Синтаксический анализ с помощью больших языковых моделей: влияние форматов генерации

Е. Д. Шамаева*, Н. В. Лукашевич‡

Статья посвящена исследованию генеративных подходов к синтаксическому анализу на основе большой языковой модели (LLM). Исследуется влияние скобочных форматов (GRCT и LCT) генерации синтаксического представления на качество анализа. Установлено, что возможно успешное дообучение LLM для каждого из форматов: по метрике LAS_LLM результаты генерации превосходят результаты классических нейросетевых синтаксических анализаторов Stanza и DeepPavlov. При этом большая языковая модель, дообученная для синтаксического анализа на текстах одного домена, корректно работает на текстах другого домена.

Ключевые слова: синтаксический анализ, дообучение, большие языковые модели, деревья зависимостей, формат скобочной последовательности.

Синтаксический анализ — важный этап лингвистического анализа текстов. В этой области актуальным является применение больших языковых моделей (Large Language Model, LLM), в частности, их дообучение (fine-tuning) для генерации синтаксического дерева зависимостей. Классический формат CoNLL-U является трудным для генерации с помощью LLM, поскольку связь между главным и зависимым выражается через их номера в предложении. Поэтому одним из важных направлений исследований является генерация дерева зависимостей в скобочном формате. На момент исследования существует небольшое количество работ про применение LLM для синтаксического анализа русского языка в скобочном формате, в частности, посвященных дообучению для формата GRCT [1].

В данной статье рассматривается дообучение LLM для генерации результата синтаксического анализа в двух разных скобочных форматах

Шамаева Елена Денисовна — сотрудник каф. алгоритмических языков ф-та ВМК МГУ, e-mail: derinhelm@yandex.ru, ORCID: 0009-0002-6233-8958.

Shamaeva Elena Denisovna — Researcher, Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Department of Algorithmic Languages.

‡*Лукашевич Наталья Валентиновна* — д.т.н., зав.каф. алгоритмических языков ф-та ВМК МГУ, e-mail: louk_nat@mail.ru, ORCID: 0000-0002-1883-4121.

Loukachevitch Natalia Valentinovna — Doctor of Technical Sciences, Researcher, Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Chair of Department of Algorithmic Languages.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

дерева зависимостей, GRCT и LCT, введенных в [3]. Синтаксическое дерево зависимостей состоит из слов и знаков препинания предложения в качестве вершин и связей между ними в качестве ребер (с метками типа связи). Для генерации синтаксического дерева зависимостей с помощью LLM дерево зависимостей преобразуется в скобочный формат GRCT (Grammatical Relation Centered Tree) или LCT (Lexical Centered Tree). Для этого метки типа связей преобразуются в вершины, затем происходит обход в глубину модифицированного дерева зависимостей. В формате GRCT вершины типа связи являются главными для вершин формы слова, в формате LCT — наоборот. Например, в формате GRCT дерево зависимостей будет выглядеть как [root[nsubj[Дети]][играют]], а в формате LCT — [играют[Дети[nsubj]]][root]].

Особенностью генерации дерева зависимостей с помощью LLM является зависимость процесса генерации от сгенерированного фрагмента. Для форматов GRCT и LCT различается порядок генерации элементов дерева зависимостей: при генерации в формате GRCT сначала фиксируется тип связи, а потом форма слова, в формате LCT — наоборот. Различается и порядок генерации вершины формы слова и ее поддеревьев: в формате LCT вершина формы слова всегда идет до описания ее поддеревьев, в формате GRCT на это влияет порядок слов в предложении.

Цель данного исследования — сравнение обобщающей способности LLM при дообучении для генерации дерева зависимостей в форматах GRCT и LCT. Дообучение производилось на наиболее крупном датасете русскоязычных деревьев зависимостей SynTagRus, а тестирование — на тестовых выборках трех датасетов деревьев зависимостей: SynTagRus (8800 предложений, v2.9), GSD (601 пред., v2.5), Taiga (881 пред., v.2.6). У этих датасетов принципиально отличаются принципы построения: SynTagRus основан на художественной прозе и публицистике, GSD — на текстах русскоязычной версии Wikipedia, Taiga — на текстах электронной коммуникации. Эксперименты по дообучению проведены для инструктивных LLM размером 0.6 и 4 миллиарда параметров(Qwen3-0.6B и Qwen3-4B). Дополнительно проведены эксперименты для модели RuadaptQwen3-4B-Hybrid, которая является адаптацией для русского языка модели Qwen3-4B [2]. Для каждой модели выполнено дообучение для генерации в формате GRCT и LCT [1]. Для дообучения выбран метод Q-LoRA (Quantized-LoRA), параметры дообучения взяты из [1]. При дообучении использовались предложения из обучающей выборки, для которых длина обучающего примера

[1] при дообучении на вход модели — текст предложения и эталонное дерево в формате GRCT или LCT, при тестировании — только текст предложения, использовалась «жадная» генерация со значением температуры 0, ограничение длины — 5000 токенов, для декодирования — стандартные алгоритмы из [1].

после токенизации не превышала 512 субтокенов. Общее количество таких предложений зависело от токенизатора LLM и составляло около 70,000 предложений (валидационная выборка — 8906 предложений). Для сравнения использовались классические нейросетевые синтаксические анализаторы DeepPavlov (модель syntax_ru_syntagrus_bert) и Stanza (модель ru_syntagrus).

Для оценки результатов дообученных моделей авторами предложены метрики UAS_LLM и LAS_LLM. Классические метрики UAS и LAS с точки зрения корректности главного слова (UAS) и главного слова с типом связи (LAS) оценивают точность (доля корректных среди предсказанных), полноту (среди эталонных) и F-меру (их гармоническое среднее). Эти метрики неприменимы для оценки результатов LLM, так как не устойчивы к изменению предложения. Модифицированные метрики UAS_LLM и LAS_LLM в числителе сравнивают ребра (а не вершины) с точки зрения совпадения форм главных и зависимых слов (и типа связи для LAS_LLM), знаменатели остаются неизменными. В данном исследовании происходит усреднение метрик по всем тестовым предложениям датасета (для предложений с неуспешным результатом — метрика 0).

Для обоих форматов для датасета SynTagRus адаптированная и неадаптированная модели размером 4 млрд по метрике LAS_LLM превзошли DeepPavlov и Stanza[2], по метрике UAS_LLM показали сопоставимые результаты. На датасетах Taiga и GSD для обоих форматов только неадаптированная модель размером 4 превзошла DeepPavlov и Stanza по метрике LAS_LLM[3]. При этом на устойчивость к смене формата влияет размер модели и не влияет ее адаптация к русскому языку (для модели размером 0.6 млрд в формате GRCT — более высокие результаты, для моделей размером 4 млрд разница не установлена).

Проведенное исследование показало устойчивость дообучения LLM к изменению порядка следования формы слова и типа связи в скобочных форматах дерева зависимостей. Кроме того, установлено, что большая языковая модель, дообученная на текстах одного домена, может качественно работать на других доменах и превосходить по качеству классические нейросетевые синтаксические анализаторы.

Список литературы

[2]По метрике LAS_LLM на SynTagRus Ruadapt_4B_GRCT: 0.92, Qwen_4B_GRCT, Qwen_4B_LCT, Ruadapt_4B_LCT: 0.91, Stanza: 0.9, DeepPavlov: 0.87.

[3]На Taiga Qwen_4B_GRCT, Qwen_4B_LCT: 0.76, Stanza: 0.73, DeepPavlov: 0.71; на GSD Qwen_4B_LCT: 0.84, Qwen_4B_GRCT: 0.81, DeepPavlov: 0.8, Stanza: 0.79.

[1] C. Hromei, C. D. Croce, R. Basili, "U-DepPLLaMA: Universal Dependency Parsing via Auto-regressive Large Language Models", *IJCoL. Italian Journal of Computational Linguistics*, 2024. DOI: 10.4000/125nm.

[2] M. Tikhomirov, D. Chernyshev, "Facilitating large language model Russian adaptation with learned embedding propagation", *Journal of Language and Education*, **10**:4. DOI: 10.17323/jle.2024.22224 (2024).

[3] D. Croce, A. Moschitti, R. Basili, "Structured lexical similarity via convolution kernels on dependency trees", *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, 1034–1046, https://aclanthology.org/D11-1096/ .

Статья поступила 9 ноября 2026 г.

LLM-based syntactic parsing: impact of generating formats

E. D. Shamaeva, N. V. Loukachevitch

This article studies generative LLM-based syntactic parsing. The impact of bracketed syntactic representation formats (GRCT and LCT) on parsing quality is investigated. Research has established that an LLM can be successfully fine-tuned for analysis on each of the formats. For both formats, according to the LAS_LLM metric, the fine-tuned Qwen3-4B model exceeds the classic neural network parsers Stanza and DeepPavlov. Moreover, the fine-tuned LLM is resistant to domain shift.

Keywords: parsing, fine-tuning, large language models, dependency tree, bracketed format.

References

[1] C. Hromei, C. D. Croce, R. Basili, "U-DepPLLaMA: Universal Dependency Parsing via Auto-regressive Large Language Models", *IJCoL. Italian Journal of Computational Linguistics*, 2024. DOI: 10.4000/125nm.

[2] M. Tikhomirov, D. Chernyshev, "Facilitating large language model Russian adaptation with learned embedding propagation", *Journal of Language and Education*, **10**:4. DOI: 10.17323/jle.2024.22224 (2024).

[3] D. Croce, A. Moschitti, R. Basili, "Structured lexical similarity via convolution kernels on dependency trees", *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, 1034–1046, https://aclanthology.org/D11-1096/.

Received on November 9, 2026