Министерство науки и высшего образования Российской Федерации
Московский физико-технический институт
(национальный исследовательский университет)
Физтех-школа прикладной математики и информатики
Магистерская программа
«Методы и технологии искусственного интеллекта»

Туеста Льаха Хуан Карлос

Интерпретация результатов дообучения языковых моделей (Трансформеров)

09.04.01 — Информатика и вычислительная техника

Магистерская диссертация

Научный руководитель:
к.п.н.
Ч доваН.В.
Конс льтант: к.ф.-м.н.
Девяткин д.А.

Москва
2024 г.

Содержание

1 Введение **4**

2 Обзорно-постановочный раздел работы **4**

2.1 Модели языка и их применение в NLP 4

2.2 Техники генерации парафраз 5

2.3 Объяснимость больших языковых моделей (LLMs) и оценка точности . 6

2.4 Синтаксическая сондировка и синтаксис в представлениях слов 7

2.5 Задачи и возможности в генерации парафраз 8

2.6 Заключение теоретической части 8

Аннотация

Модели большого языка (LLM), представленные через архитектуру Transformer в статье "Attention Is All You Need"(Vaswani et al., 2017), произвели революцию в области обработки естественного языка (NLP), демонстрируя передовые результаты в машинном переводе, генерации текста и вопросно-ответных системах. Несмотря на их успех, "черный ящик"этих моделей представляет собой значительные вызовы в плане надежности, обнаружения предвзятости и объяснимости, особенно в высокорисковых областях, таких как здравоохранение и финансы.

В данной диссертации рассматривается интерпретируемость LLM, которые были дообучены или адаптированы через подсказки. Дообучение, включающее обучение на специфических задачах, и подсказки, направляющие поведение модели с помощью тщательно разработанных входных данных, являются мощными методами для настройки LLM. Однако понимание их внутренней работы остается важным для обеспечения надежности и доверия к их предсказаниям.

Исследование изучает три основные подхода к интерпретируемости для дообученных LLM: классификаторы-зондирование, анализ активации нейронов и объяснения, основанные на концептах. Для моделей с подсказками исследуются такие методы, как анализ обучения в контексте, цепочка рассуждений и инженерия представлений. Для оценки качества этих объяснений используются такие метрики, как правдоподобие и верность.

Результаты показывают, что методы интерпретируемости предоставляют значимые инсайты в поведение LLM, хотя проблемы, такие как вычислительная эффективность и устойчивость, продолжают существовать. Решая эти проблемы, данная работа вносит вклад в развитие надежных и прозрачных NLP-систем, открывая путь для более безопасных и надежных применений LLM в критически важных областях.

1 Введение

В последние годы достижения в области обработки естественного языка (NLP) и больших языковых моделей (LLM) произвели революцию в том, как машины понимают и генерируют текст. Эти модели применяются для выполнения сложных задач, таких как автоматический перевод, генерация текста, ответы на вопросы и создание парафраз. Генерация парафраз является ключевой проблемой в NLP, так как она требует сохранения семантического содержания предложения при изменении его синтаксической структуры. В этой теоретической части мы исследуем технологии и методологии, лежащие в основе генерации парафраз, сосредоточив внимание на архитектуре Transformer, адаптивном синтаксическом контроле и объяснимости языковых моделей, а также синтаксическом представлении в больших языковых моделях.

2 Обзорно-постановочный раздел работы

2.1 Модели языка и их применение в NLP

Модели языка — это алгоритмы, обученные на предсказание вероятности последовательности слов. Эти модели обучаются на больших объемах текстовых данных и учат генерировать связный текст, основываясь на статистических паттернах, найденных в этих данных. Введение моделей, основанных на Transformers, таких как BERT (Bidirectional Encoder Representations from Transformers), позволило добиться значительных успехов в производительности задач NLP. Transformer — это архитектура, основанная на механизмах внимания (self-attention), что позволяет моделям обрабатывать информацию более эффективно, чем предыдущие архитектуры, такие как рекуррентные нейронные сети (RNN) или сверточные нейронные сети (CNN) [1].

Ключевым аспектом, который следует учитывать в языковых моделях, является то, как они обрабатывают лингвистические вариации, характерные для каждого языка. Хотя такие модели, как BERT, показывают исключительные результаты в задачах обработки естественного языка, модель BERT предобучена на данных нескольких языков, и, следовательно, её производительность может значительно различаться в зависимости от языка. В этом контексте работа Nozza, Bianchi и Hovy [3] под названием "What the [MASK]? Making Sense of Language-Specific BERT

Models"исследует, как адаптированные модели BERT могут интерпретировать и генерировать текст по-разному в зависимости от лингвистических характеристик каждого языка.

Исследование предлагает анализ многоязычных моделей BERT, подчеркивая важность специфической адаптации к языку для повышения эффективности моделей в таких задачах, как генерация парафраз. Этот подход позволяет моделям более точно захватывать грамматические структуры и лексические особенности каждого языка, что важно для задач, таких как генерация парафраз, где необходимо соблюдать языковые правила при изменении предложений.

Например, в таких языках, как испанский или французский, синтаксические структуры и формы глаголов могут значительно отличаться от английских, что означает, что модель должна быть способна адаптироваться к этим изменениям для генерации парафраз, которые будут как согласованными, так и грамматически правильными. Работа Nozza et al. подчеркивает, что модели BERT, адаптированные к одному языку, называемые "монолингвальными могут превосходить многоязычные модели, когда применяются к специфическим задачам генерации текста, таким как парафраз.

В генерации парафраз языковые модели на основе Transformers идеальны благодаря своей способности учитывать как семантические отношения, так и сложные синтаксические структуры внутри предложений. Это позволяет им генерировать парафразы, которые сохраняют тот же смысл, но с вариациями в грамматической структуре фраз.

2.2 Техники генерации парафраз

Генерация парафраз является важной задачей в обработке естественного языка, используемой в таких приложениях, как расширение данных, автоматическая переработка текста и генерация ответов в системах диалога. Генерация точных и связных парафраз требует контроля над синтаксической структурой и смыслом текста, что представляет собой значительную проблему для языковых моделей.

Работа Sun et al. [4] предлагает подход под названием AESOP, который улучшает генерацию парафраз с помощью адаптивного контроля синтаксической структуры текста. AESOP позволяет генерировать парафразы, которые не только сохраняют оригинальный смысл, но и учитывают контекст и синтаксические отношения, что

важно для поддержания согласованности и плавности в генерируемом языке. Этот подход особенно актуален для вашей диссертации, так как он рассматривает, как вариации в синтаксических представлениях могут повлиять на производительность языковых моделей.

Кроме того, Liu et al. [5] исследуют генерацию парафраз без учителя с использованием синтаксического знания. Этот подход не зависит от больших объемов размеченных данных, что делает его интересной альтернативой для улучшения качества парафраз в условиях ограниченных данных.

2.3 Объяснимость больших языковых моделей (LLMs) и оценка точности

Несмотря на достижения больших языковых моделей (LLM), таких как BERT, одной из самых значительных проблем остается их отсутствие объяснимости. Языковые модели, хотя и весьма эффективны в таких задачах, как генерация текста, могут быть непрозрачными и трудными для интерпретации, что вызывает опасения по поводу их надежности и обоснования полученных результатов. Объяснимость важна не только для обеспечения доверия к решениям моделей, но и для возможности улучшать и оптимизировать процессы генерации текста.

В контексте генерации парафраз объяснимость играет ключевую роль. Когда модель генерирует парафразу, важно понять, как и почему она сделала определенные преобразования в синтаксической структуре предложения. В этом смысле работа Zhao et al. [6] "Explainability for Large Language Models: A Survey"предоставляет общий подход для решения проблемы объяснимости в языковых моделях. Исследование этих авторов фокусируется на различных методах интерпретируемости, которые позволяют исследовать, как модели принимают решения. Среди этих методов — визуализация внимания и методы на основе значимости, которые показывают, какие части текста были наиболее важными для предсказания или генерации данного вывода.

Объяснимость также критична при сравнении разных моделей или подходов, таких как те, которые используются для генерации парафраз. Анализируя точность и синтаксические вариации в сгенерированных парафразах, эти инструменты интерпретации позволяют выявить, соблюдала ли модель семантические отношения или внесла ли предвзятость или ошибки. Это важно для того, чтобы гарантировать, что

модель не только точна, но и прозрачна и обоснована в своих решениях.

Важно отметить, что, хотя GLUE является отличным инструментом для общей оценки моделей, системы генерации парафраз также требуют специфических метрик для оценки качества синтаксических и семантических преобразований. Дополнительные метрики, такие как BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation) и METEOR, могут быть использованы для оценки качества парафраз на основе сравнения сгенерированного вывода и эталонных парафраз, учитывая как семантическое сходство, так и структурные вариации.

Работа Zhao et al. [6] подчеркивает, как подходы к объяснимости могут дополняться метриками оценки, такими как GLUE, для получения более полной картины производительности моделей. Эти метрики, в сочетании с инструментами визуализации внимания, позволяют исследователям не только оценивать точность моделей, но и понимать причины, по которым модели принимают решения в таких задачах, как генерация парафраз.

2.4 Синтаксическая сондировка и синтаксис в представлениях слов

Анализ внутренних представлений в языковых моделях можно также дополнить исследованием синтаксиса в представлениях слов, как это обсуждается в работе Hewitt и Manning (2019). Это особенно важно, когда речь идет о моделях, адаптированных к конкретным языкам, как указано в работе Nozza et al. (2020). Многоязычные модели, такие как BERT, склонны обучать представления слов, отражающие общие синтаксические структуры языков, участвующих в обучении. Однако монолингвальный BERT может лучше захватывать синтаксические зависимости, характерные для конкретного языка, что важно для генерации парафраз.

Способность синтаксической сондировки обнаруживать эти синтаксические зависимости является естественным дополнением к работе Nozza et al., так как оба подхода позволяют анализировать, как модели захватывают структурную информацию языка и как эта информация влияет на задачи генерации текста, такие как парафраз.

2.5 Задачи и возможности в генерации парафраз

Несмотря на достижения в генерации парафраз, остаются несколько проблем. Одной из главных является гарантия того, чтобы сгенерированные парафразы сохраняли тот же смысл, что и оригинальный текст, не вводя двусмысленности или изменения контекста. Кроме того, модели должны быть способны генерировать парафразы, которые будут грамматически правильными и сохранять высокую плавность.

Работа Liu et al. [5] исследует, как синтаксическое знание может улучшить генерацию парафраз, предлагая не обучаемый подход, который может быть более эффективным с точки зрения ресурсов и данных. Этот подход подчеркивает важную возможность для разработки более эффективных и точных моделей для задач генерации парафраз.

2.6 Заключение теоретической части

Таким образом, модели языка, основанные на Transformer, произвели революцию в области NLP, предоставив значительные достижения в генерации текста, машинном переводе и других задачах. Генерация парафраз и объяснимость больших языковых моделей являются важными областями, которые непосредственно влияют на производительность и надежность моделей. Анализ синтаксиса в представлениях слов, а также подходы к улучшению генерации парафраз с помощью синтаксического контроля открывают новые возможности для повышения точности и понимания языковых моделей.

Работы Vaswani et al. [1], Sun et al. [4] и Hewitt и Manning [2], среди прочих, предоставляют прочную основу для разработки метода, который позволит интерпретировать внутренние состояния сети Transformer и понимать причинно-следственные связи между ними и производительностью модели, что является основной целью данной диссертации.

Список литературы

1. *Vaswani A.* Attention is all you need // Advances in Neural Information Processing Systems. — 2017.

2. *Hewitt J.*, *Manning C. D.* A structural probe for finding syntax in word representations // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). — 2019. — С. 4129—4138.

3. *Nozza D.*, *Bianchi F.*, *Hovy D.* What the [mask]? making sense of language-specific BERT models // arXiv preprint arXiv:2003.02912. — 2020.

4. *Sun J.*, *Ma X.*, *Peng N.* AESOP: Paraphrase generation with adaptive syntactic control // Proceedings of the 2021 conference on empirical methods in natural language processing. — 2021. — С. 5176—5189.

5. Unsupervised paraphrasing under syntax knowledge / T. Liu, Y. Sun, J. Wu, X. Xu, Y. Han, C. Li, B. Gong // Proceedings of the AAAI conference on artificial intelligence. Т. 37. — 2023. — С. 13273—13281.

6. Explainability for large language models: A survey / H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, M. Du // ACM Transactions on Intelligent Systems and Technology. — 2024. — Т. 15, № 2. — С. 1—38.