Take-away from the tutorial. Attendees will gain a technical understanding and visual intuition of how debiasing techniques for word embedding work, focusing on their geometric properties. They will understand the differences and similarities between such techniques contextualized using popular examples, evaluation metrics, as well as their critiques. Finally, we hope that attendees will not only understand the technical challenges in this important area of research, but also gain experience with a visual tool that will guide their own exploration of biases and debiasing of word embeddings.

2 TUTORIAL OUTLINE

2.1 Introduction (30 minutes)

We will begin with a brief introduction of the concept of word embeddings, their computation and structure. We will motivate why understanding the bias encoded in these embeddings is important using examples of biases in embeddings. The notions of bias in word embeddings are varied, and as a result, some aspects have become contentious [1]. At this point, we will specifically address what we mean by "bias" for the purpose of this tutorial. And we will explain, and demonstrate how it manifests in word embeddings.

2.2 Debiasing word embeddings (30 minutes)

The audience will be introduced to the concept of debiasing through various methods and evaluation metrics. We will focus on methods that work on the native geometry of the word embedding space—these methods admit visual exploration, generalize to embeddings of structures other than just words, and do not require expensive retraining of the complex embedding methods. We will first visually demonstrate the concept of feature subspaces and ways to determine them, which is an important step towards understanding and debiasing word representations.

We will include the following methods of debiasing:

- Hard Debiasing [2]
- Linear Projection [5]
- Iterative Nullspace Projection [8]
- Orthogonal Subspace Correction [4]

We will also visually interpret a common method (e.g., [3]) of quantifying different types of bias based on word associations.

2.3 Overview of interactive tool (15 minutes)

We will go through the layout of the visualization tool for the exploration of debiasing techniques. We will explain various components of the tool and familiarize the audience with the various algorithmic, design and visualization choices used in the tool.

During this time, we will walk through a simple auto-install of the tool. We will also have a fallback server available that may be able to support at about 20-30 users.

2.4 Worked examples of bias and how they are mitigated and measured (45 minutes)

After an initial walk-through of all methods on a couple of standard examples (e.g., gender/occupations), we will broaden the application, and show how they work on less examined types of data and concepts (e.g., age, ethnicity). We will illustrate instances of examples of bias in word embeddings and how various techniques effect the bias. These will include many ways to examine and analyze the gender subspace, but also other biases that may arise around age, nationality, and ethnicity as encoded through words with definitions related to those concepts.

2.5 Critiques of Debiasing (30 minutes)

These debiasing techniques are not without drawbacks and tradeoffs. We will explore through various examples and analysis, the question of whether these methods are meaningfully eliminating bias or only optimizing some evaluation metric. Is (and when is) residual bias [6] retained through these methods? Is this unavoidable, or easily solvable? Are analogy-based evaluations inherently flawed, or a feasible part of the puzzle, provided a more careful standard is set? What about understanding biases expressed extrinsically in tasks or by contextual embeddings? We will examine these questions through examples in our visual system, rebuffetted with some large scale experimental numerical results. We do not plan to provide an explicit point of view to these controversial topics of active discussion, but present the audience with visual intuition and data to lead to an informed decision on their own.

2.6 Looking ahead and discussion (30 minutes)

We will conclude the tutorial by looking ahead at future directions for measuring and mitigating bias in NLP models. These include extensions to contextual embeddings, what sources of biases are captured this way and which are not, where and where-not to use these attenuation tools, how to handle multiple sources and forms of biases simultaneously. We aim to conclude with a lively discussion on this perhaps controversial topic, but posing some emerging challenges, and opportunities. [7] describes the design of the tool in detail.

REFERENCES

[1] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. arXiv:2005.14050 [cs.CL]

[2] T Bolukbasi, K W Chang, J Zou, V Saligrama, and A Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *ACM Transactions of Information Systems*.

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

[4] Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2020. OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings. arXiv:2007.00049 [cs.CL]

[5] Sunipa Dev and Jeff M. Phillips. 2019. Attenuating Bias in Word vectors. In *AISTATS (Proceedings of Machine Learning Research)*. PMLR, 879–887.

[6] 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 2019*. 609–614.

[7] Archit Rathore, Sunipa Dev, Jeff M. Phillips, Vivek Srikumar, Yan Zheng, Chin-Chia Michael Yeh, Junpeng Wang, Wei Zhang, and Bei Wang. 2021. VERB: Visualizing and Interpreting Bias Mitigation Techniques for Word Representations. arXiv:2104.02797 [cs.CL]

[8] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. *ACL 2020* (2020). arXiv:2004.07667 [cs.CL]