TOK Exhibition: Knowledge & Technology

Prompt 11: Is bias inevitable in the production of knowledge?

Object 1. BERT - language model

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

Language model pre-training has been shown to be effective for improving many natural language

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such

My first object is a language representation model I plan to use for my Extended Essay called BERT. BERT captures meanings of and relationships between words. Thanks to this capability, BERT can be used for obtaining knowledge in a variety of tasks that involve language, including but not limited to: resume scanning, categorizing articles, social media recommendation, and essay grading. However, it has been reported that the output of BERT is biased.

BERT contributes to the exhibition because it shows that human biases are inevitable in the production of knowledge. To learn meanings of words, BERT has undergone so-called pre-training which is the process of feeding large datasets composed of real-life texts to the model. However, since these training datasets were human-generated (originating from Wikipedia and BookCorpus), they contain human biases, such as ableism, racism, and sexism. Since the training datasets were biased, BERT is biased as well. For instance, by reason of the patriarchal values that have existed in many countries around the world for generations, men, rather than women, have been encouraged to pursue engineering-related careers. Therefore, the majority of the content available on the Internet on this matter links these kinds of jobs to males. As a result, when used for analyzing engineering job resumes, BERT will reflect this linguistic injustice by discrediting women’s applications.

Thus, BERT explores the exhibition's prompt by demonstrating that human biases are inescapable in the production of knowledge. The robustness of the knowledge that BERT produces is severely limited by the considerable amount of bias in the training datasets. Moreover, the situation is unlikely to change because no evaluation metrics for assessing bias in the datasets have been established due to the difficulty of choosing satisfying criteria. This goes to show that we cannot avoid bias in digital tools (such as BERT) which we use to produce knowledge.

Object 2. Braun Pulse Oximeter 1 YK-81CEU

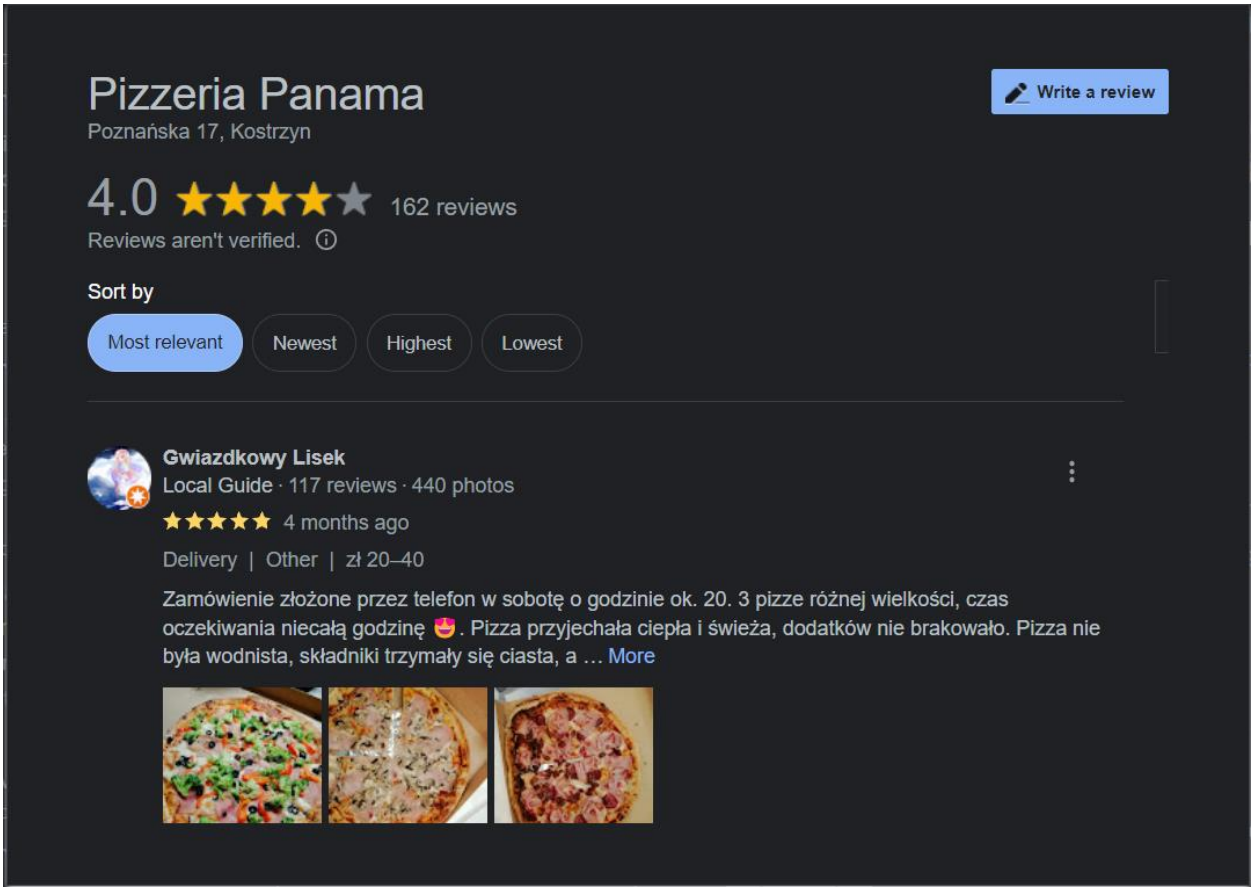
My second object is Braun Pulse Oximeter 1 YK-81CEU available from my local store. This pulse oximeter is a compact equipment that monitors oxygen saturation of the patient's blood. Because of these characteristics, the pulse oximeter is a widely-used analog tool in healthcare as it produces in-time knowledge about the patient's condition without much overhead. However, it has been shown that the readings of the pulse oximeter are racially biased.

The Braun Pulse Oximeter is relevant to the exhibition because it demonstrates that racial bias is unavoidable in the process of creating knowledge. On one end, the pulse oximeter emits a continuous beam of cold light that goes through the patient's finger and is caught by the photodetector on the other end, which measures blood oxygen level. However, some of this light is absorbed by the finger mainly owing to the patient's skin pigment. For example, hypoxemia is a condition where oxygen level in blood drops below 80% (normal range is between 95% and 100%). A recent study has suggested that "Black patients had nearly three times the frequency of occult hypoxemia that was not detected by pulse oximetry as White patients" (Sjoding et al., 2020). This is because Black patients have more melanin which is a

skin pigment that enhances the absorption of light and interferes with a pulse oximeter's readings.

Therefore, the Braun Pulse Oximeter 1 YK-81CEU enriches the exhibition because it shows that racial bias is inevitable in the production of knowledge. Due to outdated evaluation processes, the pulse oximeter has not been tested on a diverse group, which would allow for a more accurate examination of a wider range of patients with different skin colors rather than only those with a white skin color. Accordingly, knowledge the oximeter produces is racially biased as it performs better for White patients.

Object 3. Google reviews system for Pizzeria Panama restaurant



My third object is the Google reviews system for Pizzeria Panama restaurant located a few kilometers away from my home. Google reviews is a rating system that allows users to submit their reviews of a business approved by Google in the form of stars on the scale of 1 to 5 (where 5 is the best) with a short comment. Through the Google reviews, users can produce knowledge about the quality of food in Pizzeria Panama so that others can decide if they want to eat there or not. However, the vast majority of reviews come from extreme ends - they give either 1 star or 5 stars. This is known as a volunteer bias.

The Google reviews system for Pizzeria Panama restaurant is highly relevant to this exhibition because it illustrates that it is virtually impossible to generate unbiased knowledge about the tastiness of food in restaurants. As mentioned, there is the prevalence of extremely negative and extremely positive responses. The reason is that, usually, only people who are very satisfied or very dissatisfied with the meal would be motivated enough to make an effort and leave their reviews. As a

consequence, we do not receive reliable data about the whole population (all patrons who visited Panama), but about the two specific groups (lovers and haters of food there), which constitutes the aforementioned volunteer bias.

Hence, the Google reviews system for Pizzeria Panama restaurant brings value to the exhibition since it provides an excellent example that volunteer bias does appear in rating systems which produce knowledge about the possible deliciousness of food in restaurants. Such systems rely heavily on volunteer participants who are, unfortunately, very frequently driven by strong emotions such as high satisfaction or great disappointment, which makes it difficult to obtain a reliable picture of the actual situation (in this case, the quality of food in Pizzeria Panama).

Word count: 920

References

Add, edit, or delete Google Maps reviews & ratings. (n.d). *Google Maps Help*. Retrieved May 4, 2023 from: https://shorturl.at/mtuV8

Dodge, H.H., Katsumata, Y., & Zhu, J. (2014). Characteristics associated with willingness to participate in a randomized controlled behavioral clinical trial using home-based personal computers and a webcam. *Trials*. 15, 508. Retrieved May 4, 2023 from: https://doi.org/10.1186/1745-6215-15-508

Jentzsch, S., & Turan, C. (2022). Gender Bias in BERT - Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Seattle, Washington, 184–199. Association for Computational Linguistics. Retrieved May 4, 2023 from: https://aclanthology.org/2022.gebnlp-1.20/

Jubran A. (1999). Pulse oximetry. *Critical care (London, England)*, 11-17. Retrieved May 4, 2023 from: https://doi.org/10.1186/cc341

Media Expert. (n.d). Pulsoksymetr BRAUN 1 YK-81CEU Certyfikat medyczny. Retrieved May 4, 2023 from: https://www.mediaexpert.pl/agd-male/zdrowie/pulsoksymetry/pulsoksymetr-braun-1-yk-81ceu-certyfikat-medyczny

Medinaz. (2021). *Pulse oximeter: How it works and Interpretation II Pulse oximeter mechanism* [Video]. YouTube. https://www.youtube.com/watch?v=N3MCvREORVA

Pulse oximetry. (2023). *American Lung Association*. Retrieved May 4, 2023 from https://shorturl.at/inFK0

Schlessinger D. I., Anoruo M. D., & Schlessinger J. (2023). Biochemistry, Melanin. StatPearls. *Treasure Island (FL)*: *StatPearls*. Retrieved May 4, 2023 from: https://www.ncbi.nlm.nih.gov/books/NBK459156/

Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E., & Valley T. S. (2020). Racial Bias in Pulse Oximetry Measurement. *New England Journal of Medicine*. Retrieved May 4, 2023 from: https://www.nejm.org/doi/https://doi.org/10.1056/NEJMc2029240