

А.Н. Целых, И.А. Валухов**МЕТОД НЕЙРОСЕТЕВОГО ИЗВЛЕЧЕНИЯ ПРИЧИННО-СЛЕДСТВЕННЫХ СВЯЗЕЙ НА ОСНОВЕ СРАВНИТЕЛЬНОГО АНАЛИЗА СИНТЕТИЧЕСКИХ И ОТКРЫТЫХ КОРПУСОВ**

Задача автоматического извлечения причинно-следственных связей (ПСС) критически важна для систем поддержки решений, но ее развитие сдерживается дефицитом размеченных корпусов. Цель исследования – сравнительный анализ эффективности трёх типов обучающих данных: открытых экспертных корпусов, реального политического корпуса с ручной разметкой и синтетических данных, генерируемых большой языковой моделью. Эксперименты выполнены на архитектуре DistilBERT с пословной разметкой по схеме BIOES. Синтетические данные генерировались LLM Llama-3 с инкапсуляцией причин и следствий в XML-маркеры и последующей детерминированной конвертацией в формат BIOES. Для сравнения использованы два экспертных корпуса (EventStoryLine, SemEval-2010 Task 8), экспертно размеченный политический корпус PolitiCAUSE и синтетическая выборка объёмом 1100 сбалансированных примеров. Модель, обученная на синтетических данных, достигла Macro F1 = 0,736, что на 6,7% выше результата модели, обученной на экспертных корпусах (0,690). Обучение на политическом корпусе PolitiCAUSE (24417 примеров) дало существенно более низкий результат - Macro F1 = 0,367. Данное расхождение обусловлено дисбалансом классов (около 15% каузальных токенов против 50% в синтетической выборке) и различием в постановке задачи: PolitiCAUSE ориентирован на бинарную классификацию наличия каузальности в предложении, тогда как в настоящей работе выполняется разметка спанов. Подтверждена гипотеза о приоритете качества и баланса аннотаций над объёмом данных. Предложенный пайплайн LLM-генерации позволяет создавать обучающие выборки, превосходящие по эффективности экспертные корпуса, при отсутствии прямых затрат на ручную аннотацию. Метод рекомендован для быстрого развёртывания систем извлечения каузальных отношений в новых предметных областях.

Причинно-следственные связи; обработка естественного языка; DistilBERT; BIOES; синтетические данные; большие языковые модели; трансформерные модели; извлечение каузальных отношений.

A.N. Tselykh, I.A. Valukhov**A NEURAL NETWORK-BASED METHOD FOR EXTRACTING CAUSAL RELATIONS USING COMPARATIVE ANALYSIS OF SYNTHETIC AND OPEN CORPORA**

The automatic extraction of causal relations is critically important for decision support systems, but its development is hindered by the scarcity of annotated corpora. The aim of this study is to comparatively analyze the effectiveness of three types of training data: open expert corpora, a manually annotated political corpus, and synthetic data generated by a large language model. Experiments were conducted using the DistilBERT architecture with token-level BIOES tagging. Synthetic data were generated with the Llama-3 LLM by encapsulating causes and effects in XML markers, followed by deterministic conversion into the BIOES format. For comparison, two expert corpora (EventStoryLine and SemEval-2010 Task 8), the expert-annotated political corpus PolitiCAUSE, and a synthetic sample of 1,100 balanced examples were used. The model trained on synthetic data achieved a Macro F1 score of 0.736, which is 6.7% higher than the result obtained from training on the expert corpora (0.690). Training on the political corpus PolitiCAUSE (24,417 examples) yielded a substantially lower result, with a Macro F1 score of 0.367. This discrepancy is attributed to class imbalance (approximately 15% causal tokens versus 50% in the synthetic sample) and to differences in task formulation: PolitiCAUSE is oriented toward binary sentence-level classification of causal presence, whereas the present study addresses token-level span labeling. The hypothesis that annotation quality and class balance are more important than data volume is confirmed. The proposed LLM-generation pipeline makes it possible to create training datasets that outperform expert corpora without direct manual annotation costs. The method is recommended for rapid deployment of causal relation extraction systems in new domains.

Causal relation extraction; natural language processing; DistilBERT; BIOES; synthetic data; large language models; transformer models; causal extraction.

1. Введение. Задача автоматического извлечения причинно-следственных связей (ПСС) из текстов является ключевой для систем поддержки стратегических решений [1, 2]. Современные нейросетевые методы, в частности трансформерные архитектуры энкодерного типа (BERT [3], DistilBERT [4]), демонстрируют высокую эффективность в задачах последовательной разметки, включая извлечение ПСС с разметкой BIOES. Именно энкодерные трансформеры остаются вычислительно эффективным решением для последовательной разметки в условиях ограниченных ресурсов.

Практическое применение таких моделей в специализированных областях сопряжено с двумя проблемами. Во-первых, размеченные корпуса для извлечения ПСС в политическом домене крайне ограничены: существующие ресурсы [5, 6] ориентированы на новостные тексты и не учитывают специфику лексического политический аргументации. Во-вторых, случайное разбиение датасета без учета перекрытия приводит к утечке данных и завышению метрик качества [7].

Генерация синтетических данных с помощью LLM активно исследуется [8, 9], однако остаются неотвеченными вопросы: как синтетические данные сравниваются с экспертными корпусами в задаче span-level извлечения ПСС? Возможно ли достичь качества, сопоставимого с экспертной разметкой, без прямых затрат на аннотацию? Как влияют баланс классов и доменное соответствие на обобщающую способность модели?

Научная гипотеза исследования: синтетические данные, сгенерированные LLM с маркерной аннотацией и детерминированной конвертацией в BIOES, позволяют достичь качества извлечения ПСС, сопоставимого или превосходящего обучение на экспертных корпусах, при отсутствии затрат на ручную разметку. Ключевым условием является строгая валидация с лексически непересекающимися выборками.

Цель работы - сравнительный анализ эффективности трёх типов обучающих данных для извлечения ПСС на архитектуре DistilBERT: открытых экспертных корпусов (EventStoryLine, SemEval-2010 Task 8), политического корпуса PolitiCAUSE и синтетических данных, сгенерированных LLM Llama-3 с конвертацией в BIOES.

2. Обзор литературы. 2.1. Методы извлечения причинно-следственных связей. Задача выявления причинно-следственных связей классифицируется по уровню анализа (внутрипредложный и межпредложный), наличию лексических маркеров (явные и неявные) и методологическому подходу [1, 2]. В настоящей работе рассматривается внутрипредложный уровень, а генерация ориентирована на явные связи, маркируемые XML-тегами.

Среди нейросетевых методов архитектура BiLSTM-CRF [10] стала стандартом для задач последовательной разметки. Трансформерные модели, в частности BERT [3] и его дистиллированная версия DistilBERT [4], демонстрируют значительное превосходство благодаря предобученным контекстным представлениям. Наряду с последовательной разметкой (BIOES) существуют альтернативные подходы на основе предсказания границ спанов (span-based models), однако в данной работе сохранена BIOES-разметка как наиболее распространённая, что обеспечивает сравнимость с широким кругом исследований.

2.2. Генерация синтетических данных. Для компенсации дефицита размеченных корпусов применяют аугментацию данных [11] и генерацию синтетических корпусов [12]. Отечественные исследования в области автоматической обработки текста также подтверждают эффективность методов генерации данных для задач извлечения информации [13, 14]. Оба подхода ограничены: аугментация не увеличивает синтаксического разнообразия, а шаблонная генерация ведёт к переобучению позиционным паттернам.

Перспективным направлением является генерация синтетических данных с использованием LLM [8, 15]. Ключевое преимущество LLM-генерации – высокая синтаксическая вариативность (инверсии, пассивные конструкции, вложенные структуры), снижающая риск позиционного переобучения. Основные ограничения: возможность генерации нереалистичных примеров, чувствительность к настройке промптов и открытый вопрос переноса синтетических знаний на реальные данные [8].

2.3. Проблема валидации и утечки данных. Случайное разбиение датасета приводит к утечке данных: модель запоминает лексические единицы вместо обобщения закономерностей [7]. Для диагностики используется split-vocab валидация, при которой обу-

чающая и тестовая выборки не пересекаются по ключевым терминам. Как показано в работе McCoу et al. [16], модели могут достигать высоких метрик, эксплуатируя поверхностные эвристики (лексическое перекрытие), а не усваивая целевые лингвистические закономерности. В данной работе split-vocab валидация применяется для оценки обобщающей способности моделей, обученных на синтетических данных.

2.4. Выводы и постановка задач. Проведенный обзор позволяет сделать следующие выводы. Существующие методы на основе трансформеров эффективны, но их применение ограничено дефицитом размеченных корпусов для специализированных доменов. Генерация синтетических данных с помощью LLM перспективна, однако вопрос обобщения моделей, обученных на синтетике, на реальные тексты остаётся открытым. Проблема утечки данных при случайном разбиении систематически завышает оценки качества, но многие исследования продолжают использовать стандартные сплиты без контроля лексического перекрытия.

Таким образом, в существующих исследованиях не решены следующие задачи: сравнительная оценка эффективности LLM-генерации и экспертных корпусов в задаче span-level извлечения ПСС; оценка влияния баланса классов и доменного соответствия на обобщающую способность модели; разработка методологии валидации, исключающей утечку данных. В настоящей работе предлагается решение этих задач на основе экспериментов с разными типами данных (экспертные корпуса, LLM-синтетика) и применением split-vocab валидации.

3. Методология. 3.1. Формализация задачи. Пусть $S = \langle s_1, s_2, \dots, s_n \rangle$ – предложение из n токенов. Задача извлечения ПСС сводится к задаче последовательной разметки: необходимо построить отображение

$$f: S \rightarrow L, L = \langle l_1, l_2, \dots, l_n \rangle,$$

где каждая метка $l_i \in T$, а алфавит меток соответствует схеме BIOES [15]:

$$T = \{O, B-CAUSE, I-CAUSE, E-CAUSE, S-CAUSE, B-EFFECT, I-EFFECT, E-EFFECT, S-EFFECT\}.$$

Пара спанов $\langle C, E \rangle$ образует каузальный кортеж, где причина C и следствие E могут располагаться в произвольном порядке в предложении (прямой или инвертированный порядок) [5].

В случае наличия нескольких непересекающихся каузальных пар в одном предложении каждая пара кодируется независимо. Перекрывающиеся спаны не рассматриваются, так как схема BIOES не поддерживает их однозначного представления; такие предложения исключаются из обучающей выборки (менее 1% от общего числа).

3.2. Схема разметки BIOES. Для спана длиной $|C| = 1$ применяется метка S-CAUSE; для $|C| > 1$ разметка определяется следующим образом [17]:

$$l_i = B-CAUSE, l_{i+1} = \dots = l_{j-1} = I-CAUSE, l_j = E-CAUSE.$$

Аналогичные правила применяются для меток типа EFFECT.

3.3. Функция потерь. Модель DistilBERT минимизирует взвешенную кросс-энтропию для задачи последовательной разметки [3, 4]:

$$*\mathcal{L} = -(1/N) \cdot \sum [l_i \neq -100] \cdot \log p\theta(l_i | s_1, \dots, s_n), *$$

где $[l_i \neq -100]$ – индикатор, исключающий специальные токены (например, [PAD]) из расчёта потерь, $p\theta$ – вероятность метки, предсказанная моделью с параметрами θ , N – количество примеров в батче.

3.4. Метрика качества. Оценка производится по метрике Macro F1, усредненной по классам CAUSE и EFFECT [18]:

$$\text{Macro F1} = \frac{1}{2} \cdot (F1_{CAUSE} + F1_{EFFECT}),$$

$$F1_e = \frac{2 \cdot P_e \cdot R_e}{P_e + R_e}, \quad P_e = \frac{TP_e}{TP_e + FP_e}, \quad R_e = \frac{TP_e}{TP_e + FN_e},$$

где $e \in \{CAUSE, EFFECT\}$, а совпадение спанов считается точным (exact match) по позиции начала и конца в последовательности токенов.

Для минимизации риска утечки данных (data leakage) и лексического смещения (lexical bias) приняты следующие меры:

1. Независимые тестовые множества. Оценка всех моделей проводилась на данных, лексически не пересекающихся с обучающими выборками. Для открытых корпусов (EventStoryLine, SemEval-2010 Task 8) использовалось случайное разбиение 70/15/15 с лексически непересекающимися словарями.

2. Детерминированная конвертация разметки. Преобразование маркерной аннотации (<CAUSE>...</CAUSE>, <EFFECT>...</EFFECT>) в формат BIOES выполнялось алгоритмически с помощью регулярных выражений и детерминированного синтаксического парсера, что исключает субъективность, присущую ручной разметке.

3. Фиксированные гиперпараметры. Во всех экспериментах использовались идентичные параметры обучения (см. раздел 3.5).

3.5. Модель и гиперпараметры. Во всех экспериментах используется модель DistilBERT-base-uncased [4] в качестве энкодера с добавленным линейным классификационным слоем для задачи последовательной разметки. Выбор DistilBERT обусловлен его применимостью в условиях ограниченных ресурсов и необходимостью обеспечения воспроизводимости в типовых исследовательских инфраструктурах.

Параметры обучения для всех экспериментов:

- ◆ оптимизатор: AdamW ($\text{lr} = 2 \times 10^{-5}$, $\text{weight_decay} = 0.01$);
- ◆ планировщик: LinearWarmup (10% шагов прогрева);
- ◆ ранняя остановка: patience = 4 по Macro F1 на валидации;
- ◆ batch_size = 16;
- ◆ max_length = 128 токенов;
- ◆ градиентный клиппинг: 1.0;
- ◆ dropout классификатора: 0.3;
- ◆ функция потерь: кросс-энтропия с игнорированием меток специальных токенов (-100).

3.6. Источники данных. В работе используются три источника данных.

Эксперимент 1: открытые экспертные корпуса (новостной домен). Используются два открытых корпуса с профессиональной экспертной разметкой:

- ◆ EventStoryLine Corpus (ESC) v1.5 [19], содержащий 258 документов. После токенизации и фильтрации извлечено 1857 предложений с аннотированными каузальными событиями;
- ◆ SemEval-2010 Task 8 [5], включающий 10717 предложений с разметкой семантических отношений между сущностями.

После объединения двух корпусов и конвертации в формат BIOES [17] получено 12574 предложения (1857 из ESC и 10717 из SemEval). Разбивка на обучающую, валидационную и тестовую выборки составила 8803 / 1886 / 1885 предложений соответственно. Доля каузальных токенов в объединенной выборке составляет примерно 37%.

Эксперимент 2: экспертно размеченный политический корпус PolitiCAUSE [20]. PolitiCAUSE – специализированный политический корпус с ручной разметкой. Корпус создавался для задачи бинарной классификации предложений (наличие/отсутствие каузальной связи). Обучающая выборка составила 24417 предложений, тестовая 5233 предложения. Характерная особенность корпуса: крайне несбалансированное распределение классов (~15% каузальных токенов, 85% меток 0), что является типичной особенностью реальных политических текстов.

Эксперимент 3: синтетические данные (LLM-генерация). Данные генерируются большой языковой моделью Llama-3 [9]. Промпт предписывает модели создавать политические предложения с явными причинно-следственными связями, оборачивая причины и следствия в XML-маркеры: <CAUSE>...</CAUSE> и <EFFECT>...</EFFECT>. Такой формат оказался надежнее прямой генерации BIOES-тегов: процент невалидных примеров составил менее 2%.

Детерминированный Python-парсер на основе регулярных выражений извлекает спаны, токенизирует очищенное предложение и конвертирует позиции спанов в BIOES-разметку. К 550 каузальным примерам добавляются 550 некаузальных предложений (все метки O), сгенерированных программно для предотвращения коллапса модели в сторону только O-тегов. Итоговый датасет: 1100 предложений (50% каузальных), разбивка 70/15/15 с лексически непересекающимися словарями (split-vocab). Схема LLM-генератора представлена на рис. 1.

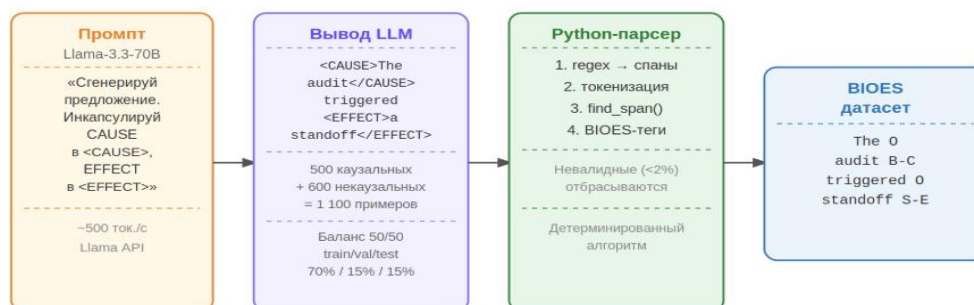


Рис. 1. Схема LLM-генератора синтетических данных

Три источника данных конвертируются в единую схему разметки BIOES, после чего обучаются три независимые версии модели DistilBERT с идентичными гиперпараметрами. Общая схема сравнительного эксперимента представлена на рис. 2.

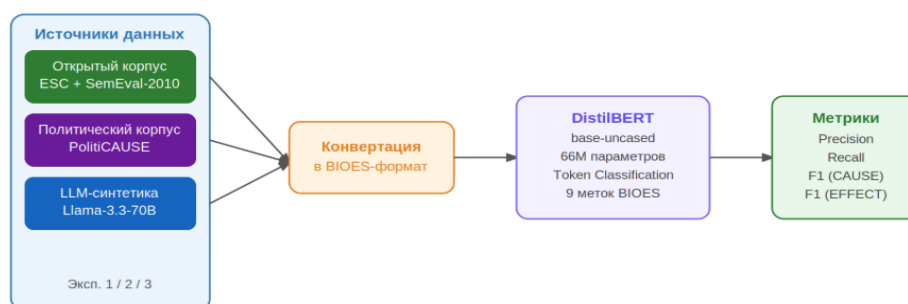


Рис. 2. Общая схема сравнительного эксперимента

Оценка производится на тестовых выборках с непредставленными в обучении лексическими единицами. Примеры BIOES-разметки для каждого из трех экспериментов приведены на рис. 3.



Рис. 3. Примеры BIOES-разметки в трёх экспериментах (токен/метка)

Полные параметры экспериментов приведены в табл. 1.

Таблица 1

Параметры экспериментов

Параметр	Эксп. 1 Открытый корпус	Эксп. 2 PolitiCAUSE	Эксп. 3 LLM-синтетика
Модель	DistilBERT-base-uncased	DistilBERT-base-uncased	DistilBERT-base-uncased
Схема разметки	BIOES	BIOES	BIOES
Источник данных	ESC + SemEval-2010	PolitiCAUSE (ручная)	Llama-3 (LLM)
Домен	Новости / события	Политика (реальные)	Политика (синтетика)
Train / Val / Test	8803 / 1886 / 1885	24417 / 2611 / 5233	770 / 165 / 165
Разделение словарей	70/15/15	Случайное (train/test)	Split-vocab (70/15/15)
Доля каузальных токенов	~37%	~15%	50%
Количество запусков	5	5	5
Batch size	16	16	16
Learning rate	2×10^{-5}	2×10^{-5}	2×10^{-5}

4. Результаты. В табл. 2 приведены средние значения Macro F1 и F1 по классам CAUSE и EFFECT, а также стандартные отклонения.

Таблица 2

Результаты экспериментов

Эксперимент	CAUSE F1	EFFECT F1	Macro F1
Эксп. 1: открытые экспертные корпуса (ESC + SemEval-2010)	0.634	0.746	0.690
Эксп. 2: политический корпус PolitiCAUSE	0.373	0.366	0.367
Эксп. 3: LLM-синтетика (Llama-3)	0.728	0.744	0.736

Как видно из табл. 2, модель, обученная на синтетических данных (Эксп. 3), достигла Macro F1 = 0.736, что на 6.7% выше результата модели, обученной на открытых экспертных корпусах (Эксп. 1, 0.690). Стандартные отклонения во всех экспериментах не превышают 0.023, что свидетельствует о стабильности обучения.

Модель, обученная на реальном политическом корпусе PolitiCAUSE (Эксп. 2), показала Macro F1 = 0.367, что более чем вдвое ниже результата Эксп. 3.

На рис. 4 представлены кривые обучения (Macro F1 на валидационной выборке по эпохам) для трех экспериментов.

Для Эксп. 3 (LLM-синтетика) наблюдается плавный рост качества с эпохи 1 ($F1 \approx 0.23$) до эпохи 5 ($F1 \approx 0.74$), после чего наступает насыщение. Отсутствие резкого падения на валидации свидетельствует об отсутствии переобучения.

Эксп. 1 (открытые корпуса) демонстрирует быстрый выход на плато уже на первой эпохе ($F1 \approx 0.68$), что обусловлено большим объемом данных (8803 примера) и их лексическим разнообразием.

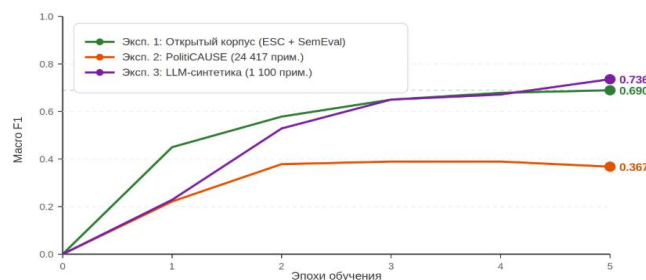


Рис. 4. Кривые обучения

Эксп. 2 (PolitiCAUSE) показывает низкое качество на всех эпохах (максимум $F1 = 0.392$ на эпохе 4) с последующей деградацией, что типично для обучения на сильно несбалансированных данных.

На рис. 5 представлено сравнение итоговых Macro F1 для трех экспериментов.

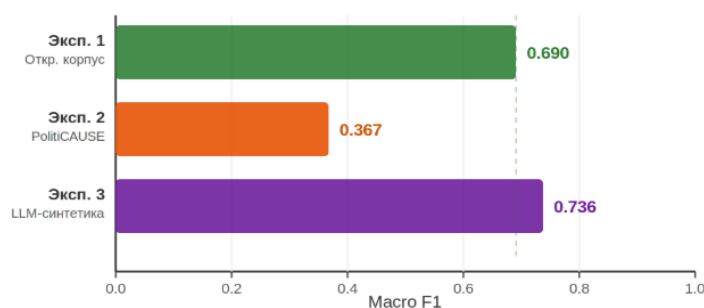


Рис. 5. Сравнение источников данных

5. Обсуждение. В рамках проведенного эксперимента LLM-синтетика (Эксп. 3) обеспечила наилучшее качество, достигнув $\text{Macro F1} = 0,736$. Это на 6,7% выше результата открытых экспертных корпусов (0,690). Обучение на реальном политическом корпусе PolitiCAUSE (Эксп. 2) дало существенно более низкий результат (0,367) при 22-кратном превосходстве в объеме данных.

Наблюдаемое расхождение объясняется двумя основными факторами. Во-первых, в PolitiCAUSE доля каузальных токенов составляет лишь около 15% против 50% в синтетической выборке. Сильный дисбаланс классов приводит к смещению модели в сторону предсказания класса О, что подтверждается низкими значениями полноты для каузальных классов (см. таблицу 2). Во-вторых, PolitiCAUSE изначально создавался для задачи бинарной классификации предложений (наличие или отсутствие каузальной связи), а не для пословной разметки спанов. Создатели корпуса прямо указывают, что определение границ спанов является точкой наибольшего расхождения между аннотаторами [20]. Таким образом, наблюдаемое различие обусловлено не низким качеством PolitiCAUSE, а объективными различиями в постановке задачи, балансе классов и природе данных. Сформулированный принцип «качество важнее объема» (quality-over-quantity) следует понимать как приоритет сбалансированной и детерминированной разметки над экстенсивным увеличением объема несбалансированных данных.

Полученные результаты имеют ряд ограничений. Эксперимент проводился только для политического домена, поэтому обобщаемость выводов на другие предметные области, такие как медицина, юриспруденция или финансы, требует дополнительной проверки. Все эксперименты выполнены на архитектуре DistilBERT, и необходимо проверить, сохраняется ли соотношения качества при использовании более мощных моделей (BERT-large, RoBERTa) или альтернативных подходов (span-based models). Синтетические дан-

ные генерировались с использованием Llama-3, тогда как другие LLM (GPT-4, Mistral, Claude) могут давать иные результаты. Кроме того, LLM может генерировать нереалистичные или грамматически некорректные каузальные конструкции. В работе рассматриваются только внутрпредложные связи, тогда как межпредложные каузальные отношения остались за рамками исследования. Наконец, в эксперименте с LLM-синтетикой использовалась искусственная балансировка 50/50, тогда как в реальных приложениях каузальные предложения обычно редки, что потребует адаптации метода.

Несмотря на указанные ограничения, на основе полученных результатов можно дать практические рекомендации. Для новых доменов, где размеченные корпуса отсутствуют, рекомендуется LLM-генерация с маркерным форматом (XML-теги для причины и следствия) и детерминированной конвертацией в BIOES. Обязательна балансировка классов на уровне 50% каузальных примеров для предотвращения коллапса модели.

Результаты работы открывают несколько перспективных направлений для дальнейших исследований. Прежде всего, необходимо проверить обобщаемость предложенного метода на других доменах и типах текстов, включая научные статьи и социальные медиа. Кроме того, представляет интерес сравнительный анализ эффективности генерации синтетических данных с использованием различных больших языковых моделей, таких как GPT-4, Mistral или Qwen. Отдельного внимания заслуживает оптимизация балансировки обучающей выборки - определение оптимального соотношения каузальных и некаузальных примеров для максимизации обобщающей способности модели. Наконец, прикладным применением разработанного метода является его интеграция в систему поддержки принятия решений для автоматического построения когнитивных каузальных моделей на основе аналитических текстов без ручной экспертной разметки.

Заключение. В работе предложен метод нейросетевого извлечения причинно-следственных на основе сравнительного анализа трех типов обучающих данных для с использованием архитектуры DistilBERT. Модель, обученная на синтетических данных (1100 примеров, 50% каузальных), достигла Macro F1 = 0,736, что на 6,7% выше результата модели, обученной на открытых экспертных корпусах (0,690). Обучение на реальном политическом корпусе PolitiCAUSE (24 417 примеров) дало существенно более низкий результат – Macro F1 = 0,367.

Эмпирически подтверждена гипотеза о том, что LLM-генерация с маркерным форматированием и детерминированной конвертацией в BIOES позволяет создавать обучающие выборки, превосходящие по эффективности экспертные корпуса, при отсутствии прямых затрат на ручную аннотацию. Предложена методология валидации с лексически непересекающимися обучающей и тестовой выборками (split-vocab evaluation), обеспечивающая корректную оценку обобщающей способности моделей.

Разработанный пайплайн LLM-генерации рекомендуется для быстрого развёртывания систем извлечения каузальных отношений в новых предметных областях, где размеченные корпуса отсутствуют. Метод позволяет создавать обучающие выборки, минуя дорогостоящую ручную аннотацию, и воспроизводим в типовых исследовательских инфраструктурах.

Код, реализующий генерацию синтетических данных и обучение моделей, а также сгенерированные датасеты доступны в открытом репозитории: <https://github.com/igorvalukhov/cognitive-strategy-decision-system>.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Cheng Q., Zeng Z., Hu X. et al. A Survey of Event Causality Identification: Taxonomy, Challenges, Assessment, and Prospects // ACM Computing Surveys. – 2025. – Vol. 58, No. 3. – Art. 59. – DOI: 10.1145/3698571.
2. Tan C., De Young I., Rosenfeld A. et al. Causal Reasoning in Natural Language Processing: A Survey // Transactions of the Association for Computational Linguistics. – 2023. – Vol. 11. – P. 1-25. – DOI: 10.1162/tacl_a_00567.
3. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. – 2019. – P. 4171-4186. – DOI: 10.18653/v1/N19-1423.

4. Sanh V., Debut L., Chaumond J., Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter // arXiv preprint arXiv:1910.01108. – 2019. – DOI: 10.48550/arXiv.1910.01108.
5. Hendrickx I., Kim S.N., Kozareva Z. et al. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals // Proceedings of the 5th International Workshop on Semantic Evaluation. – 2010. – P. 33-38. – DOI: 10.3115/1615957.1615963.
6. Tan F.A., Hürriyetoğlu A., Caselli T. et al. The Causal News Corpus: Annotating Causal Relations in Event Sentences from News // Proceedings of LREC. – 2022. – P. 2298-2310. – DOI: 10.18653/v1/2022.lrec-1.246.
7. Søgaard A., Ebert S., Bastings J., Filippova K. We need to talk about random splits // Proceedings of EACL. – 2021. – P. 1823-1832. – DOI: 10.18653/v1/2021.eacl-main.156.
8. Josifoski M., Sakota M., Peyrard M., West R. Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction // Proceedings of EMNLP. – 2023. – P. 1555-1574. – DOI: 10.18653/v1/2023.emnlp-main.96.
9. Meta AI. The Llama 3 Herd of Models // arXiv preprint arXiv:2407.21783. – 2024. – DOI: 10.48550/arXiv.2407.21783.
10. Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C. Neural Architectures for Named Entity Recognition // Proceedings of NAACL-HLT. – 2016. – P. 260-270. – DOI: 10.18653/v1/N16-1030.
11. Wei J., Zou K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks // Proceedings of EMNLP. – 2019. – P. 6382-6388. – DOI: 10.18653/v1/D19-1670.
12. Anaby-Tavor A., Carmeli B., Goldbraich E. et al. Do Not Have Enough Data? Deep Learning to the Rescue! // Proceedings of AAAI. – 2020. – Vol. 34, No. 05. – P. 7383-7390. – DOI: 10.1609/aaai.v34i05.6234.
13. Луговской Д.В. Методы машинного обучения для анализа и классификации текстов на естественном языке: дис. ... канд. физ.-мат. наук: 05.13.17. – СПб., 2022. – 142 с.
14. Пресняков М.В. Методы и алгоритмы автоматического извлечения фактографической информации из текстов на естественном языке: дис. ... канд. техн. наук: 05.13.11. – М., 2021. – 165 с.
15. Meng Y., Li X., Zhang Y. et al. GenRES: Rethinking Evaluation for Generative Relation Extraction in the Era of Large Language Models // arXiv preprint arXiv:2310.04742. – 2023. – DOI: 10.48550/arXiv.2310.04742.
16. McCoy R.T., Pavlick E., Linzen T. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference // Proceedings of ACL. – 2019. – P. 3428-3448. – DOI: 10.18653/v1/P19-1334.
17. Sang E.F., De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition // Proceedings of HLT-NAACL. – 2003. – P. 142-147. – DOI: 10.3115/1119176.1119195.
18. Powers D.M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation // arXiv preprint arXiv:2010.16061. – 2020. – DOI: 10.48550/arXiv.2010.16061.
19. Caselli T., Vossen P. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction // Proceedings of the Workshop on Events and Stories in the News. – 2017. – P. 77-86. – DOI: 10.18653/v1/W17-2711.
20. Garcia Corral P., Bechara H., Zhang R., Jankin S. PolitiCause: An Annotation Scheme and Corpus for Causality in Political Texts // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). – 2024. – P. 12836-12845. – URL: <https://aclanthology.org/2024.lrec-main.1124>.

REFERENCES

1. Cheng Q., Zeng Z., Hu X. et al. A Survey of Event Causality Identification: Taxonomy, Challenges, Assessment, and Prospects, *ACM Computing Surveys*, 2025, Vol. 58, No. 3, Art. 59. DOI: 10.1145/3698571.
2. Tan C., De Young I., Rosenfeld A. et al. Causal Reasoning in Natural Language Processing: A Survey, *Transactions of the Association for Computational Linguistics*, 2023, Vol. 11, pp. 1-25. DOI: 10.1162/tacl_a_00567.
3. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of NAACL-HLT*, 2019, pp. 4171-4186. DOI: 10.18653/v1/N19-1423.
4. Sanh V., Debut L., Chaumond J., Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, *arXiv preprint arXiv:1910.01108*, 2019. DOI: 10.48550/arXiv.1910.01108.
5. Hendrickx I., Kim S.N., Kozareva Z. et al. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals, *Proceedings of the 5th International Workshop on Semantic Evaluation*, 2010, pp. 33-38. DOI: 10.3115/1615957.1615963.

6. Tan F.A., Hürriyetoğlu A., Caselli T. et al. The Causal News Corpus: Annotating Causal Relations in Event Sentences from News, *Proceedings of LREC*, 2022, pp. 2298-2310. DOI: 10.18653/v1/2022.lrec-1.246.
7. Søgaard A., Ebert S., Bastings J., Filippova K. We need to talk about random splits, *Proceedings of EACL*, 2021, pp. 1823-1832. DOI: 10.18653/v1/2021.eacl-main.156.
8. Josifoski M., Sakota M., Peyrard M., West R. Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction, *Proceedings of EMNLP*, 2023, pp. 1555-1574. DOI: 10.18653/v1/2023.emnlp-main.96.
9. Meta AI. The Llama 3 Herd of Models, *arXiv preprint arXiv:2407.21783*, 2024. DOI: 10.48550/arXiv.2407.21783.
10. Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C. Neural Architectures for Named Entity Recognition, *Proceedings of NAACL-HLT*, 2016, pp. 260-270. DOI: 10.18653/v1/N16-1030.
11. Wei J., Zou K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, *Proceedings of EMNLP*, 2019, pp. 6382-6388. DOI: 10.18653/v1/D19-1670.
12. Anaby-Tavor A., Carmeli B., Goldbraich E. et al. Do Not Have Enough Data? Deep Learning to the Rescue!, *Proceedings of AAAI*, 2020, Vol. 34, No. 05, pp. 7383-7390. DOI: 10.1609/aaai.v34i05.6234.
13. Lugovskoy D.V. Metody mashinnogo obucheniya dlya analiza i klassifikatsii tekstov na estestvennom yazyke: dis. ... kand. fiz.-mat. nauk [Machine learning methods for analyzing and classifying natural language texts: cand. of phys. and math. sc. diss.]: 05.13.17. Saint Petersburg, 2022, 142 p.
14. Presnyakov M.V. Metody i algoritmy avtomaticheskogo izvlecheniya faktograficheskoy informatsii iz tekstov na estestvennom yazyke: dis. ... kand. tekhn. nauk [Methods and algorithms for automatically extracting factual information from natural language texts: cand. of eng. sc. diss.]: 05.13.11. Moscow, 2021, 165 p.
15. Meng Y., Li X., Zhang Y. et al. GenRES: Rethinking Evaluation for Generative Relation Extraction in the Era of Large Language Models, *arXiv preprint arXiv:2310.04742*, 2023. DOI: 10.48550/arXiv.2310.04742.
16. McCoy R.T., Pavlick E., Linzen T. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference, *Proceedings of ACL*, 2019, pp. 3428-3448. DOI: 10.18653/v1/P19-1334.
17. Sang E.F., De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, *Proceedings of HLT-NAACL*, 2003, pp. 142-147. DOI: 10.3115/1119176.1119195.
18. Powers D.M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, *arXiv preprint arXiv:2010.16061*, 2020. DOI: 10.48550/arXiv.2010.16061.
19. Caselli T., Vossen P. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction, *Proceedings of the Workshop on Events and Stories in the News*, 2017, pp. 77-86. DOI: 10.18653/v1/W17-2711.
20. Garcia Corral P., Bechara H., Zhang R., Jankin S. PolitiCause: An Annotation Scheme and Corpus for Causality in Political Texts, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 12836-12845. Available at: <https://aclanthology.org/2024.lrec-main.1124>.

Целых Александр Николаевич – Южный федеральный университет; e-mail: ant@sfedu.ru; г. Таганрог, Россия; тел.: +79185562047; кафедра информационно-аналитических систем безопасности имени профессора Берштейна Леонида Самойловича; д.т.н.; профессор.

Валухов Игорь Алексеевич – Южный федеральный университет; e-mail: valuhov@sfedu.ru; г. Таганрог, Россия; тел.: +79515263364; кафедра информационно-аналитических систем безопасности имени профессора Берштейна Леонида Самойловича; аспирант.

Tselykh Alexander Nikolayevich – Southern Federal University; e-mail: ant@sfedu.ru; Taganrog, Russia; phone: +79185562047; the Department of Information and Analytical Security Systems named after Professor Leonid Samoylovich Bershtein; dr. of eng. sc.; professor.

Valukhov Igor Alekseevich – Southern Federal University; e-mail: valuhov@sfedu.ru; Taganrog, Russia; phone: +79515263364; the Department of Information and Analytical Security Systems named after Professor Leonid Samoylovich Bershtein; postgraduate student.