Report on the First Workshop on Bias in Automatic Knowledge Graph Construction at AKBC 2020

Tara Safavi
University of Michigan
tsafavi@umich.edu

Edgar Meij
Bloomberg
emeij@bloomberg.net

Fatma Özcan
Google
fozcan@google.com

Miriam Redi
Wikimedia Foundation
miriam.redi@gmail.com

Gianluca Demartini
University of Queensland
demartini@acm.org

Chenyan Xiong
Microsoft Research
Chenyan.Xiong@microsoft.com

Abstract

We report on the First Workshop on Bias in Automatic Knowledge Graph Construction (KG-BIAS), which was co-located with the Automated Knowledge Base Construction (AKBC) 2020 conference. Identifying and possibly remediating any sort of bias in knowledge graphs, or in the methods used to construct or query them, has clear implications for downstream systems accessing and using the information in such graphs. However, this topic remains relatively unstudied, so our main aim for organizing this workshop was to bring together a group of people from a variety of backgrounds with an interest in the topic, in order to arrive at a shared definition and roadmap for the future. Through a program that included two keynotes, an invited paper, three peer-reviewed full papers, and a plenary discussion, we have made initial inroads towards a common understanding and shared research agenda for this timely and important topic.

1 Introduction

Knowledge graphs (**KGs**) store human knowledge about the world in structured relational form. Because KGs serve as important sources of machine-readable relational knowledge, extensive research efforts have gone into constructing and utilizing knowledge graphs in various areas of artificial intelligence over the past decade [Nickel et al., 2015; Xiong et al., 2017; Dietz et al., 2018; Voskarides et al., 2018; Shinavier et al., 2019; Safavi and Koutra, 2020].

However, relatively little is known about the *biases* contained within and exposed through knowledge graphs, even though acknowledging (and potentially mitigating) these biases is crucial.

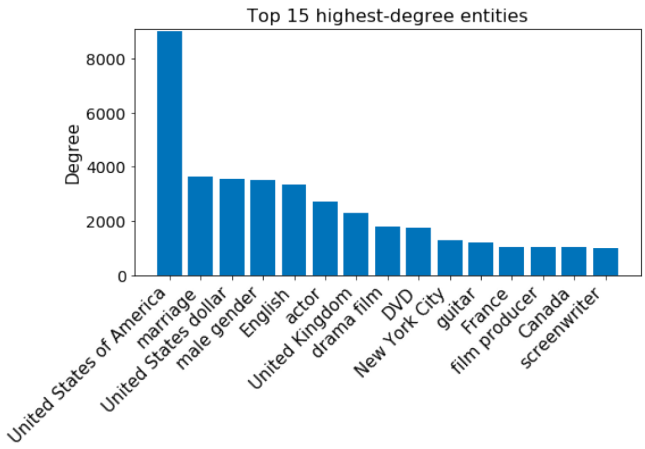

Figure 1: Top 15 highest-degree entities in the FB15K dataset, which is a widely-used benchmark in the knowledge graph construction community. One can observe various biases in the data just from these highly represented entities, i.e., a skew toward males in entertainment from Western English-speaking countries, in particular the USA.

Indeed, because knowledge graphs are often considered as "gold standard" data sources that help to safeguard the correctness of other systems, any biases within KGs are likely to become magnified and spread downstream to other systems that utilize KGs. Such biases may originate in the design of KGs, in the source data from which they are created, and in the algorithms used to sample, aggregate, and process that data. Causes of bias include systematic errors due to selecting non-random items (selection bias), misremembering certain events (recall bias), and interpreting facts in a way that affirms individuals' preconceptions (confirmation bias) [Janowicz et al., 2018; Demartini, 2019].

As an example, Figure 1 demonstrates several types of bias in the open-source FB15K benchmark knowledge graph [Bordes et al., 2013]. We see immediately that Western countries and cultures—in particular the USA and its most populous city, New York City—are highly over-represented, as are males (the entity representing the *male gender* has approximately 3.5× more occurrences in the dataset than the entity representing the *female gender*). We thus expect that AI systems that rely on datasets like FB15K may make skewed, incorrect, or potentially harmful inferences based on such biases, as knowledge graphs capture a notion of what is "important" or "noteworthy" in the world. Indeed, while there have been recent efforts to construct datasets that are less biased [Safavi and Koutra, 2020], the problem is not easily remediated, as the sources from which such datasets are drawn (e.g., Freebase [Bollacker et al., 2008], Wikidata [Vrandečić and Krötzsch, 2014]) are themselves heavily skewed.

In the first workshop on Bias in Automatic Knowledge Graph Construction (**KG-BIAS**), which was held virtually on June 25, 2020 as part of the Automated Knowledge Base Construction conference (**AKBC 2020**), we aimed to address the questions: "How do such biases originate?", "How do we define them and how can we identify them?", and "What is the appropriate way to handle them, if at all?". The goal of our workshop was to start a meaningful, long-lasting dialogue spanning researchers across a wide variety of backgrounds and communities on a timely and important but under-explored topic.

2 Workshop Aims and Scope

The scope of the workshop included the following themes:

1. Ethics, bias, and fairness as they apply to knowledge graphs and/or data sources used to construct knowledge graphs (e.g., text, images)
2. Qualitatively and quantitatively defining types of bias as they relate to knowledge graphs, for example:
    - Implicit or explicit human bias reflected in the data people generate
    - Algorithmic bias represented in learned models or rules
    - Taxonomies and categorizations of different biases
3. Empirically observing biases in KGs
    - Measuring diversity of opinions
    - Language, gender, geography, or interest bias
    - Implications of existing bias to human end-users
    - Benchmarks and datasets for bias in KGs
4. Measuring or remediating bias in KGs
    - De-biased KG construction and completion methods
    - Algorithms for making inferences interpretable and explainable
    - De-biasing or post-processing algorithms
    - Creating user awareness on cognitive biases
    - Ethics of data collection for bias management
    - Diversification of information sources
    - Provenance and traceability

3 Workshop Contributions

The KG-BIAS workshop aimed to start a conversation around a research topic where little to no prior work exists. To this end, we assembled a diverse group of around 30 participants across industry, academia, and nonprofit organizations. In keeping with the interdisciplinary themes of the workshop, we discussed biases in knowledge graphs as they relate to problems and tasks in natural language processing, human computation, information retrieval, and machine learning. In total, the half-day workshop comprised two keynote speakers, one invited paper, and three accepted papers, as well as a plenary discussion in which all participants were invited to contribute.

3.1 Keynotes

Our keynotes were delivered by Jahna Otterbacher from the Open University of Cyprus and Jieyu Zhao from the University of California, Los Angeles.

Bias in Data and Algorithmic Systems: A "Fish-Eye View" of Problems, Solutions and Stakeholders, *Jahna Otterbacher* Our first keynote focused on a cross-institution effort led by Dr. Otterbacher on surveying the state of the art in ethics, bias, and fairness across different communities of computer science (artificial intelligence, information retrieval, recommender systems, human-computer interaction, human computation, and fairness and transparency). Otterbacher emphasized that cross-communication is necessary for solving problems related to fairness because there are many stakeholders involved, from scientists and engineers to end-users and consumers to observers and regulators. However, she also argued that engineers should be the most involved with creating fair systems because they (uniquely) have direct access to such systems. Otterbacher categorized articles within the considered domains with respect to (1) problem types; (2) affected attributes, for example demographic attributes like race and gender, or information attributes like explainability and coverage; and (3) solutions such as bias detection and fairness management. She also highlighted two of her own recent research directions within the survey: Auditing image-tagging systems as well as investigating users' perceptions of fairness of image descriptions [Kyriakou et al., 2019; Barlas et al., 2019], both of which have downstream implications in knowledge graphs that deal with multimodal information such as Wikidata. In the first work, the main research finding is that proprietary image taggers do violate notions of group fairness, for example the rate of misgendering across different demographic groups. In the second work, participants rated human-generated tag sets for images as more "fair" than automatically generated tag sets, suggesting that humans tend to perceive automated image-tagging systems as biased.

Detecting and Mitigating Gender Bias in NLP, *Jieyu Zhao* Our second keynote focused on various methodologies for detecting and remediating gender bias in natural language processing tasks such as coreference resolution and generating text embeddings. Zhao began with illustrative examples where existing machine learning systems fail when it comes to gender-related NLP tasks. For example, in visual semantic role labeling, existing systems may label a man in the kitchen as a woman. She then outlined her research directions toward mitigating gender bias in NLP, starting with coreference resolution. To this end, she and her collaborators proposed the WinoBias dataset, and evaluated feature extraction and deep learning-based models on it. Showing that such models suffer from social biases in coreference resolution—for example, such a system may change an English pronoun from "her" to "his" when referring to a female in a stereotypically male profession, like a president or a lawyer—Zhao and her colleagues proposed solutions such as training models on gender-swapped datasets [Zhao et al., 2018a]. She then turned to bias in language representations and word embeddings, covering a GLoVE-based model that learns embeddings without gendered information, new resources for understanding bias in multilingual word embeddings, and gender biases in deep contextualized language models, using ELMo as a case study [Zhao et al., 2017, 2018b, 2019]. Finally, she covered how biases in text corpora are amplified by machine learning training and evaluation protocols, and explored methods

for reducing the bias within models themselves via posterior regularization.

3.2 Invited paper

We invited the paper "Measuring Social Bias in Knowledge Graph Embeddings" by Joseph Fisher, Dave Palfrey, Arpit Mittal, and Christos Christodoulopoulos; note that this work is now published as part of the 2020 Empirical Methods in Natural Language Processing (EMNLP) conference [Fisher et al., 2020]. This work, which was inspired by similar work in investigating biases in word embeddings [Bolukbasi et al., 2016], presents the first study on social biases captured by knowledge graph embeddings. Corroborating our findings in Figure 1, Fisher and his colleagues show that existing knowledge graphs like Wikidata are heavily biased toward the male gender, Western countries, and the Catholic Church. They then propose a metric to quantify social bias in knowledge graph embeddings with respect to attributes that are captured by relations, for example *profession* or *ethnicity*. To compute their metric, they take a pre-trained knowledge graph embedding and calculate an update to the embedding that increases the model's (probabilistic) prediction that the embedding has sensitive attribute a, while decreasing the model's prediction that the embedding has sensitive attribute b. They then analyze the change in the model's probabilistic predictions with respect to the relation in question after changing the embedding in the direction of attribute a. As an illustrative example, one could update an entity's embedding to be more "male" (i.e., make the model predict that the entity is male with greater probability), and then analyze the change in the model's prediction scores with regard to the *profession* relation, thus asking the question: "Does making the entity more *male* increase the model's prediction that the entity has a stereotypically male *profession*?" They show that their metric does indeed uncover social biases and stereotypes, for example that men are more likely to be bankers whereas women are more likely to be homemakers. They conclude that because the underlying data themselves are biased, care must be taken to appropriately handle biases when using knowledge graph embeddings.

3.3 Accepted papers

We solicited regular paper submissions, position papers, and demo papers of existing systems or frameworks. Submissions were peer-reviewed, and the workshop proceedings are available on arXiv [Meij et al., 2020]. We received and accepted three submissions on (1) fairness and transparency in knowledge graph construction, (2) demographic bias in named entity recognition systems, and (3) the pitfalls of personalization in conversational search.

- **From Knowledge Graphs to Knowledge Practices: On the Need for Transparency and Explainability in Enterprise Knowledge Graph Applications** by Christine Wolf. Abstract: *Transparency and explainability are important concerns in building and implementing cognitive technologies, especially those intended to augment and transform domain knowledge practices. Knowledge graphs are powerful component technologies leveraged in a number of business applications, yet relevant information about their construction, refinement, and maintenance are often not visible in resulting end-use applications. Information about a graph's training set(s), for example, can be useful to understand how it may represent incomplete coverage of a domain area; information about the strength or confidence*

of a given association, as another example, can also help signal the need for further action or investigation. In this position paper, we explore these issues and discuss an enterprise domain use case involving sales operations. Increasing transparency and explainability in knowledge graphs can help business users better assess potential biases and appropriately integrate graph outputs into their workplace knowledge practices.

- **Assessing Demographic Bias in Named Entity Recognition** by Shubhanshu Mishra, Sijun He, and Luca Belli. Abstract: *Named Entity Recognition (NER) is often the first step towards automated Knowledge Base (KB) generation from raw text. In this work, we assess the bias in various Named Entity Recognition (NER) systems for English across different demographic groups with synthetically generated corpora. Our analysis reveals that models perform better at identifying names from specific demographic groups across two datasets. We also identify that debiased embeddings do not help in resolving this issue. Finally, we observe that character-based contextualized word representation models such as ELMo results in the least bias across demographics. Our work can shed light on potential biases in automated KB generation due to systematic exclusion of named entities belonging to certain demographics.*

- **Bias in Conversational Search: The Double-Edged Sword of the Personalized Knowledge Graph** by Emma Gerritse, Faegheh Hasibi, and Arjen P. de Vries. Abstract: *Conversational AI systems are being used in personal devices, providing users with highly personalized content. Personalized knowledge graphs (PKGs) are one of the recently proposed methods to store users' information in a structured form and tailor answers to their liking. Personalization, however, is prone to amplifying bias and contributing to the echo-chamber phenomenon. In this paper, we discuss different types of biases in conversational search systems, with the emphasis on the biases that are related to PKGs. We review existing definitions of bias in the literature: people bias, algorithm bias, and a combination of the two, and further propose different strategies for tackling these biases for conversational search systems. Finally, we discuss methods for measuring bias and evaluating user satisfaction.*

4 Discussion

Our keynotes, invited paper, and accepted papers all touched on important interdisciplinary aspects of bias in knowledge graph construction. From the **human-computer interaction** and **human computation** aspects, our first keynote emphasized the importance of understanding how users perceive fairness in artificial intelligence systems. Wolf [2020] also emphasized the need for transparency and explainability in such systems from a knowledge graph-specific perspective, arguing that knowledge graphs are most useful to practitioners when they are *trustworthy*, which often translates to making metadata such as data sources, confidence levels, etc available. Indeed, in our plenary discussion, we agreed that while some kinds of bias are unavoidable in knowledge graphs or even in datasets more generally, the main types of bias that we should focus on mitigating are those that have downstream effects for *people*.

From the **natural language processing** side, our second keynote covered a wide variety of gender analysis and debiasing techniques for important NLP tasks like coreference resolution and

language representations, all of which may impact the way texts are processed in the knowledge graph construction pipeline. By contrast, Mishra et al. [2020] focused on a specific aspect of NLP that is highly related to knowledge graph construction: Named Entity Recognition or NER. In their work they showed that NER systems in English perform at different levels of accuracy for names that are thought to represent different demographic groups stratified with respect to gender and race (i.e., white males versus Hispanic females). They show that while more recent language models that rely on character-level embeddings result in the least bias, there is still plenty of room for improvement, and point out that one reason for the lack of representation of certain demographic groups in KGs may be due to exclusion by entity recognition systems. During the plenary session, the authors emphasized their choice of studying intersectional biases (i.e., among combinations of attributes), and also acknowledged that there are other attributes beyond race and gender that require more study.

From the **information retrieval** side, Gerritse et al. [2020] highlighted the benefits and drawbacks of personalization in knowledge graphs as it relates to bias. On one hand, they argue, personalization can provide help users better achieve their information needs and goals. On the other hand, personalization is inherently a type of bias, and can therefore lead to "filter bubble" effects that can be harmful to users in the long run. They present various strategies for mitigating biases introduced by personalization in knowledge graphs, and discuss the trade-offs for each strategy. In our plenary session the authors also discussed how conversational search systems that rely on personalized knowledge graphs play a unique role in mitigating biases because they have a strong interactive component. Here, we agreed that one potential bias mitigation approach taken by personalized conversational agents could be to (gently) call out potential biases and explain them to the user in order to help the user make more informed choices.

Finally, from the **machine learning** side, Fisher et al. [2020] studied social biases in knowledge graph embeddings to answer important open research questions around how to measure and quantify biases in those embeddings, and showed that biases in knowledge graphs are indeed captured by embeddings. Again, in the plenary session the authors agreed that such biases are harmful mainly because of their potential downstream effects. The next step, they argued, is to more thoroughly measure how biases in knowledge graph embeddings spread throughout other AI systems.

5 Conclusion

The KG-BIAS 2020 workshop was a successful event that kickstarted an important cross-community conversation on biases in knowledge graphs and the tasks that relate to knowledge graph construction and usage. We received input from researchers and practitioners focused on various aspects of fairness, including human-computer interaction and human computation, natural language processing, information retrieval, and machine learning. We thank AKBC 2020 for their support in organizing the event, especially given the last-minute switch to a virtual format due to the COVID-19 pandemic. We hope to expand upon the workshop in the following years.

References

Pinar Barlas, Styliani Kleanthous, Kyriakos Kyriakou, and Jahna Otterbacher. What makes an image tagger fair? In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pages 95–103, 2019.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013.

Gianluca Demartini. Implicit bias in crowdsourced knowledge graphs. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 624–630, 2019.

Laura Dietz, Alexander Kotov, and Edgar Meij. Utilizing knowledge graphs for text-centric information retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1387–1390, 2018.

Joseph Fisher, Arpit Mittal, Dave Palfrey, and Christos Christodoulopoulos. Debiasing knowledge graph embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7332–7345. Association for Computational Linguistics, November 2020. URL https://www.aclweb.org/anthology/2020.emnlp-main.595.

Emma J Gerritse, Faegheh Hasibi, and Arjen P de Vries. Bias in conversational search: The double-edged sword of the personalized knowledge graph. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 133–136, 2020.

Krzysztof Janowicz, Bo Yan, Blake Regalia, Rui Zhu, and Gengchen Mai. Debiasing knowledge graphs: Why female presidents are not like female popes. In *International Semantic Web Conference (P&D/Industry/BlueSky)*, 2018.

Kyriakos Kyriakou, Pınar Barlas, Styliani Kleanthous, and Jahna Otterbacher. Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 313–322, 2019.

Edgar Meij, Tara Safavi, Chenyan Xiong, Gianluca Demartini, Miriam Redi, and Fatma Özcan. Proceedings of the KG-BIAS Workshop 2020 at AKBC 2020, 2020. URL http://arxiv.org/abs/2007.11659.

Shubhanshu Mishra, Sijun He, and Luca Belli. Assessing demographic bias in named entity recognition, 2020. URL http://arxiv.org/abs/2008.03415.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2015.

Tara Safavi and Danai Koutra. Codex: A comprehensive knowledge graph completion benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Joshua Shinavier, Kim Branson, Wei Zhang, Shima Dastgheib, Yuqing Gao, Bogdan Arsintescu, Fatma Özcan, and Edgar Meij. Panel: Knowledge graph industry applications. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 676–676, 2019.

Nikos Voskarides, Edgar Meij, Ridho Reinanda, Abhinav Khaitan, Miles Osborne, Giorgio Stefanoni, Prabhanjan Kambadur, and Maarten de Rijke. Weakly-supervised contextualization of knowledge graph facts. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 765–774, 2018.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Christine Wolf. From knowledge graphs to knowledge practices: On the need for transparency and explainability in enterprise knowledge graph applications, 2020.

Chenyan Xiong, Russell Power, and Jamie Callan. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1271–1279, 2017.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. 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 15–20, 2018a.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, 2018b.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 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, 2019.