Identifying and Mitigating Gender Bias in Knowledge Enhanced Contextual Word Representations

Chani Jung
KAIST, Republic of Korea
1016chani@kaist.ac.kr

Dongkwan Kim
KAIST, Republic of Korea
dongkwan.kim@kaist.ac.kr

1 Introduction

Recently, contextualized language models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and GPT (Radford et al., 2019) are widely used for Natural Language Processing tasks. Together with the rise of these models, there has also been works to investigate and mitigate social bias propagated from human-generated training corpora into the sentence representations (May et al., 2019; Kurita et al., 2019; Bartl et al., 2020; Liang et al., 2020). The research on bias in language model (LM) is important because NLP systems incorporating the LMs can further discriminate against disadvantaged people by reflecting and amplifying bias.

On the other hand, there has been approaches to further develop LMs by integrating task-specific knowledge bases (KBs) with them (Yang and Mitchell, 2019; Sun et al., 2018; Chen et al., 2018) or providing entity information to them (Ahn et al., 2016; Yang et al., 2017; Ji et al., 2017). Along with the trend, a method called KnowBert (Peters et al., 2019) has gotten a lot of attention, successfully enhancing knowledge in contextual word representations by embedding KBs into large scale models. To embed KBs, it uses an integrated entity linker to retrieve relevant entity embeddings, and update the contextual word representations via a form of word-to-entity attention. In contrast to previous approaches, the entity linkers and self-supervised language modeling objective are jointly trained end-to-end in a multitask setting that combines a small amount of entity linking supervision with a large amount of raw text.

Knowledge-enhanced contextual word representations set the new direction of language models for improved NLP systems. As the use and development of them increase, the need of analyzing the bias in them also rises. One can think of several ways to mitigate the bias in KnowBert, considering the structure of the model. First, we could borrow the methods of debiasing BERT. Since KnowBert is also a language model, it makes sense to apply BERT methods such as word-swapping of training corpora (Bartl et al., 2020) or modifying result representations (Liang et al., 2020) to KnowBert in the same way. Another approach can be debiasing the knowledge bases embedded in KnowBert. Knowledge base is the differentiated part of KnowBert from BERT, so it will be a different work from previous analysis for BERT. Due to the uniqueness of the approach, I selected mitigating bias in knowledge base of KnowBert to be explored in this work.

In order to analyze social bias in KnowBert, this research (1) measures gender bias associated with *career-related attributes* in KnowBert representations using template-based mask prediction method proposed by Kurita et al. (2019), and (2) proposes a debiasing method which modifies the entity embeddings used for the entity linking of KnowBert. The experiments show that the debiasing method does not always guarantee reduced bias in KnowBert representations, but it still indicates the potential for improving the method with further refined experiments.

2 Related Work

2.1 KnowBert

KnowBert is a general method to embed knowledge bases (KBs) into large scale contextual language models, thereby enhancing their representations with human-curated knowledge (Peters et al., 2019). It has a structure that a special set of knowledge-enhancing layers named Knowledge Attention and Recontextualization component (KAR) is inserted in the middle of pretrained BERT layers. Given contextual word representations of BERT interme-

diate layer, it uses entity embeddings learned from KB for entity linking via word-to-entity attention. It is different from the previous methods in that it doesn't require fully-annotated training data and it can integrate general - not task-specific - KBs into LMs, which are capable of improving diverse downstream tasks.

In Knowledge Attention and Recontextualization component (KAR) of KnowBert, KB contributes to language model by providing two types of knowledge - entity candidate selector and entity embeddings. Entity candidate selector takes text as input and provides pairs of potential mention span and candidate entities list. Since it is often implemented using precomputed dictionaries (Spitkovsky and Chang, 2012), KB-specific rules, or other heuristics (Mihaylov and Frank, 2018), it lies beyond the scope of this work and is not treated in this work.

Entity embeddings, on the other hand, are likely to contain bias of KB, at the same time having potential to mitigate the bias in KnowBert by modification of themselves. They are used at entity linking to provide knowledge about candidate entities for specific mention span, which means they affect how KnowBert disambiguates different entities. Therefore, I selected entity embeddings as the target of debiasing in this work.

2.2 Measuring Bias in Contextual Word Representations

Recently, there has been several methods suggested for quantifying bias in contextualized word representations. May et al. (2019) suggested Sentence Encoder Association Test (SEAT), that applies Word Embedding Association Test (WEAT) to the vector representation of a sentence. Following that, Tan and Celis (2019) adopted SEAT to be applied to not only sentence representations, but also contextual word representations. Meanwhile, Kurita et al. (2019) proposed a template-based method that uses *log probability bias score* as a bias measure of BERT. They showed that the *log probability bias score* captures social bias more effectively than traditional measures based on cosine similarity. This method is applied to bias measuring of KnowBert in this work. The details of the algorithm is described in Section 3.1.

2.3 Debiasing Contextualized Word Representations

There has also been recent works to reduce bias in contextualized word representations. First, there is approaches that modifies the training data. Bartl et al. (2020) swapped gender of person-denoting word in training corpora used for fine-tuning of BERT. There are also methods that alter output vector representation of language models. Liang et al. (2020) suggested a novel method named SENT-DEBIAS, which is a post-training method that removes projection onto a bias subspace from BERT representations.

2.4 Debiasing Knowledge Graph Embeddings

Researchers have proposed several approaches to train debiased knowledge graph embeddings (KGE) using adversarial learning. Fisher et al. (2020) presented a training method that first trains KGE to be neutral to sensitive attributes using adversarial loss, and then add the attributes back on whitelisted cases. Arduini et al. (2020) suggested a novel adversarial network that filters out sensitive attributes from KGE. On the other hand, Bourli and Pitoura (2020) proposed a method that modifies pretrained KGE, not changing the training model. They reduced bias in KGE by removing bias-directional component from each embedding of entities with sensitive attributes. This method is adopted for debiasing KGE trained from KBs in this work. More details of the method is described in Section 3.2.

3 Methods

3.1 Measuring Bias in KnowBert

To measure gender bias in KnowBert representations, a template-based mask prediction method is used. To quantify gender bias related to employment, Kurita et al. (2019) measured association between gender pronouns and *career-related attributes*. Since they compute bias score from the mask prediction output of BERT given the input template sentences, it is also feasible to apply it to KnowBert, which has the same form of input and output with those of BERT.

The association between a TARGET (gender pronoun) and an ATTRIBUTE (*career-related attribute*) is computed as below:

1. Prepare a template sentence.

e.g. [TARGET] is a [ATTRIBUTE]

2. Replace [TARGET] with [MASK] and compute p_{tgt}=P([MASK]=[TARGET] | sentence)
3. Replace both [TARGET] and [ATTRIBUTE] with [MASK], and compute prior probability p_{prior}=P([MASK]=[TARGET] | sentence)
4. Compute the association as $\frac{p_{tgt}}{p_prior}$, which is referred to as *increased log probability score*.

They use the difference between *increased log probability scores* of two targets (e.g. *he*, *she*) as a bias measure, naming it *log probability bias score*. Bias contained in KnowBert before and after debiasing is measured using this method and the *career-related attributes* dataset provided by the authors.

In this work, following template sentence is used.

- "TARGET is ATTRIBUTE", where TARGET are male and female pronouns (i.e., he and she) and the ATTRIBUTE are job titles or the positive and negative trait adjectives.

3.2 Debiasing KGE of KnowBert

To reduce bias in entity embeddings, I used the method adapted from the debiasing approach of knowledge graph embeddings by Bourli and Pitoura (2020). They removed gender bias in knowledge graph embeddings by subtracting projection of themselves onto a gender-directional vector. They defined gender-directional vector as $\overrightarrow{female} - \overrightarrow{male}$, assuming that all the male- and female-related entities would be connected to the *male* and *female* entity, respectively, in the knowledge graph. The debiased embedding becomes:

$$\overrightarrow{e}' = \overrightarrow{e} - \lambda \pi_{\overrightarrow{d}}(\overrightarrow{e}) \quad (1)$$

where $\overrightarrow{e}$ is the embedding of an entity, $\overrightarrow{d}$ is the gender-directional vector, and λ is the hyperparameter that determines the strength of debiasing, which is in between 0 and 1.

A difference between the original method and this work is that there are more than one pair of *female* and *male* entities in the vocabulary of KB in this work. Therefore, I removed the projections of embeddings on each gender-directional vector, which is defined by each *female-male* entity pair in the vocabulary. Also, individual hyperparameter λ_i for ith gender-directional vector is introduced to determine the contribution of each gender-directional vector to debiasing. Therefore, if k is the number of gender-directional vectors, Eq. 1 is modified into:

$$\overrightarrow{e}' = \overrightarrow{e} - \sum_{i=1}^{k} \lambda_i \pi_{\overrightarrow{d_i}}(\overrightarrow{e}) \quad (2)$$

where

$$0 < \sum_{i=1}^{k} \lambda_i < 1$$

In order to evaluate the debiasing method, I set the debiased entity embeddings $\overrightarrow{e}$ to be the entities in the *career-related attributes*, which are used in the template sentences in bias-measuring.

4 Experimental Settings

The code and data used in this work are available in my Github repository.[1]

4.1 Model

The KnowBert model used in the experiments is KnowBert-WordNet, implemented by Peters et al. (2019), in which WordNet metadata is inserted as a KB into pretrained BERT. From the synsets, lemmas, and their relationships provided by WordNet 3.0., a knowledge graph (KG) is constructed where each node is either a synset or lemma. From the KG, the entity embeddings of synsets and lemma are extracted using TuckER method (Balazevic et al., 2019). Synset glosses are also embedded into vectors using sentence embedding method called GenSen (Subramanian et al., 2018). Then, TuckER and GenSen embeddings are concatenated to form the embeddings used for entity linking of KnowBert. In KnowBert-WordNet, Knowledge Attention and Recontextualization component (KAR) - including entity linker - is inserted between layers 10 and 11 of the 12-layer BERTBASE model.

4.2 Datasets

Career-related attributes As explained in Section 3.1, mask prediction of template sentences is performed for bias-measuring of KnowBert. For *career-related attributes*, high-paying job titles and positive/negative traits provided by Kurita et al. (2019) are used after filtering out the words not in WordNet vocabulary. Technical skills, which also were used in the previous work, are excluded since most of them are not in WordNet vocabulary. The details about the *career-related attributes* are as below:

[1] https://github.com/chanijung/debias$_k$nowbert$_k$ge.

- *Employee Salary Dataset7 for Montgomery County of Maryland*[2] - Among its first 1000 high-paying job titles, 64 instances in Wordnet vocabulary are used.
- *Positive and Negative Traits Dataset*[3] - 188 and 288 adjectives considered "positive" and "negative" traits are used.

4.3 Debiasing Details

Debiasing target As described in Section 4.1, KnowBert-WordNet is used for the model to debias. Among TuckER(Balazevic et al., 2019) and GenSen (Subramanian et al., 2018) representations which are concatenated to form the entity embeddings of KnowBert-WordNet, only TuckER embeddings are debiased in this work. This is because GenSen embeddings are not graph-based embeddings but sentence representations, so the projection method designed for knowledge graph embeddings cannot be applied to them. Among TuckER entity embeddings, those corresponding to *career-related attributes* are debiased in the experiments. Then, we use the same attributes in bias measuring with mask prediction for evaluation of debiasing method.

Gender-directional vector As mentioned in Section 3.2, five gender-directional vectors are defined in the TuckER entity embedding space. All of them are pairs of *female* and *male* entities with different definitions. A synset entity is identified with a 3-part name of the form: *word.pos.nn*. The pairs of WordNet synset entities that compose gender-directional vectors are:

- *female.a.01*, *male.a.01*
- *female.n.01*, *male.n.01*
- *female.n.02*, *male.n.02*
- *female.s.02*, *male.s.02*
- *female.s.03*, *male.s.03*

Hyperparameter λ We introduced the hyperparameters λ_i, that decide the strength of debiasing and the contribution of i-th gender-directional vector to debiasing (Eq. 2). Since the purpose of this work is to verify the debiasing effect of the proposing method, the hyperparameter λ_is are equally fixed to 0.2. Although it is not performed in this work, the hyperparameters could be tuned through experiments to achieve the best performance for a specific model.

[2]https://catalog.data.gov/dataset/employee-salaries-2017
[3]http://ideonomy.mit.edu/essays/traits.html

	Original	Debiased
High-paying jobs	**48.5%**	19.7%
Positive traits	61.3%	**48.9%**
Negative traits	**54.9%**	59.0%

Table 1: Percentage of attributes that are more strongly associated with female gender

5 Results and Discussion

5.1 Debiasing Performance

Table 1 shows the percentage of attributes that are more strongly associated with female gender, in the mask prediction test on original and debiased KnowBert. Assuming that equal distribution of the percentages in two gender (i.e., 50%) is the best result for all of the attributes, the percentage did not consistently improve with debiasing in the three attributes. It got significantly worse in high-paying job titles, better in positive traits, and slightly worse in negative traits.

I attribute the inconsistent performance of debiasing method to three factors. First of all, the training epochs of Wordnet TuckER embeddings, entity linker, and the entire KnowBert model were extremely decreased compared to that of reference papers, due to time constraint. The less-trained models and embeddings should have resulted in the inconsistent performance. Second, bias in knowledge graph embeddings themselves was not measured during the debiasing method. The debiasing process would have been more convincing and efficient if it was tested if the embeddings are correctly debiased, before incorporating them in further training step of KnowBert. It was a lot of work to train the whole model to check if debiasing the embeddings displays effectiveness in KnowBert. Third, the hyperparameter λ was not tuned. λ might have a larger effect on the debiasing performance than expected. Setting every λ_i to 0.2 might not be a reasonable choice. If there was a way to measure the bias in entity embeddings, hyperparameter tuning could also be done efficiently by quickly evaluating the debiased embeddings. These still indicate the potential of the method for mitigating bias in KnowBert with improved experiment settings and resources.

5.2 Masked Language Model Perplexity

Table 2 is the comparison of masked LM perplexity of original and debiased KnowBert. It shows that debiased KnowBert has worse perplexity than

	Original	Debiased
Perplexity	**8.409**	8.732

Table 2: Masked language model perplexity

the original KnowBert. It may indicate that the entity embeddings have lost semantic information necessary for mask prediction during debiasing. However, the necessity of examining bias in the dataset for evaulation still remains.

6 Conclusion

This paper explored the methods of identifying and mitigating gender bias in knowledge enhanced contextualized word representations. Bias in KnowBert was measured using template-based mask prediction, and was mitigated by removing gender-directional vector component from knowledge graph embeddings used in entity linker. The experiments show that the bias measuring method of BERT can also be applied to KnowBert, and debiasing entity embeddings have potential to debias KnowBert with further research. In the future, I would like to improve this work by experiments of hyperparameter tuning and estimation of bias in knowledge graph embeddings.

References

Sungjin Ahn, Heeyoul Choi, Tanel Pärnamaa, and Yoshua Bengio. 2016. A neural knowledge language model. *arXiv e-prints*, abs/1608.00318.

Mario Arduini, Lorenzo Noci, Federico Pirovano, Ce Zhang, Yash Raj Shrestha, and Bibek Paudel. 2020. Adversarial learning for debiasing knowledge graph embeddings. *CoRR*, abs/2006.16309.

Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. 2019. Tucker: Tensor factorization for knowledge graph completion. *CoRR*, abs/1901.09590.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Styliani Bourli and Evaggelia Pitoura. 2020. Bias in knowledge graph embeddings. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 6–10.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2406–2417. Association for Computational Linguistics.

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

Joseph Fisher, Arpit Mittal, Dave Palfrey, and Christos Christodoulopoulos. 2020. Debiasing knowledge graph embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7332–7345. Association for Computational Linguistics.

Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A. Smith. 2017. Dynamic entity representations in neural language models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1830–1839, Copenhagen, Denmark. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Chandler May, 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832, Melbourne, Australia. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Valentin I. Spitkovsky and Angel X. Chang. 2012. A cross-lingual dictionary for English Wikipedia concepts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3168–3175, Istanbul, Turkey. European Language Resources Association (ELRA).

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. *arXiv preprint arXiv:1804.00079*.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4231–4242, Brussels, Belgium. Association for Computational Linguistics.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *CoRR*, abs/1911.01485.

Bishan Yang and Tom M. Mitchell. 2019. Leveraging knowledge bases in lstms for improving machine reading. *CoRR*, abs/1902.09091.

Zichao Yang, Phil Blunsom, Chris Dyer, and Wang Ling. 2017. Reference-aware language models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1850–1859, Copenhagen, Denmark. Association for Computational Linguistics.