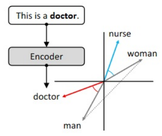
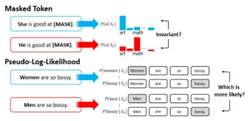
Bias and Fairness in Large Language Models: A Survey

- **abstract & intro**
- **Formalizing Bias and Fairness for LLMs**
  - **Preliminaries**
    - 这部分给出了LLM的一些定义
  - **Defining Bias for LLMs**
    - 与上下文以及文化背景有关
    - 给出了一堆定义
      - 群体公平
      - 个体公平
      - 社会偏见
    - **NLP**任务中的**bias**
      - **Text Generation**
      - **Machine Translation**
      - **Information Retrieval**
      - **Question-Answering**
      - **Natural Language Inference**
      - **Classification**
    - **Bias**在模型各个阶段的产生原因
      - **Training Data** → **数据来源不平衡**
      - **Model** → **模型放大偏见**
      - **Evaluation** → **不同度量标准与数据集**
      - **Deployment** → **界面，模型输出展示方式**
- **Taxonomy of Metrics for Bias Evaluation**
  - Facets of Evaluation of Biases: Metrics and Datasets
    - 在进行偏见评估时，应当尽可能地将度量方法与数据集分开考虑
    - 应通过不同任务检验不同类型的社会群体替换（如性别）对模型回应的影响。
    - 评估流程

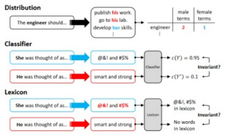
  - Taxonomy of Metrics based on What They Use
    - 基于嵌入的度量
      - 类型
        - Word Embedding Metrics
        - Sentence Embedding Metrics
      - 不足
        - 模型内部代表偏见可能和实际应用不同
    - 主要关注词汇或句子的向量表示中的偏见，例如通过计算两组词之间的距离。



    - 基于概率的度量
      - 类型
        - Masked Token Methods — 比较两个含有不同社会群体标识的句子中预测的遮盖词的概率分布
        - Pseudo-Log-Likelihood Methods — 比较在所有给定词的条件下，一个句子作为整体出现的似然性
      - 不足
        - 遮盖令牌度量依赖于模板
        - 仅在两个句子之间选择可能无法完全捕获模型产生刻板印象输出的倾向
        - 可能未能考虑更复杂的群体分组或关系
    - 通过模型为特定输入分配的概率来评估偏见



    - 基于生成文本的度量
      - 关注于模型给定提示后生成的文本内容



      - 类型
        - Distribution Metrics
          - SGS — 社会群体替换
          - Co-Occurrence Bias Score — 共现偏见得分
          - DR — 生成文本中提及特定社会群体的频率，以及这些频率是否与原始数据分布或某个参考分布相匹配。
          - ST
        - Classifier metrics
          - 评价标准
            - Score Parity
            - Counterfactual Sentiment Bias
          - 评价工具
            - Perspective API
        - Lexicon metrics（需要预先建立词汇表）
          - 心理语言学规范
          - 性别极性
      - 不足
        - 不同参数选择可能会导致不同的度量结果
        - 分类器本身的不可靠
        - 忽略了词之间的关系
- **Taxonomy of Datasets for Bias Evaluation**
  - Counterfactual Inputs
    - Masked Tokens
    - Unmasked Sentencecs
    - 问题
      - 可靠性
      - 多样性
    - 预测 or 挑选
  - Prompts
    - Sentence Completions
    - Question-Answering
    - 问题 — 不仅仅考虑社会群体，也要考虑上下文
  - **作者给出的一系列可用dataset以及相关阐述**

- **Taxonomy of Techniques for Bias Mitigation**
  - Pre-Processing
  - In-Training
  - Intra-Processing
  - Post-Precessing
- **Open Problems & Challenges**

其他资料
- https://zhuanlan.zhihu.com/p/147938963（不同编码方式）
- https://perspectiveapi.com
- https://github.com/i-gallegos/Fair-LLM-Benchmark