Mitigating Gender Bias in Knowledge-Based Graphs Using Data Augmentation: WordNet Case Study

Claudia Rosas Raya, Ana Marcela Herrera Navarro

Universidad Autónoma de Querétaro,
Mexico

crosas17@alumnos.uaq.mx, claudiarosas16@gmail.com

Abstract. WordNet ontology is examined in order to show how it reflects historical gender biases in the semantic relationships between terms in a knowledge-based graph. We define a general benchmark to diagnose the gender bias in the WordNet ontology. Subsequently, we evaluate a set of words that have the equivalent in masculine and feminine and found their semantic-related terms according to WordNet. We then propose a technique to mitigate bias by data augmentation in order to create a neutral vector that combines both features without any distinction between genders. The results were compared with the Wu-Palmer semantic metric to validate the results and corroborate that gender bias was mitigated.

Keywords: Ontology, semantic similarity, data augmentation, WordNet, gender bias.

1 Introduction

Gender bias is exhibited repeatedly in the field of Natural Language Processing (NLP), including in training data, pre-trained models and algorithms themselves. Gender bias can be defined as the unfair difference in the way women and men are treated. The propagation of gender bias in Artificial Intelligence algorithms poses a danger of perpetuating stereotypes in real-world applications. Several problems have been already reported in different fields like machine learning (ML) [1].

Specifically, machine translation [2], where problems where found when translating sentences including words that historically had belonged to men or women like “The doctor” is more likely to be interpreted as masculine when translating and “The nurse” is more likely to be feminine.

Also, working with word embeddings [3] had shown issues like in [4] where results such as “man is to computer programmer as woman is to homemaker” is one of the problems detected when working with analogies. Or in [5] where it was found that training data contains significantly more male than female entities.

Based on Crawford categorization [6], bias can include harms of allocation, harms of representation and politics of classification. In terms of NLP applications, allocation bias is reproduced when models often perform better on data associated with majority. Detection of gender bias in Artificial Intelligence applications is a nascent field, and one of the first works is [7] defining gender bias as the correlation between the magnitude of the projection onto the gender subspace of a word embedding representing a gender-neutral word and that word's bias rating.

In [8] a new method called GN-GloVe is proposed. The authors train the word embeddings by isolating gender information in specific dimensions and maintaining gender-neutral information in another dimension.

The Implicit Association Test (IAT) is applied in psychology to measure subconscious gender bias in humans. In [9] the IAT's core concept is adopted, measuring gender bias through the difference in the strength of association of concepts, to measure bias in word embeddings using the Word Embedding Association Test (WEAT).

We show with new experiments that this methodology can be used to mitigate gender bias when using an ontology and show promising results to keep working with knowledge graphs.

The paper is organized as follows: Section 2 presents a brief survey of different works related to manipulation of an ontology. Section 3 explains in detail our data augmentation approach and the validation of our method. Section 4 includes the results and their interpretation. Finally, section 5 concludes the paper.

2 Related Work

2.1 Knowledge-based Graphs

Ontologies (a type of knowledge-based graphs) are a way to systematize knowledge providing semantic context in a method that a machine can handle it. An ontology is a "specification of a conceptualization" [10]. It models the classification of entities and the relationships between said entities. They are used to attempt to understand what exists in unstructured data in order to help systems to overcome semantic heterogeneity and facilitate them to interchange knowledge. They permit to transform unstructured text to structure data for a computer to understand it as it would be processing knowledge not only symbols. Semantic models contribute to build systems with more human-like behavior.

Wordnet [11, 12] is a semantic network widely investigated for NLP because of its accessibility. It has la large number of representations of semantic relationships, making it more appropriate for natural language understanding applications.
The main relation among words in WordNet is synonymy. Synonyms refer to words that denote the same concept and are interchangeable in many contexts. Additionally, a synonym contains a brief definition.

The most frequently encoded relation among terms is the super-subordinate relation (also called hyperonymy, hyponymy or ISA relation).

It links more general concepts to increasingly specific ones. All noun hierarchies ultimately go up the root node entity. Hyponymy relation is transitive [11, 12].

2.2 Semantic Similarity Measure

In general, semantic metrics can be classified into two groups: metrics which use only a thesaurus (e.g., WordNet) and those which use a thesaurus and probabilistic information from distributions in corpora [13].

The Wu-Palmer metric (WUP) weights the edges based on distance in the hierarchy. Namely, jumping from inanimate to animate is a larger distance than jumping from Feline to Canine. Using the same logic, the word senses of love and hate, while antonyms, are very related since they essentially belong to the same semantic type. Thus, it would be expected that the metrics give a higher similarity to them, than to the tuple love-romance; romance is very similar to love, but its type is not as close as say hate or dislike [13]. E.g., *hate-love* is 0.857 and *love-romance* is 0.615.

Other metrics based on thesaurus are: Path similarity, Leacock-Chodorow Similarity that work similarly to Wu-Palmer. But for experimentation purposes Wu-Palmer was chosen.

Wu-Palmer similarity [14] proposes a measure that takes into account the position of concepts in a taxonomy relative to the position of the Least Common Subsumer. Based on the edge counting method. Assuming that the similarity between two concepts is in function of the path length and depth. The range is between 0 and 1. It is continue and it normalizes the data; the score can never be zero, it is heavily dependent on the quality of the graph. It was first created to achieve machine translation between Chinese and English. The similarity measure of Wu and Palmer is defined by the following expression:

$$Sim_{wup} = 2 * \frac{N}{N1 + N2},$$

where N represents the distance to the closest common ancestor, N1 and N2 stand for node 1 and node 2 that are being compared.

2.3 Data Augmentation

Data augmentation has been shown to be flexible. It can mitigate gender bias in different applications.

Frequently a data set has a disproportionate number of references to one gender. To reduce this, [15] proposed to create an augmented data set identical to the original data set but biased towards the opposite gender and to train on the union of the original and data-swapped sets. The augmented data set was generated using gender-swapping information. The target of data augmentation is to reduce the biased predictions by training the model on a gender-balanced data set.

In [15] when creating a gender-balanced data set, data augmentation works as follows: for every sentence in the original data set, a sentence with the gender swapped is created.

Table 1. Extract of terms after obtaining the semantic terms related to the tuple and its definition.

Term	Hypernym	Hyponym	Definition
woman	adult, person, physical entity, organism, living thing	cinderella, madame, divorcee, dominatrix, geisha, girl, belle, bimbo, maid, sex kitten, tomboy, gold digger, gravida, prostitute, trophy wife	adult, female, person, wife, mistress, girlfriend.
man	adult, person, physical entity, organism, living thing	Adonis, bachelor, boy, bull, ejaculator, gentleman, guy, iron man, patriarch, peter pan, Tarzan, womanizer, Casanova, don Juan	adult, male, person, manservant, attendant, employer

Following, name-anonymization is applied to every original sentence and its gender-swapped equivalent. Name anonymization consists of replacing all named entities with anonymized entities, such as "E1". This deletes gender associations with named entities in sentences. Next, the model is trained on the union of the original data set with name-anonymization and the augmented data set. The identification of gender-specific words and their equivalent opposite gender word requires lists of terms associated to genders.

In [8, 16] independently designed GBETs based on Winograd Schemas. The corpus consists of sentences which contain a gender-neutral occupation (e.g., doctor), a secondary participant (e.g., patient), and a gendered pronoun that refers either the occupation or the participant. For each sentence, [16] considered three types of pronouns (female, male, or neutral), and [8] considered male and female pronouns. They designed metrics to analyze gender bias by examining how the performance difference between gender with respect to each occupation.

In hate speech detection [17], data augmentation reduced the False Positive Equality Difference (FPED) and False Negative Equality Difference (FNED) between male and female predictions of a Convolutional Neural Network (CNN) by a broad margin.

Data augmentation without name-anonymization has also been used to debias knowledge graphs built from Bollywood movie scripts [18] by swapping the nodes for the lead actor and actress, but metrics evaluating the success of gender-swapping were not provided.

3 Methodology

Schiebinger [19] suggested that scientific research fails to take gender issues into account, claiming that the phenomenon of male defaults on technology enables an asymmetry and WordNet is not an exception to this.

This proposal was implemented on MacBook Pro (retina, Mid 2012), with 8 Gb 1600 MHz DDR3 RAM memory and four-core Intel i7 2.3 GHz processor. Programmed in Python 3.8.1

Table 2. General results of the dataset analysis.

Female		Male	WuP	Male Words	Female words	Augmented	Common words
1.	Woman	Man	0.667	312	149	448	13
2.	Queen	King	0.571	48	68	90	26
3.	Mother	Father	0.923	60	46	95	11
4.	Girl	Boy	0.631	42	62	84	20
5.	Aunt	Uncle	0.600	16	15	22	9
6.	Actress	Actor	0.952	45	12	47	10
7.	Princess	Prince	0.900	21	17	28	10
8.	Waitress	Waiter	0.957	17	12	20	9
9.	Hen	Rooster	0.090	12	25	37	0
10.	Mare	Stallion	0.909	16	17	23	10
11.	Spinster	Bachelor	0.571	15	12	20	7
12.	Bride	Bridegroom	0.909	13	18	18	13
13.	Sister	Brother	0.545	29	28	42	15
14.	Countess	Count	0.133	65	33	68	30
15.	Duchess	Duke	0.72	11	13	17	7
16.	Goddess	God	0.824	108	8	109	7
17.	Heroine	Hero	0.125	38	22	42	18
18.	Madam	Sir	0.600	14	14	21	7
19.	Witch	Wizard	0.667	17	31	40	8
20.	Mummy	Daddy	0.857	9	14	15	8
21.	Girl guide	Boy scout	0.944	74	30	86	18
22.	Conductress	Conductor	0.545	82	10	85	7
23.	Chairwoman	Chairman	1.000	14	11	14	11
24.	Lady	Gentleman	0.600	18	22	33	7
25.	Headmistress	Headmaster	0.917	11	11	13	9
26.	Hostess	Host	0.947	69	27	76	20
27.	Wife	Husband	0.600	20	30	39	11
28.	Handlady	Landlord	0.957	9	10	11	8
29.	Lady	Lord	0.125	47	22	61	8
30.	Nun	Monk	0.600	13	19	24	8

Based on the principle of Data Saturation [20], a selected set of 30 tuples of words, were obtained and analyzed in WordNet ontology version 3.0 The tuples were compound by one male and one female word to extract the different semantically related words contained in the ontology hierarchy: hyponyms, hypernyms and the text

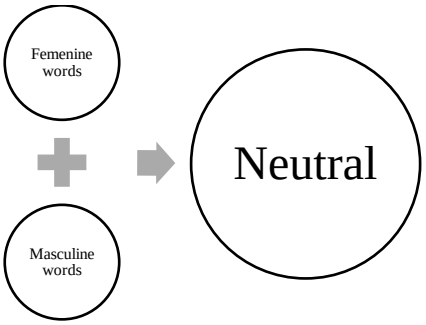

Fig. 1. The masculine terms plus the feminine terms create the gender-neutral result.

in the definition provided to every synset(synonym) of each word in the tuple. An extract of this phase is shown in Table 1.

In table 1, the example uses the tuple man-woman, getting their semantic related terms and their definitions. The text from the definition of the synonym's set were considered. Preprocessing the text included the removal of stop words (words with no semantic information like articles, prepositions, etc.), numbers and punctuation to obtain the words for further analysis.

Wu-Palmer similarity score was calculated and only the terms greater or equals that 0.5 were considered for further analysis. The automatization of noun gender detection is out of the scope of this paper. Therefore, the tuples used for experimentation were manually organized in male and female categories.

After the aforementioned analysis of the results are shown in Table 2 where seven columns can be observed. The first two contain the tuple of words, then the Wu-Palmer similarity between them is calculated. After obtaining all the semantic terms related with each word, the fourth and fifth column encloses the count of the total terms related with the tuple words.

The implementation of the Data Augmentation technique consists in adding to a neutral vector both contents of the vectors in the masculine and feminine lists of terms and deleting the duplicates to include the vocabulary that is missing and probably containing the gender bias information since the words should be describing practically the same topic (sixth column).

Once the list of words was obtained for each word, the female and the male counterpart, both lists were joined to form a neutral set of words that included both words (Fig 1). Based on the fact that both words were describing almost the exact same noun and that the definitions should be very alike, if they do not accomplish this, it may be a situation of gender bias in the description of each word regarding its counterpart.

3.1 Validation

To validate the augmented vector, a comparison between the words belonging to each original vector were compared to those contained in the neutral vector that were not covered in the original ones. Namely, if a word appeared in the neutral vector was not considered in the feminine vector, said word is compared with the feminine word in the

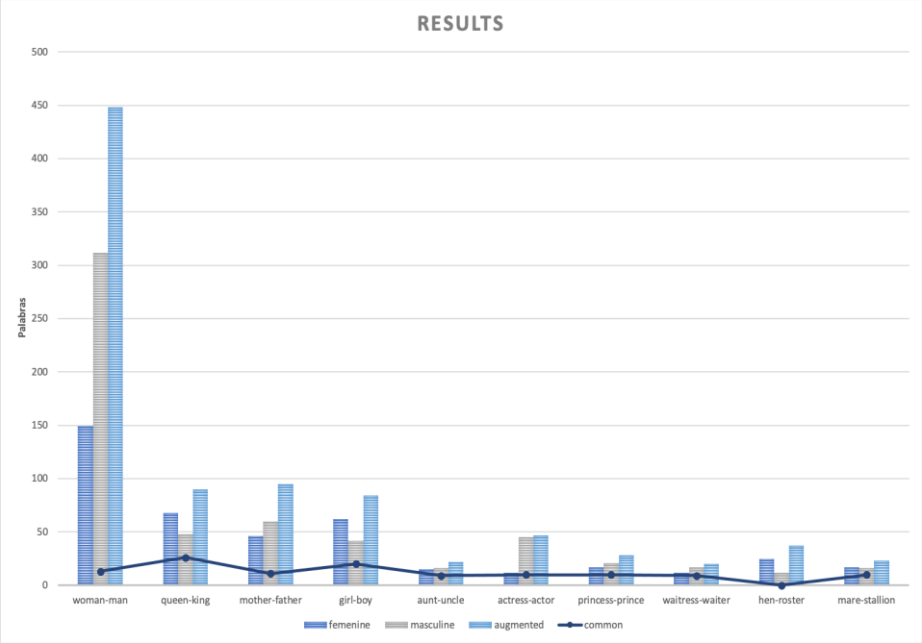

Fig. 2. Related words according to each tuple.

tuple that generated the vocabulary, e.g., the word employer in table 1 was not in the words related to woman and the neutral vector contains it then a comparison between woman and employer is applied to obtain the Wu-Palmer score.

4 Results

After extracting related words from the ontology, each tuple obtained its own list of words. In figure 2, an extract of 10 examples is plotted to show the data set size (in word quantity) individually and after the data augmentation. The number of related words between terms is also shown. The results after the process of data augmentation are shown in table 2. Where the tuples are shown individually and also, the words that previously had in common.

In table 3 it can be observed that 17 out of 30 tuples shown that the average Wu-Palmer similarity of the evaluated terms the validation process was above 0.5. That indicates that the terms contained in each word equivalent can be related to the other.

It can be interpreted as if there is a node that represents the semantic relationship between these words that were initially not considered by the ontology and thus the gender bias mitigation exists, e.g., Wu-Palmer score may be above 0.5 because the node person englobes the words woman and man and even though woman (or man) did not contain everything its equivalent counterpart had, the node *person* essentially considers the whole.

Table 3. Validation results.

	Tuple		Wu-Palmer average
1.	Woman	Man	0.50995519
2.	Queen	King	0.45074813
3.	Mother	Father	0.40816809
4.	Girl	Boy	0.60335059
5.	Aunt	Uncle	0.65689612
6.	Actress	Actor	0.73613838
7.	Princess	Prince	0.71450283
8.	Waitress	Waiter	0.51829822
9.	Hen	Roster	0.13443386
10.	Mare	Stallion	0.59177404
11.	Spinster	Bachelor	0.59299314
12.	Bride	bridegroom	0
13.	Sister	Brother	0.58704386
14.	Countess	Count	0.22238513
15.	Duchess	Duke	0.72356838
16.	Goddess	God	0.77088293
17.	Heroine	Hero	0.34991724
18.	Madam	Sir	0.64572908
19.	Witch	Wizard	0.39590476
20.	Mummy	Daddy	0.51336914
21.	Girl guide	Boy scout	0.22495529
22.	Conductress	Conductor	0.42019613
23.	chairwoman	Chairman	0
24.	lady	Gentleman	0.59448325
25.	headmistress	headmaster	0.65679842
26.	hostess	Host	0.44464883
27.	Wife	Husband	0.36695316
28.	landlady	Landlord	0.76397516
29.	lady	Lord	0.55205839
30.	nun	monk	0.22885582

4.1 Validation Results

Validation results are presented in Table 3.

5 Conclusions

Gender bias detection and mitigation are not easy tasks but the consequences of not start doing it are going to be important due to the more present artificial intelligence applications. Many of said application are based on trained data and the data is labeled somehow. In this work, the main target was one of the tools that are in charge of tagging data. In this particular scenario, we worked with the WordNet ontology, but the theoretical fundamentals can be applied to any ontology or knowledge-based graph with similar structure.

The main intention was to create a vector of neutral gender to try to stablish a start point into a space of non-binary gender approaches.

Text related bias depends not only on individual words, but also on the context in which they appear but as demonstrated, individual words are associated with historical biases posed on a binary gender reality. A possible interpretation of the biases showed by WordNet or any other model with similar problems would be that the model is no more, or less, biased than the real world. Assessing an accurate degree of prejudice of a model, requires the establishment of an ideal set of rules for language and that is a still-going discussion.

In the perspective of computing, data augmentation is easy to implement but it can be expensive if there is high variability in the data or if the data set is large. Furthermore, data augmentation tends to double the size of the training set or origin data, which can increase analysis or training time by a factor specific to the task at hand. Finally, blindly gender swapping in data augmentation can create nonsensical sentences or relationships, in a world with binary gender logic, man to godmother could obtained a high score if the graph organization admits it and therefore influence in the results of application like text clustering.

5.1 Future Work

To improve this work, random tuples could be selected from raw text. In order to prove the relevance in application-oriented works, text clustering and text indexing application may be benefited from approaches like this.

Data Augmentation techniques inevitably generate bigger sets of data and working with high-dimension vectors or datasets inevitably poses some restrictions in how to deal with it. But whenever possible, it is important that new techniques be developed because the results are increasingly having more impact in real lives.

We believe that our work can help to keep the debate going about the different machine bias. Non-binary genders as well as racial biases have largely been ignored in NLP and this may allow to perpetuate social struggles.

Gender bias in NLP is a compound issue, requiring interdisciplinary communication, teaming with social scientist and overall Computer Science must start to ask questions as NLP systems have been increasingly integrated with our daily life because it carries real consequences for people that are using the NLP developments.

References

1. Hellström, T., Dignum, V., Bensch, S.: Bias in machine learning what is it good (and bad) for? (2020)
2. Prates, M.O., Avelar, P.H., Lamb, L.C.: Assessing gender bias in machine translation: A case study with google translate. Neural Computing and Applications, pp. 1−19 (2019)
3. Kurpicz-Briki, M.: Cultural differences in bias? Origin and gender bias in pre-trained german and french word embeddings. In: Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS) (2020)
4. Nissim, M., van Noord, R., van der Goot, R.: Fair is better than sensational: Man is to doctor as woman is to doctor (2020)
5. Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., Chang, K.W.: Gender bias in contextualized word embeddings (2019)
6. Crawford, K.: The trouble with bias. Keynote at Neural Information Processing Systems (NIPS‘17) (2017)
7. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Neural Information Processing Systems (NIPS‘16) (2016)
8. Zhao, J., Zhou, Y., Li, Z., Wang, W., Chang, K.W.: Learning gender-neutral word embeddings. In: Empirical Methods of Natural Language Processing (EMNLP‘18) (2018)
9. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like bi-ases. Science, 356(6334), pp. 183–186 (2017)
10. Miller, G.: WordNet: A lexical database for english. Communications of the ACM, 38(11), pp. 39−41 (1995)
11. Fellbaum, C.: WordNet: An electronic lexical database. Cambridge, MA: MIT Press (1998)
12. Miller, G.: WordNet: A lexical database for english. In: Communications of the ACM, 38(11), pp. 39−41 (1995)
13. Jurafsky, D., Martin, J.H.: Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. Upper Saddle River, Prentice Hall, pp. 664−682 (2008)
14. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Gender bias in coreference resolution: Evaluation and debiasing methods. In: North American Chapter of the Association for Computational Linguistics (NAACL‘18) (2018)
15. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Gender bias in coreference resolution: Evaluation and debiasing methods. In: North American Chapter of the Association for Computational Linguistics (NAACL‘18) (2018)
16. May, Ch., Wang, A., Bordia, S., Bowman, S.R., Rudinger, R.: On measuring social biases in sentence encoders. In: North American Chapter of the Association for Computational Linguistics (NAACL‘19) (2019)
17. Park, J.H., Shin, J., Fung, P.: Reducing gender bias in abusive language detection. In: Empirical Methods of Natural Language Processing (EMNLP‘18) (2018)

18. Madaan, N., Mehta, S., Agrawaal, T., Malhotra, V., Aggarwal, A., Gupta, Y., Saxena, M.: Analyze, detect and re- move gender stereotyping from bollywood movies. In: Conference on Fairness, Accountability and Transparency (FAT‘18), pp. 92–105 (2018)
19. Schiebinger, L.: Scientific research must take gender into account. Nature, 507(7490), pp. 9−9 (2014)
20. Faulkner, S.L., Trotter, S.P.: Data saturation. The International Encyclopedia of Communication Research Methods, pp. 1−2 (2017)