

Investigating the Frequency Distortion of Word Embeddings and its Impact on Bias Metrics

Francisco Valentini^{1,2} Juan Cruz Sosa³ Diego Slezak^{1,3} Edgar Altszyler^{1,2}

¹ Instituto de Investigación en Ciencias de la Computación, CONICET-UBA, Argentina

² Maestría en Data Mining, Universidad de Buenos Aires (UBA), Argentina

³ Departamento de Computación, FCEyN, UBA, Argentina

fvalentini@dc.uba.ar, juan.cruz.sosa.92@gmail.com,

dfslezak@dc.uba.ar, ealtszyler@dc.uba.ar

Abstract. Recent research has shown that static word embeddings can encode words' frequencies. However, little has been studied about this behavior. In the present work, we study how frequency and semantic similarity relate to one another in static word embeddings, and we assess the impact of this relationship on embedding-based bias metrics. We find that Skip-gram, GloVe and FastText embeddings tend to produce higher similarity between high-frequency words than between other frequency combinations. We show that the association between frequency and similarity also appears when words are randomly shuffled, and holds for different hyperparameter settings. This proves that the patterns we find are neither due to real semantic associations nor to specific parameters choices, and are an artifact produced by the word embeddings. To illustrate how frequencies can affect the measurement of biases related to gender, ethnicity, and affluence, we carry out a controlled experiment that shows that biases can even change sign or reverse their order when word frequencies change.

Keywords: natural language processing, word embedding, bias.

Investigando la Distorsión de Frecuencia en Word Embeddings y su Impacto en Métricas de Sesgo

Resumen. Investigaciones recientes han demostrado que los word embeddings estáticos pueden codificar información sobre la frecuencia de las palabras. Sin embargo, poco se ha estudiado sobre este fenómeno y sus efectos en tareas posteriores. En este trabajo estudiamos sistemáticamente la asociación entre frecuencia y similitud semántica en varios word embeddings estáticos. Descubrimos que los embeddings Skip-gram, GloVe y FastText tienden a producir una mayor similitud semántica entre palabras de alta frecuencia que entre otras combinaciones de frecuencias. Demostramos que la asociación entre frecuencia y similitud también aparece cuando las palabras se mezclan aleatoriamente. Esto demuestra que los patrones encontrados no se deben a asociaciones semánti-

cas reales presentes en los textos, sino que son un artefacto producido por los embeddings. Por último, ofrecemos un ejemplo de cómo la frecuencia de las palabras puede influir mucho en la medición del sesgo de género con métricas basadas en embeddings. En particular, llevamos a cabo un experimento controlado que demuestra que los sesgos pueden incluso cambiar de signo o invertir su orden manipulando las frecuencias de las palabras.

Palabras clave: procesamiento de lenguaje, word embedding, sesgo.

Article “Journal First”. Francisco Valentini, Juan Sosa, Diego Slezak, and Edgar Altszyler. 2023. Investigating the Frequency Distortion of Word Embeddings and Its Impact on Bias Metrics. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 113–126, Singapore. Association for Computational Linguistics.