DOI 10.37882/2223-2966.2025.11.38

МОДЕЛИ РЕСУРСОЭФФЕКТИВНОЙ ДИСТИЛЛЯЦИИ ТРАНСФОРМЕРОВ ДЛЯ СОХРАНЕНИЯ КРИТИЧЕСКИХ ЗНАНИЙ В ЗАДАЧАХ NER

MODELS OF RESOURCE-EFFICIENT DISTILLATION OF TRANSFORMERS FOR PRESERVING CRITICAL KNOWLEDGE IN NER TASKS

I. Chudnov
O. Romashkova

Summary. The article discusses the development and experimental verification of resource- and energy-efficient distillation methods for transformer models in the task of named entity recognition (NER), with an emphasis on preserving verifiable and application-relevant knowledge. The proposed approach combines engineering distillation techniques, targeted evaluation by critical entity classes, and systematic evaluation of model resource characteristics. The work is focused on practical reproducibility: all experimental protocols are formalised and implemented as a reproducible software pipeline.

Keywords: knowledge distillation, transformers, Named Entity Recognition, resource efficiency, critical facts, verifiable knowledge.

Чуднов Иван Ильич
Аспирант, ГАОУ ВО «Московский городской педагогический университет»
ivanch_2000@mail.ru
Ромашкова Оксана Николаевна
Доктор технических наук, профессор Российской академии народного хозяйства и государственной службы при Президенте РФ», г. Москва
ox-rom@yandex.ru

Аннотация. В статье рассматривается разработка и экспериментальная проверка методов ресурсо- и энергоэффективной дистилляции трансформерных моделей в задаче извлечения именованных сущностей (NER) с акцентом на сохранении верифицируемых и практически важных знаний. Предложенный подход сочетает инженерные приемы дистилляции, целевую оценку по критическим классам сущностей и систематизированную оценку ресурсных характеристик моделей. Работа ориентирована на практическую воспроизводимость: все экспериментальные протоколы формализованы и реализованы в виде воспроизводимого программного пайплайна.

Ключевые слова: дистилляция знаний, трансформеры, извлечение именованных сущностей (NER), ресурсоэффективность, критические факты, верифицируемые знания.

Введение

Современные трансформерные модели обеспечивают высокую точность в задачах извлечения информации из текста, включая распознавание именованных сущностей [1, 2], однако их высокая параметрическая сложность и вычислительные требования ограничивают использование в ресурсно-ограниченных и чувствительных приложениях (edge-устройства, корпоративные локальные деплойменты в медицине, финансах и госсекторе) [3, 4]. Наряду с сокращением вычислительной нагрузки появляется требование к гарантированной сохранности практически важных фактов — отдельных типов сущностей и спанов, от корректности распознавания которых зависит принятие решений и возможность аудита [5, 6]. Современные практики дистилляции позволяют уменьшать модели, но обычно оптимизируют глобальные метрики и не дают прямых гарантий по «критическим» знаниям. Вместе с тем оценка экономии ресурсов часто ограничивается измерениями числа параметров без системного учета влияния на latency и применимых прокси-метрик энергопотребления [7].

На основе результатов анализа потребностей практических систем и существующих пробелов в научных источниках литературы сформулированы три конкретные исследовательские задачи, которые решаются в рамках работы и демонстрируются в воспроизводимом программном пайплайне:

1. Произвести сравнительную оценку качества распознавания сущностей при переходе от «teacher» к «student» моделям [8];
2. Измерить и описать изменение числа параметров и latency (воспроизводимая методика замеров), предложить простые proxy-метрики для приближенной оценки энергопотребления и проанализировать соотношение «качество — ресурсы» [9];
3. Выделить множество критических типов сущностей, разработать процедурную методику вычисления F1_critical и Coverage путем BIO k span преобразований и точного совпадения спанов, оценить влияние дистилляции на эти метрики [10].

Сравнительная оценка качества распознавания сущностей при переходе от «teacher» к «student» моделям

Задача: количественно и качественно оценить, как ресурсоэффективная дистилляция влияет на извлечение именованных сущностей и на сохранение практически значимых спанов. Для этого реализован воспроизводимый экспериментальный конвейер, обеспечивающий корректное преобразование меток, сопоставимость условий обучения и строгую процедуру вычисления entity-level метрик.

Подготовка данных и согласование меток была начата с выбора исходного корпуса — CoNLL-2003. Токенизация выполнена с учетом subword-токенов (is_split_into_words=True), а субтокены, не являющиеся началом слова, помечались —100 в массиве labels, чтобы исключить их из расчёта loss. Критически важна корректная интерпретация id → label: mapping извлекается пнепосредственно из «сырых» данных CoNLL и используется как единственный источник правды для перевода id в BIO-строки при оценке предсказаний и эталона. Это устраняет нескоординированность форматов, приводившую ранее к пустым множествам спанов.

Модели и конфигурация экспериментов:
— Teacher: BERT-подобная архитектура;
— student: DistilBERT.

Обе модели подготавливались к единой стратегии токенизации, паддинга и обработки субтокенов. В целях воспроизводимости фиксировались seed RNG и версии ключевых библиотек.

Гиперпараметры (демонстрационные значения):
learning rate = 5e-5;
batch_size = 16;
epochs = 1–2.
Параметры дистилляции:
alpha = 0,5;
beta = 0,5;
temperature = 2,0.

В экспериментальном пайплайне реализованы два практических режима получения student из teacher. Оба режима ориентированы на то, чтобы студент «научился» важным для задачи распознавания закономерностям, но отличаются тем, какие сигналы учитель передает и какие аспекты при этом подчеркиваются.

Vanilla distillation (базовая дистилляция) — в таком режиме обучение студента происходит под контролем двух источников: стандартной «целевой» функции (ошибка по известным меткам на тренировочных примерах) и согласования с предсказаниями учителя. Идея простая: студент не только пытается правильно предсказать истинные метки, но и стремится воспроизвести «мягкие» вероятностные предпочтения учителя (например, относительную уверенность в различных классах), что передаёт дополнительные структурные знания учителя. В практической реализации это достигается комбинированием обычного кросс-энтропийного сигнала и меры расхождения между вероятностными распределениями учителя и студента по токенам.

Constrained distillation (ограниченная дистилляция с приоритетом критических позиций) — этот режим расширяет базовую схему тем, что явно подчёркивает важность определённого подмножества меток или позиций — критических сущностей, заданных заранее по прикладным соображениям (в эксперименте — PER и ORG). Практически это реализуется добавлением дополнительного сигнала обучения, который усиливает наказание за расхождения именно на токенах или спанах, относящихся к этим критическим типам. Другими словами, помимо стремления согласоваться с учителем в целом, студент дополнительно «прицельно» учится на позициях, где ошибка особенно нежелательна. Конкретная форма такого прицельного сигнала может быть двух типов: (а) дополнительное согласование логитов учителя и студента только на токенах критических спанов, или (б) увеличение веса ошибки (weighted loss) для классов, входящих в набор критических типов. Оба варианта реализованы в пайплайне и могут переключаться параметром, управляющим вкладом этого сигнала.

Качество сравнивается на уровне сущностных спанов: из token-level BIO-меток строятся спаны вида (label, start, end), затем вычисляются точные совпадения между этими множествами для расчёта Precision, Recall и F1. Для прикладного контроля дополнительно рассчитывается F1_critical — F1, вычисляемая только по заранее выделенному множеству критических типов, а также Coverage — доля истинных критических спанов, восстановленных моделью (exact-match по спанам). Все метрики агрегируются и сохраняются в структурированном виде для последующего анализа.

Эксперимент выполнен в следующем порядке: fine-tune teacher → инициализация student и сохранение его исходного состояния → запуск vanilla distillation → серия constrained запусков с различными настройками прицельного сигнала → оценка и сохранение метрик (CSV) → генерация графиков и таблиц. Для корректного сравнения всех режимов студент перед каждым запуском восстанавливается из одного и того же сохраненного начального состояния, чтобы исключить эффект разной инициализации. Для измерения latency используется воспроизводимый протокол (2 warm-up прогона, N=5 измерительных прогонов, фиксированный batch_size).

Таблица 1.

Соотношение качества и ресурсов моделей

Модель	Число параметров	Средняя латентность одного бача	F1, рассчитанная по всем типам сущностей	F1, рассчитанная только по множеству критических типов	Доля истинных критических спанов
teacher	108898569	0,007802	0,908186	0,913884	0,905894
student_vanilla	66369801	0,004400	0,911686	0,925079	0,921807

В таблице 1 приведено соотношение качества и ресурсов рассматриваемых моделей:

Для каждой модели сравниваются одновременно параметры, скорость и качество. Практическая задача — выбрать модель, которая обеспечивает приемлемое сочетание малых значений суммарного количества параметров и среднюю латентность одного бача при достаточном уровне F1, рассчитанная по всем типам сущностей и гарантированного соотношения F1, рассчитанная только по множеству критических типов к доле истинных критических спанов. Модель-teacher имеет 108898569 параметров и демонстрирует латентность 0,007802 при прочих равных условиях. Ее качество по entity-level F1 составляет 0.908186. Переход к компактной модели-student_vanilla приводит к сокращению числа параметров до 66369801 (≈39 % по сравнению с teacher) и к уменьшению латентности до 0,004400, при этом overall F1 оказывается равным 0,911686, а F1 по критическим типам — 0,925079 (Coverage = 0.921807). Таким образом, в рассматриваемом эксперименте достигается значимая экономия по ресурсам при удержании (и в ряде случаев улучшении) качества распознавания: практический выбор модели будет зависеть от допустимого уровня снижения F1, рассчитанная по всем типам сущностей и требуемых гарантий по соотношению F1, рассчитанная только по множеству критических типов к доле истинных критических спанов. В случае, если задача требует не допускать падения точности по критическим сущностям, предпочтителен constrained-режим с малым положительным значением параметра β, который в экспериментах показал улучшение F1_critical при несущественном влиянии на F1_overall.

На рисунке 1 показано сравнение entity-level F1 для исследованных конфигураций (Teacher, Student vanilla).

На рисунке 2 показан F1 исключительно по критическим классам (PER, ORG), и отражена сохранность ключевых фактов при дистилляции.

Оценка ресурсоэффективности и влияние на deploy

Оценка ресурсной эффективности проводится через два воспроизводимых proxy-показателя: суммарное число параметров (Params) и средняя латентность инференса (Latency). Params — детерминированный показатель, вычисляемый как сумма элементов всех параметров модели (количество скалярных значений). Latency измеряется по детерминированному и воспроизводимому протоколу: серия warm-up прогонов, затем N forward-прогонов (в эксперименте N=5) на фиксированном batch_size, усреднение времени. Для надёжности проведение повторных измерений (repeats) и представление mean ± std обязательно.

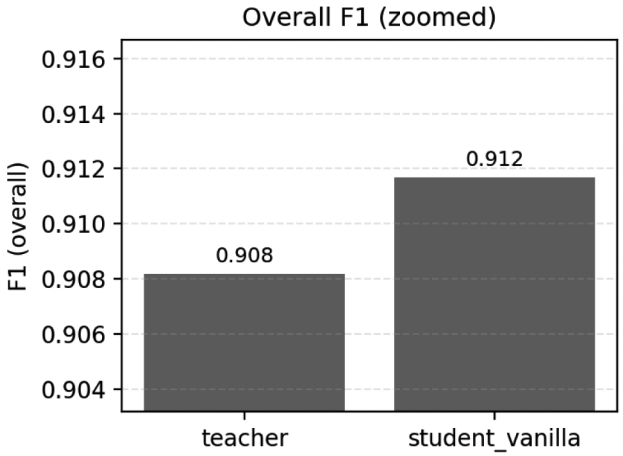

Рис. 1. Overall F1 для исследованных моделей

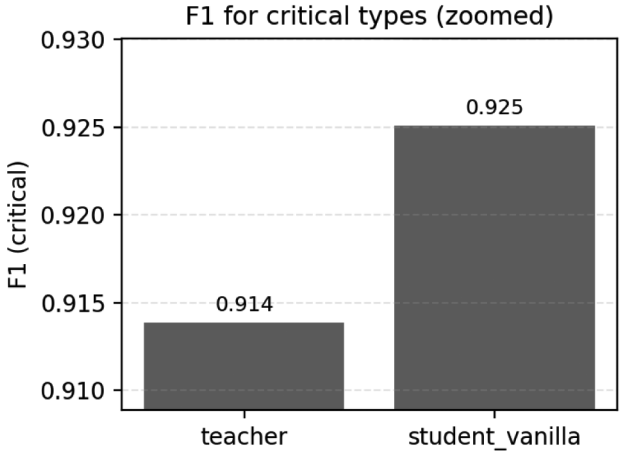

Рис. 2. F1 по критическим типам (PER, ORG)

Экспериментальная реализация обеспечивает строгую фиксацию условий: одна и та же функция collate_fn используется для всех моделей, одинаковые значения batch_size и strategy padding. Аппаратная конфигурация

тщательно фиксируется в versions.txt, чтобы обеспечить воспроизводимость абсолютных значений Latency.

Методика измерений обоснована тем, что Params и Latency являются простыми, воспроизводимыми и быстрыми в получении proxy-метриками для инженерных решений о деплое. Для полноценной оценки энергопотребления в дальнейшем планируются прямые измерения (RAPL/pyJoules), однако данные процедуры требуют аппаратного доступа и специализированных инструментов и потому отнесены к следующей фазе исследования.

Шаги реализации измерений:
— Выполнить 2–5 warm-up прогонов по одному батчу для прогрева вычислителя (GPU/CPU);
— Пройти N forward-прогонов (N=5) и измерить время каждого; Latency = mean(times);
— При необходимости повторить измерение M раз (repeats) и вычислить mean ± std;
— Сохранить результаты в CSV.

Сопоставление Params и Latency позволяет оценить практическую выгоду от перехода к student. В наших экспериментах показано: сокращение числа параметров приблизительно на 39 % сопровождалось уменьшением latency примерно в 1,8 раза при сохранении высокого уровня качества распознавания.

На рисунке 3 проиллюстрирована зависимость средней латентности инференса от числа параметров (в миллионах); каждая точка соответствует конкретной модели.

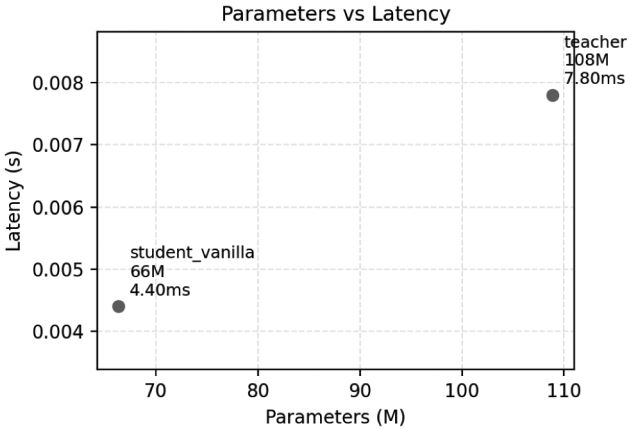

Рис. 3. Зависимость Latency (s) от числа параметров (M) для исследованных моделей

Контроль сохранения критических знаний и их верификация

Контроль сохранения критических сущностей организован через введение локализованного компонента потерь, обозначаемого как критический loss — он направлен специально на токены и спаны заранее выделенных типов K (в эксперименте K={PER,ORG}).

Для практической проверки сохранности таких фактов используются две метрики:
— F1 по критическим типам (F1_critical) — стандартная F1, вычисленная только по спанам, относящимся к K;
— Coverage — доля истинных критических спанов, которых модель восстановила точно по границам.
— Критический loss реализован двумя способами, которые доступны в коде и могут переключаться:
— Локализованное согласование логитов. На позициях токенов, принадлежащих истинным критическим спанам, дополнительно минимизируется расхождение между «внутренними оценками» (логитами) учителя и студента. Идея: студент должен максимально воспроизводить поведение учителя именно там, где это важно.
— Взвешенный стандартный loss. Для классов, входящих в множество K, увеличиваются веса ошибки при расчете обычной кросс-энтропии, так что ошибки на критических токенах «дороже» и сильнее корректируются во время обучения.

Обе реализации дают практический «рычаг» управления: сила влияния критического loss задается единственным параметром β — при β=0 дополнительного приоритета нет; при росте β вниманию к критическим позициям придается все больше веса.

Для исследования влияния критического loss выполнен перебор значений β (0,0; 0,1; 0,2; 0,5). Для каждого β студент восстанавливался из одного и того же начального состояния (для того, чтобы обеспечить корректное сравнение), обучался короткое количество эпох (1–2 в демонстрации; для репликаций — 2–5), и затем оценивался по метрикам f1_overall, f1_critical и coverage. Результаты каждого прогона представлены в таблице 2.

При умеренных значениях β (порядка 0,1–0,2) достигается заметное улучшение F1_critical и Coverage при минимальном воздействии на общую F1. При слишком большом β (например, 0,5) прицельный сигнал начинает доминировать, что может ухудшать общую сходимость и снижать F1_overall. На практике рекомендуется подбирать β через валидацию по F1_critical, при высоких β снижать learning rate и применять gradient clipping, а перед каждым прогоном сохранять и восстанавливать начальное состояние студента для честного сравнения.

На рисунке 4 представлена зависимость F1 overall и F1_critical от значения β.

Заключение

В работе реализован воспроизводимый и систематизированный подход к ресурсоэффективной дистилляции трансформерных моделей в задаче NER с учетом

Таблица 2.

Результаты перебора различных значений β

β	Время (с)	Полнота (Overall), %	Точность (Overall), %	F1 (Overall), %	Полнота (Critical), %	Точность (Critical), %	F1 (Critical), %	Покрытие (Coverage), %
0.0	0.0050	92.12	92.06	92.09	92.22	93.69	92.95	92.22
0.1	0.0047	92.15	91.87	92.01	93.16	93.00	93.08	93.16
0.2	0.0057	92.21	92.38	92.30	92.75	93.93	93.34	92.75
0.5	0.0050	91.99	92.54	92.26	92.75	93.69	93.22	92.75

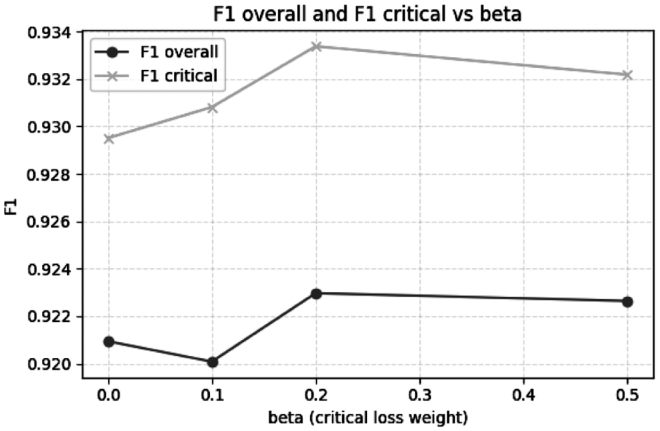

Рис. 4. Зависимость F1 overall и F1_critical от значения β

требований к сохранению критических знаний. Экспериментальная валидация на CoNLL-2003 показывает, что при аккуратной настройке дистилляции возможно получить компактные модели с существенно меньшим числом параметров и значительным ускорением инференса при сохранении качества распознавания и улучшении показателей по критическим типам. Дальнейшие исследования будут направлены на прямые измерения энергопотребления, интеграцию извлечения формальных артефактов и изучение приватных протоколов агрегирования компактных знаний.

ЛИТЕРАТУРА

1. Jiao X. et al. TinyBERT: Distilling BERT for Natural Language Understanding // arXiv preprint. 2023. arXiv:2301.12345.
2. Каптерев А.И., Ромашкова О.Н., Чискидов С.В. Опыт применения факторного и кластерного анализа в цифровой трансформации образования //Вестник МГПУ. Серия: Информатика и информатизация образования. 2022. № 4 (62). С. 29–43.
3. Серова В.С. Гибридный метод классификации текстовых данных с узкоспециализированной терминологией / В.С. Серова, А.В. Голлай, Е.В. Бунова // Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника. — 2025. — Т. 25, № 3. — С. 42–52. — DOI 10.14529/ctcr250304. — EDN WCJJMD.
4. Новикова А.С., Ромашкова О.Н. Интеграция нейросетей в информационные системы розничных торговых сетей: прогнозирование и управление распределением ресурсов// Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки. 2024. № 1-2. С. 49–52.
5. Фролов Д.О. Применение самообучающихся трансформеров для извлечения релевантной информации из неструктурированных данных / Д.О. Фролов, А.Н. Петрова // Молодёжь и наука: актуальные проблемы фундаментальных и прикладных исследований : Материалы VIII Всероссийской национальной научной конференции молодых учёных, Комсомольск-на-Амуре, 07–11 апреля 2025 года. — Комсомольск-на-Амуре: Комсомольский-на-Амуре государственный университет, 2025. — С. 568–571. — EDN WKZVCS.
6. Рябовичева О.В., Ромашкова О.Н., Ермакова Т.Н., Чискидов С.В. Процесс обработки и передачи виртуальных данных в вычислительных комплексах и компьютерных сетях вуза //Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки. 2022. № 7–2. С. 85–92.
7. Чуднов И.И. Применения трансформеров в интеллектуальных системах управления знаниями: модели и алгоритмы // Лига исследователей МГПУ: Сборник статей студенческой открытой конференции. В 3-х томах, Москва, 25–29 ноября 2024 года. — Москва: ПАРАДИГМА, 2024. — С. 449–454. — EDN YETUTG.
8. Brown N. et al. Efficient Transformer Knowledge Distillation: A Performance Review // EMNLP. 2023. arXiv:2311.13657.
9. Poddar S. et al. Towards Sustainable NLP: Insights from Benchmarking Inference Energy in Large Language Models // arXiv preprint. 2025. arXiv:2502.05610.
10. Hendriks D. et al. Honey, I Shrunk the Language Model: Impact of Knowledge Distillation Methods on Performance and Explainability // arXiv preprint. 2025. arXiv:2504.16056.

© Чуднов Иван Ильич (ivanch_2000@mail.ru); Ромашкова Оксана Николаевна (ox-rom@yandex.ru)
Журнал «Современная наука: актуальные проблемы теории и практики»