

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

DOI: 10.18137/RNU.V9187.26.03.P.4

УДК 004.8

Шишкин Максим Андреевич

аспирант, Нижневартковский государственный университет, город Нижневартовск. ORCID: 0000-0002-7249-4869.

Электронный адрес: makcim.97@inbox.ru

Maxim A. Shishkin

Postgraduate, Nizhnevartovsk State University, Nizhnevartovsk. ORCID: 0000-0002-7249-4869.

E-mail address: makcim.97@inbox.ru

ЯЗЫКОВЫЕ МОДЕЛИ КАК ОСНОВА ДЛЯ ИЗВЛЕЧЕНИЯ ЗНАНИЙ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ АРХИТЕКТУР

Аннотация. В статье представлено исследование возможностей языковых моделей как основы для извлечения знаний из текстовых данных. Основное внимание уделено сравнительному анализу архитектур моделей-трансформеров. Оценивается эффективность моделей в решении задач извлечения знаний. Практическая часть исследования включает в себя экспериментальное сравнение точности, полноты и устойчивости нескольких архитектур на датасете с акцентом на проблемы адаптации моделей к предметным областям и работы в условиях ограниченных размеченных данных. Сравнительный анализ предоставляет рекомендации по выбору языковой модели, оптимальной для решения конкретных прикладных задач извлечения знаний, таких как автоматическое пополнение онтологий или семантический поиск.

Ключевые слова: машинное обучение, обработка естественного языка, извлечение знаний, распознавание именованных сущностей, сравнительный анализ, BERT, Python.

Для цитирования: Шишкин М.А. Языковые модели как основа для извлечения знаний: сравнительный анализ архитектур // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2026. № 3. С. 4–11. DOI: 10.18137/RNU.V9187.26.03.P.4

LANGUAGE MODELS AS A FOUNDATION FOR KNOWLEDGE EXTRACTION: A COMPARATIVE ANALYSIS OF ARCHITECTURES

Abstract. This article presents a study of the potential of language models as a foundation for knowledge extraction from text data. The focus is on a comparative analysis of transform model architectures. The effectiveness of these models in solving knowledge extraction problems is assessed. The practical part of the study includes an experimental comparison of the accuracy, recall, and robustness of several architectures on a dataset, with an emphasis on the challenges of adapting models to subject areas and working with limited labeled data. The comparative analysis provides recommendations for choosing the optimal language model for solving specific applied knowledge extraction problems, such as automatic ontology enrichment or semantic search.

Keywords: machine learning, natural language processing, knowledge extraction, named entity recognition, comparative analysis, BERT, Python.

For citation: Shishkin M.A. (2026) Language models as a foundation for knowledge extraction: a comparative analysis of architectures. *Vestnik of the Russian New University. The series: Complex systems: models, analysis and management*. No. 3. Pp. 4–11. DOI: 10.18137/RNU.V9187.26.03.P.4 (In Russian).

Введение

Появление больших языковых моделей и трансформеров за последнее время начало оказывать влияние на привычные методы обработки и извлечения знаний из текстов на естественном языке. Методы стали более автоматизированными и эффективными практически во всех областях применения. В то же время происходит высокий прирост неструктурированных текстов, которые необходимо привести в структурированный формат, преобразовать их в «полезные знания» для дальнейшего использования. В таких условиях большинство известных методов испытывают сложности в масштабировании, скорости и качестве извлечения знаний. Возникает необходимость в разработке и внедрении новых моделей на основе трансформеров, которых необходимо обучить на больших текстовых корпусах, а также научить их находить неочевидные паттерны. Нахождение семантических и синтаксических паттернов в тексте расширит возможности для автоматизации модели и последующего внедрения.

Актуальность исследования заключается в сравнительном анализе существующих архитектур языковых моделей и оценке качества моделей в задачах извлечения знаний из больших текстовых данных.

Целью данного исследования является проведение сравнительного анализа архитектур языковых моделей BERT, оценивание потенциала моделей для работы с большими текстами.

Для достижения поставленной цели решаются следующие задачи:

- систематизация архитектур языковых моделей BERT;
- сравнительная оценка методов машинного обучения с учителем;
- проверка эффективности архитектур на эталонных выборках;
- разработка критериев выбора языковой модели для конкретной области применения.

Научная новизна исследования заключается в комплексном подходе к сравнительному анализу архитектур языковых моделей BERT, который будет способен учитывать ресурсоёмкость инференса и возможность адаптации модели при нехватке размеченных данных.

Теоретическая значимость исследования состоит в возможности определения границы применимости оригинальной и дистиллированной модели BERT, а также возможности описания, насколько упрощение архитектуры модели способно повлиять на качество извлеченных сущностей.

Практическая значимость исследования заключается в предоставлении критериев выбора между оригинальной и сжатой трансформерной моделью BERT, позволяя проводить предварительную оценку на этапе проектирования NLP-конвейера.

Материалы и методы

Исследование проводилось с помощью известных методов машинного обучения с учителем, специализирующихся на задачах извлечения именованных сущностей из больших текстовых массивов. В качестве датасета был выбран CoNLL-2003 на английском языке, содержащий размеченные текстовые данные с аннотациями сущностей четырех типов: персоны (PER), организации (ORG), локации (LOC) и прочее (MISC). Для обработки и анализа данных был выбран дистрибутив Anaconda с интерпретатором Python версии 3.11.9. Работа выполнялась в интегрированной среде разработки PyCharm.

Для работы с предобученными языковыми моделями использовалась библиотека Transformers [1]. В исследовании применялся сравнительный анализ двух архитектур языковых моделей: BERT-base-cased и DistilBERT-base-cased, где BERT является полноценной трансформерной архитектурой, а DistilBERT – сжатой версией оригинального BERT с сохранением качества при уменьшенной вычислительной сложности [2]. Для предварительной обработки данных использовался парсер CoNLL-2003. Проводилась токенизация входных текстов и выравнивание меток NER-сущностей после токенизации [3]. Процедура дообучения моделей проводилась на 200 примерах из тренировочной выборки CoNLL-2003 в течение 5 эпох с использованием оптимизатора AdamW. Для оценки качества использовалась валидационная выборка датасета объемом 100 примеров. Оценка эффективности моделей проводилась с использованием комплекса метрик: accuracy, precision, recall и F1-score. Измерялось время инференса для анализа производительности. Визуализация результатов осуществлялась с применением библиотек matplotlib и seaborn, отображающие матрицы ошибок и сравнительные диаграммы по типам сущностей для комплексного анализа [4]. Выбор инструментария обусловлен несколькими небольшими преимуществами. Библиотека Transformers предоставляет доступ к языковым моделям с открытыми весами. PyTorch позволяет проводить эффективные вычисления с поддержкой GPU-ускорения, также он является достаточно гибким при работе языковыми моделями [5]. Использование библиотек Scikit-learn для вычисления метрик и Matplotlib для визуализации помогает детально оценивать результаты. [6].

Результаты исследования

В данном разделе представлены результаты сравнительного анализа архитектур языковых моделей для решения задачи извлечения именованных сущностей [7]. Исследование проводилось на датасете CoNLL-2003, содержащем 14041 тренировочное предложение, 3250 предложений для валидации и 3453 тестовых предложения на английском языке. Каждое предложение размечено сущностями четырех типов: персоны (PER), организации (ORG), локации (LOC) и miscellaneous (MISC). Этапы эксперимента включали: загрузку и парсинг данных, инициализацию языковых моделей BERT-base-cased и DistilBERT-base-cased, их дообучение на 200 примерах в течение 5 эпох, оценку качества на 100 примерах валидационной выборки с вычислением метрик accuracy, precision, recall и F1-score, а также комплексную визуализацию результатов [8] (см. Таблицу 1).

Таблица 1

Сравнительные метрики языковых моделей

Модель	Accuracy	Precision	Recall	F1-Score	Время инференса, с
BERT-base	0,9126	0,9343	0,9126	0,8939	0,0050
DistilBERT-base	0,8867	0,8863	0,8867	0,8776	0,0029

Анализ результатов показывает, что модель BERT-base демонстрирует лучшие показатели по всем метрикам качества, за исключением времени инференса. Модель DistilBERT-base, несмотря на снижение метрик качества, оказывается в 1,7 раза быстрее, что делает ее предпочтительной для задач, требующих баланса между производительностью и качеством.

Языковые модели как основа для извлечения знаний:
сравнительный анализ архитектур

В Таблице 2 представлено сравнение результатов по метрике F1-score с разным объемом данных для обучения.

Таблица 2

Сравнение результатов по метрике F1-score

Модель	Метрика	50 примеров, 3 эпохи	200 примеров, 5 эпох	Увеличение точности
BERT-base	F1-score	87,55 %	89,39 %	1,84 %
DistilBERT-base	F1-score	84,47 %	87,76 %	3,29 %

После обучения модели на 200 примерах и 5 эпохах снизился параметр потерь у BERT с 0,3330 до 0,1330. У DistilBERT данный показатель снизился с 0,3466 до 1584 соответственно.

Для визуального анализа ошибок классификации построена матрица ошибок модели BERT-base (Рисунок 1). Матрица демонстрирует высокую точность распознавания сущностей PER и LOC, в то время как сущности ORG и MISC имеют более высокий уровень погрешности [9].

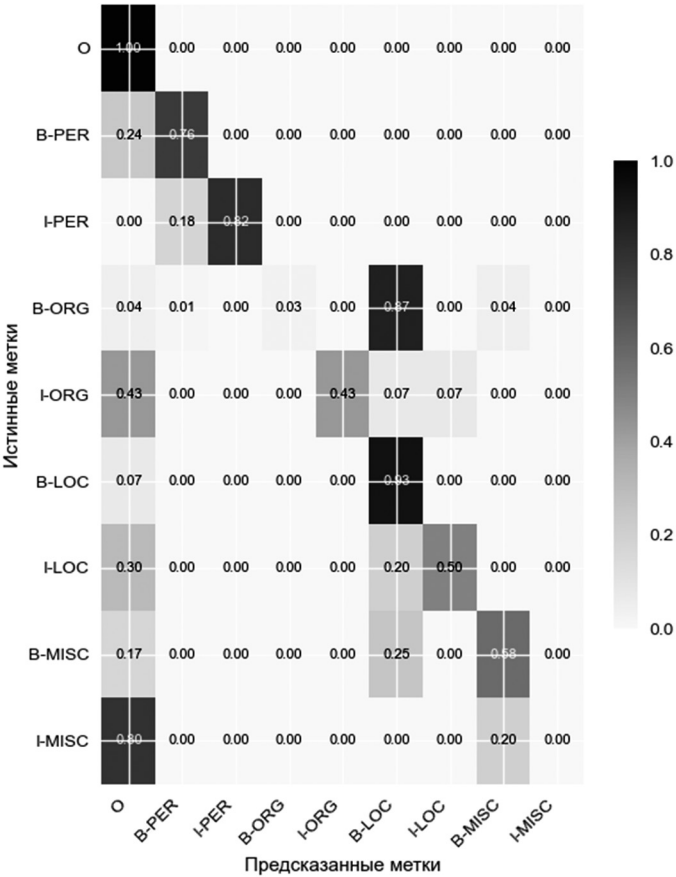


Рисунок 1. Матрица ошибок модели BERT-base

Сравнительная эффективность моделей по типам сущностей отображена на Рисунке 2. Модель BERT-base показывает лучшие результаты для сущностей PER, LOC и MISC, но теряет преимущество в распознавании ORG-сущностей.

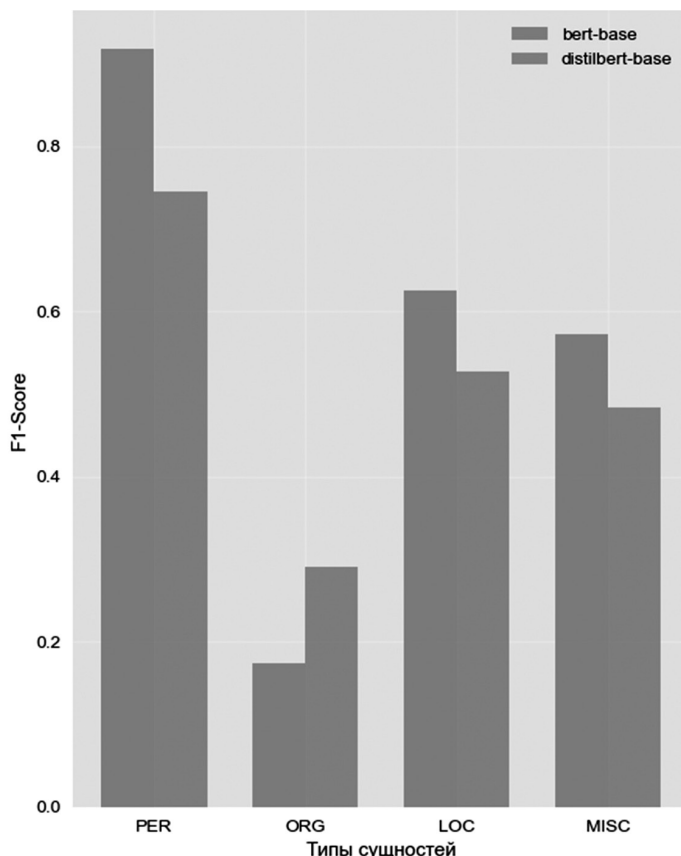


Рисунок 2. Сравнительная эффективность моделей

Также представлено распределение уверенности моделей в предсказаниях. Для каждого токена в последовательности модель вычисляет вероятностное распределение по всем классам NER [10]. Уверенность модели определяется как максимальное значение вероятности среди всех классов. При подсчете уверенностей моделей использовалась функция активации:

$$\text{Softmax}(Z_i) = \frac{e^{Z_i}}{\sum_{j=1}^K e^{Z_j}},$$

где Z_i – логит для i -го класса; K – количество классов; e^{Z_i} – экспонента логита; $\sum_{j=1}^K e^{Z_j}$ – сумма экспонент по всем классам.

Уверенность вычисляется как максимальное значение вероятности после применения функции Softmax к выходным логитам для каждого токена.

Обсуждение результатов исследования

Высокие значения метрик оценки качества делают пригодными трансформерные архитектуры для задач NER. Это подтверждается визуальным анализом матрицы ошибок, на которой видна высокая точность распознавания сущностей при незначительной погрешности в классах ORG и MISC. Модель BERT-base подчеркивает высокие показатели по всем метрикам точности. DistilBERT-base показала сопоставимые результаты, но при этом она выигрывает в скорости работы, которая крайне необходима для обработки больших объемов данных. Модель BERT-base показала положительные результаты предсказаний, модель способна корректно классифицировать сложные структуры текстов. Разработанная методология позволяет проводить комплексную оценку языковых моделей и может быть адаптирована для задач обработки естественного языка.

Заключение

В рамках данного исследования был проведен сравнительный анализ BERT-base-cased и DistilBERT-base-cased архитектур языковых моделей. Экспериментальная проверка с использованием датасета CoNLL-2003 и визуализация результатов позволила достичь поставленной цели и задач исследования. Рассмотренные языковые модели показывают высокую результативность при их использовании в сценариях с малым объемом размеченных данных. Разработанный программный комплекс является универсальным инструментом для сравнительной оценки языковых моделей. На примере извлечения именованных сущностей была проведена оценка компромисса между точностью и производительностью. Перспективы дальнейших исследований связаны с адаптацией разработанной методологии для задач автоматического подбора гиперпараметров и создания интегрированной среды для сравнительного анализа современных языковых моделей.

Литература

1. Кузьмин А.И. Обзор и сравнение популярных инструментов для обработки естественного языка // Международный журнал информационных технологий и энергоэффективности. 2021. Т. 6, № 4(22). С. 24–33. EDN LJTEW.
2. Новиков А.С., Шарлаев Е.В. Использование языковой модели BERT для анализа текстов на русском языке // Наукосфера. 2021. № 6-1. С. 200–202. EDN SUDLXE.
3. Смольянов А.Г., Клоц А.И. Анализ возможностей модели BERT в распознавании текста // Огарёвские чтения: материалы Всероссийской с международным участием научной конференции. Саранск: Национальный исследовательский Мордовский государственный университет им. Н.П. Огарёва, 2022. С. 703–707. EDN GNNTH.
4. Усенко П.В., Манцеров П.Е. Обзор библиотек Python для работы с графикой и искусственным интеллектом // Новые информационные технологии и системы (НИТиС-2023): сборник научных статей по материалам XX Международной научно-технической конференции. Пенза: Пензенский государственный университет, 2023. С. 141–143. EDN GHBBMS.
5. Лобанова П.И. Обзор библиотеки PyTorch языка программирования Python // Молодые ученые – развитию Национальной технологической инициативы (ПОИСК). 2025. № 1. С. 996–997. EDN MZKJYP.

6. Гаунов С.Р., Баймуратов У.Г., Ситников С.Ю. Машинное обучение на Python: использование библиотек Tensorflow и Scikit-Learn // Экономика и управление: проблемы, решения. 2024. Т. 8, № 12(153). С. 72–81. DOI: 10.36871/ek.up.p.r.2024.12.08.009. EDN OCAKJB.
7. Салып Б.Ю., Смирнов А.А. Анализ модели BERT как инструмента определения меры смысловой близости предложений естественного языка // StudNet. 2022. Т. 5, № 5. С. 33. EDN TDRZZC.
8. Лагутина Н.С., Васильев А.М., Зафиевский Д.Д. Задачи в области распознавания именованных сущностей: технологии и инструменты // Моделирование и анализ информационных систем. 2023. Т. 30, № 1. С. 64–85. DOI: 10.18255/1818-1015-2023-1-64-85. EDN DJMEDG.
9. Пупков Т.С. Применение матрицы ошибок для оценки качества обучения моделей классификации // Актуальные вопросы эксплуатации систем охраны и защищенных телекоммуникационных систем: сборник материалов Всероссийской научно-практической конференции. Воронеж: Воронежский институт МВД России, 2021. С. 35–36. EDN EKDEYO.
10. Кожина А.В., Бежитский С.С. Распознавание именованных сущностей в текстах // Решетневские чтения: материалы XXVIII Международной научно-практической конференции. Красноярск: Сибирский государственный университет науки и технологий им. акад. М.Ф. Решетнева, 2024. С. 61–63. EDN ETUALZ.

References

1. Kuzmin A.I. (2021) Review and comparison of popular tools for natural language processing. *International Journal of Information Technologies and Energy Efficiency*. Vol. 6, no. 4(22). Pp. 24-33. EDN LJTEW. (In Russian).
2. Novikov A.S., Sharlaev E.V. (2021) Using BERT language model for Russian text analysis. *Naukosfera*. No. 6-1. Pp. 200-202. EDN SUDLXE. (In Russian).
3. Smolyanov A.G., Klots A.I. (2022) Analysis of BERT model capabilities in text recognition. *Ogarev Readings: proceedings of the All-Russian Scientific Conference with International Participation*. Saransk: Ogarev National Research Mordovia State University. Pp. 703-707. EDN GNNTTH. (In Russian).
4. Usenko P.V., Manczerova P.E. (2023) Review of Python libraries for graphics and artificial intelligence. *New Information Technologies and Systems (NITIS-2023): collection of scientific articles of the XX International Scientific and Technical Conference*. Penza: Penza State University. Pp. 141-143. EDN GHBBMS. (In Russian).
5. Lobanova P.I. (2025) Review of PyTorch library of Python programming language. *Young Scientists for the Development of National Technology Initiative (POISK)*. No. 1. Pp. 996-997. EDN MZKJYP. (In Russian).
6. Gaunov S.R., Baimuradov U.G., Sitnikov S.Yu. (2024) Machine learning in Python: using Tensorflow and Scikit-Learn libraries. *Economics and Management: Problems, Solutions*. Vol. 8, no. 12(153). Pp. 72-81. DOI: 10.36871/ek.up.p.r.2024.12.08.009. EDN OCAKJB. (In Russian).
7. Salyp B.Yu., Smirnov A.A. (2022) Analysis of BERT model as a tool for determining semantic similarity measure of natural language sentences. *StudNet*. Vol. 5, no. 5. Pp. 33. EDN TDRZZC. (In Russian).
8. Lagutina N.S., Vasilev A.M., Zafievskiy D.D. (2023) Tasks in the field of named entity recognition: technologies and tools. *Modeling and Analysis of Information Systems*. Vol. 30, no. 1. Pp. 64-85. DOI: 10.18255/1818-1015-2023-1-64-85. EDN DJMEDG. (In Russian).
9. Pupkov T.S. (2021) Application of confusion matrix for evaluating the quality of classification model training. *Current Issues of Operation of Security Systems and Protected Telecommunication Systems: collection*

Языковые модели как основа для извлечения знаний:
сравнительный анализ архитектур

of materials of the All-Russian Scientific and Practical Conference. Voronezh: Voronezh Institute of the Ministry of Internal Affairs of Russia. Pp. 35-36. EDN EKDEYO. (In Russian).

10. Kozhina A.V. Bezhitskiy S.S. (2024) Named entity recognition in texts. *Reshetnev Readings: proceedings of the XXVIII International Scientific and Practical Conference*. Krasnoyarsk: Reshetnev Siberian State University of Science and Technology. Pp. 61-63. EDN ETUALZ. (In Russian).

Поступила в редакцию: 25.03.2026

Received: 25.03.2026

Поступила после рецензирования: 06.05.2026

Revised: 06.05.2026

Принята к публикации: 22.05.2026

Accepted: 22.05.2026