundness, or to detect bias (Zhao et al. 2019; Basta, Costa-jussà, and Casas 2019; May et al. 2019).

While research indicates that this is the direction that should be pursued to deal with bias in word embeddings, 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 ones in the right perspective.

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. More specifically, in the original proportional analogy implementation, all terms of the equation $A : B :: C : D$ are distinct (Mikolov et al. 2013). In other words, the model is *forced* to return a *different* concept than any of the original ones. Given an analogy of the form $A : B :: C : D$, the model is not allowed to yield any term D such that $D == B$, $D == A$, or $D == C$, since the code explicitly prevents this. While this constraint is helpful when all terms of the analogy are expected to be different, it becomes a problem, and even a dangerous artifact, when the terms *could* or even *should* be the same.

Second, we discuss different analogy detection strategies/measures that have been proposed, namely the original 3COSADD measure, the revised 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.

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

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; Kozłowski, Taddy, and Evans 2018; Gonen and Goldberg 2019), and we agree that this issue deserves the full attention of the field (Hovy and Spruit 2016). This is the main reason why transparency over existing and possibly future analogies is crucial.

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

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: morpho-syntactic 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 isn't. Both aspects bear important methodological consequences in the way we query and analyse (biased) analogies in word embeddings.

2.1 Should all terms be different?

Of the four constraints introduced by Turney in formally defining analogies, the last two constraints indirectly force the terms B and D to be different (Turney 2012, p. 540). In practice, 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 modelled 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, for both macro-categories (morpho-syntactic and semantic), all of the examples are structured 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 quite straightforward to assume that X should be resolved as *actress*. However, such resolution easily rescales the analogy to a morphosyntactic rather than semantic level, thereby also ensuring a factual, unbiased answer.

The morpho-syntactic and semantic levels are indeed not always distinct. When querying *man:computer_programmer :: woman:X*, or *man:doctor :: woman:X*, is one after a morphosyntactic or a semantic answer? Morpho-syntactically, 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 maybe *doctor* does here too).²

In such non factual, 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 3), 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 to the semantic analogy *man:doctor :: woman:X*?

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 (Eq. 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 (Eq. 2):

$$\operatorname{argmax}_d (\cos(d, c) - \cos(d, a) + \cos(d, b)) \quad (1)$$

$$\operatorname{argmax}_d \frac{\cos(d, c) \cos(d, b)}{\cos(d, a) + 0.001} \quad (2)$$

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

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

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. If B is chosen

² 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.

beforehand (as in classic analogies), it is more efficient to transform the formula to only search for D :

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

Note that the resulting scores for the same analogy are exactly the same as in Equation 3. This formula allows one to examine the top- N $B : D$ candidates for a specific query, while Equation 3 will only return one single $B : D$ pair for a given B .

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 definitions of analogy require all terms to be different (see Section 2), and the original analogy test set reflects this. 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.³

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.⁴ All code to reproduce our experiments is available: <https://bitbucket.org/robvander/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.⁵ We adapted all these code-bases to allow for the input words to be returned.⁶

To compare the different methods, we evaluated all of them on the test set from Mikolov et al. (2013). The results in Table 1 show a large drop in performance for 3COSADD and 3COSMUL in the unconstrained setting. In most cases, this is because the second term is returned as answer (*man is to king as woman is to king*, thus $D == B$), but in some cases it is the third term that gets returned (*short is to shorter as new is to new*, thus $D == C$). A similar drop in performance was observed before by Linzen (2016) and Schluter (2018). The Bolukbasi et al. (2016) method shows very low scores, but this

³ This was confirmed by the response we got when we uploaded a first version of the paper.

⁴ <https://code.google.com/archive/p/word2vec/>

⁵ 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.

⁶ The 3COSADD unconstrained setting can be tested in an online demo: www.robvanderoot.com/embs

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?

	3COSADD	uncon.	3COSMUL	uncon.	BOLUKBASI	uncon.
Google Analogy - micro	0.74	0.21	0.75	0.47	0.04	0.11
Google Analogy - macro	0.71	0.21	0.73	0.45	0.06	0.11

Table 1: Accuracies of the three formulas on the Google Analogy test set (Mikolov et al. 2013), comparing the constrained version (original code), with the unconstrained version (uncon.). For all formulas, unconstrained means also taking the input vectors into account. For BOLUKBASI, more constraints were removed (see Section 4.3).

	man : doctor woman : X	he : doctor she : X	man : computer_programmer woman : X
3COSADD unconstrained	gynecologist doctor	nurse doctor	homemaker computer_programmer
3COSMUL unconstrained	gynecologist doctor	nurse doctor	homemaker computer_programmer
BOLUKBASI unconstrained	midwife gynecologist	nurse nurse	schoolteacher

Table 2: Example output of the three algorithms for their regular and unconstrained implementations for three well-known gender bias analogies.

In Table 2, 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.⁷ More precisely, for 3COSADD and 3COSMUL we get word *B* as answer, and using the method described by Bolukbasi et al. (2016) we get 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.⁸

The analogy *man is to doctor as woman is to nurse* in Table 2 is a classic showcase of human bias in word embeddings. This biased analogy reflecting gendered stereotypes in our society, is however truly meaningful 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 doesn't. 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 commonly used settings, it is not exactly fair to claim the embedding space is biased because it does not return *doctor*.

⁷ This was noticed before:

<https://medium.com/artists-and-machine-intelligence/ami-residency-part-1-exploring-word-space-andprojecting-meaning-onto-noise-98af7252f749> and <https://www.youtube.com/watch?v=25nC0n9ERq4>

⁸ 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.

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 we pose and find in the literature imply human bias. People usually query for *man:doctor :: woman:X*, which in 3COSADD and 3COSMUL is different than querying for *woman:doctor :: man:X*, both in results and in assumptions (often expecting to find a biased answer). 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 morpho-syntactically (See 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 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 however appearing further down in the list. 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 unless very transparent about methodology and implementation, one can almost claim anything they desire.

In the first version of their paper (Manzini et al. 2019a), the authors accidentally searched for the inverse of the intended query: instead of asking the model *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*

	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

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

is to criminal as black is to X).⁹ What is surprising is that 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 the analogy they had in fact found was caucasian is to criminal as black is to police, so the complete opposite. This should make us extremely wary of how easy it can be to find biased analogies when specifically looking for them.**

They fixed this mistake in their second and third version (Manzini et al. 2019b,c). However, it is unclear from the text which algorithm is used to obtain these analogies. We tried the three algorithms described in Section 3, and in Table 3 we show the results of 3COSADD, for which we could most closely reproduce their results (for both versions).

For their second version, in 5 out of their 6 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 returned one. In their third version (Manzini et al. 2019c), the authors change the list of tested analogies, especially regarding the B terms. It is unclear under which assumption some of these ‘new’ terms were chosen to be tested (*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.

While 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. For example, in the top-10 results for *man is to doctor as woman is to X* using 3COSADD, some terms refer to medical professions that have only women as patients (*gynecologist, obstetrician, ob_gyn, midwife*), going to show that it is not always clear which semantic relation is implied in the queries (in this case: professions or patients?), and more than one can be present and confounded. After all, if we query the Google embeddings using unconstrained 3COSADD for *man (or woman) is to doctor as dog (or animal) is to X*, we get *veterinarian*.

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). Parameter values, together with other pre-processing choices, also affect the outcome, possibly substantially, and must be weighed in when assessing bias.

As shown in Eq. 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,

⁹ We confirmed this with the authors.

Voc. size	Threshold (δ)				
	0.8	0.9	1.0	1.1	1.2
10,000	doctors	nurse	nurse	nurse	woman
25,000	doctors	nurse	nurse	nurse	woman
50,000	doctors	nurse	<i>midwife</i>	midwife	woman
100,000	gynecologist	gynecologist	gynecologist	gynecologist	gynecologist
250,000	gynecologist	gynecologist	gynecologist	gynecologist	gynecologist
500,000	gynecologist	gynecologist	gynecologist	nurse_midwife	nurse_midwife
full vocab.	gynecologist	gynecologist	gynecologist	nurse_midwife	nurse_midwife

Table 4: 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.

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).

In order to briefly illustrate the impact of varying the values of the δ threshold and the vocabulary size when using the BOLUKBASI formula, in Table 4 we show the results when changing them for the query *man is to doctor as woman is to X*.¹⁰ 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”, and 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 AI (McQuillan 2018) and in popularised 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 while it is debatable whether we should aim at removal 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 pre-requisite to truly understand how and to what extent embeddings encode and reflect the biases of our society, and how to cope with this, both socially and computationally.

¹⁰ Given that B is known, we use the formula in Eq. 4.

References

- Basta, Christine, Marta Ruiz Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the 1st ACL Workshop on Gender Bias for Natural Language Processing*.
- 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, Association for Computational Linguistics, Minneapolis, Minnesota.
- 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, Japan.
- Fischer, Olga. 2019. *Analogy in Language and Linguistics*, Oxford Bibliographies in Linguistics.
- 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.
- Gebbru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Dauméé 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, Association for Computational Linguistics, San Diego, California.
- 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, Association for Computational Linguistics, Minneapolis, Minnesota.
- Hall Maudslay, Rowan, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's 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, Association for Computational Linguistics, Hong Kong, China.
- 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, Association for Computational Linguistics, Berlin, Germany.
- 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, Association for Computational Linguistics, Vancouver, Canada.
- Kozłowski, Austin C, Matt Taddy, and James A Evans. 2018. The geometry of culture: Analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*.
- 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, Association for Computational Linguistics, Ann Arbor, Michigan.

- 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, Association for Computational Linguistics, Berlin, Germany.
- 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, Association for Computational Linguistics, Minneapolis, Minnesota.
- May, Chandler, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Association for Computational Linguistics, Minneapolis, Minnesota.
- McQuillan, Dan. 2018. People’s councils for ethical machine learning. *Social Media+ Society*, 4(2).
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.
- 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, Association for Computational Linguistics, New Orleans, Louisiana.
- 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, Association for Computational Linguistics, New Orleans, Louisiana.
- Řehůřek, Radim and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, ELRA, Valletta, Malta.
- 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, Association for Computational Linguistics, Vancouver, Canada.
- Schlueter, 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, Association for Computational Linguistics, New Orleans, Louisiana.
- Turney, Peter D. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, 44:533–585.
- Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized 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 629–634, Association for Computational Linguistics, Minneapolis, Minnesota.
- Zou, James and Londa Schiebinger. 2018. AI can be sexist and racist-it’s time to make it fair. *Nature*, 559(7714):324.