Пробинг: локализация грамматики в нейронных сетях

Article plan

Корнилов Альберт, Степанова Ангелина, Сухарева Мария, Шумакова Лада

Постановка проблемы

Проблема 1:

- а вдруг полиперсональные токены воспринимались как ошибки? или ничего на самом деле не локализовалось? → проверить это через пробинг; четыре пути:
  - Маскирование токенов объектов и восстановление этих токенов с помощью обученных моделей;
  - Анализ матриц внимания (где применимо);

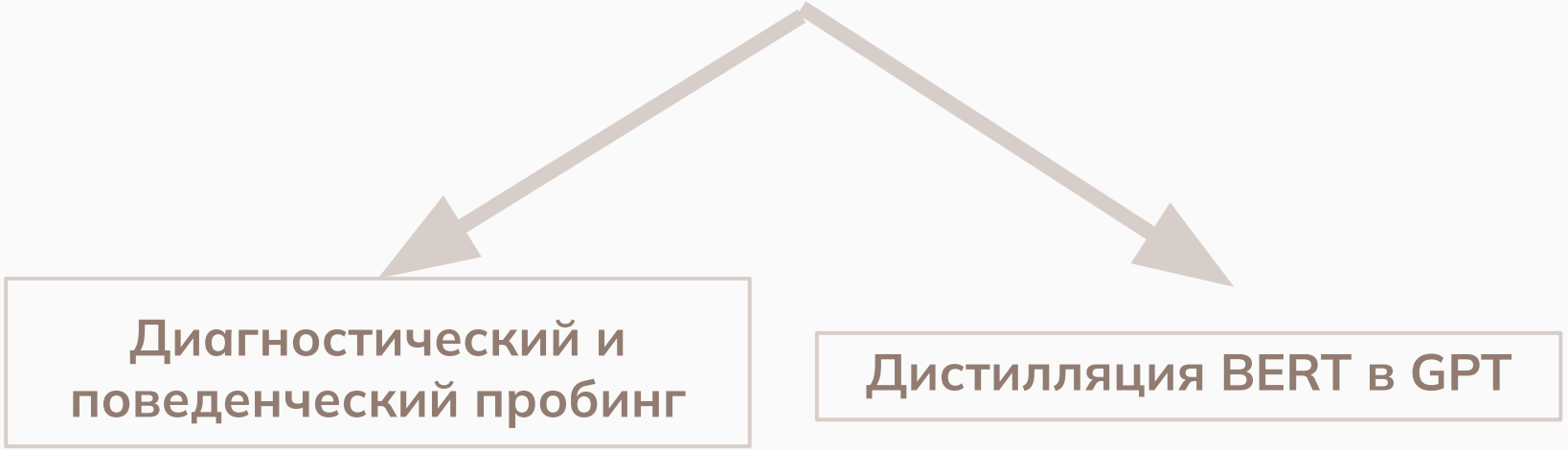
Постановка проблемы

Проблема 1:

- а вдруг полиперсональные токены воспринимались как ошибки? или ничего на самом деле не локализовалось?→ проверить это через пробинг; четыре пути:
  - Диагностический пробинг с использованием классификаторов;
  - Поведенческий пробинг → генерация текста →

    Проблема 2: make BERT генерировать текст:
    - Дистилляция

Наш план: две статьи и два направления работы



Abstract: Diagnostic probing

Deep learning models, despite their success, face criticism as "black boxes," raising reliability concerns in high-stakes domains. Interpretability methods, like probing, aim to uncover model decision-making, particularly in Large Language Models (LLMs). Recent work (Kudryashov et al., 2024) localized polypersonal agreement in Russian BERT, but it remains unclear if the model truly learned this feature or treated it as noise. We investigate whether BERT processes polypersonal tokens as grammatical markers or input errors, combining error generation, attention map analysis, and linear probing. Our approach includes: (1) developing an error generator, (2) training classifiers to distinguish errors from polypersonal tokens, and (3) evaluating BERT's representations. Findings will clarify LLM capabilities and improve probing methodologies for linguistic features.

Обзор✓

Ключевые работы:

- Tenney et al. (2020)
  - Диагностический классификатор, обученный на внутренних представлениях BERT, может выявлять лингвистические признаки
- Ravichander et al. (2021)
  - LLM часто полагаются на корреляции в данных, а не на абстрактные грамматические правила
- Clark et al. (2019)
  - Внимание BERT может указывать на значимые токены (например, полиперсональные аффиксы), но это не гарантирует их осмысленной обработки

Методы

Сделано:

- Пробинг с маскированными датасетами и с аттеншенами на моделях прошлого года (где обучался только один слой)
- Дообучение всего BERT
- Пробинг с маскированными датасетами и с аттеншенами на дообученном BERT
- Генератор ошибок

Сделать:

- Обучение двух классификаторов
- Поведенческий пробинг (после дистилляции)

Метрики

- **Пробинг с маскированными объектами:** доля предсказанных объектов в верной форме
- **Пробинг с аттеншенами на моделях:** статистически значимое отличие паттернов внимания у моделей с локализованной полиперсональностью
- **Диагностический пробинг с использованием классификаторов:** линейные классификаторы с высокой точностью отличают вектора предложений с ошибками от векторов предложений с полиперсональностью

Распределение ответственности

Мария:

- Дообучение всего BERT ✓
- Разработка пайплайна дистилляции

Лада:

- Литературный обзор ✓
- Разработка генератора ошибок ✓
- Обучение двух классификаторов:
  - классификация на тексты с ошибками и тексты без ошибок
  - классификация на полиперсональные и ошибочные токены

Распределение ответственности

Альберт:

- Оптимизация скрипта с удалением прямого объекта с помощью spaCy (3 часа -> 3.5 мин на 200,000 предложений) ✓
- Скрипт, который собирает скоры attention от маскированного токена к полиперсональным ✓
- Подбор метрик для дистилляции ✓
- Литературный обзор ✓

Ангелина:

- Разработка пайплайна пробинга с маскированными объектами ✓
- Написание класса для посчёта лосса при дистилляции
- Написание классов для оценки качества дистилляции

Конференции

- Workshop on Actionable Interpretability (2026)
- ICLR 2026

Thank you for your aTtEnTiOn!