How does Grammatical Gender Affect Noun Representations in Gender-Marking Languages?

Hila Gonen[1] **Yova Kementchedjhieva**[2] **Yoav Goldberg**[1,3]
[1]Department of Computer Science, Bar-Ilan University
[2]University of Copenhagen
[3]Allen Institute for Artificial Intelligence
hilagnn@gmail.com,yova@di.ku.dk,yoav.goldberg@gmail.com

Abstract

Many natural languages assign grammatical gender also to inanimate nouns in the language. In such languages, words that relate to the gender-marked nouns are inflected to agree with the noun's gender. We show that this affects the word representations of inanimate nouns, resulting in nouns with the same gender being closer to each other than nouns with different gender. While "embedding debiasing" methods fail to remove the effect, we demonstrate that a careful application of methods that neutralize grammatical gender signals from the words' context when training word embeddings is effective in removing it. Fixing the grammatical gender bias yields a positive effect on the quality of the resulting word embeddings, both in monolingual and cross-lingual settings. We note that successfully removing gender signals, while achievable, is not trivial to do and that a language-specific morphological analyzer, together with careful usage of it, are essential for achieving good results.

1 Introduction

Work on distributional word embeddings focuses almost exclusively on English, or on cross-lingual and language-agnostic techniques. However, languages are diverse and different languages exhibit different linguistic phenomena, which may interact with the English-centric embedding learning algorithms. In this work we look into one such phenomenon—grammatical gender—and examine its effect on the learned representation.

Many languages have rich grammatical systems, that often include a complex gender system as well (Corbett, 1991). Languages with grammatical gender assign and morphologically mark gender not only to animate nouns (which have biological sex, e.g. man, woman, mother, father), but also to inanimate nouns (e.g. dream, book). This grammatical gender assignment is mostly arbitrary: the same inanimate concept can have different gender in different languages. For example, a *flower* is masculine in Italian (*fiore*) and feminine in German (*Blume*).

Languages often maintain an *agreement system* in which certain words agree on different morphological features with other words they relate to. For example, English present-tense verbs are inflected to agree with their nominal subject on the *number* feature. In other languages the agreement system is more elaborate, and in particular verbs, adjectives, determiners and other functions agree with nouns on many features, including gender (Corbett, 2006).[1]

Such grammatical agreement affects the distributional environment of nouns, as nouns of different gender become surrounded by different word forms: feminine nouns co-occur more with the feminine forms of words, while masculine nouns with the masculine forms. For example, the Italian word *viaggio* ("*journey*"-masc) will co-occur with *durato* ("*last*"-masc) and *lungo* ("*long*"-masc), while the word *gita* ("*trip*"-fem) will co-occur with *durata* ("*last*"-fem) and *lunga* ("*long*"-fem).

Such changes in the distributional environment may bias the learned distributional representations of inanimate nouns. Indeed, we see that the majority of the top-10 nearest neighbors of the word *gita* in Italian ("*trip*"-fem) are feminine words. Also, we notice that the word *viaggio* ("*journey*"-masc) is not on the list, while in English, for comparison, we can find *journey* in the top-10 nearest neighbors of *trip*.

In this work, we are interested in investigating, demonstrating and quantifying this effect beyond

[1] As the gender of nouns is fixed, the other elements are inflected to accommodate the agreement constraint. The nouns are said to *assign gender* to the other words.

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 463–471
Hong Kong, China, November 3-4, 2019.

the anecdotal level. We also explore methods for removing such unwanted biases.

We demonstrate that both in Italian and in German, the grammatical gender affects similarities between word representations (using words from SimLex-999 (Hill et al., 2015; Leviant and Reichart, 2015)): pairs of nouns with similar gender are closer to each other while pairs of nouns with different gender are farther apart.

After quantifying the effect, we explore several methods of reducing it. A popular choice would be to simply lemmatize all the words prior to feeding them to the embedding learning algorithm. However, full lemmatization can be destructive, in the sense that it will also remove morphological distinction that we may want to keep. We thus seek more surgical approaches. Interestingly, recent embedding debiasing approaches (Bolukbasi et al., 2016) do not work well. We instead look for methods that attempt to neutralize the gender signals from the training data. We find that such methods are effective in reducing the effect, but are also language specific and tricky to get right: we rely on language specific morphological analyzers while carefully accounting for their peculiarities and adjusting our use for each language. We take this work as a reminder that (a) linguistic resources such as lexicons and morphological analyzers are still relevant and useful (cf. (Zalmout and Habash, 2017)); (b) languages are diverse and different languages require different treatments; and (c) small details may matter a lot. In particular, existing tools and resources, either learned or human curated, should not be trusted blindly, but be carefuly adapted for the problem.

Finally, we show that reducing the effect of grammatical agreement also has a positive effect on the quality of the resulting word representations, both in monolingual and cross-lingual settings. We conclude that grammatical gender indeed has its imprints on the representations of inanimate nouns, and that this should be taken into account when working with gender-marking languages. Our code and debiased embeddings are available at `https://github.com/gonenhila/grammatical_gender`.

2 Background and Related Work

Word Embeddings Word embeddings have become an important component in many NLP models and are widely used for a vast range of downstream tasks. These models are based on the distributional hypothesis according to which words that occur in the same contexts tend to have similar meanings (Harris, 1954). Indeed, they aim to create word representations that are derived from their shared contexts, where the context of a word is essentially the words in its proximity (be it according to linear order in the sentence or according to syntactic relations) (Mikolov et al., 2013; Pennington et al., 2014; Levy and Goldberg, 2014).

Gender Biases in Word Embeddings Social gender bias was demonstrated to be consistent and pervasive across different word embeddings (Caliskan et al., 2017). Bolukbasi et al. (2016) show that using word embeddings for simple analogies surfaces many gender stereotypes. In addition, they define the gender bias of a word w by its projection on the "gender direction": $\overrightarrow{w} \cdot (\overrightarrow{he} - \overrightarrow{she})$, assuming all vectors are normalized. Positive bias stands for male-bias. For example, the bias of *manager* is 0.06, while the bias of *nurse* is -0.10[2].

Recently, some work has been done to reduce social gender bias in word embeddings, both as a post-processing step (Bolukbasi et al., 2016) and as part of the training procedure (Zhao et al., 2018). Bolukbasi et al. (2016) use a post-processing debiasing method. Given a word embedding matrix, they make changes to the word vectors in order to reduce the gender bias for all words that are not inherently gendered. They do that by zeroing the gender projection of each word on a predefined gender direction.[3]

In Zmigrod et al. (2019), the authors mitigate social gender bias in gender marking languages using counterfactual data augmentation. Gender-marking languages add several interesting dimensions to the story: words relating to animate concepts such as "nurse" or "cat" may have both masculine and feminine versions; the distributional environment of a word contains many more explicit gender cues; and inanimate concepts are also assigned gender. All of these factors interact in complicated ways. In this work we focus on purely grammatical gender—the gender that is assigned to inanimate nouns—and its effect on their resulting representations.

[2] in English word2vec embeddings (Mikolov et al., 2013) trained on Wikipedia.

[3] The gender direction is chosen to be the top principal component (PC) of ten gender pair difference vectors.

Grammatical Gender Bias in Word Embeddings Grammatical gender is manifested in a similar way to social bias. For example, when projected on the Italian gender direction $\overrightarrow{lui} - \overrightarrow{lei}$ (Italian equivalents of "he" and "she"), the word "secolo" (*century*, masculine) has positive bias of 0.073, while the word "zuppa" (*soup*, feminine) has negative bias of -0.079.[4]

We attribute this behavior to grammatical agreement. Since the context of different-gender nouns is expected to be very different because of the agreement of the surrounding words, and since the resulting representations are based on the context of the word, we expect grammatical gender to play a role in the representations—nouns with the same gender are expected to be closer together than nouns with different gender. For inanimate nouns, this behavior is undesired.

Word Embeddings and Morphology Word embeddings were shown to capture grammatical and morphological properties. Avraham and Goldberg (2017) show that standard training of word embeddings in Hebrew captures also morphological properties and that using the lemmas when composing the representations helps to better capture semantic similarities. Similarly, Basirat and Tang (2018) show that typical grammatical features are captured by Swedish word embeddings.

Cotterell et al. (2016) treat the sparsity problem of morphologically rich languages in word embedding. They present a Gaussian graphical model to smooth representations of observed words and extrapolate representations for unseen words using morphological resources. With similar motivation, Vulić et al. (2017) use morphological constraints in English in order to pull inflectional forms of the same word closer together and push derivational antonyms farther apart. Finally, Salama et al. (2018) enhance Arabic word embeddings by incorporating morphological annotations.

3 Grammatical Gender Affects Word Representations

As a first step, we aim to verify that the representation of inanimate nouns in gender-marking languages is indeed affected by their grammatical gender. Since English does not have grammatical gender, a natural approach would be to use it as a reference when measuring this phenomenon.

3.1 Inanimate Noun Pairs from SimLex-999

We take the inanimate noun portion of the SimLex-999 dataset (Hill et al., 2015), a gold standard resource for evaluating distributional semantic models. This dataset has an English version, and also German and Italian versions (Leviant and Reichart, 2015), and includes both similar and dissimilar word pairs, with human-assigned similarity judgments for each pair. This gives us 529 pairs of English words, along with high quality translations to Italian and German. We manually associate the Italian and German words with their grammatical gender.

3.2 Differences in Similarities

We divide the pairs in the gender-marking (GM) language (be it German or Italian) into two sets: (1) pairs of nouns that have the same gender in the GM language; (2) pairs of nouns that have different gender in the GM language. The respective English pairs are split in the same way, according to the gender of the nouns in the GM language. Thus, we end up with two sets of pairs in a GM language and their translations to English. Note that the English sets are different when used as a reference for German and Italian, since the split depends on the gender in the respective language.

For each set we compute the average of the cosine similarity of all word pairs within it. If gender plays a role in the representation of words, and indeed brings same-gender words closer together while keeping different-gender words farther apart, we expect to see a significant difference between the average similarity of the set of same-gender nouns and the set of different-gender nouns. As mentioned above, we compute these averages for English as a reference, where we expect a low difference between the two sets. Table 1 shows the results for Italian and German, compared to English. Indeed, in both cases, the difference between the average of the two sets is much bigger.

3.3 Rank in Nearest Neighbor List

We take the same sets as before, and for each pair in them we compute the rank of the second word in the nearest neighbor list of the first word and vice versa. For example, for the pair "parola" (*word*) and "dizionario" (*dictionary*) in Italian, we compute the rank of "dizionario" in the list of nearest neighbor of "parola" and the rank of "parola" in

[4] in Italian word2vec embeddings (Mikolov et al., 2013) trained on Wikipedia.

	Italian	En	German	En
Same Gender	0.442	0.424	0.491	0.446
Different Gender	0.385	0.415	0.415	0.403
difference	0.057	0.009	0.076	0.043

Table 1: Averages of similarities of pairs with same gender vs. different gender, along with the respective averages in English. The last row (difference) is the difference between the averages of the two sets.

the list of nearest neighbors of "dizionario".

We then compare the average ranking in each set, with English as the reference. If the gender affects the similarities between words, we expect same-gender pairs to have lower average than different-gender pairs (remember that the closest word is at the lowest rank: 1). Table 2 shows the results for Italian and German, compared to English. As expected, the average ranking of same-gender pairs is significantly lower than that of different-gender pairs, both for German and Italian, while the difference between the sets in English is much smaller.

4 Debiasing Methods do not Work

As mentioned above, grammatical gender bias shares some aspects with social gender bias. Keeping that in mind we first turn to use these existing methods of gender-debiasing in English word embeddings.

Bolukbasi's method (2016) requires sets of pairs that define the gender direction. For this we use their predefined pairs, since we target grammatical gender bias, which we have demonstrated to be similar to social gender bias. In addition, a predefined set of inherently-neutral words is also needed: these are the words that will be debiased by the algorithm. As a first step, and in order to estimate the feasibility of using this method for reducing the grammatical gender bias, we use the set of the inanimate nouns from SimLex-999 as our set of inherently-neutral words.[5]

The algorithm worked well in the sense that the bias of all inanimate nouns, when measured by their projection on the gender dimension, became zero. However, it also failed: the similarities between the inanimate nouns themselves hardly changed. Table 3 depicts the average similarities in Italian before and after debiasing.

This suggests that the information about the gender is deeply embedded in the representation and is not easy to remove in a post-processing phase. Specifically, zeroing the projection of a word's vector on the gender direction is not enough in order to remove all gender information from the word's representation. The fact that similarities between words hardly change implies that the projection on the gender direction is not the only indication of gender. These results align with the findings discussed in Gonen and Goldberg (2019).

We conclude that focusing on the projection of vectors on the gender direction is not the right way to go, and we opt to removing gender inflections from the context before training. We describe this in detail in the next section.

5 Removing Gender Inflection from the Context

As mentioned above, words in the surroundings of gender-marked nouns (e.g. articles, adjectives) are often inflected to agree with the gender of the noun they relate to. As we hypothesize that most of the effect shown in Section 3 is caused by this gender agreement, we try several schemes that aim to remove gender signals from the context.

A straight-forward approach would be to lemmatize all the words, which will remove all gender signals from the context of a word. However, this approach has two main downsides: 1) We would like to have a representation for all the words in the vocabulary, but changing also the target words will reduce the vocabulary size and result with missing words (we will no longer have different masculine and feminine forms for any word); 2) Lemmatization removes not only gender information, but also additional information (such as number and tense). While gender assignment is arguably arbitrary, and does not translate to an actual physical property of inanimate nouns in reality, other properties that agree with the noun, such as number, do hold in reality and signify actual properties of the target noun, which we prefer to preserve.

Thus, a better approach would be to neutralize gender signals from the context alone, keeping the target words intact. This way we do not change the resulting embedding vocabulary. This can be done using: 1) lemmatizing all the context words, where we lose additional information, as

[5] If this method doesn't mitigate the bias we showed in the previous section, then using inherently-neutral words extracted from the vocabulary automatically cannot possibly work as well.

	Italian			German		
	Same-gender	Diff-Gender	difference	Same-gender	Diff-Gender	difference
7–10	Og: 4884 Db: 5523 En: 6978	Og: 12947 Db: 7312 En: 2467	Og: 8063 Db: 1789 En: -4511	Og: 5925 Db: 7653 En: 4517	Og: 33604 Db: 26071 En: 8666	Og: 27679 Db: 18418 En: 4149
4–7	Og: 10954 Db: 12037 En: 15891	Og: 15838 Db: 12564 En: 17782	Og: 4884 Db: 527 En: 1891	Og: 19271 Db: 24845 En: 13282	Og: 27256 Db: 22970 En: 17649	Og: 7985 Db: -1875 En: 4367
0–4	Og: 23314 Db: 26386 En: 57278	Og: 35783 Db: 28067 En: 53053	Og: 12469 Db: 1681 En: -4225	Og: 50983 Db: 60603 En: 41509	Og: 85263 Db: 79081 En: 62929	Og: 34280 Db: 18478 En: 21420

Table 2: Averages of rankings of the words in same-gender pairs vs. different-gender pairs for Italian and German, along with their differences. **Og** stands for the original embeddings, **Db** for the debiased embeddings, and **En** for English. Each row presents the averages of pairs with the respective scores in SimLex-999 (0–4, 4–7, 7–10).

	Italian			
	Original	Debiased	English	Reduction
Same Gender	0.442	0.439	0.424	–
Different Gender	0.385	0.390	0.415	–
difference	0.057	0.049	0.009	**16.67%**

Table 3: Averages of similarities of pairs with same vs. different gender in Italian compared to the debiased version using Bolukbasi's (2016) method. The last row is the difference between the averages of the two sets. "Reduction" stands for gap reduction after debiasing.

discussed above; 2) changing all the context words to the same gender, while keeping all other features of the words intact. Once the whole context is of the same gender, we essentially lose the gender information altogether as all nouns have similar context, regardless their gender.[6]

5.1 The proposed approaches

We experiment with both lemmatization of context words and gender change of context words.

Lemmatization of Context Words When training word2vec (Mikolov et al., 2013), we use a morphological analyzer to identify the lemmas of words, and replace context words, but not target words, with their lemmas.

Gender Change of Context Words When training word2vec, we choose a gender (for example, masculine) and change all context words to that gender: each word that is identified as being of a different gender (in Italian: feminine, in German: feminine or neutral), is changed to its masculine form. This is also done using a morphological analyzer: when we identify a non-masculine analysis, we search for a masculine one that shares the same lemma and features.

In general, we found Italian to work better with gender change, and German to work better with lemmatization. We report full results in Section 6.

5.2 Challenges

While conceptually simple, fully neutralizing gender information is more challenging than it initially appears, and requires careful attention to "get right". We describe some cases in which gender information can leak.

Human Curator Choices The morphological analyzer sometimes assigns different lemmas to an opposite-gender pair, as a result of human curater design choices. For example, in Italian, "delle" is the feminine of "dei", but they are assigned the lemmas "della" and "del", respectively. Such cases leak gender signal in both cases of lemmatization and gender change: (1) When lemmatizing, each of the words gets a different lemma, manifesting the gender. (2) When changing the gender, the opposite-gender form of the word is not identified as these words do not share lemma, and the words stay unchanged.

This was very prominent in some high-frequency Italian words, and dealt with by fixing the analyzer: we identified all forms without a corresponding gendered-pair, manually aligned them, and assigned each pair a shared and unique lemma. This fix dramatically improved results when using either lemmatization or gender change.

Gender-Ambiguous Word Forms Many word forms have several morphological analyses, resulting in different lemmas. Inspecting this ambiguity reveals two specific issues, in German and in

[6] Context nouns are also kept unchanged since nouns do not agree with other nouns in their context, both in Italian and in German. Notably, in German, we lose the noun-ness information when we lowercase the corpus (as all nouns in German begin with an uppercase letter).

Italian. First, many German words are ambiguous with respect to gender. For example, "eine" has a frequent feminine reading, but also a rare masculine one. When changing words to masculine, this word is identified as potentially masculine, and kept intact. The presence of the context word "eine" now leaks a feminine signal.[7]

Second, Italian has many cases of two words with a similar set of possible lemmas but with different gender. For example, "usato" and "usata" are masculine and feminine, respectively, and both have "usare" and "usato" as possible lemmas. If we select a consistent lemma for each word type, and end up selecting a different lemma for each of "usato" and "usata", we again leak signal regarding the original gender.

One solution would be to use context-sensitive lemmatization, that chooses the correct analysis in context. However, doing this accurately is still an open problem. Our proposed solution is to randomly sample a lemma per word token. This improved lemmatization results in Italian by 25%.

Multiple Opposite-gender Forms for a Word In some cases, a single word might have multiple forms in the opposite gender. For example, the Italian "delle" is the feminine form of both "dei" and "degli", depending on the phonetic context. In this case, the former is much more common than the latter. A naive approach that chooses to convert "delle" to "degli" essentially keeps the feminine signal for these cases: every instance of "delle" changes to "delgli", which marks masculine nouns in much less common cases, while most masculine nouns are usually accompanied with the more common word "dei".

Ideally, when changing the gender of a word, we want to change a word by another word with a similar frequency, otherwise, the gender signal will be manifested in the frequency mismatch, as in the example above.

We deal with this issue using the following heuristic: when changing to masculine form (or any other gender form), for each word we first find all its possible masculine forms. Then, we check the frequency of the original word in the corpus, and choose the option with the closest frequency to it. This indeed yields better results: when not addressing the frequency issue in Italian, we are able to reduce the effect only by 35.42% (compared to 91.67%, see Section 6 for more details).

6 Results

We experimented with different schemes for each language, measuring their success at removing gender bias of inanimate nouns with respect to English.[8]

For German, we found lemmatization to work better than gender change. In Italian gender change got better results. Specifically, changing to feminine got much better results than changing to masculine, probably due to less ambiguity when changing to feminine in some very common articles (see full manual mapping in the Appendix), resulting in fewer cases in which the gender signal leaks through the frequencies of the changed words, as explained above. In addition, the manual fixes to the lemmatizer were crucial to get satisfying results for both methods.

While some of these findings depend on the specific morphological analyzer in use, the challenges and issues we demonstrate are relevant in any case.

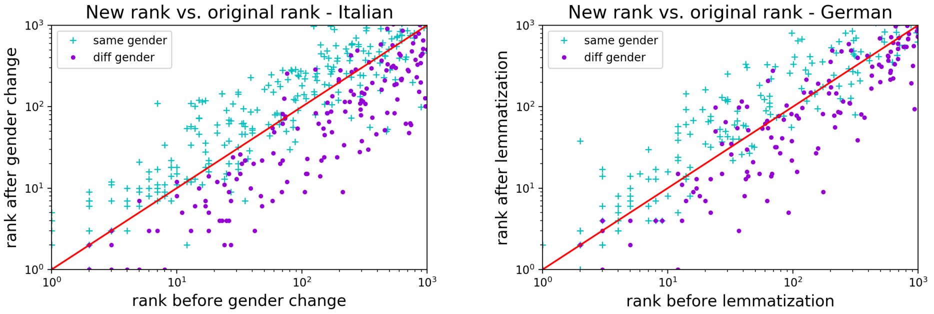
6.1 Reduction in Gender Bias

Differences in Similarities We repeat the experiment in Section 3.2—computing the average of pair similarities in each of the sets defined in Section 3, this time with the embeddings trained after removing gender signal from the context (debiasing). Table 4 shows the results for Italian and German, compared to English, both for the original and the debiased embeddings (for each language we show the results of the best performing debiased embeddings). As expected, in both languages, the difference between the average of the two sets with the debiased embeddings is much lower. In Italian, we get a reduction of 91.67% of the gap with respect to English. In German, we get a reduction of 100%. Note that for both languages, the main change is in the set of different-gender pairs, and not in the same-gender pairs. This makes sense as same-gender words have similar contexts both before and after our intervention, but different-gender words have different contexts before, but much more similar contexts after.

For comparison, in Italian we got 12.50% reduction when using the lemmatization scheme,

[7] A possible solution would be to replace words with their lemmas whenever we identify both feminine and masculine analyses. This did not improve results in practice.

[8] We used state-of-the-art morphological analyzers for both languages. Full implementation details can be found in the appendix.



(a)

(b)

Figure 1: The new rank of a word in the nearest neighbor list of its paired word. In cyan (+) – pairs with the same gender, in purple (·) – pairs with different gender. Most words with same-gender are located above $y = x$ (were drifted apart), while most words with different-gender are located below it (got closer together).

	Italian				German			
	Original	Debiased	English	Reduction	Original	Debiased	English	Reduction
Same Gender	0.442	0.434	0.424	–	0.491	0.478	0.446	–
Different Gender	0.385	0.421	0.415	–	0.415	0.435	0.403	–
difference	0.057	0.013	0.009	**91.67%**	0.076	0.043	0.043	**100%**

Table 4: Averages of similarities of pairs with same vs. different gender in Italian and German compared to English. The last row is the difference between the averages of the two sets. “Reduction” stands for gap reduction when removing gender signals from the context.

and 68.75% reduction when lemmatizing with the addition of the manual mapping. For German, the best result using gender change was a reduction of 48.48%, achieved by changing to neutral.

Rank in Nearest Neighbor List We repeat the experiment shown in Section 3.3—for each pair we compute the rank of the second word in the nearest neighbor list of the first word and vice versa. Then we compare the average ranking in each of the defined sets. Table 2 shows the results for Italian and German, both for the original and the debiased embeddings. As we expect, the difference between the average ranking of the two sets drops significantly for both languages.

In order to get a better picture of how the rankings of the different words change as a result of the gender signal removal, we take all pairs (and the inverted pairs). For each pair we plot the new rank of the second word in the nearest neighbors list of the first word as a function of its original rank before debiasing. Points above $y = x$ are of words that got a higher rank (lower in the list, farther from the first word), while points below this line are of words that got a lower rank (higher in the list, closer to the first word). Figure 1 shows these plots for Italian and German. As expected, most words of same-gender pairs are located above the line (were drifted apart), while most words of different-gender pairs are located below the line (got closer together).

6.2 Improvement in Word Similarities

Qualitative Evaluation As a qualitative evaluation, we take several words for SimLex-999 and look at their top-10 nearest neighbor lists, before and after applying our method. In Table 5 we show the top-10 lists for the words *vaso* (“*jar*”-masculine) in Italian, and *welt* (“*world*”-feminine) in German. It is evident that the words that are added to the list, are better correlated with the target word than those that are removed. Two additional words appear in the Appendix.

Evaluation on Simlex and WordSim-353 We evaluate the quality of the grammatical-gender-neutralized embeddings using two datasets for each language: SimLex-999 (Hill et al., 2015; Leviant and Reichart, 2015) and WordSim-353 (Finkelstein et al., 2002; Leviant and Reichart, 2015). Table 6 shows the results for Italian and German for both datasets, compared to the original embeddings. In both cases, the new embeddings perform better than the original ones.

Cross-lingual Word Embeddings Studies in language and cognition suggest that humans share

Italian		German	
vaso (jar-masculine)		welt (world-feminine)	
Orig	Debias	Orig	Debias
coccio	vasi	welt”	europas
recipiente	ciotola	europas	welt”
otre	bacinella (basin)	scheibenwelt	scheibenwelt
cinerario	recipiente	hässlichsten	universum (universe)
vasetto	coccio	*erde* (earth)	menschheitsgeschichte (human history)
bacile (basin)	cinerario	*weltgeschichte* (world history)	hässlichsten
kantharos	otre	*klügste* (wisest)	menschheit (mankind)
vasi	vasetto	*klügsten* (wisest)	schwarzafrikas
vassoio (tray)	brocca (pitcher)	schwarzafrikas	parallelwelten (parallel worlds)
coperchio (cover)	scodella (bowl)	*lustigsten* (funniest)	ulldart

Table 5: Examples of top-10 nearest neighbor lists for words in Italian and in German, before and after debiasing. In red (italic) are words that were removed from the list, and in blue (underlined) are words that were added to it. Translations to English (Google Translate) for the changed words are in parenthesis, when different from source.

	Italian		German	
	Orig	Debias	Orig	Debias
SimLex	0.280	**0.288**	0.343	**0.356**
WordSim	0.548	**0.577**	0.547	**0.553**

Table 6: Results on SimLex-999 and WordSim-353, in Italian and German, before and after debiasing.

	Italian		German	
	→ En	En →	→ En	En →
Orig	58.73	59.68	47.58	50.48
Debias	**60.03**	**60.96**	**47.89**	**51.76**

Table 7: Cross-lingual embedding alignment in Italian and in German, before and after debiasing.

a common semantic space, regardless of their native language (Youn et al., 2016). To the extent that embeddings capture the semantics of words, we can thus expect embedding spaces to have a similar structure across languages. Youn's statement concerns concepts and not words, however, and concepts can surface in many different forms in language, which interferes with how well embedding spaces align across languages (Søgaard et al., 2018). Thus, we expect grammatical gender to have a negative impact on alignability.

We explore this matter through the task of cross-lingual embedding alignment, wherein a cross-lingual embedding space is learned through an alignment of independently pre-trained monolingual embeddings for a directed pair of languages. The quality of cross-lingual embeddings learned this way can be evaluated intrinsically on the task of bilingual dictionary induction (BDI). BDI queries the cross-lingual embedding space with a seed of words in one language, retrieves their counterparts among the words in the other language[9] and evaluates the precision of the produced translations against a set of gold standard targets. We carry out experiments using the supervised variant of the MUSE embedding alignment system (Conneau et al., 2018) and report results on the inanimate portion of SimLex-999. We train a cross-lingual embedding alignment between English and either German or Italian, using the original and the debiased embeddings for these two languages. The results reported in Table 7 show that precision on BDI indeed increases as a result of the reduced effect of grammatical gender on the embeddings for German and Italian, i.e. that the embeddings spaces can be aligned better with the debiased embeddings.

7 Conclusion

We show that grammatical gender impacts word embeddings of inanimate nouns, both in Italian and in German, causing the similarities between words to change according to having same or different gender: the representations of same-gender words are closer together than representations of different-gender words.

We show that this effect can be almost completely removed when neutralizing gender signals in the context during training of the word embeddings. While most works in our field nowadays try to be language-independent, this is not always the right way to go: successfully removing those gender signals is not trivial to do and a language-specific morphological analyzer, to-

[9] This is done by minimizing a distance metric, most commonly, CSLS (Conneau et al., 2018).

gether with careful usage of it, are essential for achieving good results.[10]

In addition, this work serves as a reminder that languages other than English have different properties that are rarely dealt with when processing English. These aspects should be taken into account when dealing with morphologically reach languages, as not all models and algorithms for English can transfer directly to other languages.

8 Acknowledgements

The work was supported by the Israeli Science Foundation (grant number 1555/15) and by and the European Research Council (ERC Starting Grant iExtract 802774). We thank Valentina Pyatkin for the help with the Italian manual mapping.

References

Oded Avraham and Yoav Goldberg. 2017. The interplay of semantics and morphology in word embeddings. In *Proceedings of EACL*.

Ali Basirat and Marc Tang. 2018. Lexical and morphosyntactic features in word embeddings-a case study of nouns in swedish. In *ICAART*.

Tolga Bolukbasi, 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*.

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

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In *Proceedings of ICLR 2018*.

Greville G Corbett. 1991. Gender.

Greville G Corbett. 2006. *Agreement*. Cambridge University Press.

Ryan Cotterell, Hinrich Schütze, and Jason Eisner. 2016. Morphological smoothing and extrapolation of word embeddings. In *Proceedings of ACL*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on information systems*.

Hila Gonen 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 NAACL-HLT*.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3).

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4).

Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv:1508.00106*.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*.

Rana Aref Salama, Abdou Youssef, and Aly Fahmya. 2018. Morphological word embedding for arabic. In *ACLing*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of ACL*.

Ivan Vulić, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen. 2017. Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. *arXiv:1706.00377*.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F. Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. In *NIPS*.

Nasser Zalmout and Nizar Habash. 2017. Don't throw those morphological analyzers away just yet: Neural morphological disambiguation for arabic. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.

Ran Zmigrod, Sebastian J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*.

[10] Indeed, before implementing the specific fixes described in Section 5, the reduction compared to English when naively changing to masculine was substantially smaller, 35.42% reduction compared to 91.67% in Italian, and 12.12% compared to 100.00% (with lemmatization) in German.